Confidence Regions, Regularization and Non-Parametric Regression

P. L. Davies*, A. Kovac† and M. Meise‡

Abstract

In this paper we offer a unified approach to the problem of non-parametric regression on the unit interval. It is based on a universal, honest and non-asymptotic confidence region $\mathcal{A}_n$ which is defined by a set of linear inequalities involving the values of the functions at the design points. Interest will typically centre on certain simplest functions in $\mathcal{A}_n$ where simplicity can be defined in terms of shape (number of local extremes, intervals of convexity/concavity) or smoothness (bounds on derivatives) or a combination of both. Once some form of regularization has been decided upon the confidence region can be used to provide honest non-asymptotic confidence bounds which are less informative but conceptually much simpler. Although the procedure makes no attempt to minimize any loss function such as MISE the resulting estimates have optimal rates of convergence in the supremum norm both for shape and smoothness regularization. We show that rates of convergence can be misleading even for samples of size $n = 10^6$ and propose a different form of asymptotics which allows model complexity to increase with sample size.

Keywords: Nonparametric regression, shape regularization, confidence bounds

1 Introduction

Non-parametric regression on the unit interval is concerned specifying a function or functions $\tilde{f}_n$ which are a reasonable representation of a data set $\boldsymbol{y}_n = \{(t_i, y(t_i)), i = 1, \ldots, n\}$ with $0 \leq t_1 < \ldots < t_n \leq 1$. Here and below we

*Universität Duisburg-Essen, Technische Universiteit Eindhoven
†University of Bristol
‡Universität Duisburg-Essen

use lower case letters to denote generic data and upper case letters to denote data generated under a specific stochastic model such as (1) below. The first approach to the problem used kernel estimators with a fixed bandwidth (Watson, 1964) but since then many other procedures have been proposed. We mention splines (Green and Silverman, 1994; Wahba, 1990), wavelets (Donoho and Johnstone, 1994), local polynomial regression (Fan and Gijbel, 1996), kernel estimators with local bandwidths (Wand and Jones, 1995) very often with Bayesian and non-Bayesian versions.

The models on which the methods are based are of the form

$$Y(t) = f(t) + \sigma(t)\varepsilon(t), \quad t \in [0,\, 1] \tag{1}$$

with various assumptions being made about $\sigma(t)$, the noise $\varepsilon(t)$ as well as the design points $\{t_1, \ldots, t_n\}$. We shall restrict attention to the simplest case

$$Y(t) = f(t) + \sigma Z(t), \quad t \in [0,\, 1] \tag{2}$$

where Z is Gaussian white noise and the t_i are given by $t_i = i/n$. We emphasize that essentially the same ideas can be used for the more general model (1) and that robust versions are available. The central role in this paper is played by a confidence region $\mathcal{A}_n$ which is defined below. It specifies all functions $\tilde{f}_n$ for which the model (2) is consistent (in a well–defined sense) with the data $\boldsymbol{y}_n$. By regularizing within $\mathcal{A}_n$ we can control both the shape and the smoothness of a regression function and provide honest non-asymptotic confidence bounds.

The paper is organized as follows. In Section 2 we define the confidence region and show that it is honest and non-asymptotic for data generated under (2). In Section 3 we consider shape regularization and in Section 4 we show that shape regularization performs well in terms of MISE. In Section 5 regularization by smoothness is treated as well as the combination of shape and smoothness regularization. Finally in Section 6 we show how honest and non–asymptotic confidence bounds can be obtained both for shape and smoothness regularization.

2 The confidence region $\mathcal{A}_n$

2.1 Non-parametric confidence regions

Much attention has been given to confidence sets in recent years. These sets are often expressed as a ball centred at some suitable estimate (Li, 1989; Hoffmann and Lepski, 2002; Baraud, 2004; Cai and Low, 2006; Robins and

van der Vaart, 2006) with particular emphasis on adaptive methods where the radius of the ball automatically decreases if f is sufficiently smooth. The concept of adaptive confidence balls is not without conceptual difficulties as the discussion of Hoffmann and Lepski (2002) shows. An alternative to smoothness is the imposition of shape constraints such as monotonicity and convexity (Davies, 1995; Dümbgen, 1998, 2003; Dümbgen and Johns, 2004). Such confidence sets require only that f satisfy the shape constraint which often has some independent justification.

We consider data $\boldsymbol{Y}_n$ generated under (2) and limit attention to functions f in some family $\mathcal{F}_n$. We call a confidence set $\mathcal{C}_n(\alpha)$ exact if

$$P(f \in \mathcal{C}_n(\alpha)) = \alpha \quad \text{for all} \quad f \in \mathcal{F}_n, \tag{3}$$

honest (Li, 1989) if

$$P(f \in \mathcal{C}_n(\alpha)) \geq \alpha \quad \text{for all} \quad f \in \mathcal{F}_n, \tag{4}$$

and asymptotically honest if

$$\liminf_{n\to\infty} \inf_{f\in\mathcal{F}_n} P(f \in \mathcal{C}_n(\alpha)) \geq \alpha \tag{5}$$

holds but it is not possible to specify the n_0 for which the coverage probability exceeds $\alpha - \epsilon$ for all $n \geq n_0$. Finally we call $\mathcal{C}_n(\alpha)$ universal if $\mathcal{F}_n = \{f : f : [0,1] \to \mathbb{R}\}$.

2.2 Definition of $\mathcal{A}_n$

The confidence region $\mathcal{A}_n$ was first given in Davies and Kovac (2001). It is constructed as follows. For any function $g : [0,\, 1] \to \mathbb{R}$ and any interval $I = [t_j, t_k]$ of $[0, 1]$ with $j \leq k$ we write

$$w(\boldsymbol{y}_n, g, I) = \frac{1}{\sqrt{|I|}} \sum_{t_i \in I} (y(t_i) - g(t_i)) \tag{6}$$

where $|I|$ denotes the number of points t_i in I. With this notation

$$\mathcal{A}_n = \mathcal{A}_n(\boldsymbol{y}_n, \mathcal{I}_n, \sigma, \tau_n) = \big\{ g : \max_{I\in\mathcal{I}_n} |w(\boldsymbol{y}_n, g, I)| \leq \sigma\sqrt{\tau_n \log n}\, \big\} \tag{7}$$

where $\mathcal{I}_n$ is a family of intervals of $[0,\, 1]$ and for given α the value of $\tau_n = \tau_n(\alpha)$ is defined by

$$P\Big(\max_{I\in\mathcal{I}_n} \frac{1}{\sqrt{|I|}} \Big| \sum_{t_i \in I} Z(t_i) \Big| \leq \sigma\sqrt{\tau_n \log n}\, \Big) = \alpha. \tag{8}$$

If the data $\boldsymbol{y}_n$ were generated under (2) then (8) implies that $P(f \in \mathcal{A}_n) = \alpha$ with no restrictions on f so that $\mathcal{A}_n$ is a universal, exact α–confidence region. We mention that by using an appropriate norm (Mildenberger, 2006) $\mathcal{A}_n$ can also be expressed as a ball centred at the observations $\boldsymbol{y}_n$.

A function g belongs to $\mathcal{A}_n$ if and only if its vector of evaluations at the design points $(g(t_1), \ldots, g(t_n))$ belongs to the convex polyhedron in $\mathbb{R}^n$ which is defined by the linear inequalities

$$\frac{1}{\sqrt{|I|}} \Big| \sum_{t_i \in I} (y(t_i) - g(t_i)) \Big| \leq \sigma_n \sqrt{\tau_n \log n}, \quad I \in \mathcal{I}_n.$$

The remainder of the paper is in one sense nothing more than exploring the consequences of these inequalities for shape and smoothness regularization. They enforce both local and global adaptivity to the data and they are tight in that they yield optimal rates of convergence for both shape and smoothness constraints.

In the theoretical part of the paper we take $\mathcal{I}_n$ to be the set of all intervals of the form $[t_i, t_j]$. For this choice of $\mathcal{A}_n$ checking whether $g \in \mathcal{A}_n$ for a given g involves about $n^2/2$ linear inequalities. Surprisingly there exist algorithms which allow this to be done with algorithmic complexity $\mathrm{O}(n \log n)$ (Bernholt and Hofmeister, 2006). In practice we restrict $\mathcal{I}_n$ to a multiresolution scheme as follows. For some $\lambda > 1$ we set

$$\begin{aligned} \mathcal{I}_n = \big\{ [t_{l(j,k)},\, t_{u(j,k)}] \,:\, & l(j,k) = \lfloor (j-1)\lambda^k + 1 \rfloor, \\ & u(j,k) = \min\{\lfloor j\lambda^k \rfloor, n\}, j = 1, \ldots, \lceil n\lambda^{-k} \rceil, \\ & k = 1, \ldots, \lceil \log n / \log \lambda \rceil \big\}. \end{aligned} \tag{9}$$

For any $\lambda > 1$ we see that $\mathcal{I}_n$ now contains $\mathrm{O}(n)$ intervals. For $\lambda = 2$ we get the wavelet multiresolution scheme which we use throughout the paper when doing the calculations for explicit data sets. If $\mathcal{I}_n$ is the set of all possible intervals it follows from a result of Dümbgen and Spokoiny (2001) that $\lim_{n\to\infty} \tau_n = 2$ whatever the value of α. On the other hand for any $\mathcal{I}_n$ which contains all the degenerate intervals $[t_j, t_j]$ (as will always be the case) then $\lim_{n\to\infty} \tau_n \geq 2$ whatever α. In the following we simply take $\tau_n = 3$ as our default value. This guarantees a coverage probability of at least $\alpha = 0.95$ for all samples of size $n \geq 500$ and it tends rapidly to one as the sample size increases.

As it stands the confidence region (7) cannot be used as it requires σ. We use the following default estimate

$$\sigma_n = 1.48260 \,\mathrm{median}(|y(t_2) - y(t_1)|, \ldots, |y(t_n) - y(t_{n-1})|)/\sqrt{2} \tag{10}$$

which is a consistent estimate of σ for white noise data. For data generated under (2) it is clear that σ_n is positively biased and consequently the coverage probability will not decrease. Simulations show that

$$P\big(f \in \mathcal{A}_n(\boldsymbol{Y}_n, \mathcal{I}_n, \sigma_n, 3)\big) \geq 0.95 \tag{11}$$

for all $n \geq 500$ and

$$\lim_{n\to\infty} \inf_f P\big(f \in \mathcal{A}_n(\boldsymbol{Y}_n, \mathcal{I}_n, \sigma_n, 3)\big) = 1. \tag{12}$$

In other words $\mathcal{A}_n$ is a universal, honest and non-asymptotic confidence region for f. To separate the problem of specifying the size of the noise from the problem of investigating the behaviour of the procedures under the model (2) we shall always put $\sigma_n = \sigma$ for theoretical results. For real data and in all simulations however we use the σ_n of (10).

The confidence region $\mathcal{A}_n$ can be interpreted as the inversion of the multiscale tests that the mean of the residuals is zero on all intervals $I \in \mathcal{I}_n$. A similar idea is to be found in Dümbgen and Spokoiny (2001) who invert tests to obtain confidence regions. Their tests derive from kernel estimators with different locations and bandwidths where the kernels are chosen to be optimal for certain testing problems for given shape hypotheses. The confidence region may be expressed in terms of linear inequalities involving the weighted residuals with the weights determined by the kernels. The confidence region we use corresponds to the uniform kernel on [0, 1]. Because of their multiscale character all these confidence regions allow any lack of fit to be localized (Davies and Kovac, 2001; Dümbgen and Spokoiny, 2001) and under shape regularization they automatically adapt to a certain degree of local smoothness.

2.3 Confidence, approximation and truth

We use the expressions 'confidence region' or 'confidence bounds' in this paper as experience shows that our preferred expressions 'approximation region' and 'approximation bounds' cause only confusion. The term 'confidence region' usually means that the function f which generated that data belongs to the region with a specified probability. This assumes that the data were in fact generated under the model (2). This is the case for simulations where we can then talk about confidence regions. If however we have real data we do not have any reason to suppose that a generating function f exists. We practice ontological parsimony as did William of Ockham and suspend judgment on the existence of any 'true' generating function f. This requires

an interpretation of the region $\mathcal{A}_n$ which is independent of the existence of a true function f. Our interpretation is the simple one that it consists of all functions $\tilde{f}_n$ for which the model (2) with $f = \tilde{f}_n$ is consistent with the data. If we regularize for smoothness within $\mathcal{A}_n$ then this gives us the smoothest function which is consistent with the data. The definition of $\mathcal{A}_n$ and the resulting approach to non–parametric regression were motivated by the desire to provide a theory of approximation for statistics which does without an assumption of truth, approximatio sine veritate (Davies, 1995).

3 Shape regularization and local adaptivity

3.1 Generalities

In this section we consider shape regularization within the confidence region $\mathcal{A}_n$. Two simple possibilities are to require that the function be monotone or that it be convex. Although much has been written about monotone or convex regression we are not concerned with these particular cases. Given any data set $\boldsymbol{y}_n$ it is always possible to calculate a monotone regression function, for example monotone least squares. In the literature the assumption usually made is that the f in (2) is monotone and then one examines the behaviour of a monotone regression function. Although this case is included in the following analysis we are mainly concerned with determining the minimum number of local extremes or points of inflection required for an adequate approximation. This is STEP 2 of Mammen (1991). We shall investigate how pronounced a peak or a point of inflection must be before it can be detected on the basis of a sample of size n. These estimates are in general conservative but they do reflect the real finite sample behaviour of procedures. We shall also investigate rates of convergence between peaks and points of inflection. We show that these are local in the strong sense that the rate of convergence at a point t depends only on the behaviour of f in a small neighbourhood of t. Furthermore we show that in a certain sense shape regularization automatically adapts to the smoothness of f. All the calculations we perform use only the shape restrictions of the regularization and the linear inequalities which determine $\mathcal{A}_n$. The mathematics are extremely simple involving no more than a Taylor expansion and are of no intrinsic interest. We give one such calculation in detail and refer to the appendix for the remainder.

