Approximating data with weighted smoothing splines[*]

P. L. Davies
Universität Duisburg-Essen
Technical University Eindhoven

M. Meise
Universität Duisburg-Essen

October 28, 2005

Abstract

Given a data set (t_i, y_i), $i = 1, \ldots, n$ with the $t_i \in [0,\,1]$ non-parametric regression is concerned with the problem of specifying a suitable function $f_n : [0,\,1] \rightarrow \mathbb{R}$ such that the data can be reasonably approximated by the points $(t_i, f_n(t_i))$, $i = 1, \ldots, n$. A common desideratum is that the function f_n be smooth but the path towards this goal is often the indirect one of assuming a "true" data generating function f and then measuring performance by the expected mean square. The approach taken in this paper is a different one. We specify precisely what we mean by a function f_n being an adequate approximation to the data and then, using weighted splines, we try to maximize the smoothness given the approximation constraints.

Keywords: Approximation; Residuals; Smoothing Splines; Thin Plate Splines

1 Contents

In Section 2 we give a short overview of research done on non-parametric regression on the real line. An alternative approach is sketched in Section 2.3. The problem of differentiable regression functions is considered in Section 3 where the method of weighted smoothing splines is described and compared with other methods by means of examples. An application is given to a problem in thin film physics in Section 3.3. Extensions for heteroscedastic data and a robustified version are given in Section 4. In Section 5 we indicate how the idea can be applied to the determination of local bandwidths for kernel methods. A further extension to image analysis is given in Section 6 and finally in Section 7 we give some results on asymptotics.

[*] This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475

2 Non-parametric regression

2.1 Previous work

In the one-dimensional case non-parametric regression is concerned with determing functions $f_n : [0,\, 1] \rightarrow \mathbb{R}$ which adequately represent a data set $(t_i, y(t_i)), i = 1, \ldots, n$ with the t_i in $[0,\, 1]$. There are many such procedures: we mention kernel estimates with fixed and local bandwidths (Eubank (1988, 1999), Härdle (1990) and Wand and Jones (1995)), local polynomials (Fan and Gijbels (1996) and Ruppert and Wand (1994)), adaptive weights smoothing (Polzehl and Spokoiny (2000)), wavelets (Donoho et al (1995)) splines (Wahba (1990), Green and Silverman (1994) and Eubank (1988, 1999), de Boor (1978, 2001) and Schumaker (1981)) and the taut string method (Davies and Kovac (2001)). Some of these procedures are automatic in that there exists automatic choices of relevant parameters. We mention hard and soft thresholding for wavelets, cross-validation for kernel estimates, splines and local polynomials (Wahba (1977) and Craven and Wahba (1979) Härdle and Marron (1985), Härdle et al (1988),Gasser et al (1991) and Härdle et al (1992)) and plug-in methods for kernel estimates (Brockmann et al (1993) and Herrmann (1997)). Another possibility is to use model choice criteria such as AIC and MDL (Hurvich et al (1998), Rissanen (2000)). Both adaptive weights smoothing (Polzehl and Spokoiny (2000)) and the taut string method (Davies and Kovac (2001)) have default values for the relevant parameters.

2.2 Asymptotics and rates of convergence

The function f_n produced by the procedure is usually required to be sufficiently close to the data to represent it accurately and to be simple or smooth in some sense. In most of the literature mentioned above the path to attaining these goals is to assume that the data are generated as described by the model

$$Y(t) = f(t) + \sigma Z(t), \quad 0 \le t \le 1, \tag{1}$$

and then to consider rates of convergence of f_n to f as measured in some norm such as

$$\|f - f_n\|_2^2 = \int_0^1 (f(t) - f_n(t))^2 \, dt. \tag{2}$$

If the function f has a continuous derivative of order s then convergence rates of order $n^{-s/(2s+1)}$ are possible. In the case of kernel estimators this can be attained by choosing the kernel to have zero moments of order $1, \ldots, s-1$ and a non-zero moment of order s. To take advantage of it the user must know the value of s for the "true" data generating function f. Some methods such as wavelets adapt automatically to the unknown smoothness of f upto a limit depending on the procedure . For example if the wavelets have r vanishing moments and r continuous derivatives then an appropriate procedure will automatically adjust to the number of derivatives of f up a limit $s < r$ (Donoho and Johnstone (1994) and (1995)). The procedures will not only adjust in the global norm (2) but also in local versions

$$\|f - f_n\|^2_{x_0,c_n} = \frac{1}{2c_n} \int_{x_0-c_n}^{x_0+c_n} (f(t) - f_n(t))^2 \, dt, \tag{3}$$

(Cai and Low (2005)).

