The Annals of Statistics
2001, Vol. 29, No. 1, 1–65

LOCAL EXTREMES, RUNS, STRINGS AND MULTIRESOLUTION[1]

BY P. L. DAVIES AND A. KOVAC

Universität Essen

The paper considers the problem of nonparametric regression with emphasis on controlling the number of local extremes. Two methods, the run method and the taut-string multiresolution method, are introduced and analyzed on standard test beds. It is shown that the number and locations of local extreme values are consistently estimated. Rates of convergence are proved for both methods. The run method converges slowly but can withstand blocks as well as a high proportion of isolated outliers. The rate of convergence of the taut-string multiresolution method is almost optimal. The method is extremely sensitive and can detect very low power peaks.

Section 1 contains an introduction with special reference to the number of local extreme values. The run method is described in Section 2 and the taut-string-multiresolution method in Section 3. Low power peaks are considered in Section 4. Section 5 contains a comparison with other methods and Section 6 a short conclusion. The proofs are given in Section 7 and the taut-string algorithm is described in the Appendix.

1. Introduction.

1.1. *Approximating data and procedures.* Consider a data set $(t_i, y(t_i))$, $i = 1, \ldots, n$ in $[0, 1] \times \mathbb{R}$ and a family of nonparametric regression models

$$Y(t) = f(t) + \varepsilon(t), \qquad 0 \le t \le 1, \tag{1.1}$$

indexed by f and ε. We always assume that the t_i are strictly ordered, $0 \le t_1 < t_2 < \cdots < t_n \le 1$. Nonparametric regression is concerned with specifying models, that is, specifying functions f and noise ε such that typical data sets $(t_i, Y(t_i)), i = 1, \ldots, n$ generated under the model "look like" the given data set $(t_i, y(t_i)), i = 1, \ldots, n$ [Donoho (1988); Davies (1995)]. Such models will be regarded as adequate approximations to the data. They are obtained by manipulating or, more formally, applying a functional to the data and so defining a statistical procedure [Tukey (1993)]. Two such procedures, the run procedure and the taut-string multiresolution procedure, are considered in this paper. Both are based on a decomposition of the data of the form

$$y(t_i) = f_n(t_i) + r_n(t_i), \tag{1.2}$$

which is a special case of the general Tukey decomposition

$$\text{Data} = \text{Signal} + \text{Noise}. \tag{1.3}$$

Received April 1999; revised September 2000.

[1]Supported in part by Sonderforschungsbereich 475, University of Dortmund and DFG Grant 237/2-1 and 237/2-2 on time series analysis.

AMS 2000 *subject classifications*. Primary 62G07; secondary 65D10, 62G20.

Key words and phrases. Nonparametric regression, local extremes, runs, strings, multiresolution analysis, asymptotics, outliers, low power peaks.

In general Signal and Noise in (1.3) can be separated by assuming that the signal is simple and the noise complex. We measure the simplicity of a function f on $[0,1]$ by the number of local extreme values in the open interval $(0,1)$. This is not the standard definition of simplicity which usually refers to some form of smoothness. Complexity is related to randomness and we use a stochastically based concept of noise defined in terms of independently distributed random variables. The two procedures we consider depend on two different definitions of approximation to white noise. The first is based on Bernoulli sequences and can be written as

$$\text{max\_run}(\text{sgn}(r_1(t_1)), \ldots, \text{sgn}(r_n(t_n))) \leq \rho_n, \tag{1.4}$$

where $\text{sgn}(r_n(t_i))$ denotes the sign of the residual $r_n(t_i)$ and ρ_n is some given number. In other words, the residuals $r_n(t_i)$ may be adequately approximated by white noise or, more informally, look like white noise if (1.4) holds. The default value we suggest is

$$\rho_n = \lceil \log_2 n - 1.47 \rceil,$$

which is justified below. The second definition of approximation is based on Gaussian white noise. Suppose for the moment that $n = 2^m$ is a power of 2. The multiresolution coefficients $w_{i,k}$ are defined by

$$w_{j,k} = 2^{-j/2} \sum_{i=k2^j+1}^{(k+1)2^j} r_n(t_i). \tag{1.5}$$

The residuals may be adequately approximated by Gaussian white noise if

$$|w_{j,k}| \leq \sigma_n \sqrt{2.5 \log n}, \tag{1.6}$$

where σ_n is some measure of the scale of the residuals $r_n(t_i)$. The default functional we use is

$$\sigma_n = \frac{1.48}{\sqrt{2}} \text{ Median } \{|y(t_2) - y(t_1)|, \ldots, |y(t_n) - y(t_{n-1})|\}. \tag{1.7}$$

This is the same as hard thresholding for wavelets except that we use the factor 2.5 instead of the usual 2. This is for the simple pragmatic reason that 2.5 seems to give better results.

The nonparametric regression procedures we study may now be formulated as follows.

THE RUN PROBLEM. Determine the smallest integer k_n for which there exists a function f_n on $[0,1]$ with k_n local extreme values and such that the residuals $r_n(t_i)$ satisfy the run condition (1.4).

THE MULTIRESOLUTION PROBLEM. Determine the smallest integer k_n for which there exists a function f_n on $[0,1]$ with k_n local extreme values and such that the residuals $r_n(t_i)$ satisfy the multiresolution condition (1.6).

Both problems are well defined for any given data set. We give a complete solution of the run problem and make a laudable attempt at the multiresolution problem. The difficulty with the multiresolution problem is that of obtaining suitable approximating functions f_n with no unnecessary local extreme values. The method we use is that of the taut string which will be described below.

The approach given in this paper is neither Bayesian nor frequentist. It makes no assumption that the data are generated by a random mechanism. Indeed they may well be deterministic. We take the point of view that models are approximations to data. In particular we make no reference to "true" regression functions for real data as we do not think that these exist in the sense that, say, elephants exist. By formulating the problem in terms of approximation we avoid the embarrassment of using the word "true," whether in inverted commas or not. We refer to Davies (1995). Readers who do not like this may forget this paragraph, move on to the next section and assume or "know" that all data sets derive from some real existing "true but unknown" regression function f, whatever that might mean.

1.2. *Test beds.* Statistical procedures can be evaluated using real data sets and under the well-controlled conditions of a stochastic model or test bed. Tukey uses the word "challenge" [Morgenthaler and Tukey (1991)] in this context. The test beds we use are data sets generated by stochastic models of the form

$$Y(t_i) = f(t_i) + \varepsilon(t_i), \qquad i = 1, \ldots, n, \tag{1.8}$$

where the $\varepsilon(t_i)$ are generally but not always taken to be independently and identically distributed random variables. On test beds one can talk about estimating functions. They have the advantage of allowing a direct comparison of the function f used to generate the data with functions f_n yielded by the procedures. The comparisons are often visual; indeed these may be the most convincing ones, but can include, as we shall, statements on consistency and rates of convergence. Traditional confidence sets are also possible on test beds but these are of a one-sided nature [Donoho (1988)]. For any given sample size n it is possible to perturb f in such a way that it has an arbitrarily large number of local extremes without these being visible in the data. For this reason it is not possible to give a finite upper bound for the number of local extreme values of the function f. When considering rates of convergence we use the supremum norm

$$\|f - f_n\| = \sup_{0 \le t \le 1} |f(t) - f_n(t)|.$$

For certain test beds it can be shown that optimal rates of convergence exist. We show that the run-based procedure has a slow rate of convergence, namely $O\big(\frac{\log\log n}{\log n}\big)$ while the taut-string-multiresolution procedure has the optimal rate of convergence of $O\big((\frac{\log n}{n})^{1/3}\big)$.

1.3. *Smoothness.* Smoothness is not a consideration in this paper. Techniques for smoothing under shape and deviation constraints have been developed by Metzner (1997), Davies and Löwendick (1999) and Majidi (2000).

1.4. *Previous work.* Much work has been done on the problem of nonparametric regression. Of the different approaches we mention kernel estimation [Nadaraya (1964); Watson (1964)], penalized likelihood [Silverman (1985); Green and Silverman (1994)], wavelets [Donoho, Johnstone, Kerkyacharian and Picard (1995)] and local polynomials [Fan and Gijbels (1995, 1996)]. None of these methods is directly concerned with local extremes but research has been done which explicitly takes the shape of the regression function into account. Mammen (1991) uses monotone least squares fits between local extrema whilst Mammen and Thomas-Agnan (1998), Mammen, Marron, Turlach and Wand (1998), Delecroix, Simioni and Thomas-Agnan (1995) and Ramsay (1998) modify classical estimators such as spline smoothers and kernel estimators to deal with monotonicity or convexity constraints. None of these papers is directly concerned with estimating the number or the positions of the local extreme values. Work in this direction has been done by Dümbgen (1998b) who applies linear rank tests to locate local extrema. Hengartner and Stark (1995) use the Kolmogoroff ball centred at the empirical distribution function to obtain nonparametric confidence bounds for shape restricted densities. Chaudhuri and Marron (1997) assess the significance of zero crossings of derivatives and use their results to provide a graphical device for displaying the significance the local extremes. Another approach is that of mode testing. We refer to Good and Gaskins (1980), Silverman (1986), Hartigan and Hartigan (1985), Fisher, Mammen and Marron (1994). The positions of the local extreme values are considered by Minotte (1997) using a procedure which decides for each mode in a mode tree [Minotte and Scott (1993)] whether it is significant or not. Mächler (1995) presents an approach using a roughness penalty to penalize points of inflection. The taut string method was used by Davies (1995) in the context of density estimation; there are connections with the excess mass approach of Müller and Sawitzki (1991). Further work in this direction is due to Polonik (1995a, 1999). Other articles on shape restricted densities and regression functions are Groeneboom (1985), Hartigan (1987), Robertson (1967), Sager (1979, 1982, 1986) and Wegman (1970). Order restricted inference is considered in Barlow, Bartholomew, Bremner and Brunk (1972) and Robertson, Wright and Dykstra (1988). Finally in an article which appeared whilst the present paper was under revision Polzehl and Spokoiny (2000) give a method based on adaptive local means.

The run method [Davies (1995); Metzner (1997)] may be seen as the inversion of the run test for testing the independence of a sequence of observations. It yields the minimum number of local extremes consistent with the observations as well as approximation intervals for their location. Dümbgen (1998a, 1998b) inverts other tests and obtains better convergence rates on standard test beds but at the cost of greater computational complexity.

Taut strings are well understood in the context of fitting an isotone function. The greatest convex minorant of the integrated data is a taut string and its derivative is precisely the least squares isotone approximation [Barlow, Bartholomew, Bremner and Brunk (1972); Leurgans (1982)]. The idea of using taut strings for densities goes back to Hartigan and Hartigan (1985) who derived a test for the unimodality of a density. In Davies (1995) it was explicitly used to calculate approximate densities. It was first used in the general nonparametric regression problem by Mammen and van de Geer (1997). They showed that the taut string is a special case of a penalized least squares functional where the penalty is based on the total deviation norm. The derivative of the taut string has the smallest number of local extremes of all functions whose integral lies in the supremum ball. Davies (2000) gives an application of the string method to spectral density functions.

2. The run method.

2.1. *General description.* We illustrate the general method of solving the run problem of Section 1.1 using the data shown in Figure 1 with a maximum allowable run length $\rho_n = 2$. To ease the notation we set $t_i = i$.

Consider a function f_n which is initially nonincreasing. The value of $f_n(3)$ cannot exceed $\max\{y(1), y(2), y(3)\}$ as otherwise the first three residuals would be negative giving rise to a run of length 3. This gives the upper bound

$$f_n(3) \le \max\{y(1), y(2), y(3)\}. \tag{2.1}$$

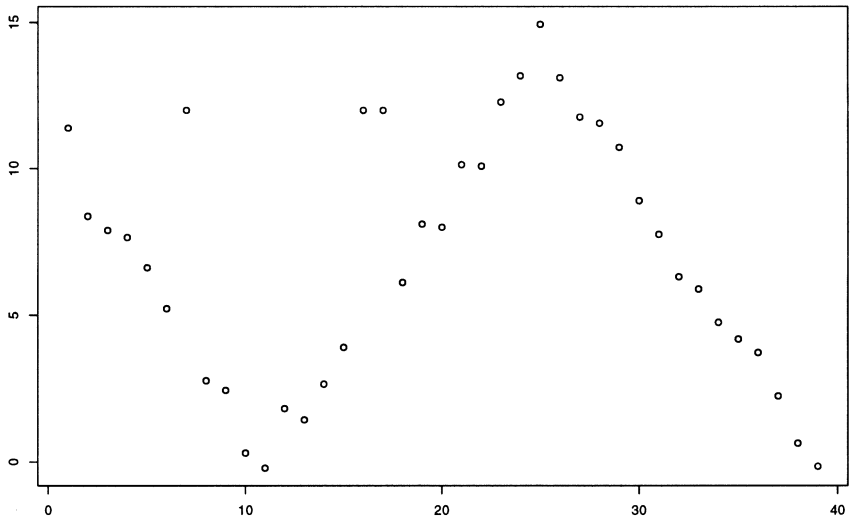

FIG. 1. *An artificial data set to illustrate the run procedure.*

No nontrivial upper bounds are available for $f_n(1)$ and $f_n(2)$. Similarly, a lower bound for $f_n(1)$ is given by

$$f_n(1) \geq \min\{y(1), y(2), y(3)\}.$$

As the function is taken to be initially nonincreasing, the above argument extended to each point gives rise to a sequence $u(j)$ of upper bounds defined by

$$u(j) = \min\big\{u(j-1), \max\{y(j-2), y(j-1), y(j)\}\big\} \tag{2.2}$$

with $u(1) = u(2) = \infty$. To see this we note that the reasoning which lead to (2.1) gives

$$f_n(j) \leq \max\{y(j-2), y(j-1), y(j)\}. \tag{2.3}$$

On the other hand if $u(j-1)$ is an upper bound for $f_n(j-1)$ then

$$f_n(j) \leq f_n(j-1) \leq u(j-1). \tag{2.4}$$

Combining (2.3) and (2.4) leads to (2.2). The corresponding lower bounds $l(j)$ are given by

$$l(j) = \max\big\{l(j+1), \min\{y(j), y(j+1), y(j+2)\}\big\}. \tag{2.5}$$

If at some point i the lower bound $l(i)$ exceeds the upper bound $u(i)$ then it is not possible for the function f_n to be nonincreasing on $1, \ldots, i+2$ and for the maximum run length not to exceed 2. Thus at the latest at the point $i+1$ we must switch from a nonincreasing to a nondecreasing function. It turns out that it is not necessary to calculate the lower bounds in order to determine the point at which the switch must be made. The point $i+2$ is also the first point at which the last three values $y(i)$, $y(i+1)$ and $y(i+2)$ lie above the upper bounds $u(i)$, $u(i+1)$ and $u(i+2)$, respectively. This can be shown by backward induction. The result can be seen in the upper panel of Figure 2. The values of y at the points 14, 15 and 16 lie above the upper bounds. Thus at the latest at the point 15 a switch must be made to a nondecreasing function. This is done as follows. We set $l(14) = l(15) = -\infty$ and calculate lower bounds for the nondecreasing section by

$$l(j) = \max\big\{l(j-1), \min\{y(j-2), y(j-1), y(j)\}\big\}, \qquad j = 16, \ldots .$$

The lower bounds are calculated until either the end of the data set is reached or at some point the last three y-values lie below the corresponding lower bounds. If the latter contingency occurs, a switch is made to a nonincreasing function and the upper bounds are calculated as before. For the data shown in Figure 1, the switch must be made at the latest at the point 28 as can be seen from the upper panel of Figure 2. The results obtained are the following. If we start with a nonincreasing function and limit the allowable run length to 2 then at least two local extremes are required. The first, a local minimum, must be attained at a point i with $i \leq 15$. The second, a local maximum, must be attained at a point j with $j \leq 28$. The process described here is called "stretching to the right" in Davies (1995). Starting with a nondecreasing

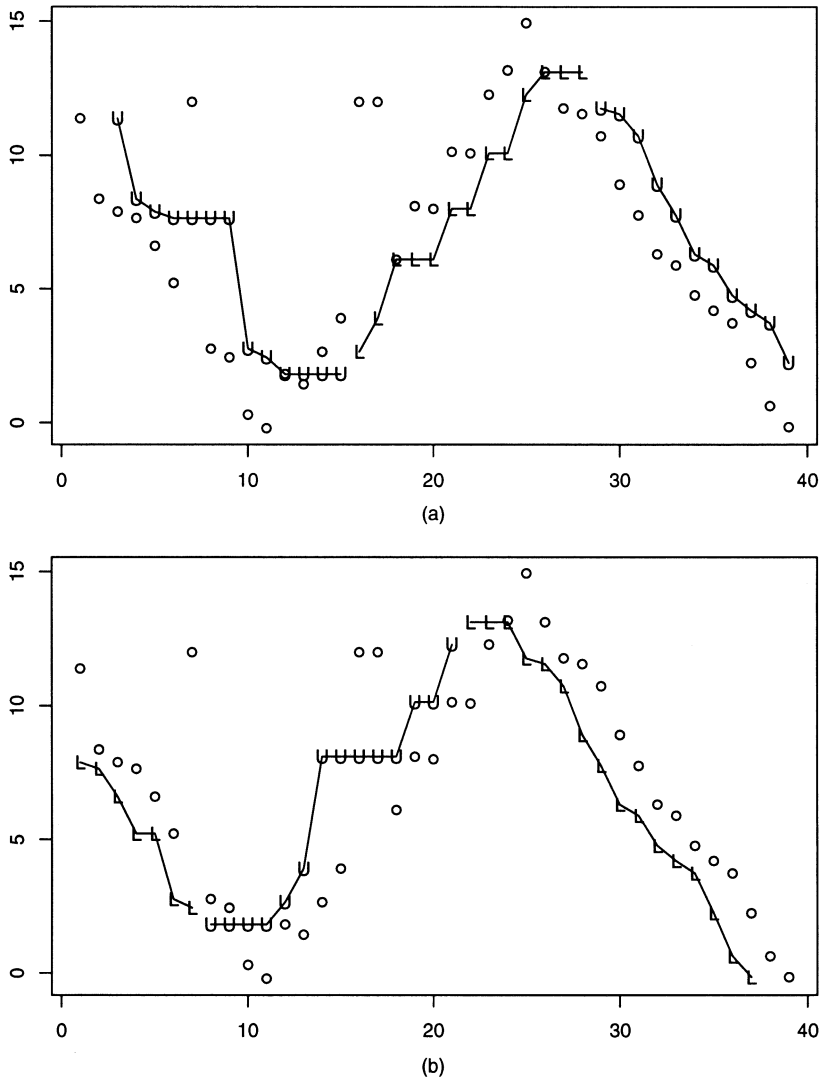

FIG. 2. *The data set of Figure* 1 *together with the bounds from stretching to the right* (*upper panel*) *and stretching to the left* (*lower panel*).

function leads to one extra local maximum: in general that initial behavior is chosen which minimizes the number of local extremes. The opposite process, stretching to the left, starts with the last point of the data set and moves to the left, calculating the bounds in the same manner. The result is shown in the lower panel of Figure 2. The number of local extremes is the same, in this case two, and the lower bounds for the positions of the local extremes are 8 and 15. The combined result for the data of Figure 1 is that the local

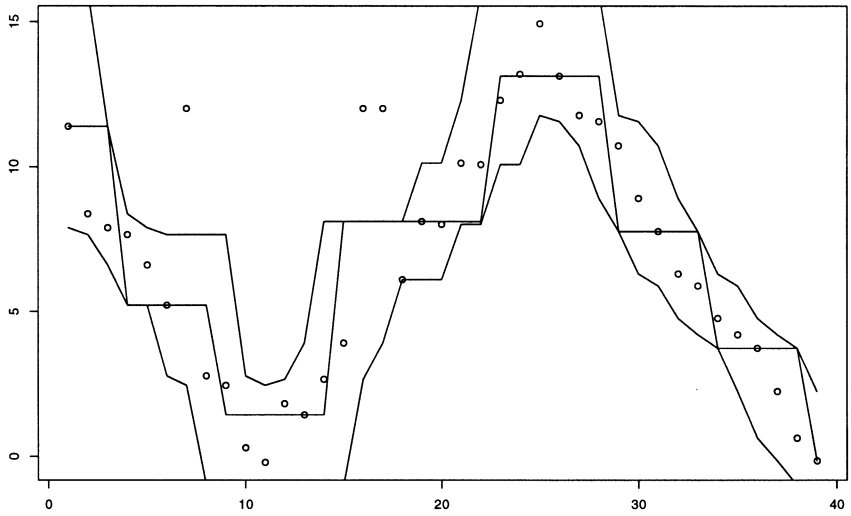

FIG. 3. *Final bounds for a run length of* 2 *together with a function satisfying the run condition and lying between the bounds.*

minimum must be attained in the interval [8, 15] and the local maximum in the interval [22, 28]. In an interval bounding the position of a local minimum there is no nontrivial lower bound for the function without further specification of the location of the local minimum. Similarly, in an interval bounding the position of a local maximum there is no nontrivial upper bound for the function. At points where only one lower bound has been calculated (either from stretching to the right or to the left), this defines the lower bound. At points where two lower bounds have been calculated, the minimum is taken. The upper bound is calculated in the corresponding manner. The final bounds are shown in Figure 3. The outliers at the points 7, 16 and 17 are seen not to have lead to additional local extremes. A more detailed analysis of the effect of outliers is given below.

The results so far show that restricting the maximal run length to 2 forces any regression function to have at least two local extreme values. If f_n is a function with two local extreme values and whose residuals have a maximum run length of 2 then the function must initially be nonincreasing, the local minimum must be attained in the interval [8, 15] and the local maximum in the interval [22, 28]. Furthermore the function must lie between the bounds at all data points.