3.2 Local extreme values

The simplest form of shape regularization is to minimize the number of local extreme values subject to membership of $\mathcal{A}_n$. We wish to determine this minimum number and exhibit a function in $\mathcal{A}_n$ which has this number of local extreme values. This is an optimization problem and the taut string algorithm of Davies and Kovac (2001) was explicitly developed to solve it. We are here not concerned with the calculation of a solution but rather with the properties of any such solution. In particular we wish to investigate the efficacy of the regularization, that is the ability to detect peaks or points of inflection. To do this we consider data generated under the model (2) and investigate how pronounced a peak of the generating function f of (2) must be before it is detected on the basis of a sample of size n. We start with the case of one local maximum and assume that it is located at $t = 1/2$. Let I_c denote an interval which contains $1/2$. For any $\tilde{f}_n$ in $\mathcal{A}_n$ we have

$$\frac{1}{\sqrt{|I_c|}} \sum_{t_i \in I_c} \tilde{f}_n(t_i) \geq \frac{1}{\sqrt{|I_c|}} \sum_{t_i \in I_c} f(t_i) - \sigma\sqrt{3 \log n} + \sigma Z(I_c)$$

and hence

$$\max_{t_i \in I_c} \tilde{f}_n(t_i) \geq \frac{1}{|I_c|} \sum_{t_i \in I_c} f(t_i) - \sigma \frac{\sqrt{3 \log n} - Z(I_c)}{\sqrt{|I_c|}} \tag{13}$$

where

$$Z(I_c) = \frac{1}{|I_c|} \sum_{t_i \in I_c} Z(t_i) \stackrel{D}{=} N(0,1).$$

Let I_l and I_r be intervals to the left and right of I_c respectively. A similar argument gives

$$\min_{t_i \in I_l} \tilde{f}_n(t_i) \leq \frac{1}{|I_l|} \sum_{t_i \in I_l} f(t_i) + \sigma \frac{\sqrt{3 \log n} + Z(I_l)}{\sqrt{|I_l|}} \tag{14}$$

and

$$\min_{t_i \in I_r} \tilde{f}_n(t_i) \leq \frac{1}{|I_r|} \sum_{t_i \in I_r} f(t_i) + \sigma \frac{\sqrt{3 \log n} + Z(I_r)}{\sqrt{|I_r|}}. \tag{15}$$

If now

$$\frac{1}{|I_c|}\sum_{t_i\in I_c} f(t_i) - \sigma\frac{\sqrt{3\log n} - Z(I_c)}{\sqrt{|I_c|}}$$
$$\geq \max\left\{\frac{1}{|I_l|}\sum_{t_i\in I_l} f(t_i) + \sigma\frac{\sqrt{3\log n} + Z(I_l)}{\sqrt{|I_l|}}, \right.$$
$$\left. \frac{1}{|I_r|}\sum_{t_i\in I_r} f(t_i) + \sigma\frac{\sqrt{3\log n} + Z(I_r)}{\sqrt{|I_r|}}\right\} \quad (16)$$

then any function in $\mathcal{A}_n$ must have a local maximum in $I_l \cup I_c \cup I_r$. The random variables $Z(I_c)$, $Z(I_l)$ and $Z(I_r)$ are independently and identically distributed $N(0,1)$ random variables. With probability at least 0.99 we have $Z(I_c) \geq -2.72, Z(I_l) \leq 2.72$ and $Z(I_r) \leq 2.72$ and hence we can replace (16) by

$$\frac{1}{|I_c|}\sum_{t_i\in I_c} f(t_i) - \sigma\frac{\sqrt{3\log n} + 2.72}{\sqrt{|I_c|}}$$
$$\geq \max\left\{\frac{1}{|I_l|}\sum_{t_i\in I_l} f(t_i) + \sigma\frac{\sqrt{3\log n} + 2.72}{\sqrt{|I_l|}}, \right.$$
$$\left. \frac{1}{|I_r|}\sum_{t_i\in I_r} f(t_i) + \sigma\frac{\sqrt{3\log n} + 2.72}{\sqrt{|I_r|}}\right\} \quad (17)$$

If we now regularize by considering those functions in $\mathcal{A}_n$ with the minimum number of local extreme values we see that this number must be at least one. As f itself has one local extreme value and belongs to $\mathcal{A}_n$ with probability rapidly approaching one we see that with high probability the minimum number is one and that this local maximum lies in $I_l \cup I_c \cup I_r$.

The condition (17) quantifies the power of the peak so that it will be detected and the precision of the location is given by the interval $I_l \cup I_c \cup I_r$. We apply this to the function

$$f_b(t) = b((t - 1/2)/0.01) \quad (18)$$

where

$$b(t) = \begin{cases} 1, & |t| \leq 1 \\ 0, & \text{otherwise.} \end{cases} \quad (19)$$

All points in the interval [0.49, 0.51] are in a sense the same local maximum. If we wish to detect the local maximum with a precision of $\delta = 0.01$ then we

require the local maximum of f_n^* to lie in the interval [0.48, 0.52]. A short calculation with $\sigma = 1$ shows that the smallest value of n for which (17) is satisfied is approximately 19500. A small simulation study using the taut string resulted in the peak being found with the prescribed accuracy in 99.6% of the 10000 simulations.

The results for a general function throw little light on the behaviour of the regularization so we now analyse the situation for a function f which has a continuous second derivative. We assume that f has exactly one local maximum situated in $t = 1/2$ and that

$$-c_2 \leq f^{(2)}(t) \leq -c_1 < 0, \quad t \in I_0, \tag{20}$$

for some open interval I_0 which contains the point $t = 1/2$. If we denote by f_n^* a function in $\mathcal{A}_n$ which minimizes the number of local extremes then f_n^* will, for large n, have exactly one local extreme value which is a local maximum situated at t_n^* with

$$|t_n^* - 1/2| = \mathrm{O}_f\left(\left(\frac{\log n}{n}\right)^{1/5}\right). \tag{21}$$

An explicit upper bound for the constant in O_f in terms of c_1 and c_2 of (20) is available. The proof is based on a Taylor expansion in (17) and is given in the appendix. We also have

$$f_n^*(t_n^*) \geq f(1/2) - \mathrm{O}_f\left(\left(\frac{\log n}{n}\right)^{2/5}\right) \tag{22}$$

with again an explicit constant available. In the other direction

$$f_n^*(t_n^*) \leq f(1/2) + \sigma(\sqrt{3\log n} + 2.4). \tag{23}$$

The proofs are given in the Appendix.

More generally suppose that f has a continuous second derivative and κ local extreme values situated at $0 < t_1^e < \ldots < t_\kappa^e < 1$ with $f^{(2)}(t_k^e) \neq 0, k = 1, \ldots, \kappa$. If $f_n^* \in \mathcal{A}_n$ has the smallest number of local extreme values of all functions in $\mathcal{A}_n$ it follows that with probability tending to one f_n^* will have κ local extreme values located at the points $0 < t_{n1}^{*e} < \ldots < t_{n\kappa}^{*e} < 1$ with

$$|t_{nk}^{*e} - t_k^e| = \mathrm{O}_f\left(\left(\frac{\log n}{n}\right)^{1/5}\right), k = 1, \ldots, \kappa. \tag{24}$$

Furthermore if t_k^e is the position of a local maximum of f then

$$f_n^*(t_{nk}^{*e}) \geq f(t_k^e) - \mathrm{O}_f\left(\left(\frac{\log n}{n}\right)^{2/5}\right) \tag{25}$$

whereas if t_k^e is the position of a local minimum of f then

$$f_n^*(t_{nk}^{*e}) \leq f(t_k^e) + \mathrm{O}_f\left(\left(\frac{\log n}{n}\right)^{2/5}\right). \tag{26}$$

In the other direction we have

$$\begin{aligned} f_n^*(t_{nk}^{*e}) &\leq f(t_k^e) + \sigma\big(\sqrt{3\log n} + \sqrt{3\log(8+\kappa)}\,\big) && (27)\\ f_n^*(t_{nk}^{*e}) &\geq f(t_k^e) - \sigma\big(\sqrt{3\log n} + \sqrt{3\log(8+\kappa)}\,\big). && (28) \end{aligned}$$

More precise bounds cannot be attained on the basis of monotonicity arguments alone.

3.3 Between the local extremes

We investigate the behaviour of f_n^* between the local extremes where f_n^* is monotone. For any function $g : [0,\, 1] \to \mathbb{R}$ we define

$$\|g\|_{I,\infty} = \sup\{|g(t)| : t \in I\}. \tag{29}$$

Consider a point $t = i/n$ between two local extreme values of f and write $I_{nk}^r = [i/n, (i+k)/n]$ with $k > 0$. Then

$$f_n^*(i/n) - f(i/n) \leq \min_{1\leq k\leq k_n^{*r}} \left\{\frac{k}{n}\|f^{(1)}\|_{I_{nk}^r,\infty} + 2\sigma\sqrt{\frac{3\log n}{k}}\right\} \tag{30}$$

where k_n^{*r} denotes the largest value of k for which f_n^* is non-decreasing on I_{nk}^r. A similar inequality holds for intervals on the left and leads to

$$|f(t) - f_n^*(t)| \leq 3^{4/3}\sigma^{2/3}|f^{(1)}(t)|^{1/3}\left(\frac{\log n}{n}\right)^{1/3}. \tag{31}$$

It follows from (30) and the corresponding inequality on the left that as long as f_n^* has the correct global monotonicity behaviour its behaviour at a point t with $f^{(1)}(t) \neq 0$ depends only on the behaviour of f in a small neighbourhood of t. If $f^{(1)}(t) = 0$ on a non-degenerate interval $I = [t_l, t_r]$ between two local extremes then for $t_l < t < t_r$ we have $I_l^* = [t_l, t]$ and $I_r^* = [t, t_r]$ which results in

$$|f(t) - f_n^*(t)| \leq \frac{3^{1/2}\sigma}{\min\{\sqrt{t-t_l}, \sqrt{t_r-t}\}}\left(\frac{\log n}{n}\right)^{1/2}. \tag{32}$$

The same argument shows that if

$$|f(t) - f(s)| \leq L|t-s|^\beta$$

with $0 < \beta \leq 1$ then

$$|f(t) - f_n^*(t)| \leq cL^{1/(2\beta+1)}(\sigma/\beta)^{2\beta/(2\beta+1)}(\log n/n)^{\beta/(2\beta+1)} \tag{33}$$

where

$$c \leq (2\beta+1)3^{\beta/(2\beta+1)}\left(\frac{1}{\beta+1}\right)^{1/(2\beta+1)} \leq 4.327.$$

Apart from the value of c this corresponds to Theorem 2.2 of Dümbgen and Spokoiny (2001).

3.4 Convexity and concavity

We now turn to shape regularization by concavity and convexity. This can be interpreted as minimizing the number of local extreme values of the first derivative rather than of the function itself. We take f to be differentiable with derivative $f^{(1)}$ which is strictly increasing on $[0, 1/2]$ and strictly decreasing on $[1/2, 1]$. We put $I_{nk}^c = [1/2 - k/n, 1/2 + k/n]$, $I_{nk}^l = [t_l - k/n, t_l + k/n]$ with $t_l + k/n < 1/2 - k/n$ and $I_{nk}^l = [t_r - k/n, t_r + k/n]$ with $t_r - k/n > 1/2 + k/n$. Corresponding to (17) we have

$$\begin{aligned}
&\max_{t\in I_{nk}^l} f^{(1)}(t)/n + \left(2\sigma(\sqrt{3\log n} + 2.72\sqrt{2})\right)/k^{3/2} \\
&\quad \geq \min\left\{\max_{t\in I_{nk}^l} f^{(1)}(t)/n + \left(2\sigma(\sqrt{3\log n} + 2.72\sqrt{2})\right)/k^{3/2},\right. \\
&\qquad \left.\max_{t\in I_{nk}^r} f^{(1)}(t)/n + \left(2\sigma(\sqrt{3\log n} + 2.72\sqrt{2})\right)/k^{3/2}\right\}.
\end{aligned} \tag{34}$$

From this it follows that with probability at least 0.99 the first derivative of every differentiable function $\tilde{f}_n \in \mathcal{A}_n$ has at least one local maximum. Let f_n^* be a differentiable function in $\mathcal{A}_n$ whose first derivative has the smallest number of local extreme values. Then as f belongs to $\mathcal{A}_n$ with probability tending to one it follows that $f_n^{*(1)}$ has exactly one local maximum with probability tending to at least 0.99. Suppose now that f has a continuous third derivative and κ points of inflection located at $0 < t_1^i < \ldots < t_\kappa^i$ with

$$f^{(2)}(t_j^i) = 0 \text{ and } f^{(3)}(t_j^i) \neq 0,\ j = 1, \ldots, \kappa.$$

If f_n^* has the smallest number of points of inflection in $\mathcal{A}_n$ then if $f \in \mathcal{A}_n$ with probability tending to one it follows that with probability tending to one f_n^* will have κ points of inflection located at $0 < t_{n1}^{*i} < \ldots < t_{n\kappa}^{*i} < 1$. Furthermore corresponding to (24) we have

$$|t_{nk}^{*i} - t_k^i| = \mathrm{O}_f\left(\left(\frac{\log n}{n}\right)^{1/7}\right),\ k = 1, \ldots, \kappa. \tag{35}$$

