Modality, Runs, Strings and Wavelets

P. L. Davies and A. Kovac

March 11, 1999

Abstract

The paper considers the problem of non-parametric regression with emphasis on controlling the number of local extrema. Two methods, the run method and the taut string-wavelet method, are introduced and analysed on standard test beds. It is shown that the number and location of local extreme values are consistently estimated.Rates of convergence are proved for both methods. The run method has a slow rate but can withstand blocks as well as a high proportion of isolated outliers. The rate of convergence of the taut string-wavelet method is almost optimal and the method is extremely sensitive being able to detect very low power peaks.
Section 1 contains a short introduction with special reference to modality. The run method is described in Section 2 and the taut string-wavelet method in Section 3. Low power peaks are considered in Section 4. Section 5 contains a short conclusion and the proofs are given in Section 6.

1 Introduction

1.1 Approximate models

Given a data set $(t_i, y(t_i)), i = 1, \ldots, n$ in $[0,1] \times \mathbb{R}$ we require a simple description of the form

$$y(t_i) = f_n(t_i) + r_n(t_i) \tag{1}$$

where f_n is a function belonging to some specified class of functions and the $r_n(t_i), i = 1, \ldots, n$ are the resulting residuals. We assume that the design points t_i are equidistant with $t_i = \frac{i}{n}$. At the expense of additional complexity more general designs can be considered. As posed the problem is ill-defined: there are many functions f_n and residuals r_i which satisfy (1). To make the problem tractable we follow Tukey and write

$$\text{Data} = \text{Signal} \oplus \text{Noise}. \tag{2}$$

In the context of non-parametric regression the decomposition (2) can be made precise using the model

$$Y(t_i) = f(t_i) + \varepsilon(t_i) \tag{3}$$

where the $\varepsilon(t_i)$ are i.i.d. random variables with a given distribution. The signal is identified with the function f and the noise with the residuals $\varepsilon(t)$. Carrying this over to real data we identify the signal with the function f_n and the noise with the residuals $r_n(t_i)$. The two can be separated by assuming that the signal is simple and that the noise is complex. In this paper the simplicity of any function g on $[0,1]$ is the number of local extreme values. We use two definitions of complexity for the noise. The first is the length of the longest run of the signs $\mathrm{sgn}(r_n(t_i))$ of the residuals and the second on a wavelet reconstruction of the residuals. They are explained in more detail below. Each of the definitions of complexity can be regarded as a definition of approximation to white noise. The main problem is to produce simple functions f_n, that is ones with low modality, such that the resulting residuals approximate white noise. The run method yields a natural lower bound for the modality as well as bounds for the approximating functions. The wavelet definition of approximation cannot be naturally inverted to give a lower bound for the modality. To produce candidate functions of low modality we use the unrelated taut string method and call the resulting procedure the taut string-wavelet procedure. The above ideas are related to a concept of approximation of stochastic models as developed in Davies (1995) to which we refer for further details.

1.2 Test beds

Non-parametric procedures may be evaluated using real data sets or under the well controlled conditions of a stochastic model .Test beds are defined by stochastic models of the form (3). They have the advantage of allowing a direct comparison of the function f used to generate the data with the function f_n which the procedure yields as an adequate approximation. Such comparisons are often based on rates of convergence of f_n to f in a norm such as

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

For certain forms of test beds it can be shown that optimal rates of convergence exist. For certain functions f we show that the run based procedure has a slow rate of convergence namely $\mathrm{O}(\frac{\log\log n}{\log n})$ whilst the taut string-wavelet procedure has the optimal rate of convergence of $\mathrm{O}((\log n/n)^{-1/3})$. In spite of this there are well defined test beds with $f = f^n$ depending on n on which the run procedure is asymptotically superior to some procedures which are optimal for fixed f.

1.3 Smoothness

Smoothness is not a consideration in this paper. The regression functions our procedures provide are piecewise constant and consequently not even continuous. Techniques for smoothing such functions under shape and deviation constraints have been developed in Metzner (1997) and Davies and Löwendick (1999).

1.4 Previous work

Much work has been done on the problem of non-parametric regression. Of the different approaches we mention kernel estimation (Nadaraya, 1964, Watson, 1964), penalised likelihood (Silverman, 1984), wavelets (Donoho et al., 1995) and local polynomials (Fan and Gijbels, 1996). None of these methods is directly concerned with modality but recently much work has been done which explicitly takes the shape of the regression function into account. One of the examples in van de Geer (1990) is concerned with monotonic regression. 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. However none of these papers is concerned with estimating modality or the positions of the local extreme values. Work in this direction has been done by Dümbgen (1998b) who applies linear rank tests simultaneously in order to find local extrema of an unknown regression function. Hengartner and Stark (1995) use the Kolmogoroff ball centred at the empirical distribution function to obtain non-parametric 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 for determining the modality of a regression function is that of mode testing. We refer to Good and Gaskins (1980), Silverman (1986), Hartigan and Hartigan (1985), Fisher, Mammen and Marron (1994). The position of the local extreme values is 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. Finally Mächler (1995) presents an approach using a roughness penalty to penalize points of inflection.

The run method (see Davies (1995) and 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 modality consistent with the observations as well as approximation intervals for the location of the extreme values. Dümbgen (1998a,1998b) inverts other tests and obtains better convergence rates under 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 et al., 1972; Leurgans, 1982). The idea of using taut strings for densities goes back to Hartigan and Hartigan (1985) who derived a test for modality of a density. In Davies (1995) it was explicitly used to calculate approximate densities. It was first used in the general non-parametric regression problem by Mammen and van de Geer (1997). They showed the taut string is a special case of a penalised least squares functional where the penalty is based on the total deviation norm. They also give a description of the taut string as well as an $O(n)$ algorithm independently of the number of local extrema. The reason the taut string is useful for providing candidate functions of low modality is that it has the smallest modality of all functions in the supremum ball.

2 The run method

2.1 Obtaining bounds for regression functions using runs

The idea of using runs to obtain bounds for adequate regression functions was introduced by Davies (1995). We give a short description of the method: for a more detailed description see Metzner (1997). Let $\Delta_i, i = 1, \ldots, n$ denote a sequence of independently distributed random variables with

$$\mathbf{P}(\Delta_i = 1) = \mathbf{P}(\Delta_i = -1) = \frac{1}{2}.$$

We denote by R_n the length of the longest subsequence of the Δ_i which is composed entirely of 1's or -1's. For any given $\alpha, 0 < \alpha < 1$ we denote the α-quantile of R_n by $qu(n, \alpha, R_n)$ that is

$$qu(n, \alpha, R_n) = \min\{m : \mathbf{P}(R_n \leq m) \geq \alpha)\}.$$

For large n we have the simple approximation

$$qu(n, \alpha, R_n) = \lceil \log_2 n - \log_2(-\log((1 + \alpha)/2)) \rceil - 1. \tag{4}$$

whic can be deduced from the results given in Section XIII.7 of Feller (1968).