It is easy to show that not every function with the correct monotonicity behavior and lying between the bounds satisfies the run condition. The construction of a function with only two local extreme values and which satisfies the run condition is not completely trivial. Suppose that it is possible, as indeed it is, to define a nonincreasing function which satisfies the run condition with $\rho_n = 2$ on the interval [1, 15]. The last two residuals are of necessity positive as 15 is an upper bound for the position of the local minimum. To prevent

a run of length 3 the next residual must be negative and the function must therefore have a value exceeding $y(16) = 12$. It is clear that this choice of f_n will lead to four local extreme values. To construct a function with only two local extreme values we proceed as follows. We take the upper bound 15 for the position of the first local extreme value. The upper and lower bounds u and l as defined by (2.2) and (2.5), respectively, are calculated for the section $[1, 15]$ of the data. The lower bound is best calculated from right to left with initial values $l(14) = l(15) = -\infty$. The first two values are defined to be

$$f_n(1) = f_n(2) = \max\{y(1), y(2)\}$$

and the others are defined as follows. If the first i values of f_n have been defined, the value at $i+1$ is defined as follows. If $f_n(i) \le u(i+1)$ then we set $f_n(i+1) = f_n(i)$. If $f_n(i) > u(i+1)$ we define f_n as follows. If $y(i+1) \ge l(i+1)$ then we set $f_n(i+1) = l(i+1)$. If not, we backtrack and determine the first j to the left of $i+1$ such that $y(j) \ge l(j)$. That is,

$$j = \max\{m\colon m \le i+1,\ y(m) \ge l(m)\}.$$

We then set $f_n(m) = l(j)$ for $j \le m \le i+1$. At the latest we must switch to a nondecreasing function at the point 15. We do this as follows. We consider the upper bound of the position of the next local extreme value. For the present data the is the point 28. We now calculate the lower and upper bounds for the section of the data on $[14, 28]$. The first two values of the function are defined by

$$f_n(14) = f_n(15) = \min\{y(14), y(15)\}.$$

The function is now defined as on the first section but with the roles of the lower and upper bounds interchanged. The function f_n is continued as a constant until the first time it drops below the lower bound. At this point a switch is made to the upper bound but only if the value of the observation at this point does not exceed the upper bound. If not, we backtrack to that first point where the upper bound exceeds the y-value. This process is continued in the obvious manner until the end of the data set is reached. The function so defined satisfies the run condition, has the correct monotonicity behavior and lies between the upper and lower bounds over the whole range. It is shown in Figure 3. This construction is due to Davies (1995) and Metzner (1997). The latter contains further details and variations on the run procedure.

The general recurrence equations for calculating the upper and lower bounds for a given run length ρ_n are as follows:

$$u_n(i) = \begin{cases} \min\{u_n(i-1), \max\{y(j), i-\rho_n \le j \le i\}\}, & \text{nonincreasing,} \\ \min\{u_n(i+1), \max\{y(j), i \le j \le i+\rho_n\}\}, & \text{nondecreasing,} \end{cases} \tag{2.6}$$

$$l_n(i) = \begin{cases} \max\{l_n(i+1), \min\{y(j), i \le j \le i+\rho_n\}\}, & \text{nonincreasing,} \\ \max\{l_n(i-1), \min\{y(j), i-\rho_n \le j \le i\}\}, & \text{nondecreasing.} \end{cases} \tag{2.7}$$

Combining these results gives the following theorem.

THEOREM 1. *For a given data set* $(t_i, y(t_i))$; $i = 1, \ldots, n$ *and a given maximal run length* ρ_n *for the residuals*

$$r_n(t_i) = y(t_i) - f_n(t_i),$$

the lower and upper bounds described above specify the minimum number k_n *of local extreme values a function must have so that the residuals satisfy the run condition. Furthermore, there exist functions with* k_n *local extreme values whose residuals satisfy the run condition. Any function with* k_n *local extreme values whose residuals satisfy the run condition must lie between the bounds over the whole range. The local extreme values of any such function must also be attained in the* k_n *intervals* $[t_i^l, t_i^r]$, $1 \le i \le k_n$, *determined by the stretching to the right and left procedures.*

The run procedure is easy to apply and requires only the specification of the maximal run length. To give a default value for the allowable run length ρ_n we consider a sequence of i.i.d. Bernoulli random variables taking the values 1 and -1 each with probability $1/2$. We denote by R_n the length of the longest subsequence which is composed entirely of 1's or -1's. For any given α, $0 < \alpha < 1$, we denote the α-quantile of R_n by $qu(n, \alpha, R_n)$; that is,

$$qu(n, \alpha, R_n) = \min\{m\colon \mathbf{P}(R_n \le m) \ge \alpha\}.$$

For large n we have the simple approximation

$$\rho_n = qu(n, \alpha, R_n) \approx \lceil \log_2 n - 2 - \log_2(-\log(\alpha)) \rceil, \tag{2.8}$$

which can be deduced from the results given in Section XIII.7 of Feller (1968) or shown directly. The default choice of run length is (2.8) with $\alpha = 0.5$ which represents a form of median for the number of local extreme values. In this case,

$$\rho_n = qu(n, 0.5, R_n) \approx \lceil \log_2 n - 1.47 \rceil. \tag{2.9}$$

Nonparametric regression is an infinite dimensional problem. This is one reason why the bounds l_n and u_n are relatively wide and why the intervals $[t_i^l, t_i^r]$, $i = 1, \ldots, k_n$, are relatively long. Another is the use of a rather crude measure of approximation to white noise, namely the run length. The effects can be seen in the upper panel of Figure 4 which shows 1000 points of a sine curve contaminated with Cauchy noise as in (2.10):

$$y(t) = 10 \sin(t) + \varepsilon(t), \qquad \varepsilon(t) \quad \text{i.i.d. standard Cauchy.} \tag{2.10}$$

The bounds are based on a run length $\rho_n = 9$ which is the default value given by (2.9) and result in $k_n = 2$ local extreme values. One method of obtaining narrower bounds is to decrease the run length whilst maintaining just two local extreme values. For the sine curve the run length can be reduced to $\rho_n = 6$ without introducing any further local extremes. The lower panel of Figure 4 shows these new bounds.

The bounding intervals based on a run length of $\rho_n = 9$ are $[141, 376]$ for the local maximum and $[600, 878]$ for the local minimum. The corresponding

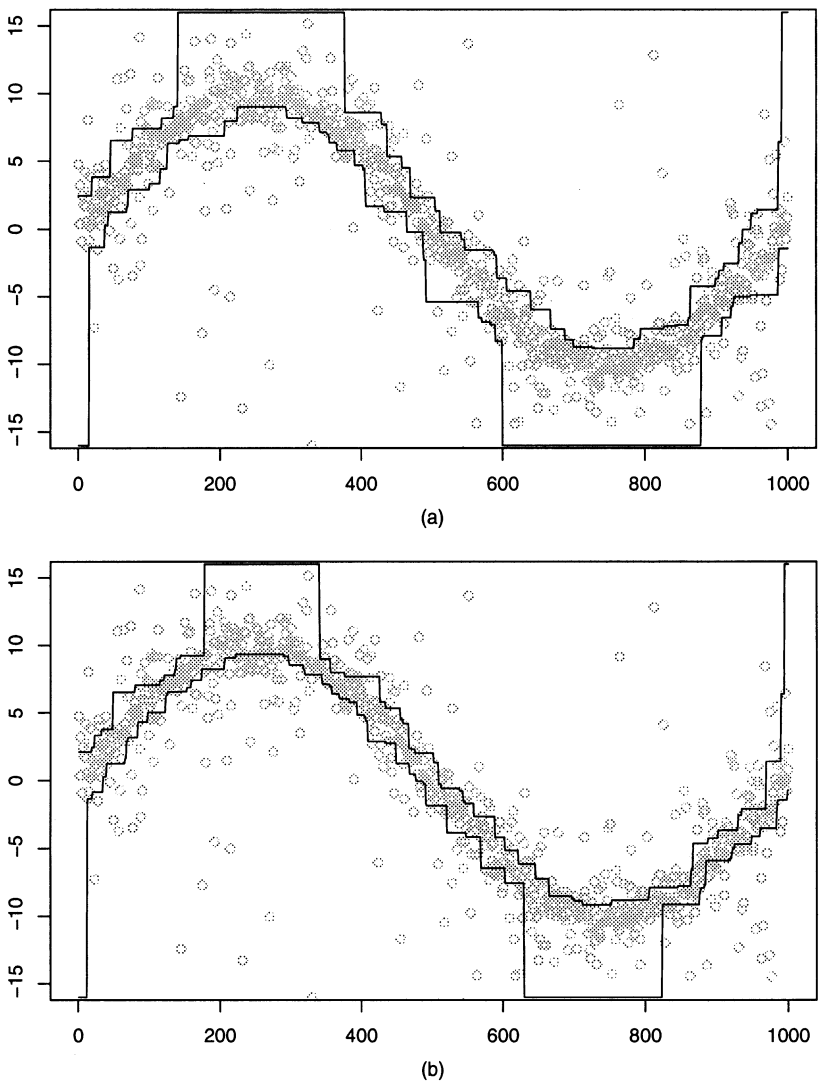

FIG. 4. 1000 *data points generated under* (2.10). *Upper panel shows bounds with default run length* $\rho_n = 9$ *given by* (2.9). *Lower panel shows bounds with minimal run length* $\rho_n = 6$ *consistent with two local extremes*.

intervals for the run length $\rho_n = 6$ are [178, 340] and [630, 823], respectively. These hold for any function which satisfies the run condition. It is however possible to specify points within these intervals near which the function is required to have a local extreme value. The data of Figure 1 show that it may not be possible to have local extremes at exactly these points and still fulfil the run condition. A modification of the construction given above of a function within the bounds which satisfies the run condition shows the following.

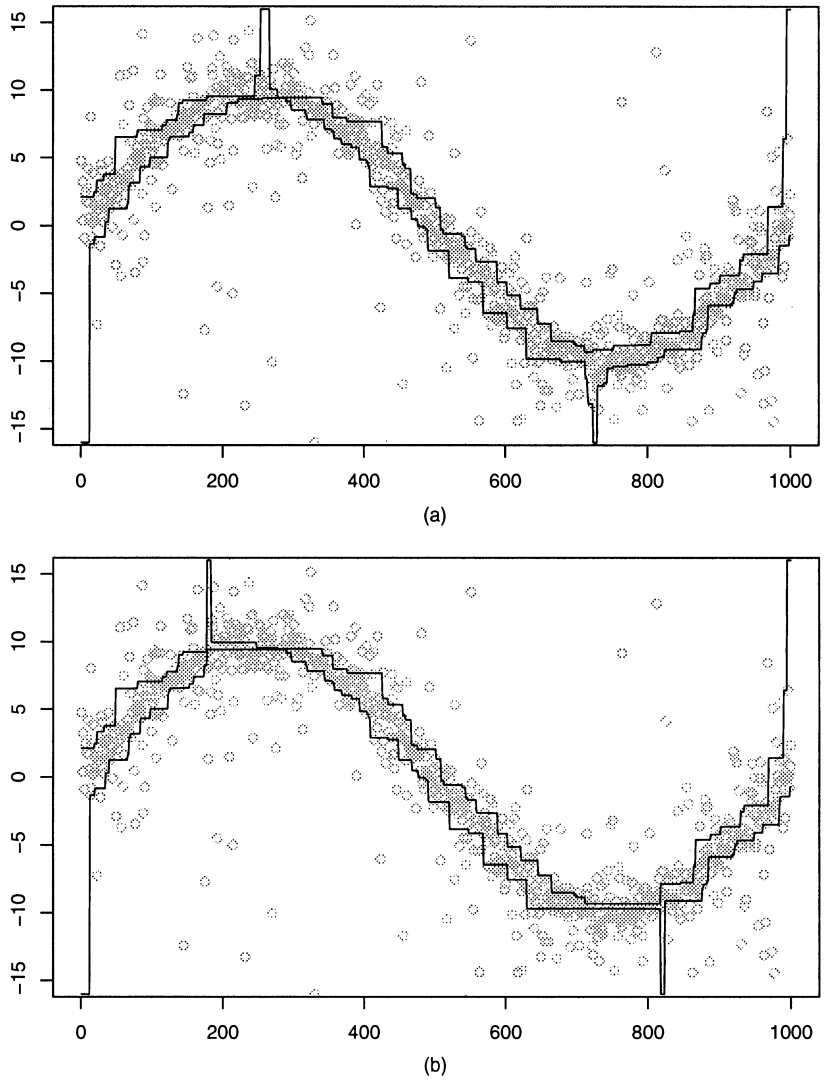

FIG. 5. *The upper panel shows bounds with run length* $\rho_n = 6$, *a local maximum near* 259 *and a local minimum near* 726. *The lower panel shows bounds with run length* $\rho_n = 6$, *a local maximum near* 180 *and a local minimum near* 820.

Given any points t_i^c in the intervals $[t_i^l, t_i^r]$ there exists a function satisfying the run condition with exactly k_n local extremes which are attained at points differing from the prescribed t_i^c by at most the run length. Different choices lead to different functions. This can be interpreted as an attempt to visualize the variability of the set of adequate functions. One obvious choice for the points t_i^c is the centers of the intervals. The top panel of Figure 5 shows the result of doing this for the Cauchy sine data using a run length

of $\rho_n = 6$. The local extreme points are required to be near the centers of the intervals $[178, 340]$ and $[630, 823]$, namely 259 and 726. The lower panel of Figure 5 shows the bounds when the local extremes are required to be near 180 (maximum) and 820 (minimum).

If a specific approximating function for the data is required, there are several alternatives. One method is to calculate the bounds either with or without minimizing the run length and with or without specifying the positions of the local extremes. From this a specific function which satisfies the run condition and lies between these bounds can be obtained from the algorithm described above. The upper panel of Figure 6 shows this function together with the underlying sine curve. It is based on the bounds shown in the upper panel of Figure 5. Although bounds are constructed using the run criterion the statistician is not obliged to consider only functions which satisfy it. The simplest manner of obtaining a function is to take the average of the lower and upper bounds. The lower panel of Figure 6 shows the average of the bounds of the upper panel of Figure 5 together with the approximating sine curve.

That no nontrivial bounds are available at the local extreme values or at the start and finish of the data set is related to the ability of the run method to withstand outliers. We analyze the situation for a block of outliers. If the observation before the block of outliers lies below the upper bound then the method can withstand a block of large positive outliers of length equal to the run length. If, however, the last s observations before the block of outliers lie above the upper bound then only a block of length $\rho_n - s$ of arbitrarily large positive outliers can be tolerated.

Similar considerations apply to the lower bounds. Figure 7 shows Gaussian noise with 9 outliers in the center. The upper panel shows that they are ignored if the run length is $\rho_n = 10$ (upper panel) but detected if the run length is $\rho_n = 9$ (lower panel). The reason is that the observation just before the block of outliers lies above the upper bound.

2.2. *Behavior on test beds.* We consider a data set generated on a test bed of the form (1.8). Let K_n^α denote the number of local extreme values determined by the run procedure using the α-quantile of the length of the longest run (2.8). The following theorem holds.

THEOREM 2. *Consider data generated on the test bed* (1.8) *where the function f has k local extremes and the errors $\varepsilon(t)$ are independently distributed with* $\mathbf{P}(\varepsilon(t) < 0) = \mathbf{P}(\varepsilon(t) > 0) = \frac{1}{2}$. *Then*

$$\mathbf{P}(k \geq K_n^\alpha) \geq \alpha.$$

The residuals $\varepsilon(t)$ fulfil the run condition with probability at least α and from Theorem 1 it follows that at least K_n^α local extreme values are required if the run length of the residuals is not to exceed $qu(n, \alpha, R_n)$.

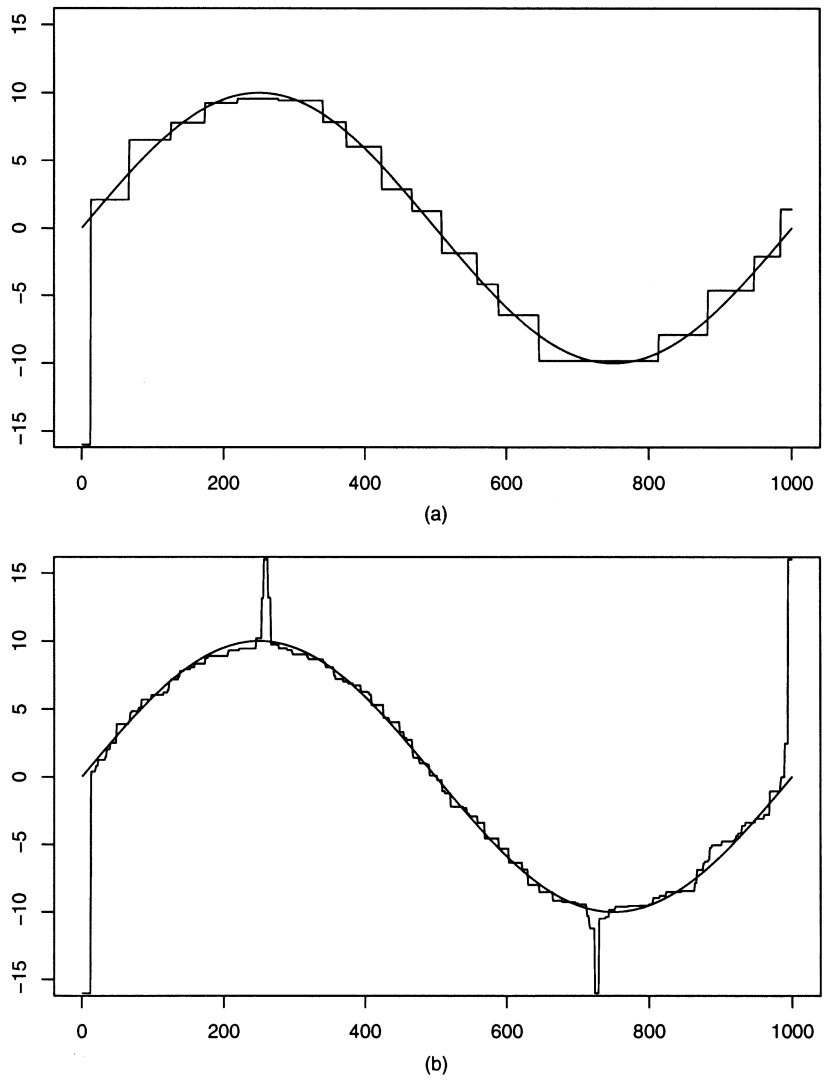

FIG. 6. *The upper panel shows a function satisfying the run condition with* $\rho_n = 6$ *and lying between the bounds of the upper panel of Figure* 5. *The lower panel shows the average of the bounds of the upper panel of Figure* 5. *The underlying sine curve is shown in both panels.*

As mentioned in Section 1.2 this form of result is best possible for fixed n. If however the sample size increases and the design points $t_i,\ 1 \le i \le n$, become dense in $[0, 1]$ then the procedure will determine the correct number of local extremes for sufficiently large n. The precise statement is contained in the next theorem where we assume $t_i = \frac{i}{n},\ 1 \le i \le n$. The theorem shows that f can be consistently estimated and gives a rate of convergence. As the rate of convergence applies to the bounds it holds for any function f_n lying

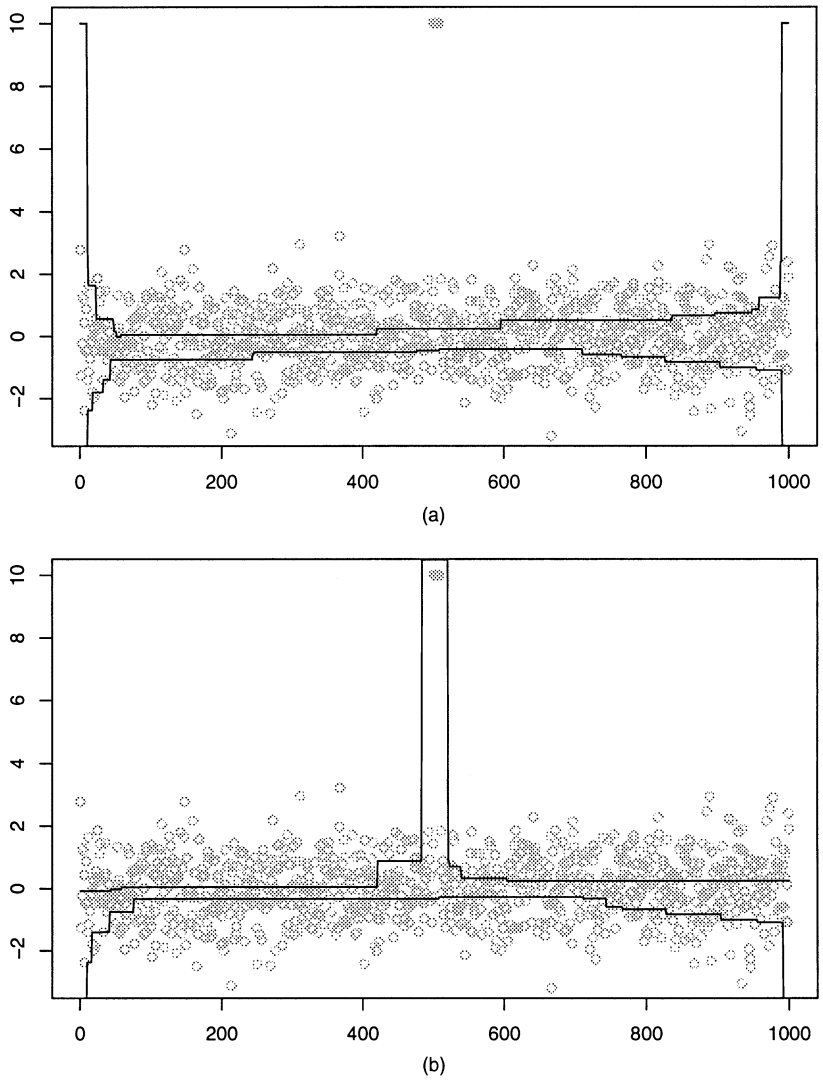

FIG. 7. *The upper panel shows the bounds for* $n = 1000$ *and run length* $\rho_n = 10$ *with Gaussian noise and a block of* 9 *outliers in the center. The lower panel shows the effect of reducing the run length to* $\rho_n = 9$.

within the bounds as long as it has the same monotonicity behavior as the bounds. It is not necessary for the function to give rise to residuals which satisfy the run condition. In particular the rate of convergence holds for the function defined to be the mean of the bounds. The intervals which contain the local extremes will be denoted by $I_i^e(n, \alpha),\ i = 1, \ldots, K_n^\alpha$ and their midpoints by $t_i^e(n, \alpha),\ i = 1, \ldots, K_n^\alpha$. The following theorem holds.

THEOREM 3. *Consider the test bed* (1.8) *where:*

(i) f *has a bounded continuous first derivative* $f^{(1)}$ *and exactly* k *local extreme values at the points* $0 < t_1^e < \cdots < t_k^e < 1$.

(ii) $f^{(1)}(t) = 0$ *only for* $t \in \{t_1^e, \ldots, t_k^e\}$.

(iii) *The* $\varepsilon(t)$ *are independently and identically distributed, have median zero and a continuously differentiable distribution function in a neighbourhood of zero.*

Then the following hold:

(a) *For all* $\delta > 0$,

$$\liminf_{n\to\infty} \mathbf{P}\Bigg(\{K_n^\alpha = k\} \cap \left\{\max_{1\le i\le k} |I_i^e(n,\alpha)| \le \delta\right\} \cap \left\{\max_{1\le i\le k} |t_i^e(n,\alpha) - t_i^e| \le \delta\right\}\Bigg) \ge \alpha.$$

(b) *For a sequence of functions* f_n *within the bounds and with the same monotonicity behavior as the bounds, the locations of the extreme values of the* f_n *converge in probability to those of* f.

(c) *There exists a constant* $b > 0$ *such that for any sequence of functions* f_n *as in* (b),

$$\lim_{n\to\infty} \mathbf{P}\left(\sup_{\{t:|f^{(1)}(t)|\ge\delta\}} |f(t) - f_n(t)| \le b\frac{\log\log n}{\log\ n}\right) = 1.$$

for all $\delta > 0$.