Similarly if t_k^i is a local maximum of $f^{(1)}$ then corresponding to (25) we have

$$f_n^{*(1)}(t_{nk}^{*e}) \geq f^{(1)}(t_k^e) - \mathrm{O}_f\left(\left(\frac{\log n}{n}\right)^{2/7}\right) \tag{36}$$

and if t_k^i is a local minimum of $f^{(1)}$ then corresponding to (26) we have

$$f_n^{*(1)}(t_{nk}^{*e}) \leq f^{(1)}(t_k^e) + \mathrm{O}_f\left(\left(\frac{\log n}{n}\right)^{2/7}\right). \tag{37}$$

3.5 Between points of inflection

Finally we consider the behaviour of f_n^* between the points of inflection where it is then either concave or convex. We consider a point $t = i/n$ and suppose that f_n^* is convex on $I_{nk}^r = [i/n, (i+2k)/n]$. Corresponding to (30) we have

$$f_n^{*(1)}(i/n) - f^{(1)}(i/n) \leq \min_{1\leq k\leq k_n^{*r}}\left\{\frac{k}{n}\|f^{(2)}\|_{I_{nk}^r,\infty} + 4\sigma n\sqrt{\frac{3\log n}{k^3}}\right\} \tag{38}$$

where k_n^{*r} is the largest value of k such that f_n^* is convex on $[i/n, (i+2k)/n]$. Similarly corresponding to (92) we have

$$f^{(1)}(i/n) - f_n^{*(1)}(i/n) \leq \min_{1\leq k\leq k_n^{*l}}\left\{\frac{k}{n}\|f^{(2)}\|_{I_{nk}^l,\infty} + 4\sigma n\sqrt{\frac{3\log n}{k^3}}\right\} \tag{39}$$

where $I_{nk}^l = [i/n - 2k/n, i/n]$ and k_n^{*l} is the largest value of k for which f_n^* is convex on I_{nk}^l. If $f^{(2)}(t) \neq 0$ we have corresponding to (31)

$$|f_n^{*(1)}(t) - f^{(1)}(t)| \leq 4.36\sigma^{2/5}|f^{(2)}(t)|^{3/5}\left(\frac{\log n}{n}\right)^{1/5}. \tag{40}$$

as n tends to infinity. If $f^{(2)}(t) = 0$ on the non-degenerate interval $I = [t_l, t_r]$ then for $t_l < t < t_r$ we have corresponding to (32)

$$|f_n^{*(1)}(t) - f^{(1)}(t)| \leq \frac{4\sqrt{3}\sigma}{\min\{(t-t_l)^{3/2}, (t_r-t)^{3/2}\}}\left(\frac{\log n}{n}\right)^{1/2}. \tag{41}$$

The results for f_n^* itself are as follows. For a point t with $f^{(2)}(t) \neq 0$ and an interval $I_{nk}^r = [t, t+2k/n]$ where f_n^* is convex we have

$$f_n^*(t) \leq f(t) + c_1(f,t)\left(\frac{k}{n}\right)\left(\frac{\log n}{n}\right)^{1/5} + \frac{k^2}{2n^2}\|f^{(2)}\|_{I_{nk}^r,\infty} + 4\sigma\sqrt{\frac{3\log n}{k}}$$

where $c_1(f,t) = 4.36\sigma^{2/5}|f^{(2)}(t)|^{3/5}$. If we minimize over k and repeat the argument for a left interval we have corresponding to (31)

$$|f_n^*(t) - f(t)| \leq 11.58\sigma^{4/5}|f^{(2)}(t)|^{1/5}\left(\frac{\log n}{n}\right)^{2/5}. \tag{42}$$

Finally if $f^{(2)}(t) = 0$ for t in the non-degenerate interval $[t_l, t_r]$ we have corresponding to (32) for $t_l < t < t_r$

$$|f_n^*(t) - f(t)| \leq \frac{14\sigma}{\min\{\sqrt{t-t_l}, \sqrt{t_r - t}\}}\left(\frac{\log n}{n}\right)^{1/2}. \tag{43}$$

If the derivative $f^{(1)}$ of f satisfies $|f^{(1)}(t) - f^{(1)}(s)| \leq L|t-s|^\beta$ with $0 < \beta \leq 1$ then corresponding to (33) we have

$$|f_n^{*(1)}(t) - f^{(1)}(t)| \leq cL^{3/(2\beta+3)}(\sigma/\beta)^{2\beta/(2\beta+3)}\left(\frac{\log n}{n}\right)^{\beta/(2\beta+3)}$$

with

$$c \leq 2^\beta\left(\frac{6\sqrt{3}}{2^\beta}\right)^{(\beta+2)/(2\beta+3)} + 4\sqrt{3}\beta\left(\frac{2^\beta}{6\sqrt{3}}\right)^{3/(2\beta+3)} \leq 8.78.$$

There is of course a corresponding result for f_n^* itself.

4 MISE, asymptotics and local adaptivity

4.1 MISE

Given data $\boldsymbol{Y}_n$ generated under (2) and an estimator $\hat{f}_n$ for f the MISE is defined by

$$MISE(\hat{f}_n) = \mathbb{E}\left(\int_0^1 (f(t) - \hat{f}_n(t))^2\, dt\right). \tag{44}$$

The MISE takes no account of the identification of local extremes or the smoothness of the estimator but, in spite of its weaknesses, it seems to be the gold standard for evaluation of estimation procedures. For a given f and n the MISE can be determined by simulations and for certain classes of functions f optimal rates of convergence are available. Other norms such as the supremum norm are also used but are not nearly as common. For a more comprehensive comparison of procedures which include the L_2– and L_∞–norms as well as peak identification we refer to Davies, Gather and Weinert (2006).

Because of the importance of the MISE this section is devoted to the performance of shape regularization in terms of MISE. The MISE performance of the taut string procedure of Davies and Kovac (2001) can be improved by the following simple modification. We use the local squeezing version of the taut string as described in Davies and Kovac (2001) with $\tau = 3$ and with $\mathcal{I}_n$ the wavelet multiresolution scheme with $\lambda = 2$. The position of each local extreme will be taken to be the midpoint of the interval which defines that local extreme. Between the local extremes we replace the taut string by monotone least squares with the proviso that the end values agree with those of the taut string. The results of Davies, Gather and Weinert (2006) show that this modification of the taut string outperforms most other procedures in terms of MISE. The final estimator is close to STEP 3 of Mammen (1991).

We compare the modified version of the taut string with a wavelet reconstruction. All wavelet calculations were done using the software provided by Nason (1998). We use Haar wavelets and also the default choice of Nason (1998) namely the Daubechies least-asymmetric orthonormal compactly supported wavelet with 10 vanishing moments. These have a second derivative which satisfies a Hölder condition of at least 0.9. We consider firstly the sine function

$$f_s(t) = \sin(2\pi t) \tag{45}$$

contaminated with Gaussian white noise with standard deviation 0.5. The MISE of the modified taut string tends to zero at a rate no better that $O(n^{-2/3})$. As the sine function is infinitely differentiable the wavelet reconstruction has a rate of convergence of at least $n^{-0.8529}$. The second and third lines of Table 1 give the $\log(MISE)$ for the two methods for the sample sizes $n = 2^k, k = 11, \ldots, 20$. The results agree more or less with the rates of convergence although for large sample sizes the MISE of the wavelet reconstruction seems to tail off to such an extent that the taut string reconstruction is just as good. This may be due to rounding errors. There are reasons for supposing that the taut string algorithm is numerically more stable than that for wavelets. The second function we consider is

$$f_{sb}(t) = \sin(2\pi t) + 0.5\exp(-5000(t - 1/2)^2) \tag{46}$$

which again is infinitely differentiable so that the rates of convergence are as before. The fourth and fifth lines of Table 1 give the corresponding results for this function. Here the taut string is superior except for the sample of size $n = 2048$. In fact the modified taut string still exhibits a smaller MISE for sample of size $n = 4 \cdot 10^6$ so it is not clear at what point the asymptotics become relevant.

n	2^{11}	2^{12}	2^{13}	2^{14}	2^{15}	2^{16}	2^{17}	2^{18}	2^{19}	2^{20}
wv	-7.91	-8.80	-8.86	-9.22	-9.42	-9.56	-9.65	-9.72	-9.72	-9.73
ts	-5.54	-5.97	-6.43	-6.90	-7.35	-7.84	-8.25	-8.70	-9.20	-9.66
wv	-5.52	-5.57	-5.57	-5.58	-5.62	-6.67	-7.27	-7.74	-8.25	-8.90
ts	-5.42	-5.76	-5.91	-6.33	-6.59	-7.72	-8.21	-8.57	-9.13	-9.57

Table 1: The second and third lines give the log(MISE) based on five simulations for the sine function $f_s(t) = \sin(2\pi t)$ contaminated with Gaussian while noise with standard deviation $\sigma = 1/2$. The fourth and fifth lines give the corresponding results for the sine curve with a bump f_{sb} as given by (46)

4.2 Local adaptivity

In order to better understand the behaviour of the wavelet and taut string reconstructions we look more closely at particular examples. The upper panel of Figure 1 shows the wavelet reconstructions of the sine curve f_s and the sine bump curve f_{sb} for samples of size $n = 2^{16} = 65536$. In each case exactly the same noise $Z(t_i), i = 1, \ldots, n$ was used so that differences in the reconstructions are entirely due to the presence of the peak at $t = 0.5$. This adversely effects the reconstruction of the sine function at points well away from the 0.5 bump. The lower panel of Figure 1 shows the modified taut string reconstruction for the same data sets. It is seen that the effect of the peak at $t = 0.5$ is limited to points very close to the peak. The upper panel of Figure 2 shows the wavelet reconstruction for t in the interval [0.2, 0.4] both with and without the peaks and for a sample size of $n = 2^{20} = 1048576$. Again it is clear that the bump still adversely effects the reconstruction of the sine curve at points well away from 0.5. In fact neither wavelet reconstruction is sufficiently close to the data to lie in $\mathcal{A}_n$. For the interval [0.21, 0.31] the normalized sum of the residuals for the sine curve f_s is -3.318 as against a threshold value of $0.5\sqrt{3 \log n} = 3.22$. For the interval [0.27, 0.33] the normalized sum of the residuals for the sine curve with bump f_{sb} is 3.92 indicating that the curve is too low. The lower panel of Figure 2 shows the modified taut string reconstruction for the same data sets. The conclusion is that the modified taut string is highly locally adaptive whereas this is not the case for wavelets.

4.3 Asymptotics

As seen above there are non–pathological examples where the rate of convergence does not reflect the true performance of a procedure even for very

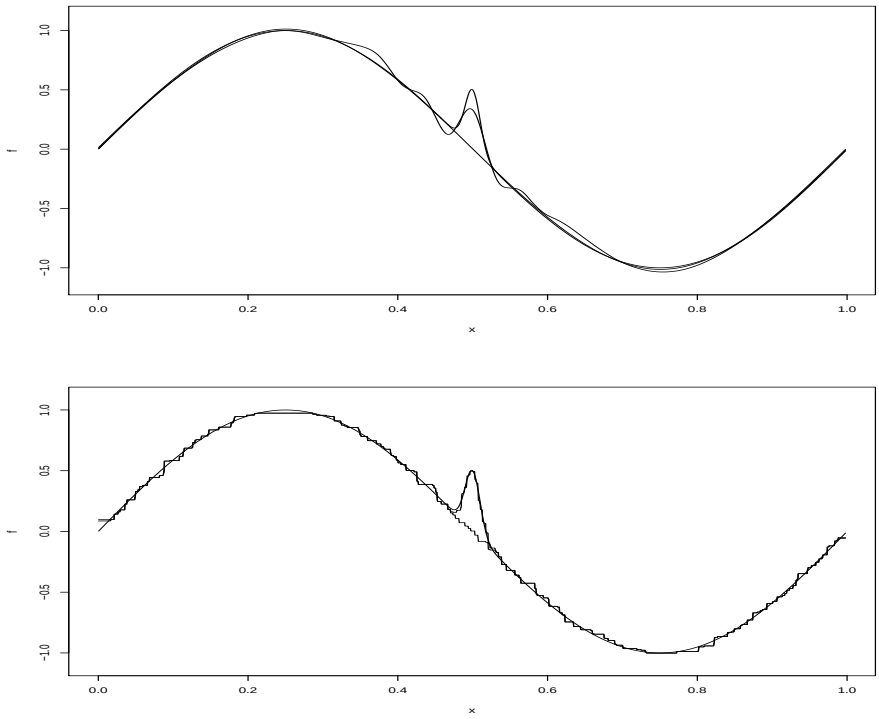

Figure 1: Reconstructions of a noisy sine curve and sine curve with bump. The upper panel shows the wavelet reconstructions and the lower panel the taut string reconstructions as described in the text. The sample size is $2^{16} = 65536$ and all curves were contaminated with exactly the same noise.

large samples. To demonstrate this effect we consider a sequence of models

$$Y_n(t) = f_n(t) + \sigma Z(t) \tag{47}$$

where the f_n become increasingly more complex as n tends to infinity. Dahlhaus (1988) considered sequences of models of increasing complexity to show the advantage of tapering in time series. Tapering was known to have good finite sample properties but asymptotically its only effect was to increase the variance. Dahlhaus was able to show that tapering can produce consistent estimates in situations where this is not possible without tapering.

We emphasize that in this section the models we use, that is the choice of the functions f_n in (47), are not chosen because of their applicability to real data but as test beds where we can investigate the performance of a procedure under well–controlled conditions. For this it is necessary that the calculations can be carried out and are applicable.