Consider the data set $((t_i, y(t_i)), i = 1, \ldots, n)$ and allow a maximal run length for the signs of the residuals of $qu(n, \alpha, R_n)$. We look for an adequate regression function which is initially non-increasing and which satisfies the run condition. An upper bound at the design point t_i is given by

$$u_i(n, \alpha) = \min\{u_{i-1}(n, \alpha), \max\{y(t_j) : i - qu(n, \alpha, R_n) \leq j \leq i\} \tag{5}$$

and a lower bound is given by

$$l_i(n, \alpha) = \max\{l_{i+1}(n, \alpha), \min\{y(t_j) : i \leq j \leq i + qu(n, \alpha, R_n)\}. \tag{6}$$

We set $u_i(n, \alpha) = \infty$ and $l_i(n, \alpha) = -\infty$ if the minimum in (5) and the maximum in (6) are not taken over $qu(n, \alpha, R_n) + 1$ observations. This results in infinite bounds at local minima and maxima and at the beginning and end of the data set. If the upper and lower bounds intersect at some point then a non-increasing function can be no longer adequate. At this point a switch is made to a non-decreasing function for which upper and lower bounds are constructed in the analogous manner. This process is continued to the end of the sample. The process is repeated this time starting with non-decreasing upper and lower bounds and those bounds are chosen which minimize the number of local extreme values. It gives an lower bound for the number of local extreme points required for an adequate regression function as well as upper bounds for the location of the local extremes. This procedure is called "stretching to the right" in Davies (1995). The reverse procedure starting with the last data point and proceeding to the first is known as "stretching to the left". It can be shown that the procedures yield the same lower bound for the number of local extreme values and that together they gives lower and upper bounds for the location of these local extremes values. At local maxima the upper bound is ∞ and at local minima the lower bound is $-\infty$. Once the lower bound for the number of local extrema is known, tighter bounds for the regression function can be obtained by reducing the allowable run length to the minimum consistent with this number of local extrema. The bounds (5) and (6) may be seen as the result of inverting a run test for the independence of the residuals. Dümbgen (1998a,1998b) has inverted other non-parametric tests and shown that more powerful tests can be made to yield optimal rates of convergence. They are computationally more exacting whereas the bounds based on the run test can be calculated very quickly.

An approximate regression function may be obtained by taking the midpoints of the bounds where these are both finite. Occasionally this may not fulfil the run condition: a somewhat more complicated procedure described by Metzner (1997) gives a regression function which always fulfils the run condition. We denote this function by f_n^R. Within the limits imposed by the bounds the user is free to specify the location and size of local extreme values as well as the boundary values. We do not pursue this any further: a discussion of the possibilities is contained in Davies and Löwendick (1999).

2.2 Behaviour on fixed function test beds

We consider a fixed function test bed of the form (3). Let k_n^α denote the number of local extreme value of f_n^R based on the α-quantile of the length of the longest run. The intervals where either the lower bound or the upper bound are infinite 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 behaviour of f_n^α is covered by the following theorem:

Theorem 2.1 *Consider the test bed (3) where*

- *f has a bounded continuous first derivative $f^{(1)}$ and exactly k local extreme values at the points $0 < t_1^e < \ldots < t_k^e < 1$*
- *the $\varepsilon(t)$ have median zero and a bounded continuous density function in the neighbourhood of zero .*

Then the following hold:

(a) there exist a constant $A > 0$ such that for all $\delta > 0$

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

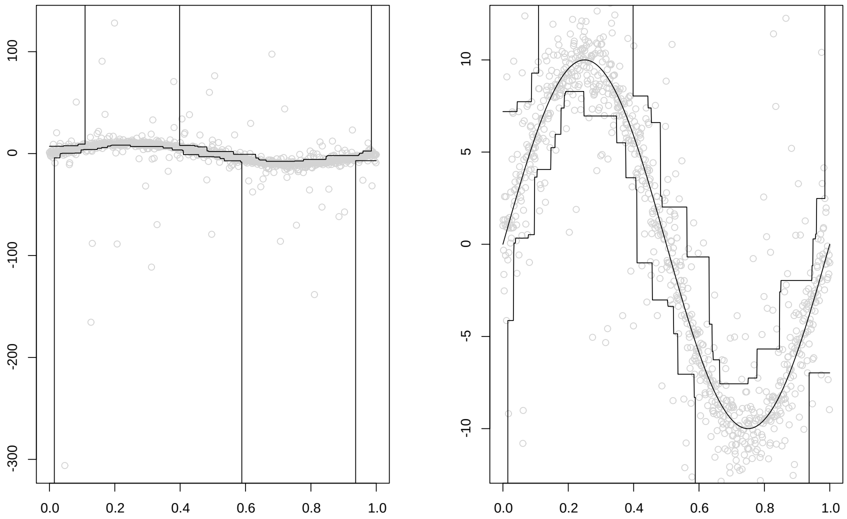

Figure 1: Bounds for the run method. The left panel shows the unscaled data. The right panel shows only part of the data scaled to reveal the underlying curve and the bounds provided by the run method.

(b) for all $\delta > 0$

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

The rate of convergence of the above procedure on fixed function test beds is very slow but in spite of this it gives surprisingly good results as we show below.

2.3 An example

Figure 1 shows the result of applying the run method to the sine curve $y = 10\sin x$ on $[0,1]$ contaminated with Cauchy noise. The sample size is $n = 1000$ and the maximal allowable run length is to 15 corresponding to the approximation (4) with $\alpha = 0.95$. It demonstrates the ability of the method to deal with many isolated outliers.

3 The taut string and wavelet method

3.1 Strings and supremum tubes

In the last section it was shown how the use of runs to define approximation to white noise leads to bounds for the modality and for the set of adequate approximating functions

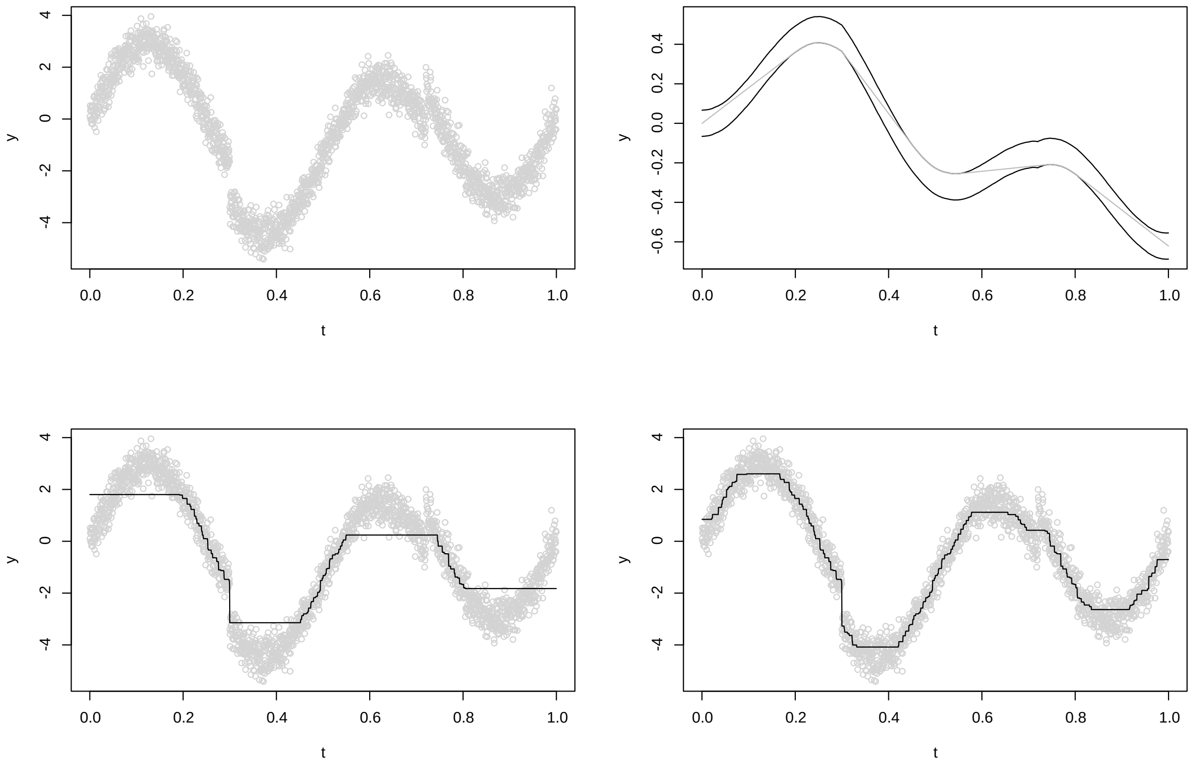

Figure 2: These figures illustrate the taut string method. In the upper left corner depicts the Heavisine function contaminated with Gaussian noise. The upper right panel shows the tube which is obtained by integrating the data points and translating by ε and $-\varepsilon$. Also shown is a taut string which is constrained to lie within the tube. The lower left panel shows the contaiminated Heavisine function and the derivative of the taut string. Finally the lower right figure shows the result of another application of the taut string method where the bandwidth is set to $1.146\sigma/\sqrt{n}$.

of this modality. No such method is available for the concept of approximation based on wavelets. The method we employ to obtain adequate functions of low modality is based on a taut string constrained to lie in a supremum neighbourhood of the integrated process y_n°

$$y_n^\circ(t_j) = \sum_{i=1}^{j} y(t_i)(t_{i+1} - t_i), \quad j = 1, \dots, n-1 \tag{7}$$

with linear interpolation between the design points The derivative of the taut string is used as a regression function. This procedure is illustrated in Figure 2. The Heavisine signal (Donoho et al, 1995) is disturbed by white noise as shown in the upper left corner. The tube obtained from the integrated process and the taut string can be seen in the upper right panel and the lower left panel shows the resulting regression function.