(d) *There exists a constant* $b > 0$ *such that*

$$\lim_{n\to\infty} \mathbf{P}\left(\inf_{\{t:|f^{(1)}(t)|\ge\delta\}} u_n(t) - f(t) \ge b\frac{\log\log n}{\log\ n}\right) = 1$$

and

$$\lim_{n\to\infty} \mathbf{P}\left(\inf_{\{t:|f^{(1)}(t)|\ge\delta\}} f(t) - l_n(t) \ge b\frac{\log\log n}{\log n}\right) = 1$$

for all $\delta > 0$.

3. The taut-string-multiresolution method.

3.1. *Multiresolution and local averages.* The multiresolution problem of Section 1.1 is based on the white noise approximation defined by (1.5), (1.6) and (1.7). If the sample size n is not a power of 2 then the interval $[(k2^j+1)/n, (k+1)2^j/n]$ of (1.5) can be replaced by $[(k2^j+1)/n, \min\{(k+1)2^j/n, 1\}]$. This is legitimate because of the subadditivity of probability measures. An alternative is to use the techniques of Kovac and Silverman (2000).

The $w_{j,k}$ can be replaced by discrete wavelet coefficients or, more generally, coefficients generated by any multiresolution scheme. Prior knowledge of

the shape of the peaks to be detected can be incorporated into the shape of the multiresolution functions. The advantages of multiresolution schemes are that they work well in practice, that they result in almost optimal rates of convergence on appropriate test beds and that the calculations are of order n. The bound (1.6) can be replaced by the more general bound

$$\sigma_n \sqrt{\tau \log n} \tag{3.1}$$

with σ_n again given by (1.7). The idea behind the bound (3.1) is that for a sequence of random variables $Z_1, Z_2 \cdots$ with a common sub-Gaussian distribution (see Theorem 6 below),

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i=1,\ldots,n} |Z_i| < \sqrt{\tau \log n} \right) = 1$$

for some $\tau > 0$. This follows from the tail estimate

$$\mathbb{P}(|Z_i| \geq x) \leq c \exp(-x^2/\tau) \tag{3.2}$$

for some $\tau > 0$ which continues to hold for sums $Z = \frac{1}{\sqrt{n}} \sum_1^n Z_i$ of i.i.d. random variables. The multiresolution coefficients are of this form and so we may use a threshold of the form (3.1).

3.2. *Strings and bounds.* Although the multiresolution problem of Section 1.1 is well posed the difficulty is that, in contrast to the run problem, there is no immediate connection between the number of local extremes on the one hand and the multiresolution definition of white noise approximation on the other. To overcome this, we use an unrelated method, the taut string, to produce candidate functions f_n^k with k local extreme values and then to take the smallest k for which the residuals

$$r_n^k\left(\frac{i}{n}\right) = y\left(\frac{i}{n}\right) - f_n^k\left(\frac{i}{n}\right)$$

approximate white noise. The success of this approach depends entirely on the efficacy of the taut string method to provide such candidate functions.

The left panel of Figure 8 shows a small simulated data set $y(\frac{i}{n})$, $i = 1, \ldots, n$: the right panel the integrated process y_n° defined by

$$y_n^\circ\left(\frac{j}{n}\right) = \frac{1}{n} \sum_{i=1}^{j} y\left(\frac{i}{n}\right), \qquad j = 0, \ldots, n. \tag{3.3}$$

Integrated or summed processes will always be distinguished by a superscript $^\circ$. Consider the lower l_n and upper bound u_n for y_n° defined by

$$l_n = y_n^\circ - \frac{C}{\sqrt{n}} \quad \text{(lower bound)}, \tag{3.4}$$

$$u_n = y_n^\circ + \frac{C}{\sqrt{n}} \quad \text{(upper bound)} \tag{3.5}$$

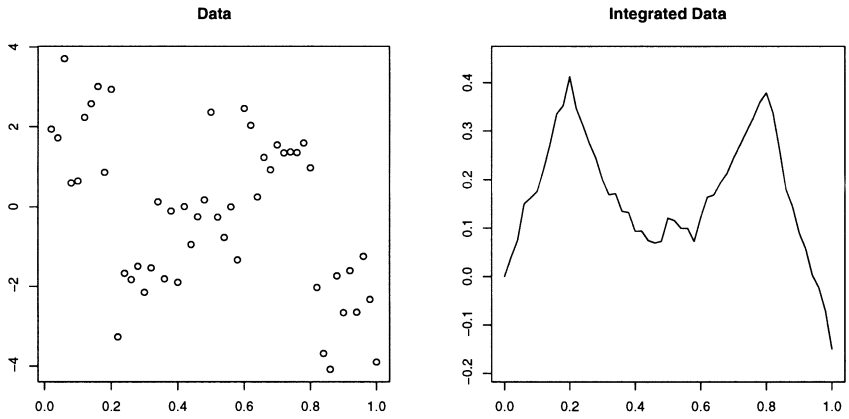

FIG. 8. *Integrating the data. The left panel shows a small simulated data set and the right panel their partial sums.*

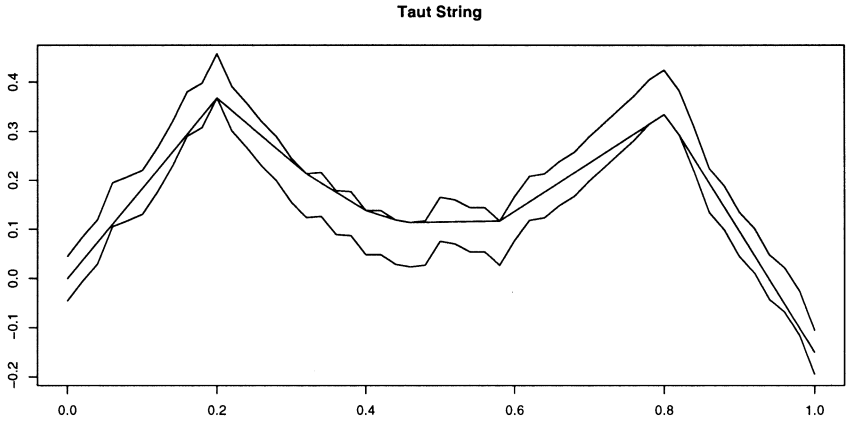

FIG. 9. *A taut string between two bounds.*

for some $C > 0$. Consider a piece of string attached with one end at the point $(0, 0)$ and the other at the point $(1, y_n^\circ(1))$ and constrained to lie between l_n and u_n. Suppose that the string is now pulled until it is taut. It defines a function s_n° on $[0, 1]$. This is shown in Figure 9. We shall use the derivative of the taut string as an approximation to the data as shown in Figure 10.

The string can be defined analytically as that function s_n° on $[0, 1]$ with the smallest length

$$\text{length}(s_n^\circ) = \int_0^1 \sqrt{1 + s_n^{\circ(1)}(t)^2}\, dt,$$

which satisfies

$$s_n^\circ(0) = 0, \quad s_n^\circ(1) = y_n^\circ(1), \quad l_n(t) \le s_n^\circ(t) \le u_n(t), \qquad 0 \le t \le 1. \tag{3.6}$$

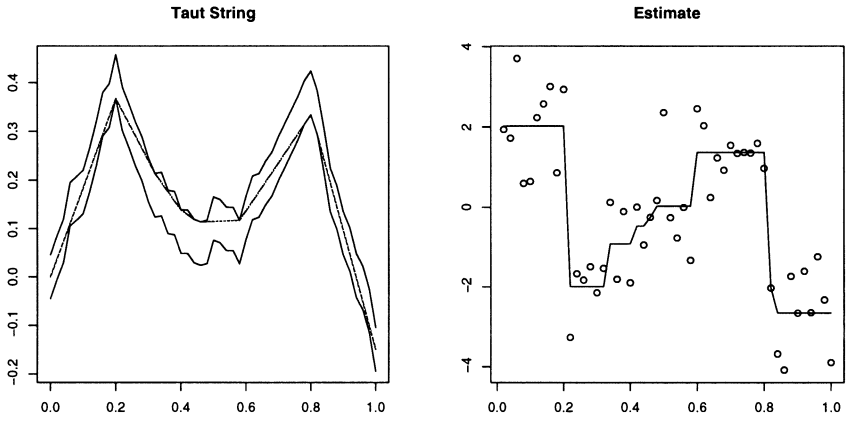

FIG. 10. *The derivative of the taut string of Figure* 9 *together with the data.*

It can be shown that s_n° also has the smallest total variation under the side conditions (3.6). That is, s° minimizes

$$\mathrm{TV}(s_n^\circ) = \int_0^1 |s_n^{\circ(1)}(t)|\, dt, \tag{3.7}$$

subject to (3.6). In practice the restrictions (3.6) only hold at a finite number of points i/n, $i = 0, \ldots, n$. The calculation of the ensuing string can be performed quickly and the complexity of the algorithm is $O(n)$. The algorithm we use is explained in the Appendix. The description of the taut string in the finite case is the following [see also Mammen and van de Geer (1997)]. The points at which the string coincides with either the lower or upper boundary are called knots. Between knots the function is linear. Between two knots x_i and x_j where the string touches the upper bound u_n and between which it does not touch the lower bound l_n the string is the largest convex minorant of the upper bound u_n. Similarly, if the knots are ones where the string touches the lower bound l_n and between which it does not touch the upper bound u_n then it is the smallest concave majorant of the lower bound l_n. At points where the string switches from the upper bound u_n to the lower bound l_n the derivative $s_n = s_n^{\circ(1)}$ has a local maximum. At points where the string switches from the lower bound l_n to the upper bound u_n the derivative s_n has a local minimum.

It is clear from the description of the taut string that its derivative s_n has the smallest number of local extremes among all functions g_n whose integral $g_n^\circ(x) = \int_0^x g_n(t)\, dt$ satisfies the conditions (3.6). It is precisely this property which makes the taut string so efficacious in controlling the number of local extremes. It is possible that the chosen initial and final conditions of (3.6) may themselves cause additional local extremes. This can be overcome by attatching the ends of string either to the upper or to the lower boundary at the start and at the end. Those of the four strings whose derivative has the smallest of local extremes minimizes the number of local extremes of all functions

whose integral lies within the bounds. Dümbgen has pointed out to us that the effect of the initial conditions disappears in the limit. The advantage of the string defined by (3.6) is that end effects are reduced. We therefore use it in the examples, although the theorems are stated in terms of the string which minimizes the number of local extremes.

Bounds of the form (3.4) and (3.5) will be referred to as a tube with center y_n° and radius $\frac{C}{\sqrt{n}}$: it will be denoted by $T\big(y_n^\circ, \frac{C}{\sqrt{n}}\big)$,

$$T\left(y_n^\circ, \frac{C}{\sqrt{n}}\right) = \{g_n\colon l_n \le g \le u_n\}. \tag{3.8}$$

Other bounds derived from local squeezing will be used later. As already mentioned, the derivative $s_n(C)$ of the taut string $s_n^\circ(C)$ is used as a candidate regression function. As $s_n^\circ(C)$ is piecewise linear, $s_n(C)$ is piecewise constant. It is perhaps worth mentioning some properties of the derivative $s_n(C)$. Consider two knots of the string at the points $\frac{i}{n} < \frac{j}{n}$ where the taut string touches the upper bound u_n and are such that the string does not touch the lower bound between the two knots. The derivative $s_n(C)$ is then the least squares monotone increasing approximation to the y-values between the knots. In other words the $s_n(C)(\frac{m}{n})$, $m = i, \ldots, j$ solve

$$\text{minimize} \quad \sum_i^j \left(y\left(\frac{m}{n}\right) - a_m\right)^2 \quad \text{subject to} \quad a_i \le \cdots \le a_j.$$

[see Barlow, Bartholomew, Bremner and Brunk (1972)]. On sections where the string is the smallest concave majorant of the lower bound l_n the derivative s_n of the string solves the corresponding problem where the values are nonincreasing. This implies the following. Between two successive knots either both on the upper bound or both on the lower bound the derivative is the mean of the observations between the knots. Between knots defining a local minimum or maximum the derivative is again the mean of the observations but increased or decreased, respectively, by the diameter of the tube at that point [see also Proposition 8 of Mammen and van de Geer (1997)]. Compared to the data, local maxima will tend to be too small and local minima too large.

3.3. *Previous work.* The first use of the taut string method in statistics would seem to be due to Barlow, Bartholomew, Bremner and Brunk (1972). Hartigan and Hartigan (1985) were the first to use it in connection with modality, it being an integral part of their dip test for unimodality. It was explicitly used by Davies (1995) to construct k-modal densities. Mammen and van de Geer (1997) extended its use to the nonparametric regression problem but do not mention the connection with the number of local extreme values.

3.4. *Behavior on test beds.* The next three theorems are concerned with the asymptotic behavior of the taut string on the test bed (1.8). In particular it is shown to have an optimal rate of convergence away from the local extremes. The first theorem corresponds to Theorem 2 for the run procedure.

We denote the number of local extremes of the derivative of the taut string in the supremum tube with upper and lower bounds given by (3.4) and (3.5) by K_n^C. Theorems 4, 5 and 6 refer to the taut string which minimizes the number of local extremes. Dümgen has proved that the theorems continue to hold with suitable modifications, for example, replacing the Brownian motion on $[0, 1]$ of Theorem 4 by the Brownian bridge on $[0, 1]$, if the initial conditions (3.6) are imposed.

THEOREM 4. *Consider the test bed* (1.8) *but where the errors* $\varepsilon(t)$ *are independently and identically distributed with mean* 0 *and finite variance* σ^2. *Then*

$$\lim_{n\to\infty} \mathbf{P}(k \geq K_n^C) = H\left(\frac{C}{\sigma}\right),$$

where H *denotes the distribution of* $\sup_{0\leq t\leq 1} |W(t)|$ *where* W *denotes a standard Wiener process.*

It is worth pointing out that the rate of convergence in Theorem 4 does not depend on the function f of the test bed but only on the rate of convergence of the partial sums

$$\sqrt{n}\varepsilon_n^\circ(t) = \frac{1}{\sqrt{n}} \sum_1^{[nt]} \varepsilon\left(\frac{j}{n}\right)$$

to a Wiener process. To make the theorem applicable without knowledge of the variance σ^2 an estimate of it is required. We use (1.7).

The taut string based on the radius $C/\sqrt{n}$ will be denoted by $S_n^\circ = S_n^\circ(C)$ with derivative $S_n = S_n(C)$. We write $I_i^e(n, C),\ 1 \leq i \leq K_n^C$, for the intervals where S_n attains it local extreme values and denote the midpoints of these intervals by $\tau_n^e(n, C),\ 1 \leq i \leq K_n^C$. The next theorem corresponds to Theorem 3(a) for the run procedure.

THEOREM 5. *Consider the test bed* (1.8) *where* f *satisfies the conditions of Theorem* 3 *and where the errors* $\varepsilon(t)$ *are identically and independently distributed with mean* 0 *and finite variance* σ^2. *Then for all* $\delta > 0$,

$$\lim_{C\to\infty} \lim_{n\to\infty} \mathbf{P}\Bigg(\{K_n^C = k\} \cap \left\{\max_{1\leq i\leq k} |I_i^e(n, C)| \leq \delta\right\} \cap \left\{\max_{1\leq i\leq k} |\tau_i^e(n, C) - t_i^e| \leq \delta\right\}\Bigg) = 1.$$

We note that the rate of convergence to the correct number of local extremes depends much more heavily on the function f than the rate of convergence in Theorem 4. It is for this reason that we state Theorem 5 separately.

In the following the length of an interval I will be denoted by $|I|$.

THEOREM 6. *Consider the test bed* (1.8) *where* f *satisfies the conditions of Theorem* 3 *and additionally:*

(i) f *has a bounded second derivative* $f^{(2)}$ *which is nonzero at the* k *local extremes.*

(ii) *The errors* $\varepsilon(t)$ *are independently and identically distributed sub-Gaussian random variables, that is,* $\mathbf{E}\exp(\lambda\varepsilon(t)) < \exp(c\lambda^2)$ *for all* λ *for some* $c > 0$.

Then

(a) $\lim_{C\to\infty}\lim_{n\to\infty}\mathbf{P}(t_i^e \in I_i^e(n,C),\ 1 \le i \le k) = 1.$

(b) $\lim_{C\to\infty}\lim_{n\to\infty}\mathbf{P}\left(1-\delta \le \dfrac{n^{1/6}|f^{(2)}(t_i^e)|^{1/3}|I_i^e(n,C)|}{(24C)^{1/3}} \le 1+\delta\right) = 1 \quad \textit{for all } \delta > 0.$

(c) *Let* x_i *and* x_{i+1} *denote successive knots which are either both on the upper or both on the lower bound. Then there exists an* $A > 0$ *such that*

$$\lim_{C\to\infty}\lim_{n\to\infty}\mathbf{P}\left(\max_i (x_{i+1}-x_i)|f^{(1)}(x_i)|^{2/3} \le A\left(\frac{\log n}{n}\right)^{1/3}\right) = 1.$$

(d) *There exists an* $A > 0$ *such that*

$$\lim_{C\to\infty}\lim_{n\to\infty}\mathbf{P}\left(\sup_{t\in A_n}\frac{|f(t)-S_n(t)|}{|f^{(1)}(t)|^{1/3}} \le A\left(\frac{\log n}{n}\right)^{1/3}\right) = 1,$$

where $A_n = \left[A\left(\frac{\log n}{n}\right)^{1/3}, 1 - A\left(\frac{\log n}{n}\right)^{1/3}\right]\setminus\cup_i^n I_i^e(n,C)$.

(e) *There exists a constant* $A > 0$ *such that*

$$\lim_{C\to\infty}\lim_{n\to\infty}\mathbf{P}\left(\sup_{t\in B_n}\frac{|f(t)-S_n(t)|}{|f^{(2)}(t)|^{1/3}} \le AC^{2/3}n^{-1/3}\right) = 1,$$

where $B_n = \cup_1^k I_i^e(n,C)$.

The definition of sub-Gaussian given above implies that the mean is zero [see Kahane (1968), page 62, Exercise 10]. We note that (d) implies that S_n is optimal at points t with $|f^{(1)}(t)| > \delta > 0$. At points t with $f^{(1)}(t) = 0$ its rate of convergence is $n^{-1/3}$ compared with the optimal rate of $n^{-2/5}$ [Leurgans (1982)]. As the length of the interval is $n^{-1/6}$ as against the optimal $n^{-1/5}$, this cannot be alleviated by replacing $S_n(t)$ for $t \in \cup_1^k I_i^e(n,C)$ by the mean of the $y-$values. Nevertheless, the change of boundaries at local extremes does alter the behavior of the multiresolution coefficients described in the following section.

Donoho, Johnstone, Kerkyacharian and Picard (1995) consider four sets of simulated data, the Blocks data, the Bumps data, the Heavisine data and the Doppler data. They were also analyzed by Fan and Gijbels (1995), (1996). Figure 11 shows the taut string method to the Blocks signal contaminated with Gaussian white noise. The tube radius is $1.149\sigma/\sqrt{n}$. The choice $C = 1.149$ corresponds to the 0.5 quantile of the maximum of the absolute value of a

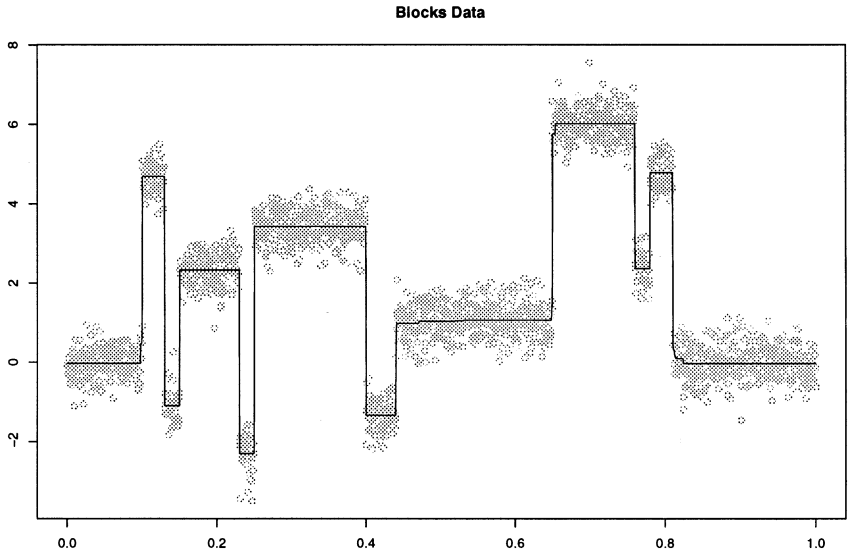

FIG. 11. *The Blocks signal. The signal was contaminated with white noise and afterwards the taut string approximation calculated using the* $C/\sqrt{n}$ *tube with* $C = 1.149$.

standard Brownian motion on $[0, 1]$ [see Freedman (1971), (34) Proposition, page 27].

3.5. *Strings and multiresolution.* As stated above, the derivative of the taut string tends to be too small at local maxima and too large at local minima. A simple analysis shows that this effect alone will cause some of the multiresolution coefficients of the residuals to exceed the bound as $n \to \infty$ regardless of the value of τ in (3.1). The problem may be rectified to some extent by the simple expedient of replacing S_n by $\widetilde{S}_n$ where $\widetilde{S}_n$ is constant between knots where it is equal to the mean of the $y-$values between the knots. The integrated $\widetilde{S}_n$ will be denoted by $\widetilde{S}_n^\circ$. It may be seen that $\widetilde{S}_n^\circ$ coincides with Y_n° at the knots and is linear in between. This alteration does not effect the number of local extremes but it improves the rate of convergence at local extremes. The behavior of the multiresolution coefficients for candidate functions based on $\widetilde{S}_n$ is covered by the next theorem.

THEOREM 7. *Suppose the assumptions of Theorem* 6 *hold and consider the multiresolution coefficients of the residuals of* $\widetilde{S}_n$, *that is,*

$$W_{i,j} = 2^{-i/2} \sum_{m=j2^i+1}^{(j+1)2^i} \left(Y\left(\frac{m}{n}\right) - \widetilde{S}_n\left(\frac{m}{n}\right) \right)$$

with supports $[j2^i + 1, (j+1)2^i]$. *Then for each constant* $A > 0$ *there exists a* $\tau > 0$ *such that the following hold:*

(a) *All multiresolution coefficients* $W_{i,j}$ *taken from intervals whose support is of length at most* $A(\log n/n)^{1/3}$ *are eventually smaller than the bound* (3.1).

(b) *All multiresolution coefficients* $W_{i,j}$ *the endpoints of whose supports do not lie in* $\cup_1^k I_i^e(n, C)$ *are eventually smaller than the bound* (3.1).