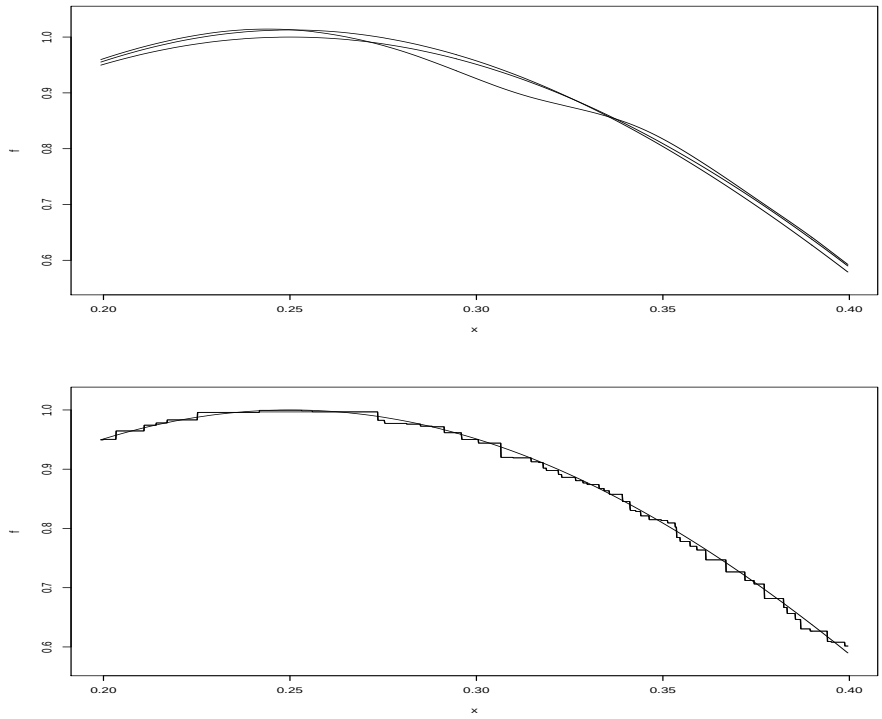

Figure 2: Reconstructions of a noisy sine curve and sine curve with bump on the interval [0.2, 0.4]. The upper panel shows the wavelet reconstructions and the lower panel the taut string reconstructions as described in the text. The sample size is $2^{20} = 1048576$ and all curves were contaminated with exactly the same noise.

4.3.1 Detecting peaks

We consider functions f_n composed of ν_n box functions (19) with the same heights δ_n and widths $2\gamma_n$ and at the same distances Γ_n apart

$$f_n(t) = \delta_n \sum_{k=1}^{\nu_n} b\Big(\frac{t - k\Gamma_n - (2k-1)\gamma_n}{\gamma_n}\Big) \tag{48}$$

where

$$\nu_n(\Gamma_n + 2\gamma_n) + \Gamma_n = 1. \tag{49}$$

Arguing as in Section 3.2 but replacing the factor 2.72 in (17) by one which takes into account the increasing number ν_n of peaks we can show the fol-

lowing. If

$$\delta_n > \sigma_n\sqrt{\frac{3\log n}{n\Gamma_n}} + 2\sigma\sqrt{\frac{\log(\nu_n+8)}{n\Gamma_n}} + \sigma_n\sqrt{\frac{3\log n}{2n\gamma_n}} + 2\sigma\sqrt{\frac{\log(\nu_n+8)}{2n\gamma_n}} \quad (50)$$

then with probability at least 0.99 and rapidly tending to 1 every function f_n^* in $\mathcal{A}_n$ which minimizes the number of local maxima will have exactly ν_n local maxima. As an example we put

$$n = 1000,\ \gamma_n = 0.005,\ \Gamma_n = 2\gamma_n \text{ and } \nu_n = 49.$$

In order for δ_n to satisfy (50) we must have $\delta_n > 5.42$. To test this we performed 500 simulations using the taut string method to identify peaks. With $\delta = 5.42$ all peaks were correctly identified in all simulations. With $\delta = 5$ and $\delta = 4.5$ all peaks were found in 99.8 and 89.8 percent of the cases respectively. This illustrates that the estimates above do reflect the actual performance of the taut string. Figure 3 shows the first two hundred points of a data set with $n = 1024$, $\gamma_n = 0.005$, $\Gamma_n = 0.01$, $\delta_n = 4.5$ and $\sigma = 1$. The top panel shows the data and the function f_n, the centre panel shows the taut string reconstruction. The bottom panel of Figure 3 shows the wavelet reconstruction using Haar wavelets.

4.4 MISE and asymptotics

We now consider a sequence of models f_n of the form

$$\begin{aligned} f_n(t) &= \exp(3t) + \delta_{n1}\sum_{k=1}^{\nu_n} b\Big(\frac{2(t-k\Gamma_n-2k\gamma_n)}{\gamma_n}+1\Big) \\ &\quad - \delta_{n2}\sum_{k=1}^{\nu_n} b\Big(\frac{2(t-k\Gamma_n-2k\gamma_n)}{\gamma_n}-1\Big) \end{aligned} \quad (51)$$

and again investigate the behaviour of the modified taut string reconstruction f_n^* and compare it with that of the Haar wavelet reconstruction. The introduction of the second sum compared with (48) is to simplify the arguments as it guarantees that the modified taut string reconstruction is non-decreasing outside of the intervals

$$I_{n,k} = [k\Gamma_n + (2k-1)\gamma_n,\ k\Gamma_n, (2k+1)\gamma_n].$$

This is not necessary for the results but simplifies the arguments. We choose δ_{n1} and δ_{n2} so that f_n^* has a local maximum in $[k\Gamma_n + (2k-1)\gamma_n,\ k\Gamma_n + 2k\gamma_n]$

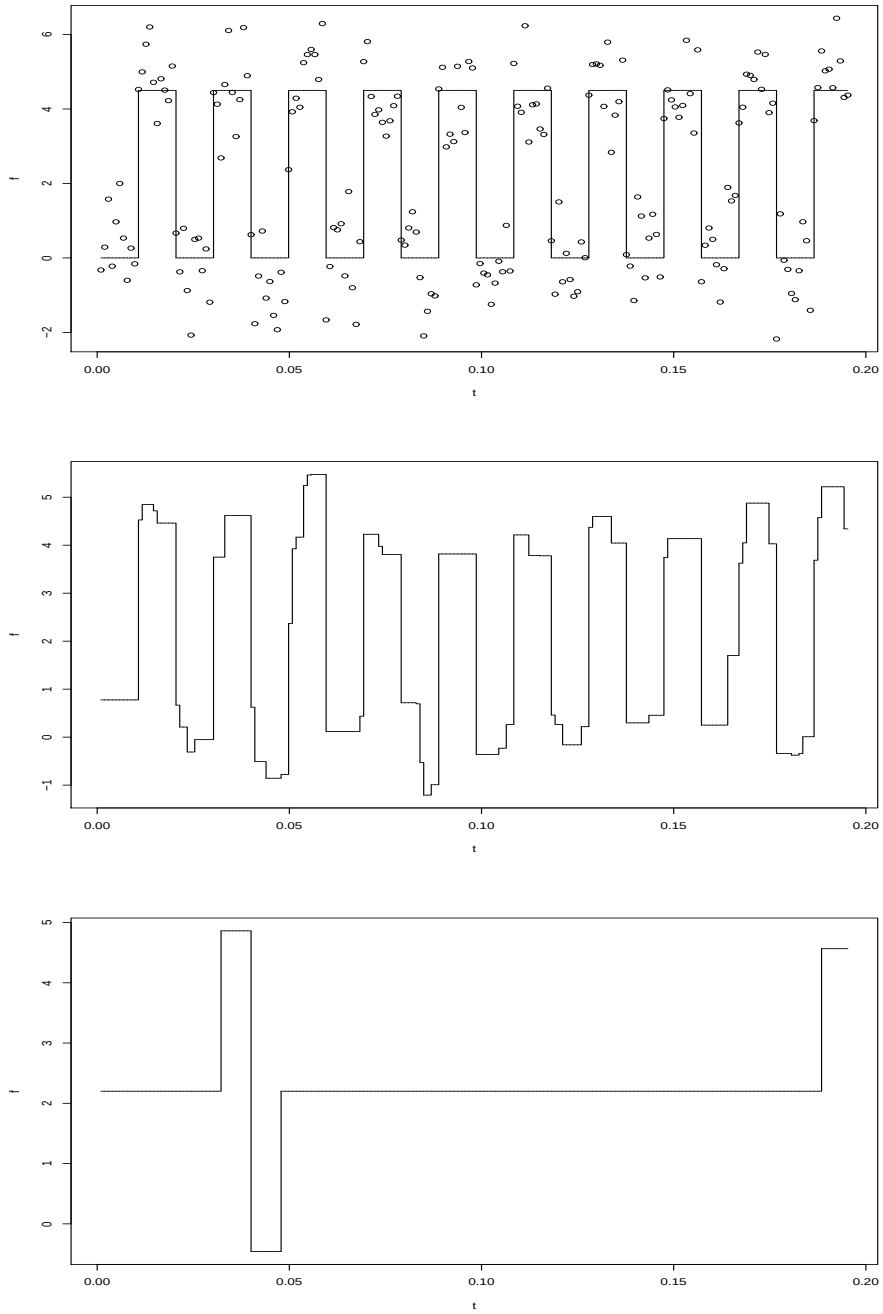

Figure 3: The first 200 data points of data generated according to (47) with $\sigma = 1$ (top panel) and the taut string reconstruction (centre panel) and the Haar wavelet reconstruction (bottom panel). The function f_n is as given by (48) with $n = 1024, \gamma_n = 0.005, \Gamma_n = 0.01$ and $\delta_n = 4.5$.

and a local minimum in $[k\Gamma_n + 2k\gamma_n,\ k\Gamma_n + (2k+1)\gamma_n]$. Sufficient conditions for this can be derived in a straightforward manner from the results given in the previous section. Again the analysis is simplified if δ_{n1} dominates δ_{n2} and so we assume $\delta_{n2} = \mathrm{o}(\delta_{n1})$. To make the situation as difficult as possible we want ν_n to be large subject to obtaining consistent estimates outside of the intervals $I_{n,k}$. This means we should take γ_n to be small and the δ_{n1} to be (arbitrarily) large. We shall be concerned with the MISE outside of the intervals $I_{n,k}$ as we are interested in local adaption and wish to investigate the degree to which the reconstruction outside of the $I_{n,k}$ is degraded by the peaks inside the $I_{n,k}$. We therefore define the restricted mean integrated square error RMISE by

$$RMISE(\hat{f}_n) = \mathbb{E}\left(\int_0^1 \{t \notin I_n\}\Big(\hat{f}_n(t) - f_n(t)\Big)^2 dt\right) \tag{52}$$

where $I_n = \cup_{k=1}^{\nu_n} I_{n,k}$. We choose the Γ_n to be as small as possible subject to $\lim_{n\to\infty} RMISE(f_n^*) = 0$. As f_n^* is the monotone least squares estimate on each of the intervals

$$I_{n,k}^c = [k\Gamma_n + (2k+1)\gamma_n,\ (k+1)\Gamma_n + (2k+1)\gamma_n] \tag{53}$$

we have with $t_i = i/n$

$$\frac{1}{n\Gamma_n}\sum_{t_i \in I_{n,k}^c}\Big(f_n^*(t_i) - f_n(t_i)\Big)^2 = \frac{1}{n\Gamma_n}\sum_{t_i \in I_{n,k}^c}\Big(f_n^*(t_i) - \exp(3t_i)\Big)^2.$$

As $\gamma_n = \mathrm{o}(1)$ and $\exp(3t)$ is essentially constant on $I_{n,k}^c$ we deduce from the behaviour of the monotone least squares reconstruction that

$$\frac{1}{n\Gamma_n}\sum_{t_i \in I_{n,k}^c}\Big(f_n^*(t_i) - f_n(t_i)\Big)^2 = \mathrm{O}\left(\frac{\log(n\Gamma_n)}{n\Gamma_n}\right)$$

from which it follows that

$$\frac{1}{n}\sum_{t_i \in I_{n,k}^c}\Big(f_n^*(t_i) - f_n(t_i)\Big)^2 = \mathrm{O}\left(\frac{\log(n\Gamma_n)}{n}\right).$$

As there are approximately $1/\Gamma_n$ such intervals we finally have

$$RMISE(f_n^*) = \mathrm{O}\left(\frac{\log(n\Gamma_n)}{n\Gamma_n}\right). \tag{54}$$

We now turn to the Haar wavelet reconstruction. Consider a wavelet

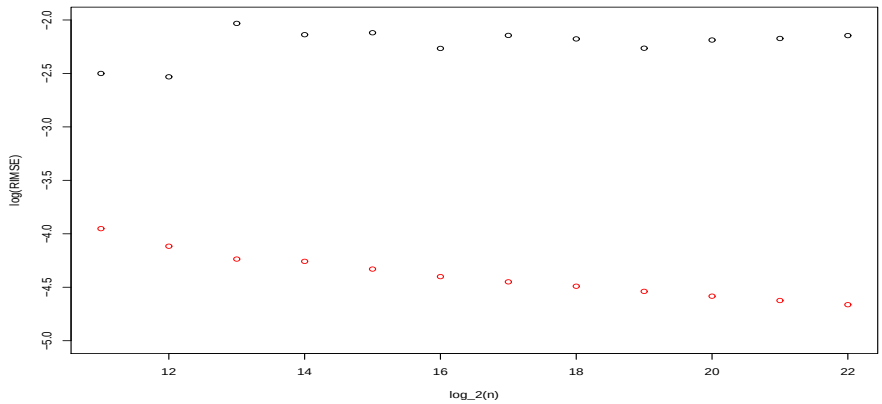

Figure 4: log(RMISE) plotted against $\log_2(n)$ for the Haar wavelets (upper points) and for the taut string (lower points). The parameter values are as given by (56). The sample sizes range from 2048 to 4194304.