We denote the taut string defined by the supremum tube $T(g, \rho)$

$$T(g, \rho) = \{h : \sup_{0 \le t \le 1} |g(t) - h(t)| \le \rho\} \tag{8}$$

with centre g and radius ρ by $S_{g\rho}$ and the derivative by $s_{g,\rho}$. The connection with modality is the following. If the supremum tube with centre g and radius ρ contains the integrated function

$$f^\circ(t) = \int_0^t f(u)du \tag{9}$$

then the modality of $s_{g,\rho}$ is at most that of f.

Mammen and van de Geer (1997) give a description of the taut string. In particular it is piecewise linear so that the resulting regression function is piecewise constant. They consider the class of all functions with finite total variation and show that a bandwidth of order $n^{-2/3}$ attains the optimal rate of convergence. In the case of white noise contamination this choice of bandwidth leads asymptotically to infinitely many local extrema. We show below that bandwidths of order $n^{-1/2}$ allow the number of local extreme values to be estimated consistently. The lower right panel in Figure 2 shows the taut string method to the Heavisine signal contaiminated with Gaussian white noise. The bandwidth is $1.146\sigma/\sqrt{n}$. The choice $C = 1.146$ corresponds to the 0.5 quantile of the maximum of the absolute value of a Brownian motion (see Freedman, (34) Proposition, page 27).

The next two theorems are concerned with the asymptotic behaviour of the taut string on the test bed (3). In particular it is shown to have an optimal rate of convergence away from the local extrema. This is not surprising as it then coincides with the least squares solution. Of more interest is its behaviour at the local extrema themselves. We denote the modality of the derivative of the taut string in the supremum tube $T(f_n^\circ, C/\sqrt{n})$ by k_n^C. The taut string based on the radius $C/\sqrt{n}$ will be denoted by $S_n^\circ = S_n^\circ(C)$ with derivativs S_n. We write $I_i^e(n, C), 1 \le i \le k$ for the intervals where S_n attains it local extreme values and denote the midpoints of these intervals by $t_n^e(n, C), 1 \le i \le k$.

Theorem 3.1 *Consider the test bed (3) where*

- *f has a bounded derivative $f^{(1)}$ with $|f^{(1)}| \neq 0$ except at the k local extrema,*
- *the errors $\varepsilon(t_i)$ have mean 0 and a bounded second moment.*

Then for all $\delta > 0$

$$\lim_{C\to\infty} \liminf_{n\to\infty} \mathbf{P}(\{k_n^C = k\} \cap \{\max_{1\le i\le k} |I_i^e(n,C)| \le \delta\} \cap \{\max_{1\le i\le k} |t_i^e(n,C) - t_i^e| \le \delta\}) = 1.$$

In the following A denotes a generic constant which depends only on f and the distribution of the $\varepsilon(t)$ and whose value may differ from appearance to appearance. The length of an interval I will be denoted by $|I|$. The knots of the taut string are the points where it touches the sides of the tube.

Theorem 3.2 *Consider the test bed (3) where*

- *f has a bounded second derivative $f^{(2)}$ which is non-zero at the k local extremes,*
- *the first derivative $f^{(1)}$ of f is non-zero except at the local extreme values,*
- *the errors $\varepsilon(t_i)$ have mean 0 and are sub-Gaussian i.e.$\mathbf{E}(\exp(\lambda\varepsilon(t_i))) < \exp(\mu\lambda^2)$ for some $\mu > 0$ and for all $\lambda > 0$.*

Then

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

(b)
$$|I_i^e(n,C)| \sim (6C)^{1/3} |f^{(2)}(t_i^e)|^{-1/3} n^{-1/6}.$$

(c) If x_i and x_{i+1} are successive knots in $[0,1] \backslash \cup_1^k I_i^e(n,C)$ then

$$x_{i+1} - x_i = O_{\mathbf{P}}\left(|f^{(1)}(x_i)|^{-2/3} \left(\frac{\log n}{n}\right)^{1/3} \right).$$

(d) At points $t \in [A\left(\frac{\log n}{n}\right)^{1/3}, 1 - A\left(\frac{\log n}{n}\right)^{1/3}] \backslash \cup_i^n I_i^e(n,C)$ we have

$$|f(t) - S_n(t)| \le O_{\mathbf{P}}\left(|f^{(1)}(t)|^{1/3} \left(\frac{\log n}{n}\right)^{1/3} \right).$$

(e) At points $t \in \cup_1^n I_i^e(n,C)$ we have

$$|f(t) - S_n(t)| \le (1 + o_{\mathbf{P}}(1)) A C^{2/3} n^{-1/3}.$$

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 $n^{-2/5}$ (see 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 $y(t_i)$. Nevertheless the change of boundaries at local extrema does alter the behaviour of the wavelet coefficients described in the following section.

3.2 Strings and wavelets

In the last section a function is regarded as an adequate approximation for the data if it has minimal modality and its integral lies in the supremum tube $T(Y_n^\circ, C/\sqrt{n})$. We introduce another more direct and also more restrictive concept of approximation based on the test bed situation (3) with the additional assumption that the errors $\varepsilon(t)$ are Gaussian white noise. A regression function f_n will be regarded as an adequate approximation to the data if the residuals

$$r_n(t) = y_n(t) - f_n(t) \tag{10}$$

"look like" Gaussian white noise in the sense that their wavelet reconstruction after using some form of thresholding is zero. Amongst all such approximations we seek the simplest f_n i.e. those with minimal modality. The wavelets we use are the Haar wavelets which are the easiest to handle. We restrict attention to the case where n is a power of 2. The general case may be handled using the techniques of Kovac and Silverman (1998). The usual threshold level is

$$\sigma_n \sqrt{2 \log n} \tag{11}$$

where σ_n is some robust measure of the scale of the highest level wavelet coefficients. A popular choice and one we shall us is $\sigma_n = 1.48MAD$ where MAD denotes the median absolute value of all wavelet coefficients. The threshold (11) is extremely tight and and can often be obtained only at the cost of increasing the modality. We use a less restrictive level

$$\sigma_n \sqrt{\tau \log n} \tag{12}$$

where reasonable values of τ can be obtained by simulations and the analysis of real data sets. The value $\tau = 2.5$ has proved to be adequate over a wide range of situations.
We consider firstly a modification of the taut string through the tube $T(Y_n^\circ, C/\sqrt{n})$. At local extreme values the taut string changes boundaries and a simple analysis shows that this effect alone will cause some of the wavelet coefficients of the residuals to exceed the threshold level in the limit regardless of the value of τ in (12). This may be rectified by the simple expedient of replacing S_n by $\tilde{S}_n$ where $\tilde{S}_n = S_n$ except at local extrema of S_n where $\tilde{S}_n$ is defined to be the mean of the $Y_n(t_i)$ values on the interval of constancy. The integrated $\tilde{S}_n$ will be denoted by $\tilde{S}_n^\circ$. This alteration neither alters the modality or the rate of convergence. The behaviour of the wavelet coefficients for candidate functions based on $\tilde{S}_n$ is covered by the next theorem.