(c) *At each local extremum* t_i^e *of* f *there exists an interval* $J_i^e(n, C) \supset I_i^e(n, C)$ *of length at most* $O(\frac{1}{\log n})$ *for which the following holds. All multiresolution coefficients* $W_{i,j}$ *with one endpoint in* $I_i^e(n, C)$ *for some* i *and the other outside of the corresponding* $J_i^e(n, C)$ *are eventually smaller than the bound* (3.1).

(d) *For each interval* $I_i^e(n, C)$ *of* (b) *there exists a multiresolution coefficient* $W_{i,j}$ *with support in* $I_i^e(n, C)$ *and whose absolute value eventually exceeds the threshold* (3.1) *whatever the value of* τ.

Theorem 7 shows that for an appropriate choice of τ in (3.1) all multiresolution coefficients will eventually be below the threshold (3.1) apart from some in shrinking intervals which contain the local extreme points of the function f. Furthermore there are multiresolution coefficients which will eventually exceed the bound (3.1) whatever the choice of τ. In other words, the suboptimal rate of convergence at local extremes will be detected by the multiresolution analysis of the residuals. Figure 12 shows the application of the taut string method to the Bumps data of Donoho, Johnstone, Kerkyachariar and Picard (1995). The left panel shows the intervals where the multiresolution coefficients of the residuals exceed the threshold (3.1). The right panel shows the effect of squeezing globally until all multiresolution coefficients are below the threshold. It results in several spurious local extreme values. In the next section we describe how local squeezing may be used to eliminate the spurious local extremes.

3.6. *Local squeezing and multiresolution.* The results of the last section show that although the taut string is locally adaptive [Mammen and van de Geer (1997)] it is not sufficiently so. Furthermore, as the Bumps data in Figure 12 show, the problem cannot be solved by decreasing the radius of the tube globally. The calculations used in the proof of Theorem 6 show that if the radius is of order $n^{-3/5}$ instead of $n^{-1/2}$ then the length of the interval which contains a local extreme value is of order $(\log n/n)^{1/5}$ and the rate of convergence is of order $(\log n/n)^{2/5}$. This is almost optimal and calculations show that all multiresolution coefficients will now be smaller than the bound (3.1). However a global radius of order $n^{-3/5}$ for the tube will eventually give rise to spurious local extremes. This may be countered by decreasing the radius of the tube only in an $n^{-1/5}$-neighborhood of those local extremes found with the tube radius $Cn^{-1/2}$. As it stands this is not feasible as the taut string with radius $n^{-1/2}$ locates the local extremes only with an accuracy of order $n^{-1/6}$. One possibility is to decrease the radius of the tube on the intervals of

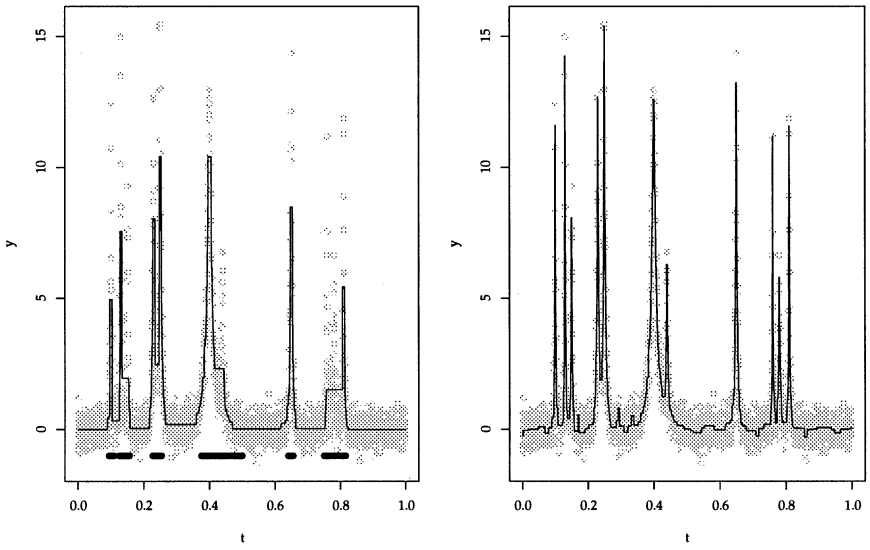

FIG. 12. *These figures show the application of the taut string methods to the Bumps signal. In the left panel the tube radius is* $1.149\sigma/\sqrt{n}$ *and the lines at the bottom indicate regions where multiresolution coefficients are too large. The right panel shows the global squeezing procedure where the tube is squeezed until all multiresolution coefficients are below the threshold value.*

order $n^{-1/6}$ which contain the local extreme points. For sub-Gaussian random variables with variance σ^2 we have

$$\mathbf{P}\left(\max_{l\le i\le l+m-1}\frac{1}{\sqrt{m}}\left|\sum_{j=l}^{i}\varepsilon\left(\frac{j}{m}\right)\right|\ge\sigma\lambda\right)\le a\exp(-c\lambda^2)$$

for all $\lambda>0$ for some constants a and c. Translating this inequality into one for the integrated process Y_n° gives the following:

$$\mathbf{P}\left(\max_{l\le i\le l+m-1}\left|Y_n^\circ\left(\frac{i}{n}\right)+\Delta_n-f^\circ\left(\frac{i}{n}\right)\right|\le\gamma(n,m)\right)\ge 1-a\,\exp(-c\lambda^2), \tag{3.9}$$

where

$$\Delta_n=f^\circ\left(\frac{l}{n}\right)-Y_n^\circ\left(\frac{l}{n}\right) \tag{3.10}$$

and

$$\gamma(n,m)=\lambda\frac{\sqrt{m}}{n}. \tag{3.11}$$

The inequality (3.9) implies that with high probability depending on λ the integrated process f° lies in the locally squeezed tube

$$T(Y_n^\circ+\Delta,\gamma,I)=\{g\colon\max_{t\in I}|Y_n^\circ(t)+\Delta-g(t)|\le\gamma\} \tag{3.12}$$

with $\Delta=\Delta_n$ given by (3.10), $\gamma=\gamma(n,m)$ given by (3.11) and $I=I(l,m,n)=\left[\frac{l}{n},\frac{l+m-1}{n}\right]$. In other words, local squeezing on the interval I is accomplished by

shifting the integrated process Y_n° by an amount Δ on I and then using the modified lower and upper bounds

$$l_n(t, I) = Y_n^\circ(t) + \Delta - \gamma \quad \text{and} \quad u_n(t, I) = Y_n^\circ(t) + \Delta + \gamma$$

for $t \in I$. The size of the shift Δ is not known in advance; (3.9) and (3.10) say that for some shift, namely that given by (3.10), the integrated process f° lies in the locally squeezed tube with high probability. This means that Δ should be chosen to minimize the number of local extreme values of the derivative of the taut string through the modified tube. This may be accomplished by calculating the taut string over a grid of the possible values of Δ.

The effect of local squeezing on an interval where the derivative of the taut string has a local extreme value may be analyzed by setting $m = m_n = cn^{5/6}$ [Theorem 6(b)] in (3.9). This results in $\gamma_n = \lambda\sqrt{c}n^{-7/12}$. The radius of the tube is now $n^{-7/12}$ compared with the optimal radius of $n^{-3/5}$. The proportional difference is $n^{-1/60}$. Repeating this procedure will always improve the asymptotic rate of convergence but no finite number of steps will result in the optimal radius of $n^{-3/5}$. Suppose however we know the points where f takes on its local extreme values. Under these circumstances it is possible to squeeze the tube locally at the local extremes over intervals which contain $cn^{4/5}$ points. The result is described by (3.9) with $\gamma_n = \lambda\sqrt{c}n^{-3/5}$. The taut string through the modified tube will behave asymptotically as before but will have an improved and indeed optimal rate of convergence at the local extreme values of f. If this is taken into account then the proof of Theorem 7 shows that *all* multiresolution coefficients will now be set to zero. The problem is that we cannot at the moment locate the local extremes of f with the required accuracy. We now show how this can be done, at least in principle, using a multiresolution analysis of the residuals.

We consider the taut string using a tube of radius $Cn^{-1/2}$ and the modified derivative $\widetilde{S}_n$. As shown above, this will, in the limit, result in the correct number of local extremes and all the extreme points will lie in the corresponding intervals of $\widetilde{S}_n$ whose lengths will be of order $n^{-1/6}$. Furthermore all multiresolution coefficients will be smaller than the bound (3.1) apart from some whose support is strictly contained in shrinking neighborhoods of the local extreme points. We now gradually squeeze the tube locally at the intervals where $\widetilde{S}_n$ has local extreme values. This is continued until all the multiresolution coefficients of the residuals drop below the threshold (3.1).

We show first that this process will not terminate too soon, that is, before the optimal rate of convergence is obtained. To demonstrate this we consider a function g defined on $[0, 1]$ and a sequence of piecewise constant functions g_n used to approximate g. The change of notation is to emphasize that the following arguments relate only to the approximation of g by the piecewise constant functions g_n. We assume the following:

1. g has a finite number of local extreme values in $(0, 1)$.
2. g has a first derivative $g^{(1)}$ which is zero only at the local extremes.

3. g has a continuous second derivative $g^{(2)}$ which is nonzero at the local extremes.
4. g_n is piecewise constant and with the same monotonicity behavior as g on each interval.
5. If g has a local extreme at t_e then g_n is constant on an interval which contains t_e and is of length at least $c\left(\frac{\log n}{n}\right)^{1/5}$ for some constant $c > 0$.

Condition 4 means that if g is monotone on an interval I then so is g_n and with the same form of monotonicity. The converse need not hold.

Suppose now that the multiresolution coefficients of the differences $g - g_n$ are all smaller than the bound (3.1). We consider what implications this has for the maximal deviation of g_n from g. Consider first a point t with $d_n = |g(t) - g_n(t)|$ and where g and g_n are nondecreasing with $g^{(1)}(t) > 0$. We write $d_n = |g(t) - g_n(t)|$ and suppose that $g(t) > g_n(t)$. The other case is treated similarly. Let the interval I_n of length λ_n be the support of some multiresolution coefficient with $I_n \subset [t - d_n/g^{(1)}(t), t]$. The multiresolution coefficient is

$$\frac{1}{\sqrt{\lambda_n}} \int_{I_n} (g(u) - g_n(u))\, du \geq \frac{1}{2} \lambda_n^{3/2} g^{(1)}(t).$$

If this coefficient does not exceed the threshold (3.1) then

$$\frac{1}{2} \lambda_n^{3/2} g^{(1)}(t) \leq \sqrt{\frac{\tau \log n}{n}},$$

which gives

$$\lambda_n \leq \left(\frac{4\tau \log n}{n g^{(1)}(t)^2} \right)^{1/3}.$$

As there exists some multiresolution coefficient with $\lambda_n \geq d_n/(2g^{(1)}(t))$ this implies

$$\frac{d_n}{2g^{(1)}(t)} \leq \left(\frac{4\tau \log n}{n g^{(1)}(t)^2} \right)^{1/3}$$

or

$$d_n \leq 2^{5/3} |g^{(1)}(t)|^{1/3} \left(\frac{\tau \log n}{n} \right)^{1/3}. \tag{3.13}$$

At the local extremes of the function g we can no longer exploit Assumption 4 but use instead Assumption 5. Let g have a local extreme value at t_e and set $d_n = |g(t_e) - g_n(t_e)|$. The function g is locally quadratic at t_e and arguments similar to the ones just used show the following. If none of the multiresolution coefficients with support close to the local extreme point t_e exceed the threshold (3.1) then

$$|d_n| \leq 2 |g^{(2)}(t_e)|^{1/5} \left(\frac{\tau \log n}{n} \right)^{2/5}. \tag{3.14}$$

We now apply the above results to the nonparametric regression problem. The residuals are

$$r_n(t) = Y(t) - f_n(t) = f(t) - f_n(t) + \varepsilon(t).$$

On setting $g = f$ and $g_n = f_n$, we note first that if the $\varepsilon(t)$ are i.i.d. sub-Gaussian then their contribution to the multiresolution coefficients will lie below the threshold for some value of τ. If f_n and f satisfy Assumptions 1 to 5 above and if all the multiresolution coefficients of the residuals r_n do not exceed the threshold, it follows that multiresolution coefficients of the differences $f - f_n$ will also lie below the threshold. This implies that the rate of convergence of the f_n to f will be optimal away from the local extremes in the sense of uniform convergence. At the local extremes themselves the rate of convergence will be at least $\left(\frac{\log n}{n}\right)^{2/5}$ as against the optimal rate of $n^{-2/5}$. Let now $f_n = \widetilde{S}_n$ where $\widetilde{S}_n$ denotes the modified taut string through the tube which is now locally squeezed at the local extremes with $m_n = C(\log n)^{1/5} n^{4/5}$ in (3.6). The proof of Theorem 7 shows that all multiresolution coefficients of the residuals will now be below the threshold. This implies the following. Suppose we start with a tube of radius $C/\sqrt{n}$ and then gradually locally squeeze the tube at the local extremes of $\widetilde{S}_n$. As the lengths of the intervals where $\widetilde{S}_n$ has its local extreme values are initially of order $n^{5/6}$ (Theorem 6) we see that this process will eventually lead to local squeezing over intervals containing $m_n = C(\log n)^{1/5} n^{4/5}$ observations. At this point all multiresolution coefficients will be below the threshold. This implies that if we stop squeezing at the local extremes of $\widetilde{S}_n$ as soon as all multiresolution coefficients drop below the threshold then $g = f$ and $g_n = \widetilde{S}_n$ will satisfy Conditions 1 to 5 above. The resulting rate of uniform convergence will be $\left(\frac{\log n}{n}\right)^{1/3}$ away from the local extreme values of f. At the local extreme values of f the rate of convergence will be $\left(\frac{\log n}{n}\right)^{2/5}$. We have therefore shown that in principle local squeezing based on a multiresolution analysis of the residuals results in almost optimal rates of convergence at the local extreme values. In practice local squeezing must be done by an algorithm and to complete the proof it must be shown that the algorithm attains what is in principle possible. We do not pursue this any further and we simply describe the algorithm we use without a proof that it is asymptotically correct. In practice it works very well.

3.7. *A practical implementation.* Local squeezing is described by (3.12) and requires the specification of the shift Δ, the radius of the tube γ and the interval I. The intervals on which local squeezing is performed are determined by the multiresolution analysis of the residuals. As mentioned in the discussion of local squeezing given above, the choice of shift Δ seems not to be critical and is set to zero. Finally the radius of the tube γ must be specified and this is critical. It is implemented by reducing the radius of the tube locally by a constant factor $\rho, 0 < \rho < 1$. If ρ is small, say $\rho = 0.5$, then the procedure terminates quickly as all multiresolution coefficients soon drop below the threshold. This occasionally leads to too many local extreme values.

If ρ is nearly 1, say $\rho = 0.95$, then the calculations require more time but this choice of ρ has in our experience not led to superfluous local extreme values. The following simple procedure turned out to be very reliable and was used for all the examples.

First, the scale of the noise is calculated using (1.7). The next step is to choose a large initial global tube radius γ_0. A reasonable choice is such that the straight line connecting $(0, 0)$ and $(1, y_n^\circ(1))$ lies inside the tube. An alternative is to use Theorem 4 and use an initial tube radius $\gamma_0 = C/\sqrt{n}$ where C is set to 2.242, the 0.95-quantile of the distribution of the maximum of a Brownian motion. Given the initial radius the taut string is calculated as described in the Appendix.

We use the modified version $\tilde{s}_n$ of the derivative s_n as described just before Theorem 7. If all multiresolution coefficients of the residuals $\tilde{r}_n\left(\frac{i}{n}\right) = y\left(\frac{i}{n}\right) - \tilde{s}_n\left(\frac{i}{n}\right)$ are smaller than the bound (3.1) the algorithm terminates. If not we define local tube radii γ_i^1 $(i = 0, \ldots, n)$ by $\gamma_i^1 = \rho\gamma_0$ if a multiresolution coefficient which depends on y_i or y_{i+1} exceeds the bound (3.1) and by $\gamma_i^1 = \gamma_0$ elsewhere. The tube is squeezed centrally, that is, $\Delta = 0$. The squeezing factor ρ is set to 0.95. A new candidate function is finally defined by the result of the taut string procedure.

The algorithm proceeds by calculating the multiresolution analysis of the new residuals. If again some coefficients are still too large, further squeezing is applied by reducing the tube radius in the relevant intervals by a factor of ρ until eventually all coefficients are below the threshold. As indicated in Section 3.1 the value of τ we use is 2.5.

3.8. *Examples.* The effect of local squeezing on the bounds is shown in Figure 13. Displayed are the final bounds for the Bumps signal of Figure 12 in the left panel and the reconstruction in the right panel. This should be compared with Figure 12.

4. Low power peaks. Most work on nonparametric regression and density problems evaluates procedures in terms of rates of convergence on test beds of the form (1.8). The distribution of the errors $\varepsilon(t)$ is specified and the regression function f is kept fixed while the number n of observations is increased. In this situation there exist optimal rates of convergence [Khas'minski (1978); Ibragimov and Khas'sminski (1980); Stone (1982)]. As shown, the taut string method based on a ball of radius $Cn^{-1/2}$ attains the optimal rate away from the local extremes for functions with a specified number of local extremes. The run method falls well short with a rate of convergence of order $(\log\log n)/\log n$. In spite of this the run-based procedure can give better results than an optimal taut string method based on a tube of constant radius. To investigate this phenomenon we consider a different form of test bed. Let f be a continuous function with k peaks. We consider an interval $[a_n, b_n]$ which does not contain a peak and graft a peak onto the function f. The height of the peak is h_n and

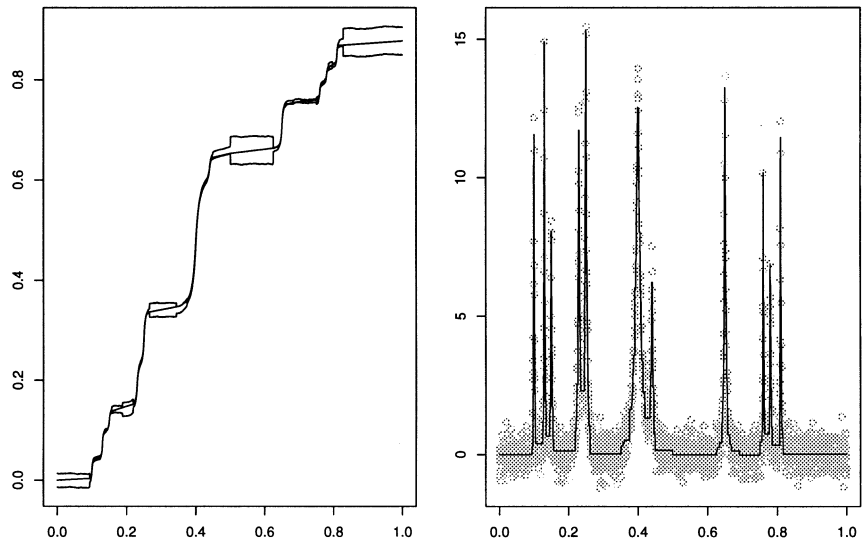

FIG. 13. *Local squeezing and the Bumps data. The left panel shows the final upper and lower bounds together with the taut string. The Bumps data and the derivative of the taut string are plotted in the right panel.*

we denote the new function by f^n. The power of the peak is defined to be

$$P_n = \int_0^1 |f - f^n| = \int_{a_n}^{b_n} |f - f^n|. \tag{4.1}$$

Consider now the asymptotic test bed

$$Y_{n,i} = f^n\left(\frac{i}{n}\right) + \varepsilon_{n,i} \tag{4.2}$$

with $P_n = o(n^{-1/2})$ and $\lim_{n\to\infty} h_n = \infty$. For large C the tube $T(f^{n\circ}, C)$ will contain f° with large probability. As f has k local extremes so will the taut string through the ball. From this it follows that Theorems 5 and 6 will continue to hold for the asymptotic test bed (4.2). In other words the taut string method will fail to identify the peak at $[a_n, b_n]$. Similar considerations apply to kernel estimation. If the optimal global window is used, a low powered peak will not be detected if $P_n = o(n^{-1/5})$.

In the case of the run method on the test bed (4.2) the peak will be detected if

$$\liminf \frac{n(b_n - a_n)}{\log_2 n} > 1.$$

If

$$\liminf \frac{n(b_n - a_n)}{\log_2 n} < 1,$$

it will not be detected. Figure 7 shows this effect.

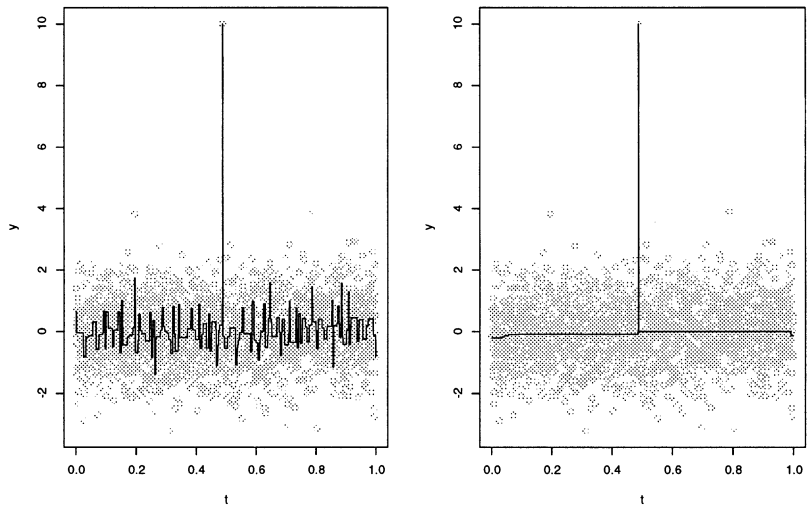

FIG. 14. *A regression function corrupted with Gaussian noise and the reconstructions based on* $n = 2048$ *observations. The left panel shows the effect of globally squeezing until the peak near* 0.5 *is detected. The right panel shows the result of the local squeezing procedure.*

The local squeezing method using local averages will clearly identify the peak of f^n. Indeed the method will pick up a signal confined to one single observation as long as $h_n \geq B\sqrt{\log n}$ where B depends on f and the bound parameter τ. This is demonstrated by Figure 14 where a very low power peak defined by a single observation is detected without introducing any spurious peaks.

5. Examples and comparisons.

5.1. *Artificial data.* Figure 15 shows the results of applying the taut-string-multiresolution procedure to the data sets of Donoho, Johnstone, Kerkyacharian and Picard (1994). The Doppler signal is displayed in the upper left corner. Local squeezing detects the oscillations near the origin very well without introducing spurious extreme values at other positions. The main feature of the Heavisine signal in the upper right panel is the presence of discontinuities near 1/3 and 2/3. The taut string reproduces them, again without introducing spurious extreme values elsewhere. The Blocks data signal is piecewise constant so it is not surprising that the taut string performs very well. Indeed the reconstruction can hardly be distinguished from the original signal. Finally the Bumps data can be compared with Figure 12. Local squeezing has identified all the relevant peaks without introducing spurious ones.