In the approach to non-parametric regression described above the goals of closeness to the data and smoothness of the regression function attained only indirectly. Although some procedures such as wavelets work well under certain circumstances others do not. If we consider kernel estimators and suppose that f has a continuous second derivative then the optimal global bandwidth h_n is of the order of $n^{-2/5}$ and gives rise to the optimal rate of convergence in integrated mean square error of $n^{-4/5}$. The optimal bandwidth depends on the second derivative of f but if the bandwidth is chosen by cross-validation then it is asymptotically optimal (Stone (1982)). It follows that for a sufficiently large sample size the function f_n will be close to the generating function f. Moreover it will be smooth and the quantitative smoothness will improve as the sample size becomes larger. Similar considerations apply to the choice of the penalizing factor λ in the maximum penalized likelihood procedure where the function f_n is the solution of

$$\text{minimize} \quad S_\lambda(f) := \sum_{i=1}^{n} \lambda(y_i - f(t_i))^2 + \int_0^1 f^{(2)}(x)^2 \, dx. \tag{4}$$

Although asymptotically everything is satisfactory it is the case that for finite samples and in particular where the data show large local variation it is not possible to simultaneously be close to the data and to be smooth as there is no choice of global bandwidth which the function f_n is satisfactory in both senses.

2.3 Approximation and regularization

The approach we take in this paper is a different one. We give a precise definition of what is meant by a function f_n being sufficiently close to the data and then, given these side conditions, we maximize the simplicity or the smoothness of the functions which satisfy them. These leads to an optimization problem for the data at hand. The choice of side conditions and the defintion of simplicity or smoothness will depend on the data and the questions to be answered. In this we follow Davies and Kovac (2001) and (2004). Given data $(t_i, y(t_i)), i = 1, \ldots, n$ and a function f_n we consider the residuals

$$r(t_i, f_n) = y(t_i) - f_n(t_i), \quad i = 1, \ldots, n \tag{5}$$

and their normalized sums over intervals $I \subset [0,\ 1]$

$$w(I, f_n) = \frac{1}{\sqrt{|I|}} \sum_{t_i \in I} r(t_i, f_n), \tag{6}$$

where $|I|$ denotes the number of points t_i in I. The intervals will be restricted to a family $\mathcal{I}_n$ of intervals of [0, 1] which may be the set of all intervals but for large n will be a family of size of order n such as ones defined by multiresolution schemes. The idea is to give an upper bound for the $w(I, f_n)$ and so to force the function f_n to be close to the data. The upper bound we use is based on the model (1) and is based on the behaviour of the maximum of gaussian random variables. It leads to

$$\max_{I \in \mathcal{I}_n} |w(I, f_n)| \leq \sigma \sqrt{\tau \log(n)}. \tag{7}$$

for some $\tau > 2$. These bounds hold asymptotically for every $\tau > 2$ if the data were generated according to (1) with $f_n = f$. To be of use we must quantify the noise level corresponding to σ. We put

$$\sigma_n = \frac{1.483}{\sqrt{2}} \text{Median}\{|y(t_2) - y(t_1)|, \ldots, |y(t_n) - y(t_{n-1})|\} \tag{8}$$

(see Section 5.4 of Donoho and Johnstone (1995)) which together with the default value $\tau = 2.3$ leads to the approximation conditions

$$\max_{I \in \mathcal{I}_n} |w(I, f_n)| \leq \sigma_n \sqrt{2.3 \log(n)} \tag{9}$$

(see Davies and Kovac (2001)).

The second part of the approach is to regularize the function f_n subject to the bounds (9). In Davies and Kovac (2001) it was proposed to minimize the number of local extremes of f_n subject to (9) and the taut string method was developed to accomplish this. A smooth function can be obtained by minimizing for example

$$\sum_{i=3}^{n} (f_n(t_{i-2}) + f_n(t_i) - 2f_n(t_{i-1}))^2$$

subject to (9). Here for the sake of simplicity we have assumed that the t_i lie on a grid. We note that the monotonicity constraints derived from the taut string solution can be incorporated if required. This gives rise to a quadratic programming problem and the solution gives a very smooth regression function which is both close to the data and has the correct monotonicity behaviour. The main problem is the lack of numerical stability when the data show large variations in local behaviour and unfortunately this makes the method unsuitable for general applications. We refer to Majidi (2003) for further details.