Theorem 3.3 *For each constant $A > 0$ there exist a $\tau > 0$ such that under the thresholding (12) the following hold.*

(a) The coefficients of all wavelets whose support is of length at most $A(\log n/n)^{1/3}$ are eventually set to zero.

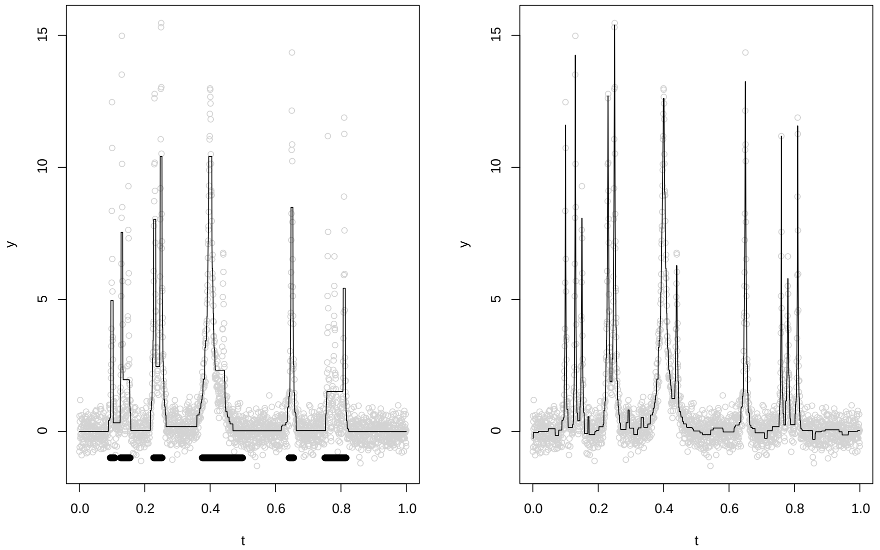

Figure 3: These figures show the application of the taut string methods to the Bumps signal. In the left panel the bandwidth is $1.146\sigma/\sqrt{n}$ and the lines at the bottom indicate regions where wavelet coefficients are not set to zero. The right panel shows the global squeezing procedure where the band is squeezed until all wavelet coefficients are set to zero.

(b) *At each local extremum t_i^e of f there exists an interval $J_i^e(n, C)$ of length at most $A \cdot C^2/\log n$ and which contains t_i^e such that the coefficients of all wavelets whose support is not contained in one of the $J_i^e(n, C)$ are eventually set to zero.*

Theorem 3.3 indicates that the suboptimal rate of convergence at local extrema will be detected by the wavelet decomposition of the residuals. This is shown using the Bumps data. Figure 3 shows the intervals where the wavelet coefficients of the residuals are not set to zero after thresholding. It also shows the effect of squeezing globally until all wavelet coefficients are zero: it results in several spurious local extreme values. In the next section we describe how local squeezing may be used to improve the rate of convergence at local extrema and with the result that the correct modality is obtained.

3.3 Local squeezing and wavelets

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 the problem cannot be solved by decreasing the width of the tube globally. The calculations used in the proof of Theorem 2.4 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 wavelet coefficients will now be set to zero. Figure 3 demonstrates this effect. It the right panel the tube has been squeezed until all wavelet coeffcients are zero resulting in many spurious local extreme values. This may be countered by decreasing the width of the tube only in an $n^{-1/5}$-neighbourhood of the local extrema. This is as it stands not feasible as the taut string with radius $n^{-1/2}$ locates the local extrema only with an accuracy of order $n^{-1/6}$. The remaining possibility is to decrease the width of the tube on the intervals of order $n^{-1/6}$ which contain the local extreme points. We have

$$\mathbf{P}\left(\max_{1\le i\le m}\frac{1}{\sqrt{m}}|\sum_{j=1}^{i}\varepsilon(t_j)|\ge\lambda\right)\le\frac{\sigma^2}{\lambda^2}$$

which implies the following. Suppose a local extreme of f is located at $t^e=\frac{i_e}{n}$. On putting $m=m_n=Cn^{5/6}$ we obtain

$$\mathbf{P}\left(\max_{i_e-m_n\le i\le i_e+m_n}\left|\frac{1}{n}\sum_{j=i_e-m_n}^{i}\left(Y_n\left(t_i\right)-f\left(t_i\right)\right)\right|\ge\lambda\frac{\sqrt{C}}{n^{7/12}}\right)\le\frac{\sigma^2}{\lambda^2}.$$

The width of the tube is now $n^{-7/12}$ compared with the optimal width 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 width of $n^{-2/5}$. We now show that decreasing the width until the wavelet coefficients are set to zero after thresholding using (12) does result in an almost optimal rate of convergence.

We consider the taut string using a tube of radius $Cn^{-1/2}$ and the modified derivative $\tilde{S}_n$. As shown above with high probability this will result in the correct modality and all the extreme points will lie in the corresponding intervals of $\tilde{S}_n$ whose lengths will be of order $n^{-1/6}$. Furthermore after thresholding all wavelet coefficients will be set to zero apart from some in shrinking neighbourhoods of the local extreme points. We now gradually squeeze the tube locally at the intervals where $\tilde{S}_n$ has local extreme values. The squeezing is continued until after thresholding all wavelet coefficients are set to zero.

To analyse this procedure consider a piecewise constant function f_n with the correct number of local extreme values and such that the local extreme points of f are contained in the corresponding intervals where f_n has its local extreme values. We suppose that after thresholding the wavelet coefficients of the residuals are set to zero and consider what implications this has for the maximal deviation of f_n from f. Consider firstly a point t where both f and f_n are monotone increasing and let $d_n=f_n(t)-f(t)>0$. A short calculation shows that

$$\int_{I_n(t)}(f_n(u)-f(u))du\ge A\cdot d_n^2/f^{(1)}(t).$$

This implies that there exists a wavelet with support of length $\frac{1}{2}d_n \le l \le 2d_n$ and whose coefficient is at least $A \cdot l^2/f^{(1)}(t)$. This will be set to zero after thresholding if and only if

$$l^2/f^{(1)}(t) \le \sqrt{\frac{l \log n}{n}}$$

which is equivalent to

$$d_n \le A|f^{(1)}(t)|^{2/3} \left(\frac{\log n}{n}\right)^{1/3}. \tag{13}$$

To investigate the behaviour at local extremes we consider an interval I of length l. A short calculation shows

$$\min_{I,a}\{\int_I \max\{0, t^2 - a\}dt, \int_I \max\{0, a - t^2\}dt\} \ge A \cdot l^3$$

for some universal constant A. Using this we see that there exists a wavelet whose support is of length h and whose coefficient is at least

$$A_1 \cdot l^3 - A_2\sqrt{\frac{-l \log l}{n}}. \tag{14}$$

If

$$h \ge A \left(\frac{\log n}{n}\right)^{1/5}.$$

then (14) is at least

$$A \left(\frac{l \log n}{n}\right)^{1/2}$$

which means that this wavelet will not be set to zero after thresholding. Putting all this together gives the following theorem:

Theorem 3.4 *Consider the model (3) as in Theorem 3.2 and let f_n denote a sequence of piecewise constant functions for which the following hold:*

- *the local extremes of f are contained in the local extremes of f_n*
- *between local extremes f_n has the same monotonic behaviour as f*
- *the lengths of the intervals where f_n has its local extreme values are of order at least $(\log n/n)^{1/5}$*
- *after thresholding the wavelet coefficients of the residuals $Y_n - f_n$ are all set to zero.*

Then at points t outside the intervals where f_n attains its local extreme values

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

At points t in the intervals where f_n attains its local extreme values

$$|f(t) - f_n(t)| = O\left(\left(\frac{\log n}{n}\right)^{2/5}\right).$$

Theorems 3.2 and 3.4 taken together suggest the following procedure. Choose a $C > 0$ and consider the resulting functions $\tilde{S}_n$. Apply local squeezing at the local extreme points of $\tilde{S}_n$ until after thresholding all the wavelet coefficients are set to zero. The resulting regression function will exhibit the optimal rates of convergence at all points outside the intervals containing the local extrema. At such points the rate of convergence will differ only by a term of order $(\log n)^{2/5}$ from the optimal rate. A practical implementation of this procedure will be described below.