5.2. *Real data.* Figure 16 shows the results of a spectroscopical analysis of a gallstone. The chemists involved informed us that for this particular data

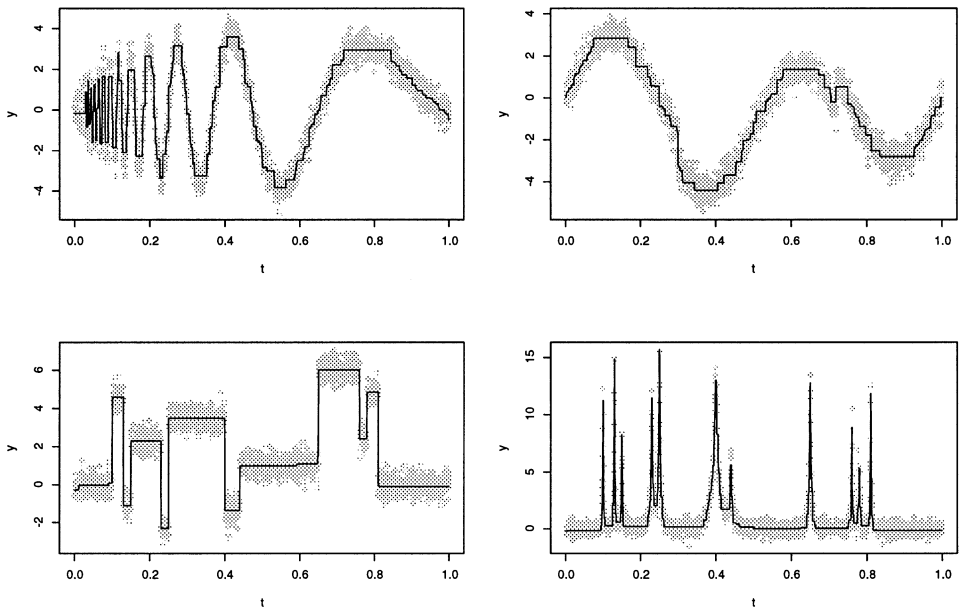

FIG. 15. *Four noisy test signals and their reconstructions using the taut string procedure with local squeezing.*

all the peaks found were real ones and no spurious peaks were introduced. We note that between peaks the reconstruction is almost constant. This data set, although real, is not much different from the Bumps data sets of Figure 13.

5.3. *Comparison with other methods.* Comparison with other methods is always difficult especially when the aims may be different. This paper is concerned with the number of local extremes, only secondarily with rates of convergence and not at all with L^2 errors. Other methods are concerned with smoothness and rates of convergence in spaces of smooth functions. The problem of comparison is made more difficult by the lack of available software. We therefore restrict the comparisons to wavelets for which standard software is available. Even here many options are open, such as the choice of the wavelet and the threshold level. The left panel of Figure 17 shows the block data and the wavelet reconstruction. The Haar wavelet was used, which is presumably the best for this data. Nevertheless, pseudo-Gibbs effects are apparent and lead to 47 local extreme values. The corresponding result for the taut-string-multiresolution method is shown in the right panel of Figure 17. It gives the correct number of local extremes, namely nine.

6. Conclusion. We have introduced two methods for obtaining regression functions while keeping the number of local extremes under control. Each method has advantages and disadvantages. The run method can be calculated quickly in $O(n)$ operations and it has certain desirable robustness properties.

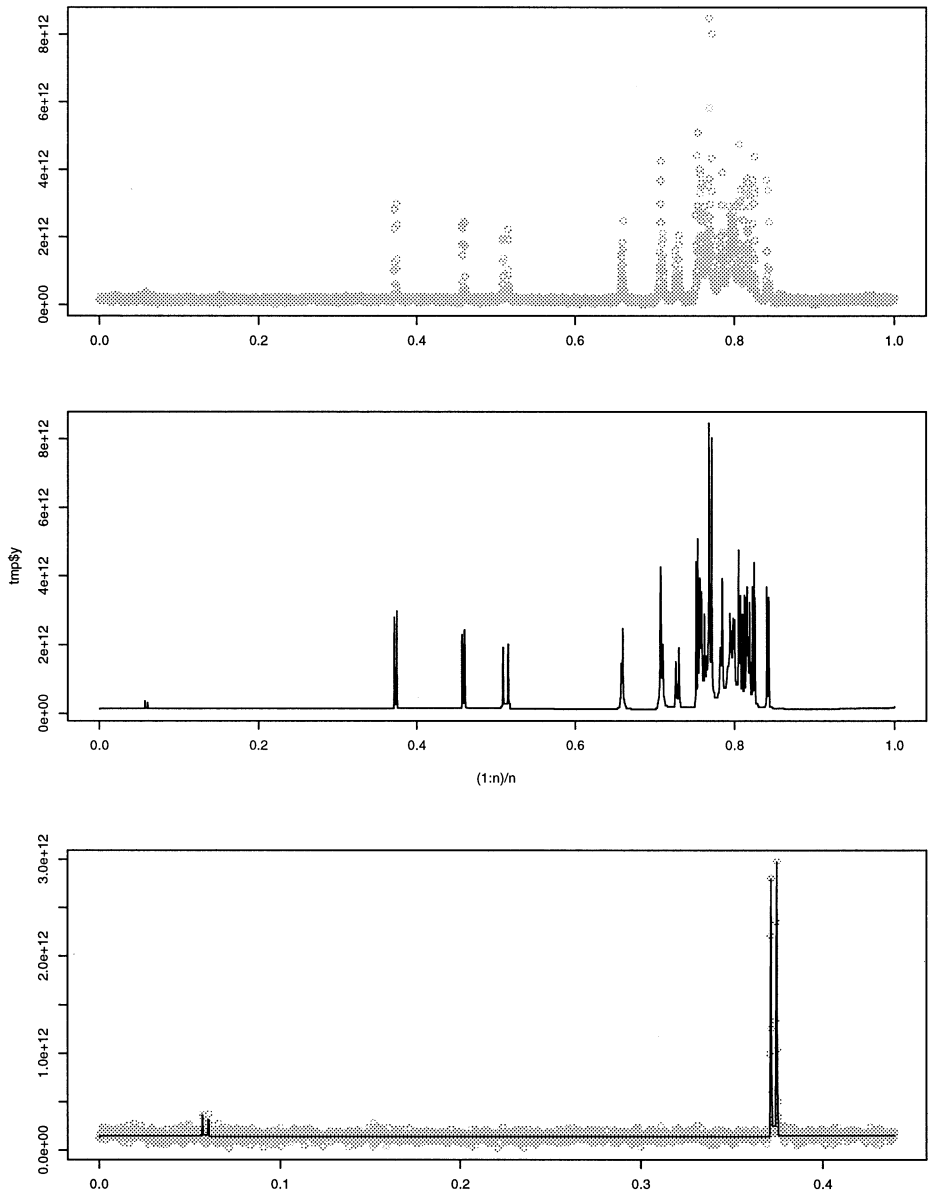

FIG. 16. *Data from spectroscopy. The upper panel shows data that were gathered during a spectroscopical analysis of a gallstone. The reconstruction is shown in the second panel and an excerpt in the bottom panel.*

It can withstand many isolated outliers and can also be tuned to detect blocks of outliers of a specified length. The disadvantage is a slow rate of convergence. The taut-string-multiresolution method has a complexity of $O(n \log n)$ and has almost optimal rates of convergence on standard test beds. It is extremely

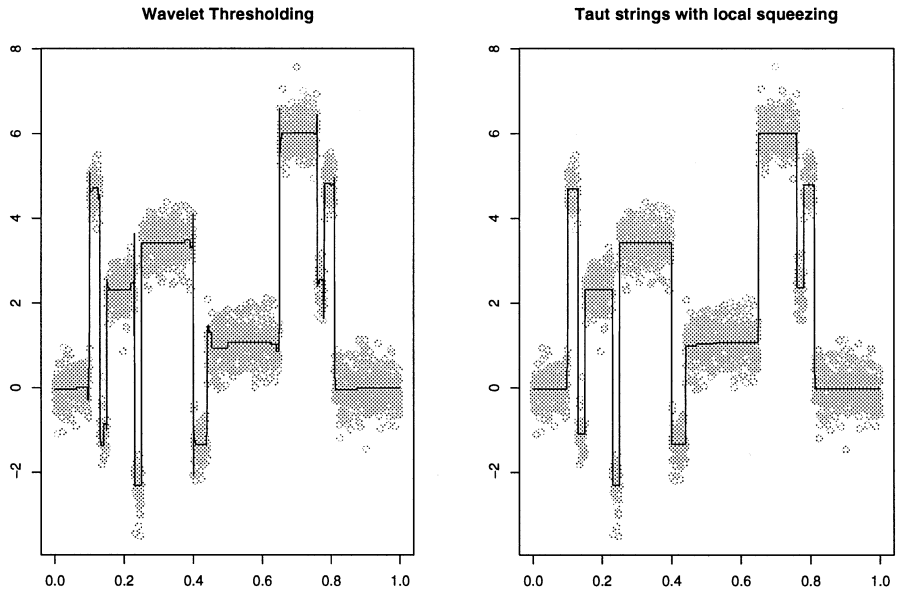

FIG. 17. *Wavelet thresholding and local squeezing. Both methods were applied to the Blocks data. Wavelet thresholding tends to produce pseudo-Gibbs effects near discontinuities leading to* 47 *local extreme values. The taut string method with local squeezing attains the correct modality of nine.*

sensitive, able to identify peaks of very low power. This very sensitivity makes it susceptible to outliers. Implementations are available from our home page at http://www.stat-math.uni-essen.de.

7. Proofs.

7.1. *Proof of Theorem* 3. To prove (c) we analyze the behavior of the upper bound $u_n(\cdot) = u_n(\cdot, \alpha)$ defined by (2.6). We have from (2.7),

$$q_n = qu(n, \alpha, R_n) = \log_2 n + O(1).$$

Let

$$\Gamma_i = \max_{(i-1)q_n+1 \le j \le iq_n} \varepsilon(t_j), \qquad i = 1, \ldots, \left\lceil \frac{n}{q_n} \right\rceil.$$

If F denotes the common distribution function of the $\varepsilon\left(\frac{i}{n}\right)$ then the common distribution function of the Γ_i is F^{q_n}. We set $a_n = \lceil \frac{n \log\log n}{(\log_2 n)^2} \rceil$ and define

$$\Theta_i = \min_{(i-1)a_n+1 \le j \le ia_n} \Gamma_j, \qquad i = 1, \ldots, \left\lfloor \frac{\log n}{\log\log n} \right\rfloor.$$

The common distribution function of the Θ_i is $1-(1-F^{q_n})^{a_n}$. On using $F(x) = \frac{1}{2} + ax$ for small x with $a > 0$ it follows that

$$\mathbf{P}\left(\Theta_i \geq \frac{b \log\log n}{\log n}\right) = O\Big(\exp(-c(\log n)^{b'})\Big),$$

where $c > 0$ and where b may be chosen so that $b' > 0$. For this choice of b we have

$$\mathbf{P}\left(\max_{1\leq i\leq \lfloor \log n/\log\log n\rfloor} \Theta_i \geq \frac{b \log\log n}{\log n}\right) = O\Big((\log n)\exp(-c(\log n)^{b'})\Big).$$

The upper bound u_n is nonincreasing and we first analyze its behavior on an interval where $f^{(1)}(t) > \delta > 0$. Without loss of generality we set $I = [0, \frac{j}{n}]$. The above estimates show that

$$u_n(t) \leq f(0) + \frac{A \log\log n}{\log n}, \qquad t = \frac{\log\log n}{\log n}. \tag{7.1}$$

The corresponding result for the lower bound l_n is

$$l_n(t) \geq f\left(\frac{j}{n}\right) - \frac{A \log\log n}{\log n}, \qquad t = \frac{j}{n} - \frac{\log\log n}{\log n}. \tag{7.2}$$

As both the upper and lower bounds are nonincreasing, (7.1) and (7.2) imply

$$A\frac{\log\log n}{\log n} + f(0) \geq f\left(\frac{j}{n}\right) - A\frac{\log\log n}{\log n}. \tag{7.3}$$

As $f^{(1)}(t) > \delta > 0$ on this interval we have

$$\frac{j}{n} \leq A\frac{\log\log n}{\log n}. \tag{7.4}$$

It follows that if the bounds have the wrong monotonic behavior this will be detected at latest on an interval of length $O(\frac{\log\log n}{\log n})$ where the constant depends on the size of the derivative of f on the interval.

The case where the bounds have the same monotonic behavior as f is as follows. Because of the construction of the intervals, it is clear that f will remain below the upper bound and above the lower bound with probability at least α. On combining these two results we see the monotonic behavior of the bounds which minimizes the number of local extreme values that will coincide with the monotonic behavior of f with probability at least α as n tends to infinity. In other words the number of local extreme values will be determined correctly for large n with probability at least α. The reasoning also shows that the lengths of the intervals $I_i^e(n, \alpha)$ tend to zero with n and that the midpoints converge to the local extreme points of f. Finally the reasoning which lead to (7.1) and (7.2) implies the rate of convergence of (b) of the theorem.

To prove (d) of the theorem it is sufficient to consider the upper bound on a stretch where f is monotone decreasing. On writing $t = \frac{i}{n}$ and $h = \frac{j}{n}$ we have

$$u_n(t) = \min\{u_n(t-h), M(i), \ldots, M(i-j+1)\} \tag{7.5}$$

where

$$M(m) = \max\left\{Y\left(\frac{m}{n}\right), \ldots, Y\left(\frac{m-q_n}{n}\right)\right\}.$$

As $Y(t) = f(t) + \varepsilon(t)$ and f is nonincreasing it follows that

$$M(m) \geq f(t) + \Lambda(m),$$

where

$$\Lambda(m) = \max\left\{\varepsilon\left(\frac{m}{n}\right), \ldots, \varepsilon\left(\frac{m-q_n}{n}\right)\right\}. \tag{7.6}$$

On substituting this into (7.5) we obtain

$$u_n(t) \geq \min\{u_n(t-h), f(t) + \Lambda(i-j+1), \ldots, f(t) + \Lambda(i)\}.$$

As $f(t-h) \leq u_n(t-h)$ and $f(t-h) = f(t) - hf^{(1)}(t)(1+o(1))$ this implies

$$u_n(t) \geq f(t) + \min\{-hf^{(1)}(t)(1+o(1)), \Lambda(i-j+1), \ldots, \Lambda(i)\}. \tag{7.7}$$

We have

$$\mathbf{P}\left(\Lambda(m) \geq \frac{b\log\log n}{\log n}\right) = 1 - \mathbf{P}\left(\Lambda(m) \leq \frac{b\log\log n}{\log n}\right) = 1 - F\left(\frac{b\log\log n}{\log n}\right)^{q_n},$$

which implies

$$\begin{aligned}\mathbf{P}\left(\Lambda(m) \geq \frac{b\log\log n}{\log n}\right) &\approx 1 - \left(\frac{1}{2} + \frac{ab\log\log n}{\log n}\right)^{q_n} \\ &\approx 1 - 2^{-q_n}\exp\left(q_n\frac{2ab\log\log n}{\log n}\right) \\ &\geq 1 - \frac{1}{n}\exp(3ab\,\log\log n)\end{aligned}$$

for large n. From this it follows that

$$\mathbf{P}\left(\bigcup_{m=i-j+1}^{i}\left\{\Lambda(m) \leq \frac{b\log\log n}{\log n}\right\}\right) \leq \frac{j}{n}\exp(3ab\,\log\log n).$$

If we set $j = \lceil \frac{n\log\log n}{\log n} \rceil$ and choose b so that $3ab < \frac{1}{2}$ it follows that $h \approx \frac{\log\log n}{\log n}$ and

$$\mathbf{P}\left(\min_{i-j+1\leq m\leq i}\Lambda(m) \geq \frac{b\log\log}{\log n}\right) \geq 1 - O\left(\frac{\log\log n}{(\log n)^{1/2}}\right).$$

On using this in (7.7) it follows that

$$u_n(t) \geq f(t) + \min\left\{-f^{(1)}(t)\frac{\log\log n}{\log n}, \frac{b\log\log n}{\log n}\right\}$$

with $b > 0$ and $f^{(1)}(t) < 0$. This completes the proof for the upper bound on stretches where f is monotone decreasing. The other cases follow by appropriate changes of sign. □

7.2. *Proof of Theorem* 4. The assumptions of the theorem imply that the integrated error process

$$\varepsilon_n^\circ(t) = \frac{1}{n}\sum_1^{[nt]} \varepsilon\left(\frac{j}{n}\right)$$

converges weakly to a Wiener process on $[0, 1]$:

$$\sqrt{n}\varepsilon_n^\circ \Rightarrow \sigma W,$$

where W denotes the standard Wiener process. In particular we have

$$\lim_{n\to\infty} \mathbf{P}\left(\max_{0\le t\le 1} |\sqrt{n}\varepsilon_n^\circ(t)| \le x\right) = \mathbf{P}\left(\max_{0\le t\le 1} |W(t)| \le \frac{x}{\sigma}\right) = H\left(\frac{x}{\sigma}\right).$$

It follows that on the test bed (1.8),

$$\lim_{n\to\infty} \mathbf{P}\left(\max_{0\le t\le 1} |Y_n^\circ(t) - f(t)| \le \frac{C}{\sqrt{n}}\right) = H\left(\frac{C}{\sigma}\right).$$

As n tends to infinity the probability that the function f lies in the tube $T(Y_n^\circ, C/\sqrt{n})$ tends to $H(\frac{C}{\sigma})$. As the taut string minimizes the number of local extremes in $T(Y_n^\circ, C/\sqrt{n})$ we see that

$$\lim_{n\to\infty} \mathbf{P}(K_n^C \le k) = H\left(\frac{C}{\sigma}\right). \qquad \square$$

7.3. *Proof of Theorem* 5. We note that for f satisfying the assumptions of the theorem

$$\inf_{g\in M_{k-1}} \sup_{0\le t\le 1} |f^\circ(t) - g^\circ(t)| > 0,$$

where M_j denotes the set of functions on $[0, 1]$ with at most j local extremes. This implies

$$\lim_{n\to\infty} \mathbf{P}(K_n^C < k) = 0 \quad \text{for all } C > 0.$$

In the other direction Theorem 4 implies

$$\lim_{C\to\infty} \lim_{n\to\infty} \mathbf{P}(K_n^C \le k) = 1$$

and hence

$$\lim_{C\to\infty} \lim_{n\to\infty} \mathbf{P}(K_n^C = k) = 1.$$

The other claims are proved similarly. If the lengths of the intervals $I_i^e(n, C)$ do not converge in probability to zero then

$$\max_{t\in I_i^e(n,C)} |S_n(t) - f(t)|$$

does not converge in probability to zero and this carries over to the integrated functions. A similar argument applies to the convergence of the location of the local extreme points. $\square$

7.4. *Proof of Theorem* 6. The proof of Theorem 6 relies on the modulus of continuity of the integrated process ε_n° as expressed by

$$\lim_{\delta\to 0}\mathbf{P}\left(\sup_{0\le t,\, t+h\le 1,\, h\le\delta}\sqrt{n}|\varepsilon_n^\circ(t+h)-\varepsilon_n^\circ(t)|\le A\sqrt{-\delta\,\log\delta}\right)=1 \tag{7.8}$$

for some $A>0$. This follows from the sub-Gaussian form of the $\varepsilon(t)$. This will be used for $\delta=\delta_n$ of the form $(\log n)^\alpha n^{-\beta}$ with $\alpha\ge 0$ and $\beta>0$ in which case we have

$$\lim_{n\to\infty}\mathbf{P}\left(\sup_{0\le t,\, t+h\le 1,\, h\le\delta_n}\sqrt{n}|\varepsilon_n^\circ(t+h)-\varepsilon_n^\circ(t)|\le A\sqrt{-\delta_n\,\log\delta_n}\right)=1. \tag{7.9}$$

The proofs given below require that the taut string through the tube of radius $C/\sqrt{n}$ has the correct shape; that is, it has the same number of local extremes as f. In order for this to happen in probability as n tends to infinity we have to let C tend to infinity. The order is first n to infinity and then C to infinity. As the distribution of the maximum of a Brownian motion on the interval $[0,1]$ behaves exactly like the tails of the standard normal distribution, the probabilities tend very quickly to 1 as C increases. In practice $C=3\sigma_n$ with σ_n a robust scale functional such as (1.7) is a large value of C.

The proofs can probably be best understood by taking a probability space $(\Omega,\mathscr{F},\mathbf{P})$ and considering a sequence of events $\Omega^0(C,n)$ which are chosen as the need arises but which satisfy

$$\lim_{C\to\infty}\lim_{n\to\infty}\mathbf{P}(\Omega^0(C,n))=1.$$

We can, for example, put

$$\Omega^1(C,n)=\left\{\omega:\sup_{0\le t\le 1}|Y_n^\circ(t)-f^\circ(t)|\le\frac{C}{\sqrt{n}}\right\}.$$

Then clearly

$$\lim_{C\to\infty}\lim_{n\to\infty}\mathbf{P}(\Omega^1(C,n))=1.$$

From Theorem 5 it follows that there exists a sequence $(\delta_n)_1^\infty$ tending to zero such that

$$\lim_{C\to\infty}\lim_{n\to\infty}\mathbf{P}(\Omega^2(C,n))=1,$$

where

$$\Omega^2(C,n)=\left\{\omega\colon K_n^C=k,\max_{1\le i\le k}|I_i^e(n,C)|\le\delta_n,\max_{1\le i\le k}|\tau_i^e(n,C)-t_i^e|\le\delta_n\right\}.$$

Similarly for another given sequence $(\delta_n)_1^\infty$ tending to zero we may define

$$\Omega^3(A,n)=\left\{\omega:\sup_{0\le t,\, t+h\le 1,\, h\le\delta}\sqrt{n}|\varepsilon_n^\circ(t+h)-\varepsilon_n^\circ(t)|\le A\sqrt{-\delta_n\,\log\delta_n}\right\}.$$

For some appropriate $A > 0$ we have $\lim_{n\to\infty} \mathbf{P}(\Omega^3(n)) = 1$. Such choices are made according to our m needs and finally we set

$$\Omega^0(C,n) = \Omega^1(C,n) \cap \Omega^2(C,n) \cap \cdots \cap \Omega^m(C,n).$$

For all $\omega \in \Omega^0(C,n)$ all relevant properties hold such as, for example, $\widetilde{S}_n$ having the same monotonicity behavior as f. The terms O and o used below can be translated into inequalities that will hold for each n and for all $\omega \in \Omega^0(C,n)$.