$w_{lk'}^{(n)}$ whose support contains the interval $I_{n,k}$ but is disjoint to the intervals $I_{n,j}, j \neq k$. As δ_{n1} is large the coefficient of this wavelet will exceed the threshold $\sqrt{2\log n}$ and it will consequently be included in the reconstruction. The coefficient includes a part due to the noise and this contribution of this part to the RIMSE will be of order $1/n$. For each interval $I_{n,k}$ there are approximately $\log(1/\Gamma_n)$ such wavelets $w_{lk}^{(n)}$ and their total contribution to the RMISE is therefore at least $\log(1/\Gamma_n)/n$. As we have approximately $1/\Gamma_n$ intervals $I_{n,k}$ we have

$$RMISE(f_n^{hw}) \geq c\left(\frac{-\log(\Gamma_n)}{n\Gamma_n}\right) \tag{55}$$

for some $c > 0$. In particular if $\Gamma_n \asymp \log(n)/n$ then

$$RMISE(f_n^*) = \mathrm{O}\left(\frac{\log\log n}{\log n}\right)$$

so that f_n^* is consistent whereas

$$\liminf_{n\to\infty} RMISE(f_n^{hw}) > 0$$

so that the Haar wavelet reconstructions are not consistent. Figure 4 shows the results of a small simulation study with $\log RMISE$ plotted against the $\log_2 n$. The parameter values are given by

$$\gamma_n = 20/n,\ \Gamma_n = 50\log n/n,\ \delta_{n1} = (\log n)^4,\ \delta_{n2} = \log\delta_{n1}. \tag{56}$$

The results seem to confirm the analysis given above for the taut string which are consistent with (54). The wavelet reconstruction however is worse than (55) would indicate. One possibility is the sensitivity of the wavelet reconstruction to the exact position of the peaks but we do not analyse this any further.

5 Regularization by smoothness

Sections 3 and 4 were concerned with shape regularization. We now turn to regularization by smoothness.

5.1 Minimizing total variation

We define the total variation of the kth derivative of a function g evaluated at the design point $t_i = i/n$ by

$$TV(g^k) := \sum_{i=k+2}^{n} \left|\Delta^{(k+1)}(g(i/n))\right|, \quad k \geq 0 \tag{57}$$

where

$$\Delta^{(k+1)}(g(i/n)) = \Delta^{(1)}(\Delta^{(k)}(g(i/n))) \tag{58}$$

with

$$\Delta^{(1)}(g(i/n)) = n(g(i/n) - g((i-1)/n)).$$

Similarly the supremum norm $\|g^{(k)}\|_\infty$ is defined by

$$\|g^{(k)}\|_\infty = \max_i \left|\Delta^{(k)}(g(i/n))\right|. \tag{59}$$

Minimizing either $TV(g^k)$ or $\|g^{(k)}\|_\infty$ subject to $g \in \mathcal{A}_n$ leads to a linear programming problem. Minimizing the more traditional measure of smoothness

$$\int_0^1 g^{(k)}(t)^2\, dt$$

subject to $g \in \mathcal{A}_n$ leads to a quadratic programming problem which is numerically much less stable (cf. Davies and Meise, 2005). The results for data sets for which the solution can be computed are not essentially different from those obtained from minimizing $TV(g^{(k)})$ or $\|g^{(k)}\|_\infty$ so we restrict attention to these latter cases.

Minimizing the total variation of g itself, $k = 0$, leads to piecewise constant solutions which are very similar to the taut string solution. In most

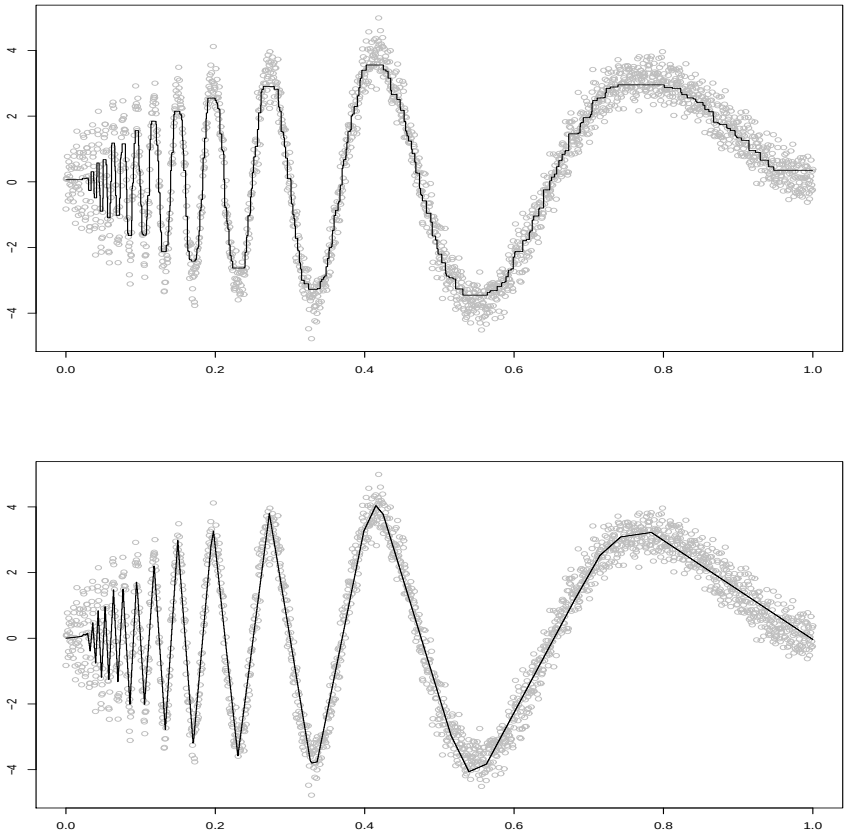

Figure 5: Minimization of $TV(\tilde{f}_n)$ (upper panel) and $TV(\tilde{f}_n^{(1)})$ (lower panel) subject to $\tilde{f}_n \in \mathcal{A}_n$ for a noisy Doppler function.

cases the solution also minimizes the number of local extreme values but this is not always the case. The upper panel of Figure 5 shows the result of minimizing $TV(\tilde{f}_n)$ for the Doppler data of Donoho and Johnstone (1994). It has the same number of peaks as the taut string reconstruction. The lower panel of Figure 5 shows the result of minimizing $TV(\tilde{f}_n^{(1)})$. The solution is a linear spline. Just as minimizing $TV(\tilde{f}_n)$ can be used for determining the intervals of monotonicity so we can use the solution of minimizing $TV(\tilde{f}_n^{(1)})$ to determine the intervals of concavity and convexity. Minimizing $TV(g^{(k)})$ or $\|g^{(k)}\|_\infty$ for larger values of k leads to very smooth functions but the numerical problems increase as rounding errors become more and more important. Figure 6 shows data generated according to

$$Y(t) = \sin(4\pi t) + 0.1Z(t) \tag{60}$$

with $n = 1024$. The top left panel shows the data, the function $\sin(4\pi t)$ and the reconstruction obtained by minimizing $\|\tilde{f}_n^{(4)}\|_\infty$. The remaining panels in the left column show from top to bottom the first four derivatives of $\sin(4\pi t)$ and the derivatives of the reconstruction. The panels in the right column show the corresponding results of a wavelet approximation with Daubechies least-asymmetric orthonormal compactly supported wavelet with 10 vanishing moments. Simply minimizing $\|\tilde{f}_n^{(4)}\|_\infty$ is in general not to be recommended as there is insufficient control of the solution at points t with $|\tilde{f}_n^{(4)}(t)|$ much less than $\|\tilde{f}_n^{(4)}\|_\infty$. This can be avoided by minimizing $TV(\tilde{f}_n^{(4)})$ in a second step subject to the bound already obtained for $\|\tilde{f}_n^{(4)}\|_\infty$. This in turn is not always satisfactory as the solution will in general not satisfy obvious shape constraints as we now show.

5.2 Smoothness and shape regularization

Regularization by smoothing will in general lead to solutions which are not acceptable as they do not fulfill obvious shape constraints. Figure 7 shows the effect of minimizing the total variation of the second derivative without further constraints and the minimization with the imposition of the taut string shape constraints. The minimization of $TV(g^{(2)})$ without shape constraints shows a behaviour similar to that of the wavelet reconstruction of Figure 1, namely the effect of the peak on the reconstruction well away from the peak. It seems plausible that if we consider a sequence of models as in Section 4.4 then simple smoothness regularization will be inconsistent in situations where shape regularization is consistent. We have no theoretical results in this direction.

5.3 Rates of convergence

In Section 3 we showed how rates of convergence could be obtained by shape regularization within $\mathcal{A}_n$. We now show how rates of convergence can be obtained by smoothness regularization. We exemplify the method by minimizing the supremum norm of the second derivative of all functions in $\mathcal{A}_n$. As already mentioned this should be combined with a total variation minimization. This does not alter the rate of convergence. Let $f_n^\#$ be such that

$$\|f_n^{\#(2)}\|_\infty \leq \|g^{(2)}\|_\infty \quad \forall g \in \mathcal{A}_n. \tag{61}$$

For data generated under (2) with f satisfying $\|f^{(2)}\|_\infty < \infty$ it follows that with probability rapidly tending to one

$$\|f_n^{\#(2)}\|_\infty \leq \|f^{(2)}\|_\infty. \tag{62}$$

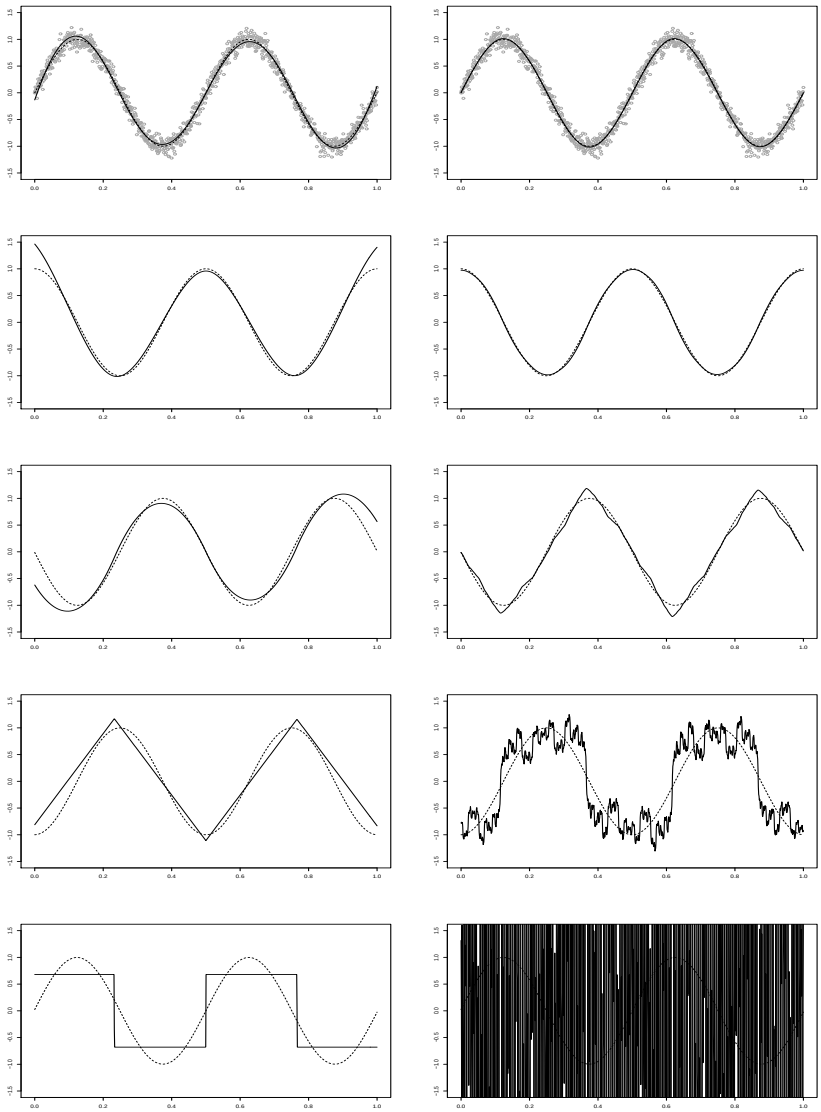

Figure 6: The reconstruction of a noisy sine function $Y(t) = \sin(4\pi t) + 0.1Z(t)$ and its first four derivatives from a sample of size $n = 1024$. The lines show from top to bottom the reconstruction of the function itself and its first four derivatives. The left column shows the reconstructions obtained by minimizing $\|\tilde{f}_n^{(4)}\|$. The right column shows the wavelet reconstructions using the Daubechies least-asymmetric orthonormal compactly supported wavelet with 10 vanishing moments. Superimposed are the corresponding derivatives of $\sin(4\pi t)$.

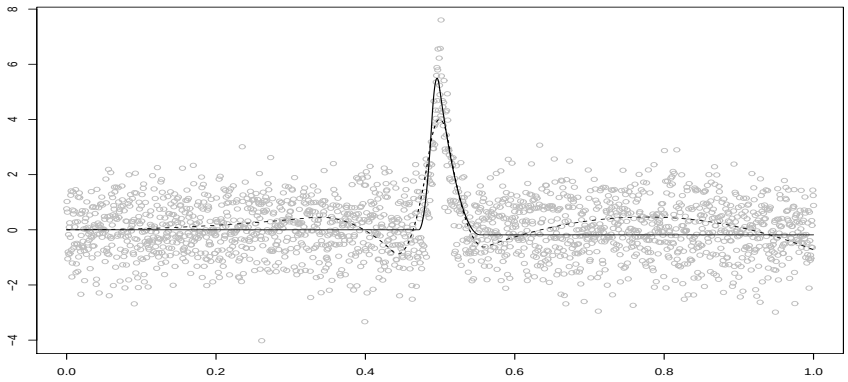

Figure 7: The minimization of the total variation of the second derivative with (solid line) and without (dashed line) the shape constraints derived from the taut string. The solution subject to the shape constraints was also forced to assume the same value at the local maximum as the taut string solution.