3.4 A practical implementation

The taut string-wavelet method may be implementated as follows. Firstly, the standard deviation of the noise is estimated by taking the median absolute deviation of the finest scale wavelet coefficients of the raw data divided by 0.6745 (see Donoho et al.). The next step is to choose a large initial global bandwidth γ_0. A reasonable choice is such that the straight line connecting $(0,0)$ and $(1,(1/n)\sum y_i$ lies inside the tube. An alternative is to use Theorem 2.2 and use an initial bandwidth $\gamma_0 = C/\sqrt{n}$ where C is chosen to be the 0.95-quantile of the distribution of the maximum of a Brownian motion. Given the initial width the taut string is calculated. The idea is to construct simultaneously the greatest convex minorant and the smallest concave majorant of the first m points of the lower and the upper bounds respectivelywhere both curves are forced to start in $(0,0)$. The number m is successively increased until the minorant and majorant intersect at a point (u_1, v_1). The procedure is repeated with starting point (u_1, v_1) instead of $(0,0)$ and continued until the right end of the interval $[0,1]$ is reached. The complexity is $O(n)$.

Having calculated the taut string we use the modified version $\tilde{s}^0$of the derivative s^0 as described just before Theorem 3.3. If the wavelet reconstruction of the residuals $r^*(t_i) = y(t_i) - \tilde{s}^0(t_i)$ the algorithm terminates. If not the intervals I_j are determined where $r_i^* \neq 0$. More precisely $r^*(t_i) \neq 0$ if and only if $t_i \in \cup_j I_j$. Each interval is extended until its end points are knots of the taut string, ie for each j we determine the largest knot u_j less than or equal to the minimum of I_j and the smallest knot v_j greater than or equal to the maximum of I_j. This defines subintervals $J_j \supset I_j$ and for each j the taut string procedure applied to the data on the interval J_j gives a local model $\tilde{s}_j^0$, this time using the smaller bandwidth $\gamma_1 = q\gamma_0$. The squeezing factor q can be chosen as 0.9 of 0.95 if one is cautious or 0.5 if one wants to speed up the calculations at the risk of overestimating the modality. A new candidate function s^1 is finally defined by the results of the taut string procedure, ie $s^1(t_i)$ is set to $\tilde{s}_j^0(t_i)$ if $t_i \in J_j$ and $s^1(t_i) = s^0(t_i)$ if $t_i \in [0,1] \setminus \cup_j J_j$. Simple considerations show that this definition is sensible near at the end points of each interval J_j. This is made more precise by the following lemma.

Lemma 3.1 *If the taut string touches the upper (lower) bound in u_j, then $s^1(u_j) \geq s^0(u_j)$ ($s^1(u_j) \leq s^0(u_j)$). If the taut string touches the upper (lower) bound in v_j, then $s^1(v_j) \leq s^0(v_j)$ ($s^1(v_j) \geq s^0(v_j)$).*

The algorithm proceeds by calculating the wavelet reconstruction of the new residuals. If again some wavelet coefficients are still not set to zero further squeezing is applied, this time with band width $\gamma_2 = q^2\gamma_0$, until eventually the reconstruction of the residuals is identically zero.

An implementation of this procedure requires that the threshold constant τ of (12) be specified. Simulations and evaluation of real data sets indicate that $\tau = 2.5$ is a satisfactory choice.

3.5 Examples

Donoho and Johnstone (1994) use four test signals to demonstrate the behaviour of wavelet thresholding. The results of applying the taut string-wavelet procedure to these functions are shown in Figure 4
The Doppler signal is displayed in the upper left corner and shows a sinusoidal curve which is smooth in the right half and very rough in the left half of the data. 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 panels 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 signal is a 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 noisy bumps signal is to be compared with Figure 3. Local squeezing has identified all the relevant peaks without introducing spurious ones. Between the relevant peaks the taut string-wavelet function is almost zero.

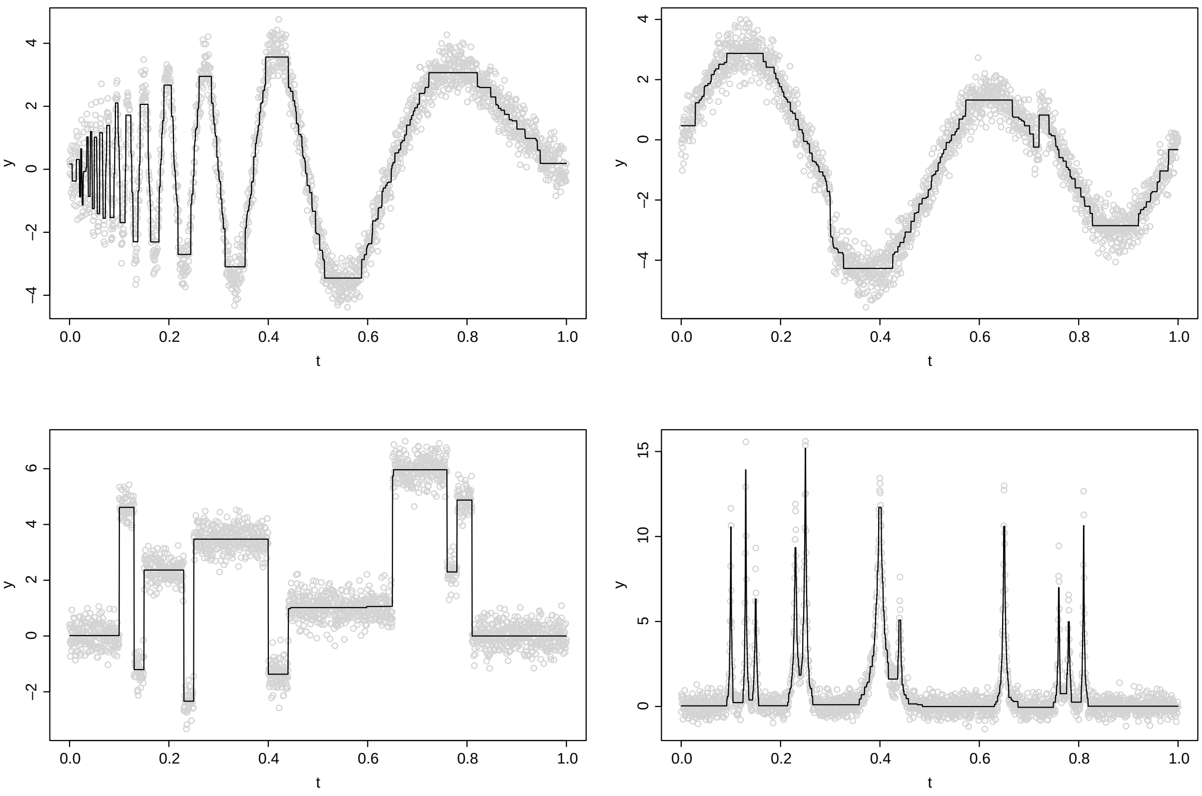

Figure 4: Four noisy test signals and their reconstructions using the taut string procedure with local squeezing.