Proof of (a). Suppose S_n° is initially convex. Then S_n° is the largest convex minorant of $Y_n^\circ + C/\sqrt{n}$ until it reaches the left endpoint $t_1^l(n,C)$ of $I_1^e(n,C) = (t_1^l(n,C), t_1^r(n,C))$. Then with $h_0 = t_1^r(n,C) - t_1^l(n,C)$ we have

$$h_0 = \operatorname*{argmax}_{0\le h\le \delta} \frac{Y_n^\circ(t_1^l(n,C)+h) - Y_n^\circ(t_1^l(n,C)) - \frac{2C}{\sqrt{n}}}{h} \tag{7.10}$$

for arbitrarily small δ as n tends to infinity. On writing $t_1^l(n,C) = t_1^e - \kappa$ we may rewrite (7.10) to obtain

$$h_0 = \operatorname*{argmax}_{0\le h\le \delta} \frac{Y_n^\circ(t_1^e - \kappa + h) - Y_n^\circ(t_1^e - \kappa) - \frac{2C}{\sqrt{n}}}{h}.$$

A Taylor expansion together with the modulus of continuity of ε_n° gives

$$h_0 = \operatorname*{argmax}_{0\le h\le \delta} \left(-\frac{1}{6} h(3\kappa - h) f^{(2)}(t_1^e)(1+o(1)) - \frac{2C}{\sqrt{n}h}(1+o(1)) \right).$$

This implies

$$-\frac{1}{6}(3\kappa - 2h_0) f^{(2)}(t_1^e)(1+o(1)) = -\frac{2C}{\sqrt{n}h_0^2}(1+o(1)) \tag{7.11}$$

and as $f^{(2)}(t_1^e) < 0$ we may conclude $3\kappa \le 2h_0(1+o(1))$ which implies $t_1^e < t_1^r(n,C)$. In the other direction we have

$$h_0 = \operatorname*{argmax}_{0\le h\le \delta} \frac{Y_n^\circ(t_1^r(n,C)) - Y_n^\circ(t_1^r(n,C) - h) - \frac{2C}{\sqrt{n}}}{h}.$$

On writing $t_1^r(n,C) = t_1^e + \kappa^*$ we obtain

$$-\frac{1}{6}(3\kappa^* - 2h_0) f^{(2)}(t_1^e)(1+o(1)) = -\frac{2C}{\sqrt{n}h_0^2}(1+o(1)). \tag{7.12}$$

This implies $3\kappa \le 2h_0(1+o(1))$ and hence $t_1^e > t_1^l(n,C)$ so that t_1^e lies in the interval $[t_1^l(n,C), t_1^r(n,C)]$ as was to be proved.

Proof of (b). On adding (7.11) and (7.12) we obtain

$$-(3(\kappa+\kappa^*) - 2h_0) f^{(2)}(t_1^e)(1+o(1)) = -\frac{24C}{\sqrt{n}h_0^2}(1+o(1))$$

and as $\kappa + \kappa^* = h_0$ this implies

$$h_0 \sim (24C)^{1/3} |f^{(2)}(t_1^e)|^{-1/3} n^{-1/6}.$$

Proof of (c). It is sufficient to consider x_1 and x_2 and to suppose that f° and S_n° are both convex on (x_1, x_2). On writing $x_1 = i_1/n$ and $x_2 = (i_1 + l_1)/n$ we see that l_1 is the local argmin of

$$\frac{Y_n^\circ(x_1 + \frac{l}{n}) - Y_n^\circ(x_1)}{\frac{l}{n}}.$$

A Taylor series expansion gives

$$\frac{Y_n^\circ(x_1 + \frac{l}{n}) - Y_n^\circ(x_1)}{\frac{l}{n}} = f(x_1) + \frac{1}{2} f^{(1)}(x_1) \frac{l}{n} + \frac{\varepsilon_n^\circ(x_1 + \frac{l}{n}) - \varepsilon_n^\circ(x_1)}{\frac{l}{n}} + O\left(\left(\frac{l}{n}\right)^2\right).$$

The modulus of continuity (7.8) of $\sqrt{n}\varepsilon_n^\circ$ implies

$$\sqrt{n}\left|\varepsilon_n^\circ\left(t + \frac{l}{n}\right) - \varepsilon_n^\circ(t)\right| \le A\sqrt{\frac{l \log n}{n}}$$

uniformly in t and l, $1 \le l \le n^{9/10}$, say. On setting

$$l = a|f^{(1)}(x_1)|^{-2/3} n^{2/3} (\log n)^{1/3}$$

we have

$$\begin{aligned}\frac{Y_n^\circ(x_1 + \frac{l}{n}) - Y_n^\circ(x_1)}{\frac{l}{n}} \ge f(x_1) &+ \frac{1}{2} a |f^{(1)}(x_1)|^{1/3} \left(\frac{\log n}{n}\right)^{1/3} \\ &- \frac{A}{\sqrt{a}} |f^{(1)}(x_1)|^{1/3} \left(\frac{\log n}{n}\right)^{1/3} \\ &+ O\left(|f^{(1)}(x_1)|^{-4/3} \left(\frac{\log n}{n}\right)^{2/3}\right).\end{aligned}$$

From (b) of the theorem $|f^{(1)}(x_1)| \ge An^{-1/6}$ for $C \ge C_0$ and consequently the term

$$O\left(|f^{(1)}(x_1)|^{-4/3} \left(\frac{\log n}{n}\right)^{2/3}\right)$$

may be neglected. This implies that for a sufficiently large,

$$\frac{Y_n^\circ(x_1 + \frac{l}{n}) - Y_n^\circ(x_1)}{\frac{l}{n}} \ge f(x_1) + \frac{1}{4} a |f^{(1)}(x_1)|^{1/3} \left(\frac{\log n}{n}\right)^{1/3}. \tag{7.13}$$

The lower bound

$$\frac{Y_n^\circ(x_1+\frac{l}{n})-Y_n^\circ(x_1)}{\frac{l}{n}} \le f(x_1)+a|f^{(1)}(x_1)|^{1/3}\left(\frac{\log n}{n}\right)^{1/3} \tag{7.14}$$

is obtained analogously. On putting $a = a_1$ in (7.13) and $a = a_2$ in (7.14) with $a_1 = 4a_2$ with a_2 sufficiently large it follows that the local minimum is attained at a point $x_1 + \frac{l}{n}$ with

$$\frac{l}{n} = O\left(|f^{(1)}(x_1)|^{-2/3}\left(\frac{\log n}{n}\right)^{1/3}\right).$$

This proves (c) of the theorem.

Proof of (d). We consider first the case where $t = x_i$ is a knot which does not delimit the position of a local extreme value of S_n. We take S_n° to be convex at x_i. We have

$$S_n(x_i) \le \frac{Y_n^\circ(x_i+\frac{l}{n})-Y_n^\circ(x_i)}{\frac{l}{n}}.$$

A Taylor expansion of order two combined with (7.8) gives for

$$\begin{aligned} l &= Af^{(1)}(x_i)^{-2/3}n^{2/3}(\log n)^{1/3}, \\ S_n(x_i) &\le f(x_i)+A|f^{(1)}(x_i)|^{1/3}\left(\frac{\log n}{n}\right)^{1/3}. \end{aligned} \tag{7.15}$$

Using

$$S_n(x_i) \ge \frac{Y_n^\circ(x_i)-Y_n^\circ(x_i-\frac{l}{n})}{\frac{l}{n}},$$

a similar argument gives

$$S_n(x_i) \ge f(x_i)-A|f^{(1)}(x_i)|^{1/3}\left(\frac{\log n}{n}\right)^{1/3},$$

which when combined with (7.15) gives

$$|f(x_i)-S_n(x_i)| = O\left(|f^{(1)}(x_i)|^{1/3}\left(\frac{\log n}{n}\right)^{1/3}\right)$$

at all knots x_i which do not delimit a local extreme value of S_n. For a point t not in $[A(\frac{\log n}{n})^{1/3},\ 1-A(\frac{\log n}{n})^{1/3}]\backslash\cup_{i=1}^k I_i^e(n,C)$ we have

$$\begin{aligned} |f(t)-S_n(t)| &= |f(t)-S_n(x_i)| \\ &\le |f(x_i)-S_n(x_i)|+|f(t)-f(x_i)| \\ &\le |f(x_i)-S_n(x_i)|+A|f^{(1)}(x_i)||f^{(1)}(x_i)|^{-2/3}\left(\frac{\log n}{n}\right)^{1/3} \\ &\le A|f^{(1)}(t)|^{1/3}\left(\frac{\log n}{n}\right)^{1/3}, \end{aligned}$$

where we have used

$$\sup_{x_i \le t \le x_{i+1}} \left| \frac{f^{(1)}(x_i)}{f^{(1)}(t)} - 1 \right| \le A$$

for all intervals $[x_i, x_{i+1}]$ which do not delimit a local extreme value. This follows from (b) of the theorem.

Proof of (e). This follows as in the other cases but using the next term of the Taylor expansion as $f^{(1)}(t) = 0$ for some point in the interval. □

7.5. *Proof of Theorem* 7.

Proof of (a). The multiresolution coefficient $w_{j,k}$ is given by

$$w_{i,k} = \frac{\sqrt{n}}{2^{j/2}} \cdot \left(Y_n^\circ\left(\frac{k2^j}{n}\right) - \widetilde{S}_n^\circ\left(\frac{k2^j}{n}\right) - Y_n^\circ\left(\frac{(k+1)2^j}{n}\right) + \widetilde{S}_n^\circ\left(\frac{(k+1)2^j}{n}\right)\right). \tag{7.16}$$

To show that all multiresolution coefficients of the noise are smaller than the bound (3.1) for some $\tau > 0$ it is sufficient to show that each coefficient $w_{i,k}$ is at most $\sigma\sqrt{\tau \log n}$. This in turn follows from the inequality

$$|Y_n^\circ(t_l) - \widetilde{S}_n^\circ(t_l) - (Y_n^\circ(t_i) - \widetilde{S}_n^\circ(t_i))| \le A\sqrt{|t_l - t_i|}\sqrt{\frac{\log n}{n}}, \tag{7.17}$$

where (t_i, t_l) is one of the pairs $((k2^j - 2^{j-1})/n, k2^j/n), (k2^j/n, (k2^j + 2^{j-1})/n)$. We have

$$Y_n^\circ - \widetilde{S}_n^\circ = \varepsilon_n^\circ + f^\circ - \widetilde{S}_n^\circ.$$

and on using the modulus of continuity of $\sqrt{n}\varepsilon_n^\circ$ as given by (7.8) we obtain for any points t_i and t_l with $|t_l - t_i| \ge 1/n$,

$$\begin{aligned} &\left|Y_n^\circ(t_l) - \widetilde{S}_n^\circ(t_l) - (Y_n^\circ(t_i) - \widetilde{S}_n^\circ(t_i))\right| \\ &\quad \le A\sqrt{|t_l - t_i|}\sqrt{\frac{\log n}{n}} + \left|f^\circ(t_l) - \widetilde{S}_n^\circ(t_l) - (f^\circ(t_i) - \widetilde{S}_n^\circ(t_i))\right|. \end{aligned} \tag{7.18}$$

As

$$\begin{aligned} |f^\circ(t_l) - \widetilde{S}_n^\circ(t_l) - (f^\circ(t_i) - \widetilde{S}_n^\circ(t_i))| &= \left|\int_{t_i}^{t_l} (f(t) - \widetilde{S}_n(t))\,dt\right| \\ &\le |t_l - t_i| \sup |f(t) - \widetilde{S}_n(t)|, \end{aligned}$$

(c) and (d) of Theorem 6 imply

$$\begin{aligned} |f^\circ(t_m) - \widetilde{S}_n^\circ(t_m) - (f^\circ(t_i) - \widetilde{S}_n^\circ(t_i))| &\le A|t_m - t_i|\left(\frac{\log n}{n}\right)^{1/3} \\ &\le A\sqrt{|t_l - t_i|}\sqrt{\frac{\log n}{n}} \end{aligned}$$

if

$$|t_l - t_i| \le A\left(\frac{\log n}{n}\right)^{1/3}. \tag{7.19}$$

Thus (7.17) holds for all intervals satisfying (7.19).

Proof of (b). This follows from (a) and Theorem 6(c).

Proof of (c). Consider a multiresolution coefficient whose support is $[t_1, t_2]$ with $t_1 \in I_i^e(n, C)$ and $t_2 \notin I_i^e(n, C)$ for some i, the other outside of the same interval. Let x_1 be the right endpoint of $I_i^e(n, C)$. It follows from the definition of $\widetilde{S}_n^\circ$ that $|Y_n^\circ(t) - \widetilde{S}_n^\circ(t)| \le C/\sqrt{n}$ for all $t \in I_i^e(n, C)$. In particular we have

$$|Y_n^\circ(x_1) - \widetilde{S}_n^\circ(x_1) - (Y_n^\circ(t_1) - \widetilde{S}_n^\circ(t_1))| \le \frac{2C}{\sqrt{n}}.$$

This is also the worst case for the other point t_2 and the result is the inequality

$$|Y_n^\circ(t_2) - \widetilde{S}_n^\circ(t_2) - (Y_n^\circ(t_1) - \widetilde{S}_n^\circ(t_1))| \le \frac{4C}{\sqrt{n}}.$$

The multiresolution coefficient will satisfy (7.17) if

$$t_2 - t_1 \ge \frac{16C^2}{\tau \log n}$$

as was to be shown.

Proof of (d). Consider a multiresolution coefficient whose support is contained in $I_1^e(n, c)$ with length $h \ge \frac{1}{2}|I_1^e(n, C)|$. From Theorem 6 we have $h \approx c_1 n^{-1/6}$ and $|f(t) - \tilde{s}_n(t)| \ge c_2 n^{-1/3}$ for some constants $c_1 > 0$ and $c_2 > 0$. Without loss of generality we may assume that $f(t) - \tilde{s}_n(t) \ge c_2 n^{-1/3}$ on the support of the multiresolution coefficient. This implies that the absolute value of the multiresolution coefficient is at least $c_3 n^{-1/2}$ for some $c_3 > 0$ which eventually exceeds the bound

$$\sqrt{\frac{\tau h \log n}{n}} \le c_4 n^{-7/12}\sqrt{\tau \log n},$$

whatever the value of τ. □

APPENDIX

The taut string algorithm. We give a short description of the taut string algorithm. Without loss of generality we suppose the data points are $(i, y(i))$, $i = 0, \ldots, n$. Let l and u denote the lower and upper bounds for the integrated process y° given by

$$y^\circ(k) = \frac{1}{n}\sum_{i=1}^{k} y(i), \qquad 0 \le k \le n.$$

We assume that the endpoints of the string are fixed, that is, $l(0) = u(0)$ and $l(n) = u(n)$. Starting from the point $(0, l(0))$ and for given i the greatest convex minorant sx_i of the upper bound $(j, u(j))$, $j = 0, \ldots, i$ and the smallest concave majorant sv_i of the lower bound $(j, l(j))$, $j = 1, \ldots, i$ are both calculated. The greatest convex minorant sx_i may be calculated as follows [see also Barlow, Bartholomew, Bremner and Brunk (1972)]. Suppose sx_i has been calculated. It is defined by knots $(k(j), u(k(j)))$, $1 \leq j \leq K(i)$ where it touches the upper bound and it is linear in between. The first and last points are knots. The line joining $(i + 1, u(i + 1))$ is now included as a knot. If the resulting linear interpolation is convex then this is sx_{i+1}. If not, then the knot to the left of $(i+1, u(i+1))$ is permanently eliminated. If the resulting linear interpolation is convex this is sx_{i+1}. If it is not convex, then again the knot to the left of $(i + 1, u(i + 1))$ is eliminated. This process is continued until a convex interpolation is attained and this defines sx_{i+1}. This algorithm has complexity $O(i)$ for calculating all the functions sx_j and sv_j $0 \leq j \leq i$. It is clear that the taut string must lie between the sx_i and sv_i. The functions sx_i and sv_i are linear between the knots where they touch the upper and lower bounds, respectively. After sx_i and sv_i have been calculated it is checked whether

$$sx_i^{(1)}(0+) > sv_i^{(1)}(0+) \tag{A.1}$$

holds where $sx_i^{(1)}(0+)$ and $sv_i^{(1)}(0+)$ denote respectively the right hand derivatives at 0 of sx_i and sv_i. These are nothing more than the gradients of the first sections of the functions sx_i and sv_i. If (A.1) holds then sx_{i+1} and sv_{i+1} are calculated. This process is continued until for some i (A.1) does not hold. At this point the leftmost knot of the set of knots of both the sx_i and sv_i is determined whereby the very first knot $(0, u(0))$ is not counted. This leftmost knot is the first knot of the taut string and we denote it by $(k_1, s(k_1))$. It may be either a knot from the upper bound or from the lower bound. Figure 18 demonstrates this procedure.

The process is now repeated. The origin is moved to the point $(k_1, 0)$ with translated upper bounds $(j, u(k_1 + j))$, $j = 0, \ldots, n - k_1$ and lower bounds $(j, l(k_1 + j))$, $j = 0, \ldots, n - k_1$. Starting from the point $(0, s(k_1))$ the greatest convex minorant of the upper bounds and the smallest concave majorant of the lower bounds are calculated as before. The important point to notice is that these calculations have already been done for the first $i - k_1$ points. If the first knot of the string is on the upper bound, that is, $(k_1, s(k_1)) = (k_1, u(k_1))$, then the convex minorant for the section $0 \leq j \leq i - k_1$ of the translated bounds coincides with the translation of convex minorant already calculated. The concave majorant is simply the straight line joining $(0, s(k_1))$ and $(i - k_1, l(i))$. A corresponding statement holds if the first knot is on the lower bound. It is this which makes the calculation of the taut string $O(n)$.

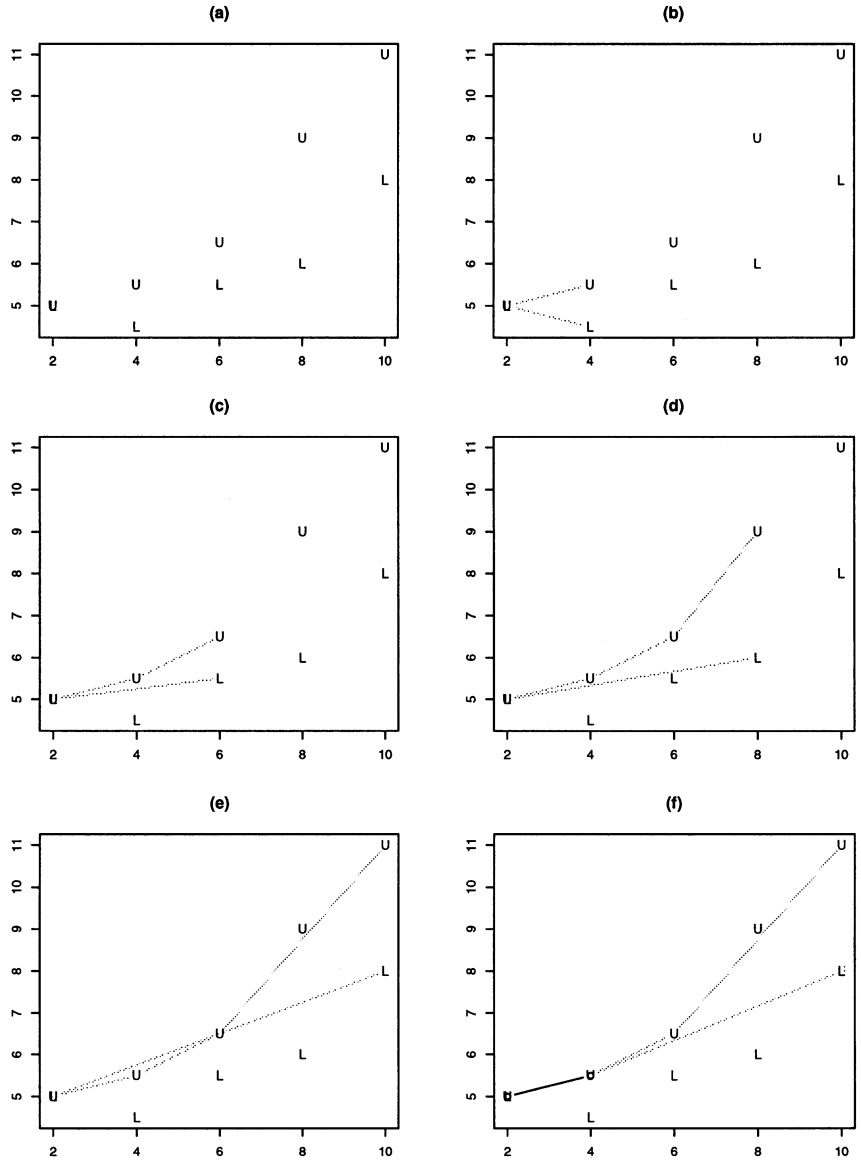

FIG. 18. *Example of the taut string method. Moving from the left to the right, the greatest convex minorants of the upper bounds and the smallest concave majorant of the lower bounds up to the current position are calculated until both curves intersect. The leftmost knot is added to the taut string.*

The calculations just described are continued until the last point is reached. The taut string is now determined by linear interpolation between its knots.

Acknowledgments. We are indebted to Lutz Dümbgen and Martin Löwendick for useful comments. We also thank two anonymous referees and in particular an Associate Editor whose comments resulted in an increase in accuracy and clarity.

REFERENCES

BARLOW, R., BARTHOLOMEW, D., BREMNER, J. and BRUNK, H. (1972). *Statistical Inference under Order Restrictions.* Wiley, New York.

CHAUDHURI, P. and MARRON, J. S. (1999). Sizer for exploration of structures in curves. *J. Amer. Statist. Assoc.* **94** 807–823.

DAVIES, P. (1995). Data features. *Statist. Neerlandica* **49** 185–245.

DAVIES, P. L. (2000). Hidden periodicities and strings. Dept. Mathematics and Computuer Science, Univ. Essen, Germany.

DAVIES, P. L. and LÖWENDICK, M. (1999). On smoothing under bounds and geometric constraints. Technical report, Univ. Essen.

DELECROIX, M., SIMIONI, M. and THOMAS-AGNAN, C. (1995). A shape constrained smoother: simulation study. *Comput. Statist.* **10** 155–175.

DONOHO, D. (1988). One-sided inference about functionals of a density. *Ann. Statist.* **16** 1390–1420.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc.* **57** 371–394.

DÜMBGEN, L. (1998a). Application of local rank tests to nonparametric regression. Medical Univ., Lübeck.

DÜMBGEN, L. (1998b). New goodness-of-fit tests and their application to nonparametric confidence sets. *Ann. Statist.* **26** 288–314.

FAN, J. and GIJBELS, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaption. *J. Roy. Statist. Soc. Ser. B* **57** 371–394.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications* **1**, 3rd ed. Wiley, New York.

FISHER, N. I., MAMMEN, E. and MARRON, J. S. (1994). Testing for multimodality. *Comput. Statist. Data Anal.* **18** 499–512.

FREEDMAN, D. (1971). *Browniam Motion and Diffusion*, 3rd. ed. Holden-Day, San Francisco.

GOOD, I. and GASKINS, R. (1980). Density estimating and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75** 42–73.

GREEN, P. and SILVERMAN, B. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman and Hall, London.