A Taylor expansion and a repetition of arguments already used lead to

$$|f_n^{\#}(i/n) - f(i/n)| \leq 3.742 \|f^{(2)}\|_\infty^{1/5} \sigma^{4/5} \left(\frac{\log n}{n}\right)^{2/5} \tag{63}$$

on an interval

$$\left[0.58\sigma^{2/5}(\log n)^{1/5}/\left(\|f^{(2)}\|_\infty^{2/5} n^{1/5}\right),\ 1 - 0.58\sigma^{2/5}(\log n)^{1/5}/\left(\|f^{(2)}\|_\infty^{2/5} n^{1/5}\right)\right]$$

with a probability rapidly tending to one. A rate of convergence for the first derivative may be derived in a similar manner and results in

$$|f_n^{\#(1)}(i/n) - f^{(1)}(i/n)| \leq 4.251 \|f^{(2)}\|_\infty^{3/5} \sigma^{2/5} \left(\frac{\log n}{n}\right)^{1/5} \tag{64}$$

on an interval

$$\left[2.15\sigma^{2/5}(\log n)^{1/5}/\left(\|f^{(2)}\|_\infty^{2/5} n^{1/5}\right),\ 1 - 2.15\sigma^{2/5}(\log n)^{1/5}/\left(\|f^{(2)}\|_\infty^{2/5} n^{1/5}\right)\right].$$

6 Confidence bands

6.1 The problem

Confidence bounds have the advantage of simplicity and we show how they can be constructed from the confidence region $\mathcal{A}_n$. For each point t_i we require a lower bound $lb_n(\boldsymbol{y}_n, t_i) = lb_n(t_i)$ and an upper bound $ub_n(\boldsymbol{y}_n, t_i) = ub_n(t_i)$ such that

$$\mathcal{B}_n(\boldsymbol{y}_n) = \{g : lb_n(\boldsymbol{y}_n, t_i) \leq g(t_i) \leq ub_n(\boldsymbol{y}_n, t_i), i = 1, \ldots n\} \tag{65}$$

is an honest non-asymptotic confidence region

$$P(f \in \mathcal{B}_n(\boldsymbol{Y}_n)) \geq \alpha \quad \text{for all} \quad f \in \mathcal{F}_n \tag{66}$$

for data $\boldsymbol{Y}_n$ generated under (2). In a sense the problem has a simple solution. If we put

$$lb_n(t_i) = y(t_i) - \sigma_n\sqrt{3\log n}\,, \quad ub_n(t_i) = y(t_i) + \sigma_n\sqrt{3\log n}\,, \tag{67}$$

then $\mathcal{A}_n \subset \mathcal{B}_n$ and (66) for all holds with $\mathcal{F}_n = \{f \,|\, f : [0,1] \to \infty\}$. Such universal bounds are too wide to be of any practical use and are consequently not acceptable. The problem is that they can only be made tighter by restricting $\mathcal{F}_n$ by imposing shape or quantitative smoothness constraints. Of the two, shape constraints are the more appealing as they often have some independent justification, for example, the concavity of utility functions. Nevertheless most of the literature is concerned with smoothness constraints and these are always of a qualitative form such as a continuous second derivative. However the qualitative assumption

$$\mathcal{F}_n = \{f : \|f^{(2)}\|_\infty < \infty\} \tag{68}$$

does not lead to any improvement of the bounds (67). They can only be improved by replacing (68) by a quantitative assumption such as

$$\mathcal{F}_n = \{f : \|f^{(2)}\|_\infty < 60\}. \tag{69}$$

However the refusal to replace (68) by (69) translates into a refusal to state the n_0 for which say

$$P(f \in \mathcal{B}_n(\boldsymbol{Y}_n)) \geq \alpha - 0.01 \quad \forall n \geq n_0 \tag{70}$$

holds so nothing is gained. There are of course problems with bounds of the form (69). Firstly they may be inconsistent with the data in that $\mathcal{F}_n \cap \mathcal{A}_n = \emptyset$. Secondly if the chosen upper bound is too large the confidence bounds may also be too large to be acceptable. One can of course specify $\mathcal{F}_n$ post factum as we do in Section 6.3 (see also Figure 11) but the best approach is to give up the idea of truth and use the idea of approximation as in Section 2.3.

6.2 Shape regularization

6.2.1 Monotonicity

As an example of a shape restriction we consider bounds for non-decreasing approximations. If we denote the set of non-increasing functions on [0, 1] by

$$\mathcal{M}^+ = \{g : \ g : [0,\, 1] \to \mathbb{R},\ g \ \text{ non-decreasing}\}$$

then there exists a non-decreasing approximation if and only if

$$\mathcal{M}^+ \cap \mathcal{A}_n \neq \emptyset. \tag{71}$$

This is the case when the set of linear inequalities which define $\mathcal{A}_n$ together with $g(t_1) \leq \ldots \leq g(t_n)$ are consistent. This is once again a linear programming problem. If (71) holds then the lower and upper bounds are given respectively by

$$lb_n(t_i) = \min\{\tilde{f}_n(t_i) : \tilde{f}_n \in \mathcal{M}^+ \cap \mathcal{A}_n\}, \tag{72}$$
$$ub_n(t_i) = \max\{\tilde{f}_n(t_i) : \tilde{f}_n \in \mathcal{M}^+ \cap \mathcal{A}_n\}. \tag{73}$$

The calculation of $lb_n(t_i)$ and $ub_n(t_i)$ requires solving a linear programming problem and although this can be done it is practically impossible for larger sample sizes. If the family of intervals $\mathcal{I}_n$ is restricted to a wavelet multiresolution scheme then samples of size $n = 1000$ can be handled. For larger sample sizes fast honest bounds can be attained as follows. If $\tilde{f}_n \in \mathcal{M}^+ \cap \mathcal{A}_n$ then for any i and k with $i + k \leq n$ it follows that

$$\sqrt{k+1}\ \ \tilde{f}_n(t_i) \geq \frac{1}{\sqrt{k+1}} \sum_{j=0}^{k} Y_n(t_{i+j}) - \sigma\sqrt{3\log n}\,.$$

From this we may deduce the lower bound

$$lb_n(t_i) = \max_{0\leq k\leq i-1} \left(\frac{1}{k+1} \sum_{j=0}^{k} Y_n(t_{i-j}) - \sigma\sqrt{\frac{3\log n}{k+1}} \right) \tag{74}$$

with the corresponding upper bound

$$ub_n(t_i) = \min_{0\leq k\leq n-i} \left(\frac{1}{k+1} \sum_{j=0}^{k} Y_n(t_{i+j}) + \sigma\sqrt{\frac{3\log n}{k+1}} \right). \tag{75}$$

Both these bounds are of algorithmic complexity $O(n^2)$. Faster bounds can be obtained by putting

$$lb_n(t_i) = \max_{0\le\theta(k)\le i-1}\left(\frac{1}{\theta(k)+1}\sum_{j=0}^{\theta(k)} Y_n(t_{i-j}) - \sigma\sqrt{\frac{3\log n}{\theta(k)+1}}\right) \quad (76)$$

$$ub_n(t_i) = \min_{0\le\theta(k)\le n-i}\left(\frac{1}{\theta(k)+1}\sum_{j=0}^{\theta(k)} Y_n(t_{i+j}) + \sigma\sqrt{\frac{3\log n}{\theta(k)+1}}\right) \quad (77)$$

where $\theta(k) = \lfloor \theta^k - 1 \rfloor$ for some $\theta > 1$. The fast bounds are not necessarily non–decreasing but can be made so by putting

$$\begin{aligned} ub_n(t_i) &= \min\left(ub_n(t_i), ub_n(t_{i+1})\right), \quad i = n-1,\ldots,1, \\ lb_n(t_i) &= \max\left(lb_n(t_i), lb_n(t_{i-1})\right), \quad i = 2,\ldots,n. \end{aligned}$$

These latter bounds are of algorithmic complexity $O(n\log n)$. The upper panel of Figure 8 shows data generated by

$$Y(t) = \exp(5t) + 5Z(t) \quad (78)$$

evaluated on the grid $t_i = i/1000, i = 1,\ldots,100$ together with the three lower and three upper bounds with σ replace by σ_n of (10). The lower bounds are those given by (72) with $\mathcal{I}_n$ a dyadic multiresolution scheme, (74) and (76) with $\theta = 2$. The times required for were about 12 hours, 19 seconds and less than one second respectively with corresponding times for the upper bounds (73), (75) and (77). The differences between the bounds are not very large and it is not the case that one set of bounds dominates the others. The methods of Section 3 can be applied to show that all the uniform bounds are optimal in terms of rates of convergence.

6.2.2 Convexity

Convexity and concavity can be treated similarly. If we denote the set of convex functions on [0, 1] by $\mathcal{C}^+$ then there exists a convex approximation if and only if

$$\mathcal{C}^+ \cap \mathcal{A}_n \neq \emptyset.$$

Assuming that the design points are of the form $t_i = i/n$ this will be the case if and only if the set of linear constraints

$$g(t_{i+1}) - g(t_i) \ge g(t_i) - g(t_{i-1}),\ i = 2,\ldots,n-1,$$

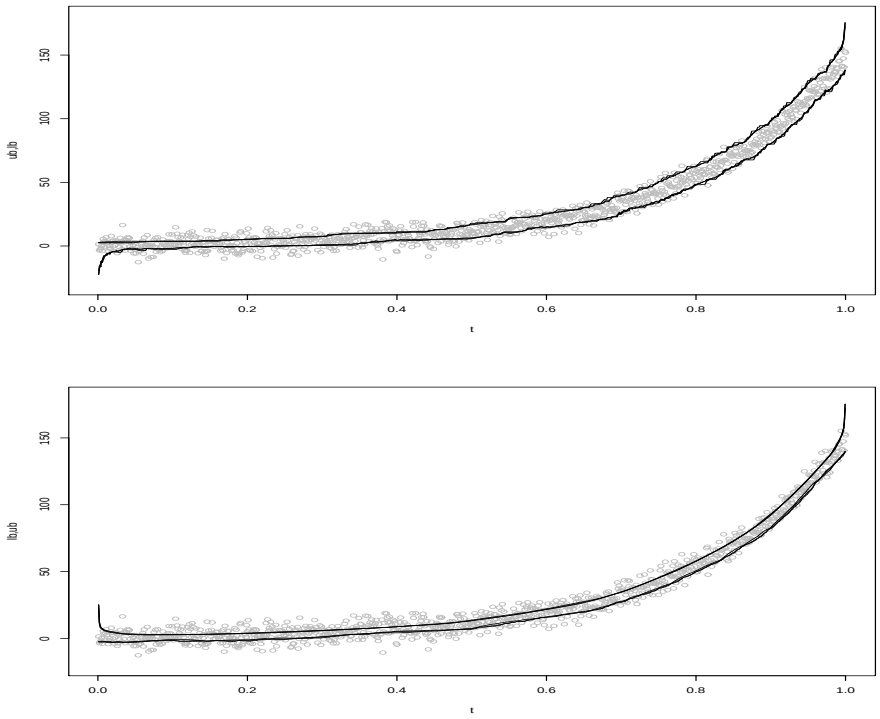

Figure 8: The function $f(t) = \exp(5t)$ degraded with $N(0, 25)$ noise together with monotone confidence bounds (upper panel) and convex confidence bounds (lower panel). The three lower bounds in the upper panel are derived from (72), (74) and (76) and the corresponding upper bounds are (73), (75) and (77). The lower bounds for the lower panel are (79), (83) and (85) and the corresponding upper bounds (80), (81) and (84)

are consistent with the linear constraints which define $\mathcal{A}_n$. Again this is a linear programming problem. If this is the case then lower and upper bounds are given respectively by

$$lb_n(t_i) = \min\{\tilde{f}_n(t_i) : \tilde{f}_n \in \mathcal{C}^+ \cap \mathcal{A}_n\}, \quad (79)$$

$$ub_n(t_i) = \max\{\tilde{f}_n(t_i) : \tilde{f}_n \in \mathcal{C}^+ \cap \mathcal{A}_n\} \quad (80)$$

which again is a linear programming problem which can only be solved for relatively small values of n. An honest but faster upper bound can be obtained by noting that

$$g(i/n) \leq \frac{1}{2k+1} \sum_{j=-k}^{k} g((i+j)/n), \quad k \leq \min(i-1, n-i)$$

which gives rise to

$$ub_n(t_i) = \min_{0 \le k \le \min(i-1, n-i)} \left(\frac{1}{2k+1} \sum_{j=-k}^{k} Y_n(t_{i+j}) + \sigma \sqrt{\frac{3 \log n}{2k+1}} \right). \tag{81}$$

A fast lower bound is somewhat more complicated. Consider a function $\tilde{f}_n \in \mathcal{C}^+ \cap \mathcal{A}_n$ and two points $(i/n, \tilde{f}_n(i/n))$ and $((i+k)/n, ub_n((i+k)/n))$. As $\tilde{f}_n((i+k)/n) \le ub_n((i+k)/n)$ and $\tilde{f}_n$ is convex it follows that $\tilde{f}_n$ lies below the line joining $(i/n, \tilde{f}_n(i/n))$ and $((i+k)/n, ub_n((i+k)/n))$. From this and $\tilde{f}_n \in \mathcal{A}_n$ we may derive a lower bound by noting

$$\begin{aligned} &lb_n(t_i) \le lb_n(t_i, k) := \\ &\max_{1 \le j \le k} \left(\frac{1}{j} \sum_{l=1}^{j} Y_n(t_{i+j}) - ub_n(t_{i+k})(j+1)/(2k) - \sigma \sqrt{3 \log n / j} \right) \end{aligned} \tag{82}$$

for all k, $-i+1 \le k \le n-i$. An honest lower bound is therefore given by

$$lb_n(t_i) = \max_{-i+1 \le k \le n-i} lb_n(t_i, k). \tag{83}$$

The algorithmic complexity of ub_n as given by (81) is $\mathrm{O}(n^2)$ whilst that of the lower bound (83) is $\mathrm{O}(n^3)$. Corresponding to (77) we have

$$ub_n(t_i) = \min_{0 \le \theta(k) \le \min(i-1, n-i)} \left(\frac{1}{2\theta(k)+1} \sum_{j=-\theta(k)}^{\theta(k)} Y_n(t_{i+j}) + \sigma \sqrt{\frac{3 \log n}{2\theta(k)+1}} \right) \tag{84}$$

and to (76)

$$lb_n(t_i) = \max_{-i+1 \le \theta(k) \le n-i} lb_n(t_i, \theta(k)). \tag{85}$$

where

$$\begin{aligned} &lb_n(t_i) \le lb_n(t_i, \theta(k)) := \\ &\max_{1 \le \theta(j) \le \theta(k)} \Bigg(\frac{1}{\theta(j)} \sum_{l=1}^{\theta(j)} Y_n(t_{i+j}) - ub_n(t_{i+\theta(k)})(\theta(j)+1)/(2\theta(k)) - \\ &\qquad\qquad \sigma \sqrt{3 \log n / \theta(j)} \Bigg) \end{aligned} \tag{86}$$

with $\theta(k) = \lfloor \theta^k \rfloor$ for some $\theta > 1$. The algorithmic complexity of (84) is $\mathrm{O}(n \log n)$ and that of (85) is $\mathrm{O}(n (\log n)^2)$.