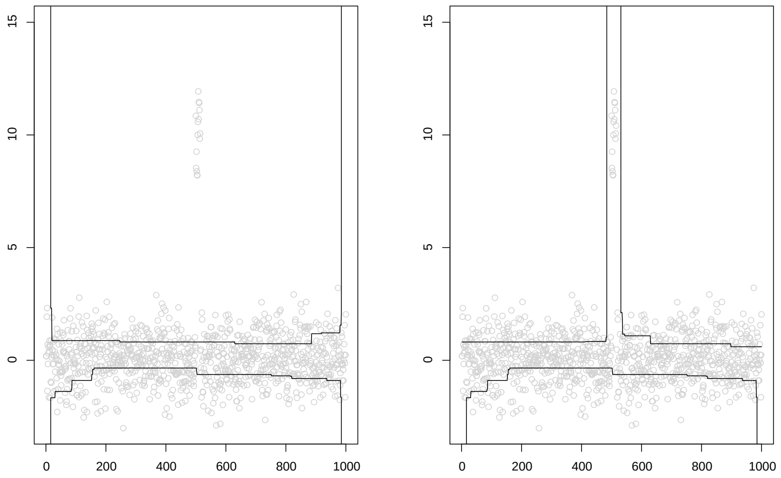

Figure 5: The data show the function $f = 0$ with a small peak of height $h_n = 10$ locates at 0.5 contaimnated with Gaussian noise. The sample size is $n = 100$. In the left panel the peak is defined by 15 observations and it is not detected by the run method with $qu(n, 0.95, R_n) = 15$. The left panel shows the same data but with one extra point in the peak. It is now detected by the run method.

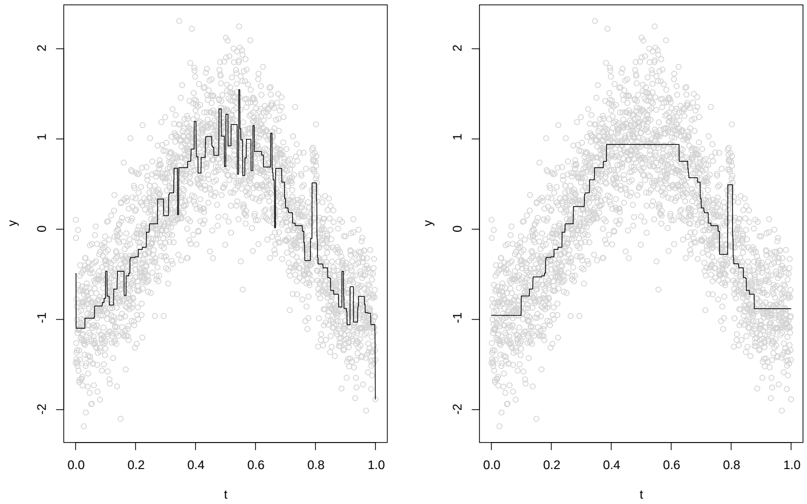

Figure 6: A regression function corrupted with Gaussian noise and the reconstructions based on $n = 2048$ observations. The left panel shows the effect of globally squeezing until the peak near 0.8 is detected. The right panel shows the result of the local squeezing procedure.

4 Low power peaks

Most work on non-parametric regression and density problems evaluates procedures in terms of rates of convergence on test beds of the form (3). The distribution of the errors $\varepsilon(t)$ is specified and the regression function f is kept fixed whilst the number n of observations is increased. In this situation there exist optimal rates of convergence (Khas'minski (1978), Ibragimov and Khas'sminski (1980) and 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 of specified modality. The run method falls well short with a rate of convergence of order $(\log\log n)/\log n$. In spite of this the run base 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 we denote the new function by f^n. The power of the peak is defined to be

$$L_n = \int_0^1 |f - f^n| = \int_{a_n}^{b_n} |f - f^n|. \tag{15}$$

Consider now the asymptotic test bed

$$Y_{n,i} = f^n\left(\frac{i}{n}\right) + \varepsilon_{n,i} \tag{16}$$

with $L_n = \mathrm{o}\left(n^{-1/2}\right), \lim_{n\to\infty}\sqrt{n}(b_n - a_n) = 0$ 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 modality k so will the taut string through the ball. From this it follows that Theorems 3.1 and 3.2 will continue to hold for the asymptotic test bed (16). In other words the taut string method will fail to identify the peak at $[a_n, b_n]$. If $L_n = \mathrm{o}\left(n^{-2/3}\right)$ then the Mammen and van de Geer taut string method will not be able to distinguish between the test beds (3) and (16).

The behaviour of the run method on the test bed (16) is as follows: if

$$n(b_n - a_n) \geq qu(n, \alpha, R_n)+$$

then the peak will be detected whereas if

$$n(b_n - a_n) \leq qu(n, \alpha, R_n)$$

it will not be detected. Figure 5 shows this effect. The allowable run length is 15 ((4) with $\alpha = 0.95$). The left panel shows a low power peak defined by 15 observations which is ignored. If one additional observation is included in the peak then it is detected as shown by the left panel. This property of the run method is useful as it means that blocks of outliers can be dealt with. Indeed the length of the allowable run can be specified not in terms of α but in terms of the minimum block length of possible outliers or in terms of the required modality.

The local squeezing method using wavelets will clearly identify the peak of f^n. This is demonstrated by Figure 6. Indeed the method will pick up a signal confined to one single observation as long $h_n \geq B\sqrt{\log n}$ where B depends on f and the threshold parameter τ. This is demonstrated by Figure 7 where a very low power peak defined by a single observation is detected without introducing any spurious peaks. Whether a low power peak represents a signal or the effect of outliers is a question that cannot be decided by statistics alone.

5 Conclusion

We have introduced two methods for obtaining regression functions whilst keeping the modality under control. Each method has advantages and disadvantages. The run method can be calculated quickly in $\mathrm{O}(n)$ operations and it has certain desirable robustness properties. It can withstand many islolated 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

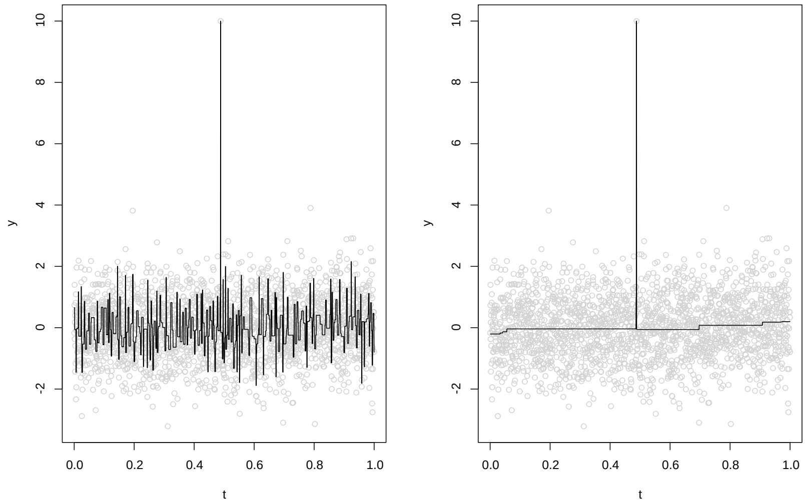

Figure 7: A regression function corrupted with Gaussian noise and the reconstructions based on $n = 2048$ observations. The left panel shows the effect of globally squeezing until the peak near 0.8 is detected. The right panel shows the result of the local squeezing procedure.

string-wavelet method has a complexity of $\mathrm{O}(n \log n)$ and has almost optimal rates of convergence on standard test beds. It is extremely sensitive being 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://wwwstat.mathematik.uni-essen.de`.
We are indebted to Lutz Dümbgen and Martin Löwendick for useful comments.