GROENEBOOM, P. (1985). Esimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* **2** (L. Le Cam and R. Olshen, eds.) Wadsworth, Monterey, CA.

HARTIGAN, J. A. (1987). Esimation of a convex density contour in two dimensions. *J. Amer. Statist. Association* **82** 267–270.

HARTIGAN, J. A. and HARTIGAN, P. (1985). The dip test of unimodality. *Ann. Statist.* **13** 70–84.

HENGARTNER, N. and STARK, P. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* **23** 525–550.

IBRAGIMOV, I. and KHAS'MINSKII, R. (1980). On nonparametric estimation of regression. *Soviet Math. Dockl.* **21** 810–814.

KAHANE, J.-P. (1968). *Some Random Series of Functions*. Heath, Lexington, MA.

KHAS'MINSKII, R. (1978). A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory Probab. Appl.* **23** 794–798.

KOVAC, A. and SILVERMAN, B. W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Amer. Statist. Assoc.* **95** 172–183.

LEURGANS, S. (1982). Asymptotic distributions of slope-of-greatest-convex-minorant estimators. *Ann. Statist.* **10** 287–296.

MAJIDI, A. (2000). Nichtparametrische regression unter modalitätskontrolle. Diplomarbeit, Fachbereich Mathematik und Informatik, Univ. Essen.

MAMMEN, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759.

MAMMEN, E., MARRON, J., TURLACH, B. and WAND, M. (1998). A general framework for constrained smoothing. Unpublished manuscript.

MAMMEN, E. and THOMAS-AGNAN, C. (1998). Smoothing splines and shape restrictions. Unpublished manuscript.

MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413.

MÄCHLER, M. (1995). Variational solution of penalized likelihood problems and smooth curve estimation. *Ann. Statist.* **23** 1496–1517.

METZNER, L. (1997). Facettierte nichtparametrische Regression. Ph.D. thesis, Univ. Essen, Germany.

MINOTTE, M. C. (1997). Nonparametric testing of the existence of modes. *Ann. Statist.* **25** 1646–1660.

MINOTTE, M. C. and SCOTT, D. W. (1993). The mode tree: a tool for visualization of nonparametric density features. *J. Comput. Graph. Statist.* **2** 51–68.

MORGENTHALER, S. and TUKEY, J. (1991). *Configural Polysampling: A Route to Practical Robustness*. Wiley, New York.

MÜLLER, D. and SAWITZKI, G. (1991). Excess mass estimates and tests of multimodality. *J. Amer. Statist. Assoc.* **86** 738–746.

NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **10** 186–190.

POLONIK, W. (1995a). Density estimation under qualitive assumptions in higher dimensions. *J. Mulitivariate Anal.* **55** 61–81.

POLONIK, W. (1995b). Measuring mass concentrations and estimating density contour clusters: an excess mass approach. *Ann. Statist.* **23** 855–881.

POLONIK, W. (1999). Concentration and goodness of fit in higher dimensions: (aymptotically) distribution-free methods. *Ann. Statist.* **27** 1210–1229.

POLZEHL, J. and SPOKOINY, G. (2000). Adaptive weights smoothing with applications to image restoration. *J. Roy. Statist. Soc. Ser. B* **62** 335–354.

RAMSAY, J. (1998). Estimating smooth monotone functions. *J. Roy. Statist. Soc.* **60** 365–375.

ROBERTSON, T. (1967). On estimating a density measurable with respect to a σ-lattice. *Ann. Math. Statist.* **38** 482–493.

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

SAGER, T. W. (1979). An iterative method for estimating a multivariate mode and isopleth. *J. Amer. Statist. Assoc.* **74** 329–339.

SAGER, T. W. (1982). Nonparametric maximum likelihood estimation of spatial patterns. *Ann. Statist.* **10** 1125–1136.

SAGER, T. W. (1986). An application of isotonic regression to multivariate density estimation. In *Advances in Order Restricted Statistical Inference* (R. L. Dykstra, T. Robertson and F. T. Wright, eds.) Springer, New York.

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

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

STONE, C. (1982). Optimal rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

TUKEY, J. (1993). Exploratory analysis of variance as providing examples of strategic choices. In *New Directions in Statistical Data Analysis and Robustness* (S. Morgenthaler, E. Ronchetti and W. A. Stahel, eds.) Birkhäuser, Basel.
WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā* **26** 101–116.
WEGMAN, E. J. (1970). Maximum likelihood estimation of a unimodal density. *Ann. Math. Statist.* **41** 457–471.

UNIVERSITÄT GESAMTHOCHSCHULE ESSSEN
FB06 MATHEMATIK UND INFORMATIK
D-45117 ESSEN
GERMANY
E-MAIL: arne.kovac@uni-essen.de

DISCUSSION

RUDOLF BERAN

University of California, Berkeley

> Today's understanding of how well elementary techniques work owes much to both mathematical analysis and experimentation by computer.—John Tukey (1977)
>
> The history of the arts and sciences could be written in terms of the continuing process by which new technologies create new environments for old technologies.—H. Marshall McLuhan (1964)

Man makes tools whose use then reshapes his life. Technological advances in statistical computing and in empirical process theory have swept away statistics as a normative mathematical philosophy. Our subject is under reconstruction on a more scientific foundation of computational experiments linked to mathematical probes of statistical methods. Environmental stimuli, notably competition from data-analytic techniques that fall outside the statistical canon, favor the evolution of statistics towards empirically tested, falsifiable theory.

In presenting their innovative nonparametric regression estimators, Davies and Kovac carefully distinguish among data, statistical procedure and probability model. Computational experiments have brought this fundamental distinction to the forefront of statistical thought. Tukey's (1977) *Exploratory Data Analysis* made the point dramatically by not using probability models at all in the exposition.

1. Experimental and theoretical statistics. Davies and Kovac remark that "statistical procedures can be evaluated using real data sets and under the well-controlled conditions of a stochastic model or test bed." Their figures

illustrate another possibility, trying a procedure on artificial data. We delve a little further into the matter.

Until recent decades, the primary tools available to a statistician were mathematics and logic. A powerful technology, measure-theoretic probability theory, directed statistics toward probability models for data, toward discussions of abstract principle and toward test statistics or pivots whose distributions could be tabulated because they did not depend on nuisance parameters. Simple probability models can motivate promising classes of statistical procedures. The rules (1.4) and (1.6) that Davies and Kovac propose as part of their nonparametric regression technology are in this tradition. More complex probability models can be used in studying mathematically the performance of statistical procedures. The asymptotic theorems of Section 3 reflect this second tradition.

Advances in computing and graphical output have brought about rapid development of experimental statistics. One type of computational experiment elicits finer details of a procedure's performance in repeated pseudorandom realizations from probability models. These details may supplement results from asymptotic theory, such as rates of convergence and asymptotic minimax bounds, whose implications for statistical practice are less than clear. Of course, pseudorandom realizations constitute a mathematical model for data that deserves study in its own right. Such artificial data mimics only some features of the motivating probability model and has properties not envisaged by that probability model. This point tends to be minimized in the statistical literature.

Another type of experiment consciously investigates performance of procedures outside probability models. Case studies on actual data or on idealizations of actual data are important. So are comparative analyses with perturbations of real or artificial data. One might change signal-to-noise ratio or sample size in the Blocks example in Figure 11. One might add a high frequency sine wave to that example, increasing the frequency of the sinusoid in order to discover at what point the regression estimators in this paper first lose sight of its regularity and at what point they fail completely to detect it. Such perturbation experiments provide interpretable ways of comparing competing nonparametric regression estimators. Though the human eye often fails to spot high-frequency patterns amidst noise, good regression techniques do better.

Assessing procedures under a probability or other mathematical model is a bold speculation that relies on two hopes: the ability of the model to mimic observational data and the ability of the mathematical analysis to address questions of data-analytic interest. Greater interplay between computational experiments and mathematical probes of procedures has replaced speculation about data analysis with the rudiments of scientific method. Nevertheless, reproducibility of published computational experiments still has low priority in core statistical journals. In an unpublished technical report, Buckheit and Donoho (1995) discussed what is required to make computational portions of statistical research reproducible.

2. The one-way layout in nonparametric regression. A regression model that motivates parts of the paper is

$$y(t_i) = f(t_i) + \varepsilon(t_i), \qquad 1 \le i \le n, \tag{1}$$

where the values $0 < t_1 < t_2 < \cdots t_n < 1$ are strictly ordered. Estimation of the function f on the basis of the observed $\{(y_i, t_i)\}$ is the task undertaken. In theoretical study of function estimators f_n under probability models, it is customary to compare f_n with f through quantities such as the supremum norm and to impose conditions on f such as differentiability (cf. Section 3 of the paper).

Nonparametric regression combines two distinguishable problems, each of which may be studied constructively on its own. The first problem is estimation of the values $\{f(t_i)\colon 1 \le i \le n\}$. This amounts to estimation of the values $\{\mu_i\}$ in the one-way layout

$$y_i = \mu_i + \varepsilon_i, \qquad 1 \le i \le n, \tag{2}$$

where $y_i = y(t_i)$, $\mu_i = f(t_i)$ and $\varepsilon_i = \varepsilon(t_i)$. That the least squares estimator $y = \{y_i\}$ of $\mu = \{\mu_i\}$ is not a good answer to the problem was pointed out by Stein (1956). This naive estimator can have relatively high risk under a probability model.

Once we have devised a more efficient estimator of μ, the second problem is interpolation among its components so as to estimate the function f. This is essentially a problem in approximation theory and is considerably more sensitive to assumptions on the nature of f than is the estimation of the $\{f(t_i)\}$. Because the data will not tell us how many derivatives f has, we might settle, in the absence of strong prior information, for linear or spline interpolation among the estimated components of μ. Instead, the Davies and Kovac estimators of f simultaneously determine the estimator of μ and the interpolation scheme by minimizing the number of local extrema in f_n, subject to achieving residuals at the $\{t_i\}$ that behave like white noise. Their idea is refreshingly novel. Comparing the performance of their estimators with linearly interpolated thresholding competitors when f is very wiggly seems to be a natural question.

To consider separately the estimation and interpolation aspects of nonparametric regression clarifies what we can achieve in each respect. In particular, handling the practically important case where the $\{t_i\}$ are not all distinct can begin with treating an unbalanced one-way layout.

3. Loss estimates as diagnostic tool. A statistical practitioner needs credible indications of a procedure's success or failure in analyzing the data at hand. Ensemble results such as minimaxity of a procedure or asymptotic rate of convergence under a probability model do not diagnose adequacy of a procedure applied to specific data, though they may suggest instructive experiments with competing procedures in worst-case scenarios. The residency system for practical training of physicians arose in the second half of the nineteenth century, replacing a system whereby the training of medical practitioners was

mostly theoretical. Surgeons and army doctors who worked with their hands long had lower social status, even as they pioneered important medical procedures. Improved computing environments now encourage developments in statistical diagnostics and statistical training that may parallel, at an abstract level, the evolution of modern scientific medicine.

Feedback about which nonparametric regression procedure to use in a particular data analysis can come from estimated performance summaries, from diagnostic plots and from the substantive field in which the observations were obtained. A broadband diagnostic approach is surely more effective than any single tool. For example, estimated losses sharpen visual scrutiny of competing regression estimators and their residuals by prompting one to *see* why one estimated loss is larger than another. Estimated losses are especially helpful when the domain of the regression function is not one- or two-dimensional.

We discuss here technical aspects of estimating the quadratic loss

$$L_n(\hat{\mu}, \mu) = n^{-1}|\hat{\mu} - \mu|^2 \tag{3}$$

for the estimator $\hat{\mu} = \hat{\mu}(y)$. Let $g(y) = \hat{\mu}(y) - y$. If the errors $\{\varepsilon_i\}$ in the one-way layout are i.i.d. $N(0, \sigma^2)$ and g satisfies assumptions detailed in Stein (1981), then the risk of $\hat{\mu}$ is

$$R_n(\hat{\mu}, \mu, \sigma^2) = \sigma^2 + E\left[2\sigma^2 n^{-1} \sum_{i=1}^{n} \partial g_i(y)/\partial y_i + n^{-1}|g(y)|^2\right]. \tag{4}$$

Let U be an orthogonal matrix whose columns from a basis for R^n and are increasingly wiggly as column index increases. For example, U might be the orthogonal polynomial basis or the discrete cosine basis for R^n. Let $z = U'y$. Some approaches to estimating σ^2 are typified by the formulas,

$$\hat{\sigma}_D^2 = [2(n-1)]^{-1} \sum_{i=2}^{n} (y_i - y_{i-1})^2 \quad \text{and} \quad \hat{\sigma}_H^2 = (n-q)^{-1} \sum_{i=q+1}^{n} z_i^2 \tag{5}$$

and robustifications like (1.7) in the paper. Consistency theorems for these estimators of σ^2 and their variants suggest diagnostics that guide their use.

Let

$$\widehat{L}_n = \hat{\sigma}^2 + 2\hat{\sigma}^2 n^{-1} \sum_{i=1}^{n} \partial g_i(y)/\partial y_i + n^{-1}|g(y)|^2. \tag{6}$$

For consistent $\hat{\sigma}^2$ and and certain classes of estimators $\hat{\mu}$, the loss $L_n(\hat{\mu}, \mu)$ and the risk $R_n(\hat{\mu}, \mu, \sigma)$ converge together as n tends to infinity; and $\widehat{L}_n$ also converges, in an L_1-norm sense, to the common asymptotic value of loss and risk. Further details are given in Beran (2000) and references cited there.

The proposal being made is to consider $\widehat{L}_n$ as a diagnostic tool for assessing nonparametric regression estimators. When $\hat{\mu}(y)$ lacks a tractable closed form, the partial derivatives needed in (6) may be approximated numerically. Let u_i

denote the vector in R^n whose ith component is 1 and whose other components are 0. Then, for small real values of δ,

$$\partial g_i(y)/\partial y_i \approx \delta^{-1}[g_i(y+\delta u_i) - g_i(y)], \qquad 1 \le i \le n. \tag{7}$$

Computing these difference quotients requires computing $\hat{\mu}(y) = y + g(y)$ and the n perturbed estimators $\{\hat{\mu}(y+\delta u_i)\colon 1 \le i \le n\}$.

Like other statistical procedures, diagnostic tools need experimental testing. In trials on pseudo-random artificial data, we may compare the actual loss of a regression estimator with loss estimates such as $\widehat{L}_n$. In the author's experiments, approximate evaluations of $\widehat{L}_n$ on pseudo-Gaussian data have yielded respectable estimates of loss for the James–Stein estimator of μ and for more efficient linear shrinkage and soft-thresholding estimators. The findings suggest trial of $\widehat{L}_n$ as a performance diagnostic for the Davies and Kovac regression estimators and further theoretical investigation of loss estimators outside Gaussian models.

REFERENCES

BERAN, R. (2000). REACT scatterplot smoothers: superefficiency through basis economy. *J. Amer. Statist. Assoc.* **95** 155–171.

BUCKHEIT, J. B. and DONOHO, D. L. (1995). WaveLab and reproducible research. Technical report 474, Dept. Statistics, Stanford Univ. Available at www-stat.stanford.edu/~donoho/.

MCCLUHAN, E. and ZINGRONE, F. (1995). *Essential McLuhan.* House of Anansi Press, Concord Ontario, Canada.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Third Berkeley Symp. Math. Statist. Probab.* **1** 197–208. Univ. California Press, Berkeley.

STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151.

TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

UNIVERSITY OF CALIFORNIA
DEPARTMENT OF STATISTICS
367 EVANS HALL #3860
BERKELEY CALIFORNIA 94720-3860
E-MAIL: beran@stat.berkeley.edu

DISCUSSION

LUTZ DÜMBGEN

Medical University at Lübeck

Laurie Davies and Arne Kovac have written a nice and very stimulating paper on nonparametric regression. They propose a novel approach to curve estimation, combining the traditional notion of estimation error (measured by supremum norm) and the complexity of the fitted function (measured by its modality). My comments are formulated in terms of a "true" underlying regression function f.

Despite its slow rate of convergence, the runs method has its own merits. For if one fixes an upper bound for the number of modes, this method yields a confidence band for the underlying curve f, while the taut string methods provide only point estimators. In addition, at jump points of f the confidence band's horizontal width may become as small as $O_p(\log(n)/n)$. Here one could bring up the multiresolution idea by replacing the runs test statistic (1.4) with a multiscale sign statistic such as

$$\max_{0\le a<b\le n}\left((b-a)^{-1/2}\left|\sum_{i=a+1}^{b}\operatorname{sign}(r_n(t_i))\right|-C_n(b-a)\right),$$

where $C_n(d) := (2\log(n/d))^{1/2}$. This statistic yields confidence bands for monotone median functions with the correct asymptotic width [cf. Dümbgen and Johns (2000)]. My guess is that the authors' method of "stretching to the right (or left)" can be adapted to this test statistic, yielding an algorithm with running time $O(n^2)$. While the bands produced by the runs method are okay in situations with low noise level, the alternative statistic should yield better results if the sample size is large and the data are rather noisy.

As for the taut string method, Theorem 3.1 is a nice example for a nonparametric lower confidence bound for the modality of a curve. In the context of density estimation this procedure was proposed by Donoho (1988), but here is an explicit algorithm for computing this bound together with a candidate $\hat{f}$ for f. I tried this method for density estimation and the results look very promising indeed. The true modality tends to be underestimated quite often, so that some version of local squeezing seems to be appropriate.

The taut string method with local squeezing can be viewed as an ad hoc method for (almost) solving the following optimization problem: find a curve $\hat{f}$ whose total variation is as small as possible (or such that the graph of the function $t\mapsto\int_0^t\hat{f}(x)\,dx$ has minimal length) under the constraint that

$$(*)\qquad \left|\sum_{i=a+1}^{b}(y(t_i)-\hat{f}(t_i))\right|\le c_n(a,\,b)$$

for all integers $0\le a<b\le n$. The authors' multiresolution analysis corresponds to

$$c_n(a,b) := \hat{\sigma}_n(2.5\log(n))^{1/2}(b-a)^{1/2}$$

if $a=j2^d, b=k2^d$ for some integers j,k and $d\ge 0$, and $c_n(s,\,t) := \infty$ else. Now I wonder whether this optimization problem can be solved directly. Furthermore, one might replace the constraints $(*)$ with

$$\left|\sum_{i=a+1}^{b}\operatorname{sign}(y(t_i)-\hat{f}(t_i))\right|\le c_n(a,b).$$

This would lead to a really distribution-free nonparametric method for estimating a median curve which is consistent with the data and as "simple" as possible.

REFERENCES

DÜMBGEN, L. and JOHNS, R. B. (2000). Confidence bands for isotonic median curves via sign tests. Preprint.
DONOHO, O. (1988). One-sided inference about functionals of a density. *Ann. Statist* **16** 1390–1420.

MATHEMATICAL INSTITUTE
MEDICAL UNIVERSITY AT LÜBECK
WALLSTRASSE 40
23560 LÜBECK
GERMANY
E-MAIL: duembgen@math.mu-luebeck.de

DISCUSSION

J. A. HARTIGAN

Yale University

This paper contains interesting new theoretical and practical results for approximating a data set by a function with few extreme values. Davies and Kovac consider three technologies: runs, taut strings and multiresolution evaluation of residuals. I will confine my remarks to the taut strings part of the paper, since that method (referred to as *stretched* strings) was used in Hartigan and Hartigan (1981). Some work on stretched strings for multimodal density estimation appears in Hartigan (2000).

I believe the stretched string fit of a function $f(t_i)$ to data $y(t_i)$ is best expressed by finding the function f that minimizes the criterion

$$\sum_i (y(t_i) - f(t_i))^2 + \lambda \sum_{i=2}^{n} |(f(t_i) - f(t_{i-1})|.$$

Perhaps the penalty function could be the weighted function

$$\lambda \sum_{i=2}^{n} |(f(t_i) - f(t_{i-1})/(t_i - t_{i-1})|.$$

Using the weighted penalty function gives essentially the same mathematics.

It is quite common in smoothing to use instead a squared penalty function such as

$$\lambda \sum_{i=2}^{n} [(f(t_i) - f(t_{i-1})]^2/(t_i - t_{i-1}),$$

which specifies that f be Brownian motion, in a Bayesian framework. Linear kernel estimates of the unknown f follow. So it is interesting to discover the effect of replacing the squared difference by the absolute difference; an obvious effect is that we will be much quicker to accept sharp changes in f.

A probability model that would justify the above criterion would be independent normal errors and a Laplace process for the unknown f. Perhaps if we wish to be insensitive to a few outliers, we should use

$$\sum_i |(y(t_i) - f(t_i))| + \lambda \sum_{i=2}^{n} |(f(t_i) - f(t_{i-1})|.$$

Davies and Kovac define their taut string fit through an optimization on the summed data

$$0, y\left(\frac{1}{n}\right)\Big/n, \left[y\left(\frac{1}{n}\right) + y\left(\frac{2}{n}\right)\right]\Big/n, \ldots, \left[y\left(\frac{1}{n}\right) + \cdots + y\left(\frac{n}{n}\right)\right]\Big/n.$$

The taut string is a curve of minimum length passing between the graph of these sums plus or minus a constant (I call these upper and lower limit graphs the *jaws*) and tied down at the endpoints (0,0), $(1,\bar{y})$. The taut string is a straight line when the constant is large, having slope (the fitted value of f) equal to the mean of the observations.

I agree with the general idea of local squeezing, which is analogous to the idea in kernel density estimation that the kernel width should be varied according to discovered behavior in the underlying density. However, I am uneasy about the multiresolution route for analyzing residuals, because, as the authors acknowledge, it is not connected in any way to the taut string fits.

We are well set to consider the fits for all *operators* λ. We begin with a single straight line stretched string between (0,0) and $(1, \bar{y})$. As the jaws close around the graph of cumulative sums, one or other jaw touches the string and it begins to bend, at each touch adding a new straight line segment to the stretched string. When the aperture is zero, there will be a perfectly accurate fit to the data. Considering all the straight line segments formed along the way, we see that they form a hierarchical tree with a total of $2n - 1$ segments, and taking $O(n \log n)$ to compute.

Our fitting problem is to decide which *partition* of segments selected from that hierarchical tree is the appropriate one to use in the fit. The particular set of segments selected will consist of segments from different stretched string fits. Corresponding to each segment, the fitted value is the mean of the k observations in a segment, adjusted down by λ/k at a maximum, and adjusted up by λ/k at a minimum. Thus we want λ/k to be small to prevent bias, but we want λ large to encourage long segments with small variance.

In Hartigan (2000), for multimodal densities I use the analogue of the following rule to select the partition of segments within each of which the fit is constant: Consider a segment whose endpoints are on the upper cumulative jaws at some λ. A similar process is followed for segments with endpoints on the lower jaws.