The lower panel of Figure 8 shows the same data as in the upper panel but with the lower bounds given by (79) with $\mathcal{I}_n$ a dyadic multiresolution scheme,

(83) and (85) and the corresponding upper bounds (80), (81) and (84). The calculation of each of the bounds (79) and (80) took about 12 hours. The lower bound (83) took about 210 minutes whilst (85) was calculated in less than 5 seconds. The lower bound (79) is somewhat better than (83) and (85) but the latter two are almost indistinguishable.

6.2.3 Piecewise monotonicity

We now turn to the case of functions which are piecewise monotone. The possible positions of the local extremes can in theory be determined by solving the appropriate linear programming problems. The taut string methodology is however extremely good and very fast so we can use this solution to identify possible positions of the local extremes. The confidence bounds depend on the exact location of the local extreme. If we take the interval of constancy of the taut string solution which includes the local maximum we may calculate confidence bounds for any function which has its local maximum in this interval. The result is shown in the top panel of Figure 9. Finally if we use the mid-point of the taut string interval as a default choice for the position of a local extreme we obtain confidence bounds as shown in the lower panel of Figure 9. The user can of course specify these positions and the programme will indicate if they are consistent with the linear constraints which define the approximation region $\mathcal{A}_n$.

6.2.4 Piecewise concave–convex

We can repeat the idea for functions which are piecewise concave–convex. There are fast methods for determining the intervals of convexity and concavity based on the algorithm devised by Groeneboom (1996) but in this section we use the intervals obtained by minimizing the total variation of the first derivative. The upper panel of Figure 10 shows the result for convexity/concavity which corresponds to Figure 9. Finally the lower panel of Figure 10 shows the result of imposing both monotonicity and convexity/concavity constraints.

6.3 Smoothness regularization

We turn to the problem of constructing lower and upper confidence bounds under some restriction on smoothness. For simplicity we take the supremum norm $\|g^{(2)}\|_\infty$ to be the measure of smoothness for a function g. The discussion in Section 6.1 shows that honest bounds are attainable only if we restrictn f a set $\mathcal{F}_n = \{g : \|g^{(2)}\|_\infty \leq K\}$ with a specified K. If we

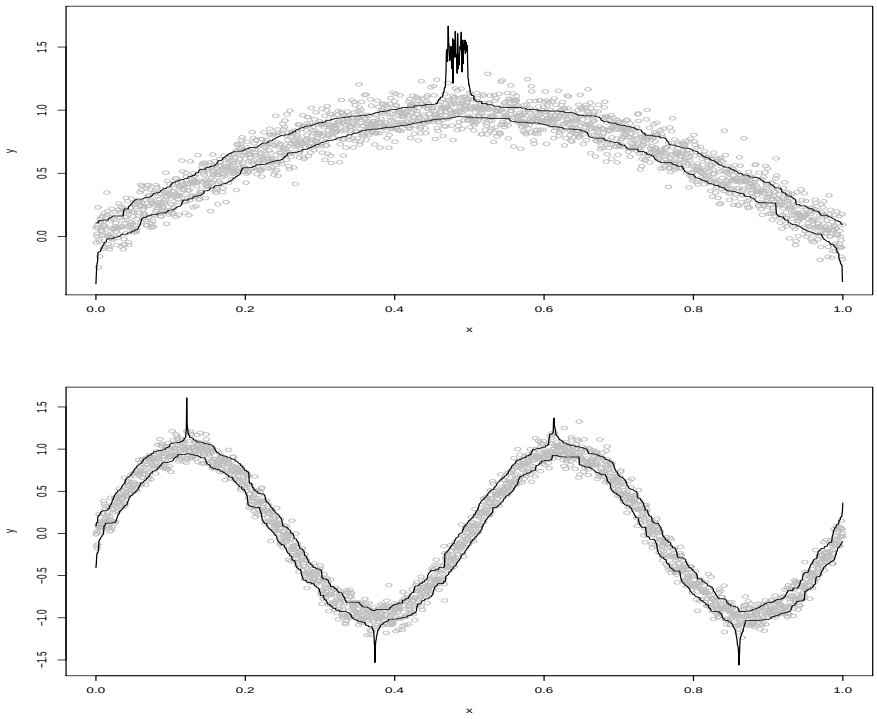

Figure 9: Confidence bounds without (upper panel) and with (lower panel) the specification of the precise positions of the local extreme values. The positions in the lower panel are the default choices obtained from the taut string reconstruction.

choose K a priori then the set $\mathcal{F}_n$ may be inconsistent with the data in that $\mathcal{A}_n \cap \mathcal{F}_n = \emptyset$. On the other hand if we take K too large then the confidence bounds may also be too large. The approach we take is firstly to minimize $\|\tilde{f}_n^{(2)}\|_\infty$ subject to $\tilde{f}_n \in \mathcal{A}_n$ and then to calculate the bounds for values of K which are two or three times this minimum value. We illustrate the procedure using data generated by (2) with $f(t) = \sin(4\pi t)$ and $\sigma = 1$. The minimum value of $\|\tilde{f}_n^{(2)}\|_\infty$ is 115.0 which compares with $16\pi^2 = 157.9$ for f itself. The upper panel of Figure 11 shows the data together with the resulting function f_n^*. The bounds under the restriction $\|\tilde{f}_n^{(2)}\|_\infty \leq 115.0$ coincide with the function f_n^* itself. The middle panel of Figure 11 show the bounds based on $\|\tilde{f}_n^{(2)}\|_\infty \leq K$ for

$$K = 136.5(= (115.0 + 157.9)/2),\ 157.9 \quad \text{and} \quad 315.8(= 2 \times 157.9).$$

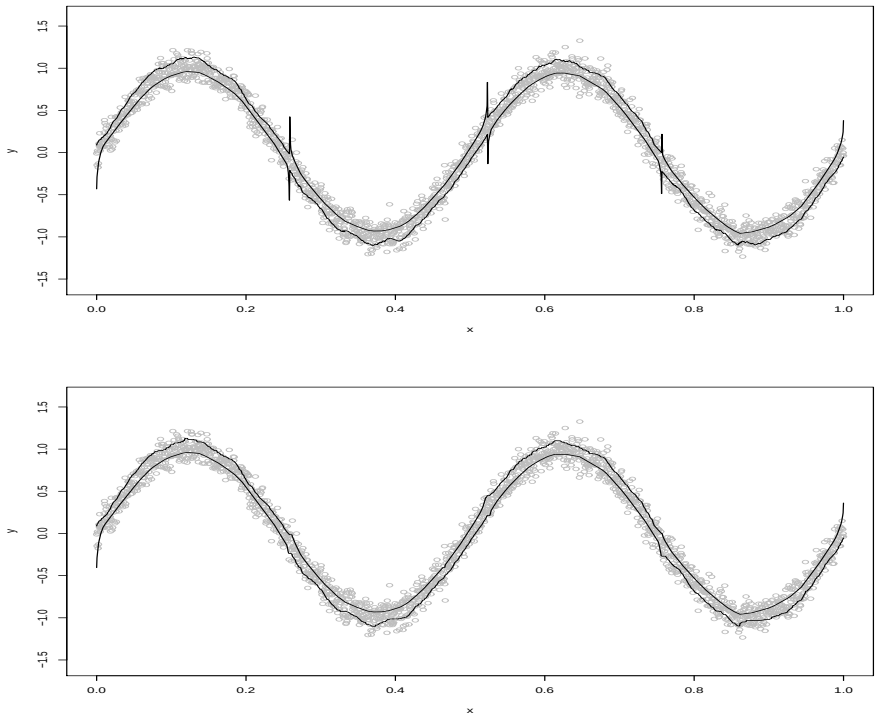

Figure 10: Confidence bounds with default choices for the intervals of convexity/concavity (upper panel) and combined confidence bounds for default choices of intervals of monotonicity and convexity/concavity.

Just as before fast bounds are also available. We have for the lower bound for given K

$$lb(i/n) \leq \min_k \left(\frac{1}{2k+1} \sum_{j=-k}^{k} Y((i+j)/n) + \left(\frac{k}{n}\right)^2 K + \sigma \sqrt{\frac{3 \log n}{2k+1}} \right) \quad (87)$$

and for the upper bound

$$ub(i/n) \geq \max_k \left(\frac{1}{2k+1} \sum_{j=-k}^{k} Y((i+j)/n) - \left(\frac{k}{n}\right)^2 K - \sigma \sqrt{\frac{3 \log n}{2k+1}} \right). \quad (88)$$

As it stands the calculation of these bounds is of algorithmic complexity $\mathrm{O}(n^2)$ but this can be reduced to $\mathrm{O}(n \log n)$ by restricting k to be of the form θ^m. The method also gives a lower bound for $\|\tilde{f}_n^{(2)}\|_\infty$ for $\tilde{f}_n$ to be consistent with the data. This is the smallest value of K for which the lower bound lb

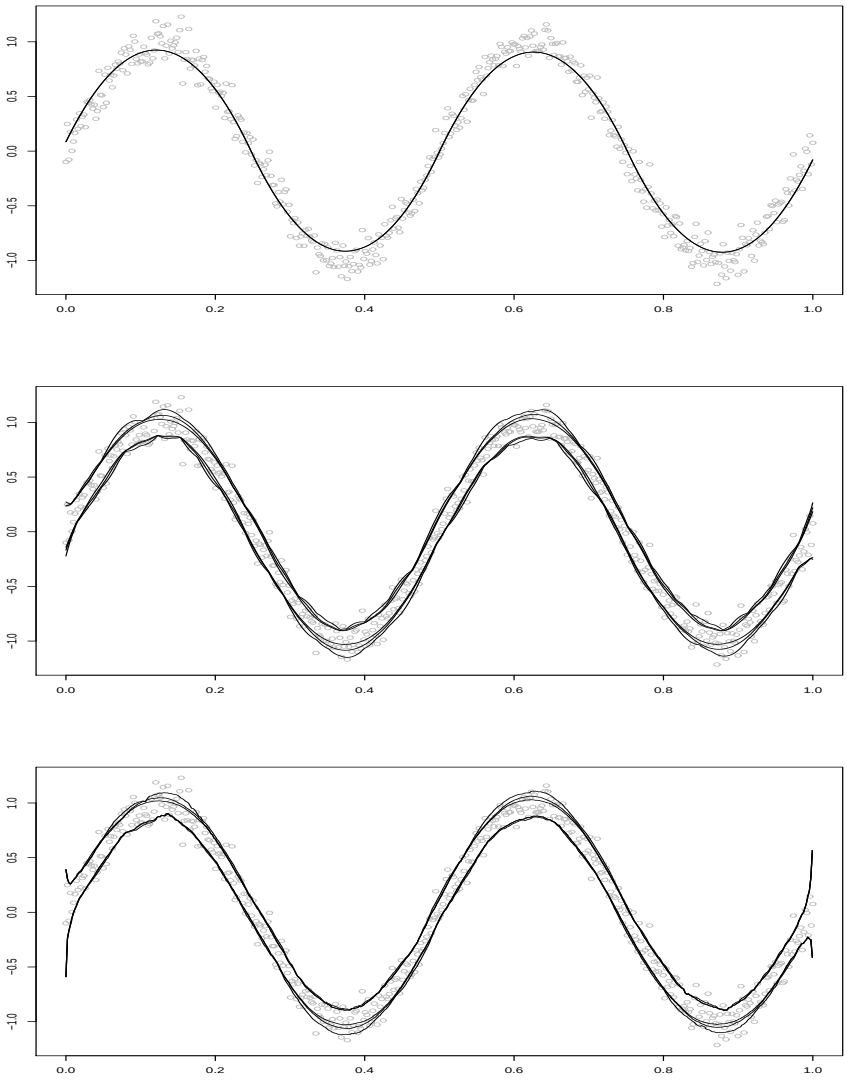

Figure 11: Smoothness confidence bounds for $f \in \mathcal{F}_n = \{f : \|\tilde{f}_n^{(2)}\|_\infty \leq K\}$ for data generated according to (2) with $f(t) = \sin(4\pi t)$, $\sigma = 1$ and $n = 500$. The top panel shows the function which minimizes $\|\tilde{f}_n^{(2)}\|_\infty$. The minimum is 115 compared with $16\pi^2 = 157.9$ for $f(t)$. For this value of K the bounds are degenerate. The centre panel shows the confidence bounds for $K = 136.5$, 157.9 and 315.8 The bottom panel shows the corresponding fast bounds (87) and (88) for the same values of K.

lies beneath the upper bound ub. If we do this for the data of Figure 11 then the smallest value is 97.03 as against the correct bound of 115.0. The lower panel of Figure 11 shows the fast bounds for the same data and values of K.