6 Proofs

6.1 Proof of Theorem 2.1

We analyse the behaviour of the upper bound $u_i(n, \alpha)$ defined by (5). We have from (4)

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

Let

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

If F denotes the common distribution function of the $\varepsilon(t_j)$ 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

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

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

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

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

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

The upper bound $u_i(n, \alpha)$ is non-increasing and we firstly analyse its behaviour on an interval where f where $f^{(1)}(t) > \delta > 0$. Without loss of generality we set $I = [0, \frac{j}{n}]$. The above estimates show that

$$u_i(n, \alpha) \le f(0) + \frac{A \log\log n}{\log n}, i = \frac{\log\log n}{\log n}. \tag{17}$$

The corresponding result for the lower bound $l_i(n, \alpha)$ is

$$l_i(n, \alpha) \ge f\left(\frac{j}{n}\right) - \frac{A \log\log n}{\log n}, i = j - \frac{\log\log n}{\log n}. \tag{18}$$

As both the upper and lower bounds are non-increasing (17) and (18) imply

$$A\frac{\log\log n}{\log n} + f(0) \ge f\left(\frac{j}{n}\right) - A\frac{\log\log n}{\log n}. \tag{19}$$

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

$$\frac{j}{n} \le A\frac{\log\log n}{\log n}. \tag{20}$$

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

The case where the bounds have the same monotone behaviour 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 behaviour of the bounds which minimizes the number of local extreme values will coincide with the monotonic behaviour 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 (17) and (18) implies the rate of convergence of (b) of the theorem. □

6.2 Proof of Theorem 3.1

Let σ denote the standard deviation of the $\varepsilon(t)$. The assumptions of the theorem imply that

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

where $\Rightarrow$ denotes weak convergence and W denotes the standard Brownian motion on $C[0,1]$. In particular we have

$$\lim_{n\to\infty} \mathbf{P}\left(\max_{0\le t\le 1} |\sqrt{n}\varepsilon_n^\circ(t)| \le x\right) = \mathbf{P}\left(\max_{0\le t\le 1} |W(t)|\right) = 2\Phi\left(\frac{x}{\sigma}\right) - 1$$

where Φ denotes the distribution function of a standard Gaussian random variable. It follows that on the test bed (3)

$$\lim_{n\to\infty} \mathbf{P}\left(\max_{0\le t\le 1} |Y_n^\circ(t) - f(t)| \le \frac{C}{\sqrt{n}}\right) = 2\Phi\left(\frac{C}{\sigma}\right) - 1.$$

As n tends to infinity the probability that the function f lies in the tube $T(Y_n^\circ, C/\sqrt{n})$ tends to one. As the taut string minimises the modality in $T(Y_n^\circ, C/\sqrt{n})$ we see that

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

In the other direction 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 $[0,]$ of modality at most j. This implies

$$\lim_{n\to\infty} \mathbf{P}(k_n^C < k) = 0$$

so that

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

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

6.3 Proof of Theorem 3.2

The proof of Theorem 2.3 relies on the modulus of continuity of the integrated process ε_n° as expressed by

$$\lim_{\delta\to 0}\mathbf{P}\left(\sup_{0\le t,t+h\le 1,h\le\delta}\sqrt{n}|\varepsilon_n^\circ(t+h)-\varepsilon_n^\circ(t)|\le A\sqrt{-\delta\log\delta}\right)=1 \tag{21}$$

for some $A>0$. This follows from the sub-Gaussian form of the $\varepsilon(t)$.

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

$$h_0 = \text{argmax}_{0\le h\le\delta}\frac{Y_n^\circ(t_1^l(n,C)+h) - Y_n^\circ(t_1^l(n,C)) - \frac{2C}{\sqrt{n}}}{h} \tag{22}$$

for arbitrarily small δ as n tends to infinity. On writing $t_1^l(n,C) = t_1^e - \kappa$ we may rewrite (22) to obtain

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

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

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

This implies

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

and as $f^{(2)}(t_1^e)<0$ we may conclude $3\kappa\le 2h_0(1+\mathrm{o}(1))$ which implies $t_1^e < t_1^r(n,C)$. Similarly $t_1^e \ge t_1^l(n,C)$ which proves that t_1^e lies in $I_1^e(n,C)$. The proofs for the other intervals are analogous.

Proof of (b):
We suppose that S_n has a local maximum on $I_1^e(n,C) = [t_1^l(n,C), t_1^r(n,C)]$. As

$$|t_1^l(n,C) - t_1^e| = \mathrm{o}(1)$$

the modulus of continuity of ε_n° implies that we may replace (22) by

$$h_0 = \text{argmax}_{0\le h\le\delta}\frac{Y_n^\circ(t_1^e+h) - Y_n^\circ(t_1^e) - \frac{2C}{\sqrt{n}}(1+\mathrm{o}(1))}{h}.$$

A Taylor expansion and the modulus of continuity of ε_n° shows that this reduces to

$$h_0 = \text{argmax}_{0\le h\le\delta}\left(\frac{1}{6}h^2 f^{(2)}(t_1^e)(1+\mathrm{o}(1)) - \frac{2C}{\sqrt{nh}}(1+\mathrm{o}(1))\right)$$

and hence

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

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 $x_2 = (i_1 + l_1)/n$ and $x = l/n$ for $x_1 \le x \le x_2$ 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

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

The modulus of continuity (21) of $\sqrt{n}\varepsilon_n^\circ$ implies that large n

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

uniformly in t and $l \ge 1$. On setting $l = a|f^{(1)}(x_1)|^{-2/3} n^{2/3} (\log n)^{1/3}$ we have for

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

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

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

may be neglected. This implies that for a sufficiently large

$$\frac{Y_n^\circ(x_1 + \frac{l}{n}) - Y_n^\circ(x_1)}{\frac{l}{n}} \ge f(x_1) + \frac{1}{4}a\left|f^{(1)}(x_1)\right|^{1/3}\left(\frac{\log n}{n}\right)^{1/3}. \tag{23}$$

The lower bound

$$\frac{Y_n^\circ(x_1+\frac{l}{n})-Y_n^\circ(x_1)}{\frac{l}{n}} \le f(x_1)+a|f^{(1)}(x_1)|^{1/3}\left(\frac{\log n}{n}\right)^{1/3}. \tag{24}$$

is obtained analogously. On putting $a=a_1$ in (23) and $a=a_2$ in (24) 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

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

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 (21) gives for $l=Af^{(1)}(x_i)^{-2/3}n^{2/3}(\log n)^{1/3}$

$$S_n(x_i)\le f(x_i)+A|f^{(1)}(x_i)|^{1/3}\left(\frac{\log n}{n}\right)^{1/3}. \tag{25}$$

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 (25) gives

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

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

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

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. $\square$

6.4 Proof of Theorem 3.3

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

$$\begin{aligned} w_{j,k} = \frac{\sqrt{n}}{2^{j/2}} \cdot \Bigg(& -Y_n^\circ\left(\frac{k2^j - 2^{j-1}}{n}\right) + \tilde{S}_n^\circ\left(\frac{k2^j - 2^{j-1}}{n}\right) + Y_n^\circ\left(\frac{k2^j}{n}\right) - \tilde{S}_n^\circ\left(\frac{k2^j}{n}\right) \\ & - Y_n^\circ\left(\frac{(k+1)2^j}{n}\right) + \tilde{S}_n^\circ\left(\frac{(k+1)2^j}{n}\right) + Y_n^\circ\left(\frac{k2^j + 2^{j-1}}{n}\right) - \tilde{S}_n^\circ\left(\frac{k2^j + 2^{j-1}}{n}\right)\Bigg). \end{aligned} \tag{26}$$

To show that the wavelet reconstruction of the noise is zero using (12) for some $\tau > 0$ it is sufficient to show that each wavelet coefficient $w_{i,k}$ is at most $\sigma\sqrt{\tau \log n}$. This in turn follows from the inequality

$$|Y_n^\circ(t_l) - \tilde{S}_n^\circ(t_l) - (Y_n^\circ(t_i) - \tilde{S}_n^\circ(t_i))| \le A\sqrt{|t_l - t_i|}\sqrt{\frac{\log n}{n}} \tag{27}$$