The data within that segment defines a cumulative sum process that begins at one endpoint of the segment, terminates at the other endpoint, and always remains above the segment. If indeed the unknown function f were constant in the segment, the asymptotic distribution of this process would be that of an excursion of a Brownian motion. We estimate the variance for the Brownian

motion by the mean squared difference of the successive sums in the segment; we will split the segment (rejecting its inclusion in the final partition) if the maximum excursion is surprisingly large, indicating a nonconstant f within the segment. The maximum of a Brownian excursion that passes through (1,0) [from Kennedy (1976)] has distribution function

$$F(x) = \sum_{-\infty < j < \infty} \exp(-2j^2x^2)(1 - 4x^2j^2).$$

So the proposal for selecting the final partition is as follows:

1. Construct all $2n - 1$ segments for all apertures λ.
2. Eliminate all segments that touch the same jaw twice if the corresponding excursion of partial sums is too large, and eliminate all segments that include them.
3. Select the final partition to consist of the maximal remaining segments.

I offer no theoretical or empirical comparisons of this selection method with the multiresolution method of Davies and Kovac; I advance it because the selection process analysis follows simply from the fitting method.

REFERENCES

KENNEDY, D. P. (1976). The distribution of the Maximum Brownian Excursion. *J. Appl. Probab.* **13** 371–376.

HARTIGAN, J. A. (2000). Testing for antimodes. In *Data Analysis, Scientific Modeling and Practical Application* (W. Gaul, O. Opitz and M. Schader, eds.) 169–181. Springer, New York.

HARTIGAN, J. A. and HARTIGAN, P. (1985). The dip test of unimodality. *Ann. Statist.* **13** 70–84.

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
P.O. BOX 208290
NEW HAVEN, CONNECTICUT 06520-8290
E-MAIL: hartigan@stat.yale.edu

DISCUSSION

ENNO MAMMEN AND SARA VAN DE GEER

Ruprecht-Karls-Universität Heidelberg and University of Leiden

1. This is a nice paper, which we enjoyed reading. As J. Tukey does, the authors follow the practice of decomposing data into two parts. The first part is the signal which is defined as having a simple structure. The second one, denoted "noise," is the remaining part having high complexity and nearly no structure. The crucial point of this approach is to define a simple structure. In this paper, for a one-dimensional function, simplicity is defined as having a small number of modes. This is a very appealing notion of simplicity. However, there also exist other definitions which are more appropriate in certain

applications. For example, in the last century a related notion was introduced in Sprague (1887). In the analysis of mortality tables he defines smoothness (or in the terminology of the paper of P. L. Davies and A. Kovac, simplicity) of a function as the number of modes of the first derivative (or equivalently, the number of concave or convex pieces of the function). For more history on smoothing (in actuarial sciences) and this notion of Sprague, see also Diewert and Wales (1993). By the way, Sprague was also one of the first researchers who considered smoothing as construction of roads through rough terrain. Removing hills and valleys as done in restricting the number of modes could also be motivated by this image. This is nicely illustrated in this paper [see also Mammen and van de Geer (1997)]: taut strings are constructed by moving earth from (local) hills to valleys. Also other measures of smoothness implicitly used in smoothing methods could be used in the approach of Davies and Kovac. The notion of simplicity that is used should depend on the aims of the statistical application.

2(a). In many applications some a priori knowledge on the structure of the signal is available that might be incorporated into the model. Then a simple criterion like the number of modes of the signal is no longer appropriate. Moreover, it may not be the aim of the statistical analysis to find a simple model for the signal.

2(b). For higher-dimensional data sets it is not immediately clear which criterion can be used. If no structure on the high-dimensional signal is assumed, clearly one can also look at the number of modes of the signal. However, for high-dimensional functions the geometric shape is not fully described by the number of modes. An informative qualitative description of the shape would need other shape characteristics.

2(c). The second approach of the paper is based on two steps. In the first step a scale of candidate regression functions is calculated (having different amounts of simplicity). The generation of these functions is done by the taut string method (and modifications). Also, for multivariate signals one could consider a generalization of the taut string method. This could be done if one looks for another more appropriate notion of a simple signal or even if one uses the criterion of number of modes. For univariate data the taut string estimate is given by the minimizer of

$$\sum_{i=1}^{n}[Y_i - f(t_i)]^2 + \lambda \mathrm{TV}(f), \tag{1}$$

where $\mathrm{TV}(f)$ is the total variation of the function f and λ is a smoothing parameter. A possible generalization of the taut string method to multivariate data is given by the minimizer of (1) where now $\mathrm{TV}(f)$ is the higher-dimensional total variation of the function f. A discussion of this estimator can be found in Mammen and van de Geer (1997). A related approach has

been proposed by Kuensch (1994) in the context of image restoration. He looks at (monotone transforms of) the absolute differences between values of f at neighboring points and uses their sum as penalty term. The question remains how to define what is a simple signal in higher dimensions [see 2(b)].

3. The taut string method is based on minimization of penalized least squares. Other losses could be used as well. Then one would look at the minimizers $\hat{f}_\lambda$ of

$$G(\lambda, f) = \sum_{i=1}^{n} \rho[Y_i - f(t_i)] + \lambda \mathrm{TV}(f), \tag{2}$$

where ρ is an appropriate loss function. For a large class of functions ρ the minimizer $\hat{f}_\lambda$ can be taken as a piecewise constant function with number of modes decreasing in λ. At a piece $J = \{k, \cdots, k+m\}$ the value $\hat{\gamma}$ of $\hat{f}_\lambda$ is defined as the minimizer of $\sum_{i \in J} \rho[Y_i - \gamma] + c\gamma$ where the constant c depends on the value of $\hat{f}_\lambda$ in the neighboring pieces. If $\hat{f}_\lambda$ is monotone c is equal to 0, if $\hat{f}_\lambda$ has a local extreme in the piece then c is equal to $+2\lambda$ or -2λ. For convex ρ, the value of $\pm 2\lambda$ leads to a down shifting at local maxima and to up shifting at local minima. This coincides with the behavior of the taut string estimate. A popular choice of ρ would be $\rho(x) = |x|$. Then $\hat{f}_\lambda$ is a local median that is down- or up-shifted at local extremes. The estimate can be easily calculated by using standard L_1 fitting routines. This can be done by minimizing the L_1 norm between $(\theta_1, \ldots, \theta_n, \lambda[\theta_2 - \theta_1], \ldots, \lambda[\theta_n - \theta_{n-i}]$ and the artificial observation vector $(Y_1, \ldots, Y_n, 0, \ldots, 0)$. Here we write $\theta_i = f(t_i)$. Alternatively, one could directly apply linear programming methods; see also the next point. It would be natural to use these L_1 fits instead of taut strings in case one has a model (a test bed) with median zero errors. In particular, it would be more appropriate for heavy tailed errors like Cauchy errors, therefore a good alternative to the run method of the paper.

4. Minimizers of (2) have been proposed in the literature on regression quantiles for the modified penalty $\mathrm{TV}(f')$. With the total variation of the derivative of f instead of f, the resulting minimizer is now a piecewise linear function. This penalty corresponds to Sprague's definition of smoothness. The estimates can easily be calculated by linear programming; see Koenker, Ng and Portnoy (1994). For asymptotic theory see also Portnoy (1997) and van de Geer (2000). Using this estimate in the approach of Davies and Kovac would be more satisfactory than Sprague's method. He proposed just drawing curves by hand.

REFERENCES

DIEWERT, W. E. and WALES, T. J. (1993). A "new" approach to the smoothing problem. Technical report.

KOENKER, R., NG, P. T. and PORTNOY, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680.

KUENSCH, H. R. (1994). Robust priors for smoothing and image restoration. *Ann. Inst. Statist. Math.* **46** 1–19.

MAMMON, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist* **25** 387–413.

PORTNOY, S. (1997). Local asymptotics for quantile smoothing splines. *Ann. Statist.* **25** 414–434.

SPRAGUE, T. B. (1887). The graphic method of adjusting mortality tables (with discussion). *J. Inst. Actuaries* **26** 270–285.

VAN DE GEER, S. (2000). *M*-estimation using penalties or sieves. *J. Statist Plann. Inference.* To appear.

INSTITUT FÜR ANGEWANDTE MATHEMATIK
RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELS 294
69120 HEIDELBERG
GERMANY
E-MAIL: mammen@statlab.uni-heidelberg.de

MATHEMATICAL INSTITUTE
UNIVERSITY OF LEIDEN
P.O. BOX 9512
2300 RA LEIDEN
NETHERLANDS

DISCUSSION

J. S. MARRON

University of North Carolina, Chapel Hill

There are at least two levels on which this paper is very interesting to read. First the underlying philosophical view is not standard in the mathematical statistics community. Second, a fascinating alternative paradigm for developing smoothing methods is presented.

I agree that the underlying concept "there is no true underlying regression curve, only data," is not discussed much in the mathematical statistics literature. Why is this? Perhaps it comes from the fact that mathematics requires assumptions, and a "true curve" is an assumption that leads to a lot of useful and insightful mathematics. In particular, this assumption, labelled as "test bed," leads to both the interesting mathematics and the convincing simulations of the present paper.

However, the idea of "what we are doing is only an approximation" does deserve the highlighting it has gotten here. How can this important concept be harnessed, and used to address serious problems? One approach is the scale space view of kernel smoothing, discussed in Chaudhuri and Marron (2000) and used for statistical inference in Chaudhuri and Marron (1999). Assuming the existence of a true underlying curve for a moment, an important part of the scale space view is that there is not enough information in the data to completely estimate the true curve. Hence, one should instead focus on other "targets" which reflect what aspects of the true curve are obtainable from the data. This seems to be a useful position that lies in between the two sides of the debate as to whether or not a "true underlying curve exists."

The specifics on the interesting new approaches to curve estimation are connected to some ideas of Tukey. The "Tukey terminology" can be carried one

step further, in that Tukey sometimes used the terms "smooth and rough" for what is presently called "signal and noise." The new approach seems to be unique in that it strives to control the properties of the rough (e.g., the sign change run length), and then optimize the smooth subject to that constraint. This viewpoint is reminiscent of smoothing splines [see, e.g., Wahba (1990) or Green and Silverman (1994)], where an approach is to control the smoothness of the smooth and then to minimize the rough subject to that constraint. The former is in some sense an unexpected and deep reversal of the latter.

Can something be gained from combining these ideas? Section 1.3 stresses that smoothness is not an issue presently under discussion, which is fair enough for the present first pass at this approach. However, in future work it may be sensible to try to put these together.

A potentially exciting future area is suggested in Section 1.2: "confidence sets are also possible." Development of new smoothing methods, and calculating their rates of convergence is fun. However, the real value added by statistical methods in the actual analysis of data comes not from suggesting a smoothing method, but instead from providing meaningful inference. For example, an often central issue is which features (e.g., bumps or spikes) in a smooth are really there (as opposed to being sampling artifacts)? The present approach has potential in this direction, and it would be very good see it developed in future work.

The one perhaps unsettling point of the proposed methodology is the choice of the threshold 2.5 in (1.6). It seems clear that 2 will create difficulties, but why 2.5 as opposed to 2.2 or 3? The "it worked well in the few examples that we tried" suggestion is not very satisfying. Perhaps this problem could be addressed using an analog of the scale space approach to kernel smoothing: instead of focussing on a single threshold, present the family of curve estimates for a range of thresholds.

REFERENCES

CHAUDHURI, P. and MARRON, J. S. (1999). Sizer for exploration of structure in curves. *J. Amer. Statist. Assoc.* **94** 807–823.

CHAUDHURI, P. and MARRON, J. S. (2000). Scale space view of curve estimation. *Ann. Statist.* **28** 408–428.

GREEN, P. and SILVERMAN, B. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman and Hall, London.

WAHBA, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-3260
E-MAIL: marron@stat.unc.edu

REJOINDER

P. L. DAVIES AND A. KOVAC

Universität Essen

We thank all the discussants for their contributions and the interest shown in our work.

The main motivation for this paper was to put forward a view of statistics based on the idea of approximation and exemplify this in the context of an interesting statistical problem. Although most if not all statisticians regard probability models as useful pragmatic approximations, this is not reflected in statistical theory or terminology or, to a large extent, in statistical practice. As an example we mention that the additivity of beliefs in Bayesian statistics is based on truth. If we accept the approximate nature of statistical models then it seems quite natural to use it in a consistent manner. However, as far as we are aware, this idea seems to be new. In some unpublished comments on an earlier version of Davies (1995), Tukey (1993b) states that the "emphasis on approximation is well-chosen and surprisingly novel." We have no closely knit paradigm to offer such as the Bayesian one and no "principles" of statistical inference. Indeed, we do not think there are any apart from a critical "distanced" view of one's own statistical theory and practice. This is not to say that the approach has no consequences. We think it has many, although they may not be obvious at first sight. One immediate consequence is the strict separation of data and model mentioned by Beran. Given that the model is only an approximation to the data, the two must be strictly separated. This separation is not new. In the context of the two-way table, we find on page 254 of Tukey (1993a) the following equations:

$$y_{ij} = t + a_i + b_j + c_{ij}, \tag{*}$$

and

$$y_{ij} = \tau + \alpha_i + \beta_j + \gamma_{ij}. \tag{**}$$

Tukey writes

> ... there is a great danger, almost an overriding certainty, that the conventually trained will look at (∗) and see, or think seriously about (∗∗) with conventional probabilistic assumptions about the Greek "population" quantities.

Earlier he writes

> The conventional approach would have been unchallengeable IF, (a) IT WOULD INDEED BE CARRIED THROUGH, which may or may not be possible, and (b) ITS ASSUMPTIONS WERE SECURELY CORRECT, which they never are. One ... but also with an unmentioned (unmentionable) logical gap between the assumptions and the data.

We also emphasize the role of procedures (again advanced by Tukey) by which we probably mean algorithms with default values for all nonobvious parameters. It makes no sense to restrict a statistical procedure to data which fulfil certain assumptions such as the ubiquitous "i.i.d." triad. There is simply no way in which such an assumption can be verified. Instead we should use both empirical experiments and theoretical investigations to understand the strengths and weaknesses of a procedure. We use here the words "test bed" which correspond to Tukey's "challenge." We again quote Tukey (1993a), page 250:

> Theoretical results typically have assumptions. However applicable procedures, even those suggested by theoretical results, are typically *never* used where theory's assumptions apply exactly and in detail. Thus applicable procedures do not themselves have assumptions—only some circumstances in which they work (i.e. serve our purposes) better, and some in which they work less well. ... I know some will think these statements heretical, but I find no escape from them and their implications.

Beran makes a strong case for analyzing procedures on test beds where they begin to break down. He has made explicit suggestions for such experiments in the context of nonparametric regression and we intend to carry them out. He has made further interesting suggestions about other more theoretical investigations of the performance of a procedure. They could be perhaps be so tuned to measure the accuracy with which peaks are identified and not only whether the correct number is found.

Beran remarks quite correctly that our simulated data sets use pseudorandom variables. With the possible exception of the spectroscopy data they are all deterministic. This raises the interesting question as to the status of randomness. On the one hand there are deterministic sequences such as the prime numbers which can, in a well-defined sense, be *proved* to behave randomly. We refer to Kac (1959) who writes "the primes, indeed, play a game of chance." On the other hand the Bohmian version of quantum mechanics opens up the possibility that the universe is a super-deterministic system. We refer to Bell (1987), Albert (1992), Berndl, Daumer, Dürr, Goldstein and Zanghi (1995) and Goldstein (1998a,b). None of these pose any difficulties of interpretation if one abandons the concept of truth for statistical models and replaces it by approximation. We very much welcome Beran's contribution, which is in the same spirit as our paper.

Marron mentions the usefulness of the idea of a "true" regression function f. There is no doubt that stochastic models give useful insights for analyzing data. As Beran mentions, both (1.4) and (1.6) are based on stochastic models. They can also suggest concrete procedures such as using the mean squared error to derive a procedure for the choice of the local bandwidth for kernel estimates. We do not think that it is necessary even in this case to assume the truth of a model. Everything can be formulated in terms of approximation. We mention in passing that local bandwidth selection for kernel estimators

can also be done using a multiresolution analysis of the residuals. We are not sure what a position between a "true" f and an "approximate" f would look like. Marron calls the reversing of the roles of smooth and rough "unexpected." From the point of view of approximation this approach seems quite natural. Our attitude is the following. Given a data set, a fully or sometimes a partially specified model and a concept of approximation it should be decidable whether or not the model is an approximation to the data in the stated sense. Suppose that the model is

$$Y(t) = f(t) + \sigma\varepsilon(t)$$

with $\varepsilon(t)$ denoting standard Gaussian white noise. We now specify a function f and ask whether this is an adequate approximation to the given data $(x(t_i), y(t_i))$, $1 \le i \le n$. To do this we form the residuals $r(t_i) = y(t_i) - f(t_i)$ and check whether they satisfy the multiresolution condition (1.6) with σ_n given by (1.7). This can be done and checked for any given function f. Indeed, one can bet on whether a given f is a good approximation to the data set or not. Moreover, these bets can be called in as they are decidable. The odds however do not give rise to a probability measure over the parameter space. If required, the concept of approximation can be made to include the value of σ. The requirement of specifying exactly what is meant by an approximation includes a specification of the approximation for the random part of the model. From this point of view it seems natural to define an approximation to the noise and then to maximize simplicity for the signal.

The choice of 2.5 was pragmatic. For some data sets in the area of spectroscopy a choice of 1.8 led to results more in line with the expectations of the chemists. It is unlikely that theoretical investigations will lead to a correct value of the threshold and so it will probably remain a pragmatic choice.

Dümbgen restricts his remarks to the two approximization problems. As he notes, the run method provides bounds for regression functions rather than an explicit candidate. The run bounds are obtained by inverting a run test for independence. He suggests another possibility which will in general lead to tighter bounds and which has a complexity of $O(n^2)$. For many data sets this will be sufficient and it may be possible to obtain approximations for larger data sets. He points out correctly that the taut string method is an ad hoc attempt to solve the multiresolution approximation problem and suggests that it may be possible to solve it explicitly. Since writing the paper we have occasionally had simulated data sets where the taut string method, even with local squeezing, has produced a superfluous local extreme. It is therefore of interest to solve the problem exactly. It may also be easier to analyze the behavior of the exact procedure rather than that of the taut string procedure.

Mammen and van de Geer discuss the simplicity criterion. There were two reasons for choosing the number of local extremes. First, we wanted to remove small local fluctuations which many methods exhibit. One can overlook these with the eye but we felt it was a challenge to remove them. Second, there are data sets where the detection of peaks is of prime interest. One example is spectroscopy data. In a forthcoming paper we discuss the problem of densities.

Much of the work done in this area is based on L^2 errors although the ensuing examples are discussed in terms of detecting, or more often, not detecting peaks. Other choices of simplicity are clearly possible. One is to minimize the number of intervals where the function takes on different values. In higher dimensions some thought will have to be given to the form of the geometric constraints. These will depend on the type of data which are to be analyzed.

Mammen and van de Geer and also Hartigan point out that the taut string with a tube of constant diameter is the solution of the following minimization problem:

$$\sum_1^n (y(t_i) - f(t_i))^2 + \lambda \sum_2^n |f(t_i) - f(t_{i-1})|, \tag{1}$$

where λ is related to the diameter of the tube. As mentioned by Dümbgen, the taut string is an ad hoc method for solving the multiresolution problem and in this respect solutions of (1) are not good enough. This is shown, for example, by the Bumps data. It was this failure of the raw taut string method which lead us to look for improvements. The result was the local squeezing method and it was only then that we felt the results were worth publishing. Dümbgen has pointed out to us that local squeezing as implemented in the algorithm may be seen as the solution of the minimization problem

$$\sum_1^n (y(t_i) - f(t_i))^2 + \sum_2^n \lambda_i |f(t_i) - f(t_{i-1})| \tag{2}$$

with data driven λ_i.

Mammen and van de Geer discuss other variants such as

$$\sum_1^n |y(t_i) - f(t_i)| + \lambda \sum_2^n |f(t_i) - f(t_{i-1})|. \tag{3}$$

We have already experimented with solutions of (3) and the results are very promising. What we would like to do is to solve the multiscale sign statistic problem posed by Dümbgen; it seems plausible that solutions of (3) will be related to it. However, it may again be the case that (3) must be replaced by

$$\sum_1^n |y(t_i) - f(t_i)| + \sum_2^n \lambda_i |f(t_i) - f(t_{i-1})| \tag{4}$$

before acceptable results are obtained. Mammen and van de Geer say that solutions of (3) may provide an alternative to the run method. One has to be careful about this. Solutions of (3) can be used as a data analytical tool by varying λ and this is indeed suggested by Mammen and van de Geer (1997) in the context of solutions of (1). If one is interested in a well-defined statistical procedure then some form of automatic choice of λ must be given and the performance of the procedure will be determined by this choice. If the choice is the maximal run length of the residuals then solutions of (1) will have at least as many local extremes as the run method provides. Thus solutions of (1) can only be judged in combination with some automatic choice of λ.

Hartigan points out that the taut string method is not related to the multiresolution scheme. This is indeed the case. As mentioned above we regard the taut string with local squeezing as an ad hoc method for solving at least approximately the multiresolution problem. Our interpretation of his method is that he suggests a different definition of approximation to white noise which is based on the partial sums process of the residuals, the behavior of which is approximated by that of the limiting Brownian motion. He then suggests a scheme for splitting the intervals defined by the taut strings. We are not sure whether this will control the number of local extremes of the resulting regression function but it is certainly an interesting and novel idea which deserves further investigation.

We would like to thank all the discussants for their contributions. We must also thank Dümbgen and Hartigan for pointing out an error in our initial definition of the taut string. This has been corrected in the paper.

REFERENCES

ALBERT, D. (1992). *Quantum Mechanics and Experience*. Harvard Univ. Press.

BELL, J. (1987). *Speakable and Unspeakable in Quantum Mechanics*. Cambridge Univ. Press.

BERNDL, K., DAUMER, M., DÜRR, D., GOLDSTEIN, S. and ZANGHI, N. (1995). A survey of Bohmian mechanics. *Nuovo Cimento B* **110** 737–750.

DAVIES, P. L. (1995). Date features. *Statist. Neerlandica* **49** 184–245.

GOLDSTEIN, S. (1998a). Quantum theory without observers I. *Physics Today* 42–46.

GOLDSTEIN, S. (1998b). Quantum theory without observers II. *Physics Today* 38–42.

KAC, M. (1959). *Statistical Independence in Probability, Analysis and Number Theory*. Wiley, New York.

MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413.

TUKEY, J. (1993a). Exploratory analysis of variance as providing examples of strategic choices. In *New Directions in Statistical Data Analysis and Robustness* (S. Morgenthaler, E. Ronchetti and W. A. Stahel, eds.) Birkhäuser, Basel.

TUKEY, J. (1993b). Issues relevant to an honest account of data-based inference, partially in the light of Laurie Davies' paper. Princeton Univ.

UNIVERSITÄT GESAMTHOCHSCHULE ESSEN
FB06 MATHEMATIK UND INFORMATIK
D-45117 ESSEN
GERMANY
E-MAIL: arne.kovac@uni-essen.de