7 Model choice and conclusions

The choice of a regression function $\tilde{f}_n$ for a data set $\boldsymbol{y}_n$ can be seen as a problem of model choice. The approach taken here is to define what is meant by an adequate approximation and then to choose the model by regularizing for shape or smoothness or both. There is no explicit penalty term for complexity and no smoothing parameter to be chosen. In all examples we take $\tau = 3$ and $\mathcal{I}_n$ to be a wavelet multiresolution scheme. Different regression functions result from different choices of regularization which in turn depend on the problem at hand. Our approach also has the advantage that models can be rejected outright For example the data shown in Figure 11 cannot be modelled by any $\tilde{f}_n$ with $\|\tilde{f}_n^{(2)}\|_\infty \leq 100$. It seems to us that any satisfactory procedure for model choice must be able to reject all models on offer as without this there is no motivation for looking for something better.

Finally we point out that our approach is general and tight. It can deal both with shape and smoothness regularization or a combination of both and at the same time it gives optimal rates of convergence for both types of regularization.

8 Acknowledgment

The authors gratefully acknowledge talks with Lutz Dümbgen which in particular lead to the smoothness regularization described in Section 5.

We also gratefully acknowledge the financial support of the German Science Foundation (Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475, Reduction of Complexity in Multivariate Data Structures).

References

[Baraud, 2004] Baraud, Y. (2004). Confidence balls in Gaussian regression. *Annals of Statistics*, 32(2):528–551.

[Bernholt and Hofmeister, 2006] Bernholt, T. and Hofmeister, T. (2006). An algorithm for a generalized maximum subsequence problem. In *LATIN*

2006: Theoretical informatics, volume 3887 of *Lecture Notes in Computer Science*, pages 178–189. Springer, Berlin.

[Cai and Low, 2006] Cai, T. T. and Low, M. G. (2006). Adaptive confidence balls. *Annals of Statistics*, 34(1):202–228.

[Dahlhaus, 1988] Dahlhaus, R. (1988). Small sample effects in time series analysis: a new asymptotic theory and a new estimate. *Annals of Statistics*, 16(2):808–841.

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

[Davies et al., 2006] Davies, P. L., Gather, U., and Weinert, H. (2006). Nonparametric regression as an example of model choice. Technical Report 24/06, Sonderforschungsbereich 475, Fachbereich Statistik, University of Dortmund, Germany.

[Davies and Kovac, 2001] Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution (with discussion). *Annals of Statistics*, 29(1):1–65.

[Davies and Meise, 2005] Davies, P. L. and Meise, M. (2005). Approximating data with weighted smoothing splines. Technical Report 48/05, Sonderforschungsbereich 475, Fachbereich Statistik, University of Dortmund, Germany.

[Donoho and Johnstone, 1994] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455.

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

[Dümbgen, 2003] Dümbgen, L. (2003). Optimal confidence bands for shape-restricted curves. *Bernoulli*, 9(3):423–449.

[Dümbgen and Johns, 2004] Dümbgen, L. and Johns, R. (2004). Confidence bands for isotonic median curves using sign-tests. *J. Comput. Graph. Statist.*, 13(2):519–533.

[Dümbgen and Spokoiny, 2001] Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics*, 29(1):124–152.

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

[Green and Silverman, 1994] Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and Generalized Linear Models: a roughness penalty approach*. Number 58 in Monographs on Statistics and Applied Probabality. Chapman and Hall, London.

[Groeneboom, 1996] Groeneboom, P. (1996). Inverse problems in statistics. In *Proceedings of the St. Flour Summer School in Probability*, number 1648 in Lecture Notes in Mathematics 1648, pages 67–164. Springer Verlag, Berlin.

[Hoffmann and Lepski, 2002] Hoffmann, M. and Lepski, O. (2002). Random rates in anisotropic regression. *Annals of Statistics*, 30(2):325–396.

[Li, 1989] Li, K.-C. (1989). Honset confidence regions for nonparametric regression. *Annals of Satistics*, 17:1001–1008.

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

[Mildenberger, 2006] Mildenberger, T. (2006). A note on the geometry of the multiresolution criterion. Technical Report 36/06, Sonderforschungsbereich 475, Fachbereich Statistik, University Of Dortmund, Germany.

[Nason, 1998] Nason, G. (1998). *WaveThresh3 Software*. Department of Mathematics, University of Bristol, Bristol, UK.

[Robins and van der Vaart, 2006] Robins, J. and van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *Annals of Statistics*, 34(1):229–253.

[Wahba, 1990] Wahba, G. (1990). *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

[Wand and Jones, 1995] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Monographs on Statistics and Applied Probability. Chapman and Hall, London.

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

9 Appendix

9.1 Proofs of Section 3.2

9.1.1 Proof of (21)

Let k be such that $I_c = [1/2 - k/n,\, 1/2 + k/n] \subset I_0$. A Taylor expansion together with (20) implies after some manipulation

$$\frac{1}{2k+1}\sum_{t_i \in I_c} f(t_i) - \sigma\frac{\sqrt{3\log n} + 2.72}{\sqrt{2k+1}} \geq f(1/2) - \frac{k^2}{2n^2}c_2 - \sigma\frac{\sqrt{3\log n} + 2.72}{\sqrt{2k}}$$

and on minimizing the right hand side of the inequality with respect to k we obtain

$$\frac{1}{|I_c|}\sum_{t_i \in I_c} f(t_i) - \sigma\frac{\sqrt{3\log n} + 2.72}{\sqrt{|I_c|}} \geq f(1/2) - 1.1 c_2^{1/5}\sigma^{4/5}\Big(\sqrt{3\log n} + 2.72\Big)^{4/5}/n^{2/5}. \tag{89}$$

This inequality holds as long as $I_c = [1/2 - k_n/n,\, 1/2 + k_n/n] \subset I_0$ with

$$k_n = \left\lfloor 0.66 c_2^{-2/5}\sigma^{2/5} n^{4/5}\Big(\sqrt{3\log n} + 2.72\Big)^{2/5} \right\rfloor. \tag{90}$$

If we put $I_l = [1/2 - (\eta+1)k_n/n,\, 1/2 - \eta k_n/n]$ similar calculations give

$$\frac{1}{2k+1}\sum_{t_i \in I_l} f(t_i) + \sigma\frac{\sqrt{3\log n} + 2.72}{\sqrt{2k+1}} \leq f(1/2) - \frac{k^2}{2n^2}c_1 + \sigma\frac{\sqrt{3\log n} + 2.72}{\sqrt{2k}}$$

and hence

$$\frac{1}{|I_l|}\sum_{t_i \in I_l} f(t_i) + \sigma\frac{\sqrt{3\log n} + 2.72}{\sqrt{|I_l|}} \geq f(1/2) - \frac{c_2^{1/5}\sigma^{4/5}\Big(\sqrt{3\log n} + 2.72\Big)^{4/5}}{n^{2/5}}\Big[0.2178\eta^2 c_1/c_2 - 1.23\Big]$$

with the same estimate for $I_r = [1/2 + \eta k_n/n,\ 1/2 + (\eta+1)k_n/n]$. If we put $\eta = 3.4\sqrt{c_2/c_1}$ and

$$I_n := \big[1/2 - (\eta+1)k_n/n,\ 1/2 + (\eta+1)k_n/n\big] \subset I_0 \tag{91}$$

then all estimates hold. Because of (90) this will be the case for n sufficiently large. This implies that (17) holds for sufficiently large n and in consequence any function $\tilde{f}_n \in \mathcal{A}_n$ has a local maximum in I_n.

9.1.2 Proofs of (22) and (23)

From (13) and (89) we have

$$f_n^*(t_n^*) \geq f(1/2) - 1.1c_2^{1/5}\sigma^{4/5}\Big(\sqrt{3\log n} + 2.72\Big)^{4/5}/n^{2/5}$$

which is the required estimate (22). To prove (23) we simply note

$$f_n^*(t_n^*) \leq f(t_n^*) + \sigma Z(t_n^*) + \sigma\sqrt{3\log n} \leq f(1/2) + \sigma(\sqrt{3\log n} + 2.4).$$

9.1.3 Proof of (30) and (31)

As $f_n^* \in \mathcal{A}_n$ by definition and $f \in \mathcal{A}_n$ with probability tending to one we have for the interval $I_{nk}^r = [i/n, (i+k-1)/n]$

$$\frac{1}{\sqrt{k}}\sum_{j=0}^{k-1} f_n^*((i+j)/n) \leq \frac{1}{\sqrt{k}}\sum_{j=0}^{k-1} f((i+j)/n) + 2\sigma\sqrt{3\log n}$$

from which it follows that

$$f_n^*(i/n) \leq f(i/n) + \frac{k}{n}\|f^{(1)}\|_{I_{nk}^r,\infty} + 2\sigma\sqrt{\frac{3\log n}{k}}$$

which proves (30). Similarly for the intervals $I_{nk}^l = [(i-k+1)/n,\ i/n]$ we have

$$f(i/n) - f_n^*(i/n) \leq \min_{1\leq k\leq k_n^{*l}}\left\{\frac{k}{n}\|f^{(1)}\|_{I_{nk}^l,\infty} + 2\sigma\sqrt{\frac{3\log n}{k}}\right\}. \tag{92}$$

We note that (30) and (92) imply that f_n^* adapts automatically to f to give optimal rates of convergence. If $f^{(1)}(t) \neq 0$ then it may be checked that the lengths of the optimal intervals I_{nk}^{r*} and I_{nk}^{l*} tend to zero and consequently

$$\|f^{(1)}\|_{I_{nk}^{l*},\infty} \approx |f^{(1)}(t)| \approx \|f^{(1)}\|_{I_{nk}^{r*},\infty}.$$

The optimal choice of k is then

$$k_n^{*l} \approx \left(\frac{3\sigma^2 n^2 \log n}{|f^{(1)}(t)|^2} \right)^{1/3} \approx k_n^{*r}$$

which gives

$$\lambda(I_{nk}^{l*}) \approx \frac{3^{1/3}\sigma^{2/3}}{|f^{(1)}(t)|^{2/3}} \left(\frac{\log n}{n} \right)^{1/3} \approx \lambda(I_{nk}^{l*})$$

from which (31) follows.

9.2 Proofs of Section 3.4

9.2.1 Proof of (34)

Then adapting the arguments used above we have for any differentiable function $\tilde{f}_n \in \mathcal{A}_n$

$$\begin{aligned} \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \left(\tilde{f}_n(1/2 + i/n) - \tilde{f}_n(1/2 - k/n + i/n) \right) \\ \geq \frac{1}{\sqrt{k}} \sum_{i=1}^{k} (f(1/2 + i/n) - f(1/2 - k^c/n + i/n)) \\ -2\sigma(\sqrt{3\log n} + \sqrt{2} Z(I_{nk}^c)) \end{aligned}$$

which implies

$$\max_{t \in I_{nk}^c} f_n^{*(1)}(t)/n \geq \min_{t \in I_{nk}^c} f^{(1)}(t)/n - \left(2\sigma(\sqrt{3\log n} + \sqrt{2} Z(I_{nk}^c)) \right)/k^{3/2}. \quad (93)$$

Similarly if $I_{nk}^l = [t_l - k/n, t_l + k/n]$ with $t_l + k/n < 1/2 - k/n$ we have

$$\min_{t \in I_{nk}^l} f_n^{*(1)}(t)/n \leq \max_{t \in I_{nk}^l} f^{(1)}(t)/n + \left(2\sigma(\sqrt{3\log n} + \sqrt{2} Z(I_{nk}^l)) \right)/k^{3/2} \quad (94)$$

and for $I_{nk}^l = [t_r - k/n, t_r + k/n]$ with $t_r - k/n > 1/2 + k/n$ we have

$$\min_{t \in I_{nk}^r} f_n^{*(1)}(t)/n \leq \max_{t \in I_{nk}^r} f^{(1)}(t)/n + \left(2\sigma(\sqrt{3\log n} + \sqrt{2} Z(I_{nk}^r)) \right)/k^{3/2}. \quad (95)$$

Again following the arguments given above we may deduce from (93), (94) and (95), that for sufficiently large n it is possible to choose I_{nk}^l, I_{nk}^c and I_{nk}^r so that (34) holds.

9.2.2 Proof of (38)

We have

$$\begin{aligned}&\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\big(f_n^*(k/n+i/n)-f_n^*(i/n)\big)\\ &\quad\leq\quad\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\big(f(k/n+i/n)-f(i/n)\big)+2\sigma\sqrt{3\log n}\,.\end{aligned}$$

and $f_n^{*(1)}$ is non-decreasing on I_{nk}^r we deduce

$$\frac{k^{3/2}}{n}f_n^{*(1)}(t)\leq\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\big(f(k/n+i/n)-f(i/n)\big)+2\sigma\sqrt{3\log n}\,.$$

A Taylor expansion for f yields

$$f_n^{*(1)}(t)\leq f^{(1)}(t)+\frac{k}{n}\|f^{(2)}\|_{I_{nk}^r,\infty}+2\sigma n\sqrt{\frac{3\log n}{k^3}}$$

from which (38) follows.