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

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

and on using the modulus of continuity of $\sqrt{n}\varepsilon_n^\circ$ as given by (21) we obtain for any points t_i and t_l with $|t_l - t_i| \ge 1/n$

$$\begin{aligned} |Y_n^\circ(t_l) - \tilde{S}_n^\circ(t_l) - (Y_n^\circ(t_i) - \tilde{S}_n^\circ(t_i))| \ \le\ & A\sqrt{|t_l - t_i|}\sqrt{\frac{\log n}{n}} \\ & + |f^\circ(t_l) - \tilde{S}_n^\circ(t_l) - (f^\circ(t_i) - \tilde{S}_n^\circ(t_i))|. \end{aligned} \tag{28}$$

As

$$
\begin{aligned}
|f^\circ(t_l) - \tilde{S}_n^\circ(t_l) - (f^\circ(t_i) - \tilde{S}_n^\circ(t_i))| &= |\int_{t_i}^{t_l} (f(t) - \tilde{S}_n(t))dt| \\
&\leq |t_l - t_i| \sup |f(t) - \tilde{S}_n(t)|
\end{aligned}
$$

(c) and (d) of Theorem 3.2 imply

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

if

$$
|t_l - t_i| \leq A \left(\frac{\log n}{n}\right)^{1/3}. \tag{29}
$$

Thus (27) holds for all intervals satisfying (29).
Proof of (b):
Because of (a) and Theorem 3.2 (c) all wavelets whose supports are contained between two knots which do not delimit a local extreme values will be set to zero. It remains to consider the case where none of the points defining the wavelet lies in an interval $I_i^e(n, C)$. Consider two such points $t_i < t_l$ and let x_i denote the first knot to the right of t_i and x_l the knot to the left of t_l. As

$$
Y_n^\circ(x_l) - Y_n^\circ(x_i) = \tilde{S}_n^\circ(x_l) - \tilde{S}_n^\circ(x_i) \tag{30}
$$

it follows from (28) that it is sufficient to prove

$$
|f^\circ(x_i) - \tilde{S}_n^\circ(x_i) - (f^\circ(t_i) - \tilde{S}_n^\circ(t_i))| \leq A|x_i - t_i|^{1/2}\sqrt{\frac{\log n}{n}}. \tag{31}
$$

As

$$
|f^\circ(t_m) - \tilde{S}_n^\circ(t_m) - (f^\circ(t_i) - \tilde{S}_n^\circ(t_i))| \leq \int_{t_i}^{t_m} |f(x) - \tilde{S}_n(x)|dx
$$

it follows from Theorem 3.2 and the fact that x_i and t_i are not separated by one of the intervals where $\tilde{S}_n$ has a local extreme value that

$$
|f^\circ(x_i) - \tilde{S}_n^\circ(x_i) - (f^\circ(t_i) - \tilde{S}_n^\circ(t_i))| \leq A|f^{(1)}(x_i)|^{1/3}|x_i - t_i| \left(\frac{\log n}{n}\right)^{1/3}. \tag{32}
$$

Theorem 3.2 (c) implies

$$|x_i - t_i| \le A|f^{(1)}(x_i)|^{-2/3}| \left(\frac{\log n}{n}\right)^{1/3}$$

and hence

$$|x_i - t_i| \left(\frac{\log n}{n}\right)^{1/3} \le A|f^{(1)}(x_i)|^{1/3}|x_i - t_i|^{1/2}\sqrt{\frac{\log n}{n}}.$$

On substituting this in inequality (32) we obtain (28) which in turn implies (27).

Finally suppose that two of the points defining the wavelet are both contained in one of the intervals $I_i^e(n, C)$. Using (b) and (e) of Theorem 3.2 and the arguments just given we see that the wavelet coefficient $w_{j,k}$ will be set to zero if

$$Cn^{-1/3}n^{-1/6} = n^{-1/2} \le A\sqrt{\frac{h \log n}{n}}$$

where h denotes the length of the wavelet. This reduces to $h \le A \cdot C^2 / \log n$. $\square$

References

[1] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions.* Wiley, New York, 1972.

[2] P. L. Davies and M. Loewendick. On smoothing under bounds and geometric constraints. Technical report, University of Essen, 1999.

[3] P.L. Davies. Data features. *Statistica Neerlandica*, 49:185–245, 1995.

[4] M. Delecroix, M. Simioni, and C. Thomas-Agnan. A shape constrained smoother: simulation study. *Computational Statistics*, 10:155–175, 1995.

[5] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Pic ard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society*, 57:371–394, 1995.

[6] L. Dümbgen. Application of local rank tests to nonparametric regression. Medical University, Lübeck, 1998.

[7] L. Dümbgen. New goodness-of-fit tests and their application to nonparametric confidence sets. *Annals of Statistics*, 26:288–314, 1998.

[8] J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications.* Chapman and Hall, London, 1996.

[9] W. Feller. *An Introduction to Probability Theory and its Applications: Volume 1.* Wiley, New York, 3 edition, 1968.

[10] N. I. Fisher, E. Mammen, and J. S. Marron. Testing for multimodality. *Computational Statistics and Data Analysis*, 18:499–512, 1994.

[11] D. Freedman. *Browniam Motion and Diffusion.* Holden-Day, San Francisco, 3 edition, 1971.

[12] I.J. Good and R.A. Gaskins. Density estimating and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, 75:42–73, 1980.

[13] J.A. Hartigan and P.M. Hartigan. The dip test of unimodality. *Annals of Statistics*, 13:70–84, 1985.

[14] N.W. Hengartner and P.B. Stark. Finite-sample confidence envelopes for shape-restricted densities. *Annals of Statistics*, 23:525–550, 1995.

[15] I.A. Ibragimov and R.Z. Khas'minskii. On non-parametric estimation of regression. *Soviet Math. Dockl.*, 21:810–814, 1980.

[16] R.Z. Khas'minskii. A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory Prob. Appl.*, 23:794–798, 1978.

[17] A. Kovac and B. W. Silverman. Extending the scope of wavelet regression methods by coefficient-dependent thresholding. Technical report, Sonderforschungsbereich 475, 5/1998, 1998.

[18] S. Leurgans. Asymptotic distributions of slope-of-greatest-convex-minorant estimators. *Annals of Statistics*, 10:287–296, 1982.

[19] E. Mammen. Nonparametric regression under qualitative smoothness assumptions. *Annals of Statistics*, 19:741–759, 1991.

[20] E. Mammen, J.S. Marron, B.A. Turlach, and M.P. Wand. A general framework for constrained smoothing. unpublished, 1998.

[21] E. Mammen and C. Thomas-Agnan. Smoothing splines and shape restrictions. unpublished, 1998.

[22] E. Mammen and S. van de Geer. Locally adaptive regression splines. *Annals of Statistics*, 25:387–413, 1997.

[23] M. Mächler. Variational solution of penalized likelihood problems and smooth curve estimation. *Annals of Statistics*, 23:1496–1517, 1995.

[24] L. Metzner. *Facettierte nichtparametrische Regression*. PhD thesis, Universität Essen, Essen, Germany, 1997.

[25] M. C. Minotte. Nonparametric testing of the existence of modes. *Annals of Statistics*, 25:1646–1660, 1997.

[26] M. C. Minotte and D. W. Scott. The mode tree: a tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics*, 2:51–68, 1993.

[27] E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 10:186–190, 1964.

[28] J. Ramsay. Estimating smooth monotone functions. *Journal of the Royal Statistical Society*, 60:365–375, 1998.

[29] B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regre ssion curve fitting. *Journal of the Royal Statistical Society*, 47:1–52, 1985.

[30] B. W. Silverman. *Density estimation for statistics amd data analysis*. Chapman and Hall, London, 1986.

[31] C.J. Stone. Optimal rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–1053, 1982.

[32] S. van de Geer. Estimating a regression function. *Annals of Statistics*, 18:907–924, 1990.

[33] G. S. Watson. Smooth regression analysis. *Sankhyā*, 26:101–116, 1964.