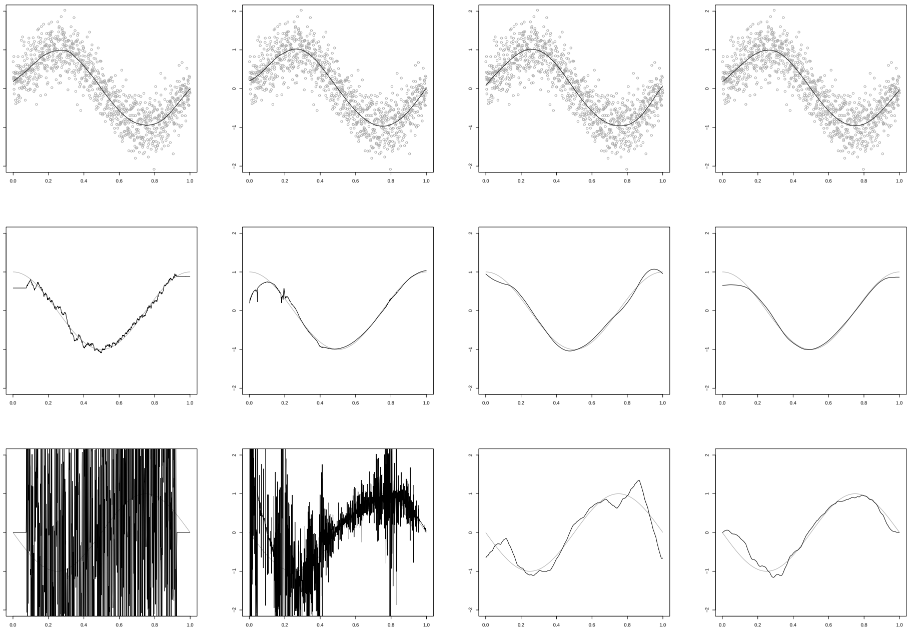

Figure 1: From top to bottom the non-parametric approximation and its first and second derivatives. From left to right kernel approximation with local bandwidths, AWS, wavelets and a smoothing spline.

3 Weighted splines and differentiable approximations

3.1 Weighted splines

The standard cubic smoothing spline is the unique solution of (4) for a given choice of λ and as mentioned above there may well be no value of λ which gives a satisfactory representation of the data. This reflects the inability of the smoothing spline to adapt locally to the smoothness of the function. To overcome this problem we propose the use of weighted smoothing splines defined as the solution of

$$\text{minimize } S(f, \boldsymbol{\lambda}) := \sum_{i=1}^{n} \lambda_i (y(t_i) - f(t_i))^2 + \int_0^1 f^{(2)}(t)^2 \, dt \tag{10}$$

for a given $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$. If all the $\lambda_i > 0$ then the unique solution of the minimization problem of (10) is a natural cubic spline which we denote by $f_n(\cdot : \boldsymbol{\lambda})$. We now explain how the weights λ_i are determined. We initialize $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda}_1$ by setting the λ_i to a small common value such that $f_n(\cdot : \boldsymbol{\lambda}_1)$ is approximately linear. We then check whether $f_n(\cdot : \boldsymbol{\lambda}_1)$ is an adequate approximation in the sense of (9). If it is then we terminate the procedure and the regression function is simply $f_n(\cdot\, : \boldsymbol{\lambda}_1)$. If some of the inequalities of (9) are not satisfied we note all the values of i with t_i in such an interval and for these i we increase the λ_i by a factor q. Our default value for q is 2. We denote this new value of $\boldsymbol{\lambda}$ by $\boldsymbol{\lambda}_2$. If $f_n(\cdot\, : \boldsymbol{\lambda}_2)$ is an adequate approximation then the procedure terminates and $f_n(\cdot\,, \boldsymbol{\lambda}_2)$ is the regression function. Otherwise the procedure is continued until the first adequate approximation is found. It is clear that the procedure will terminate because the function $f_n(\cdot\, : \boldsymbol{\lambda})$ will almost interpolate the data for large λ_i. For a more detailed description of the procedure see Meise (2004).

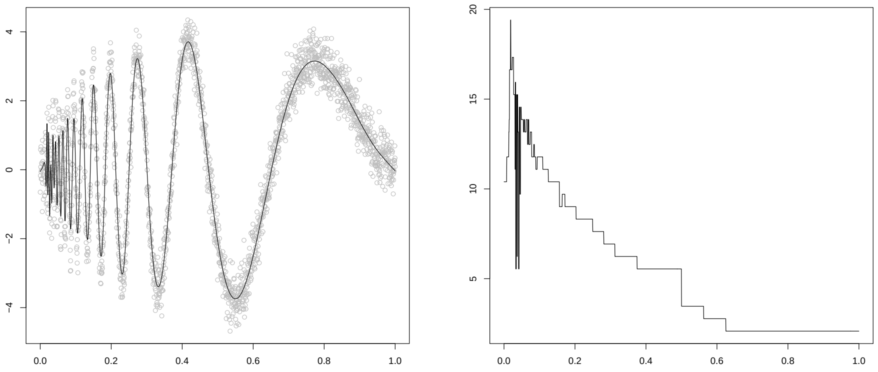

Figure 2: The left panel shows the weighted smoothing spline regression function for the Doppler data. The right panels show the final values of the $\log(\lambda_i)$.

3.2 Examples

Figure 1 shows a sine curve contaminated by Gaussian noise together. The top row shows, from left to right, a kernel estimator with local bandwidths, the AWS estimator of Polzehl and Spokoiny, wavelets and the weighted smoothing spline. The centre row shows the first derivatives and the bottom row the second derivatives.

It is evident from this figure and it is indeed the general case that wavelet methods and weighted smoothing splines perform best so in future we shall only compare these two. This and each of the following examples was computed using the statistics software R and additional available packages as aws, lokern and WaveThresh3 (Nason (1998)).

The second example is the Doppler data of Donoho and Johnstone (1995). Figure 2 shows the weighted smoothing reconstruction (left panel) and the final values of

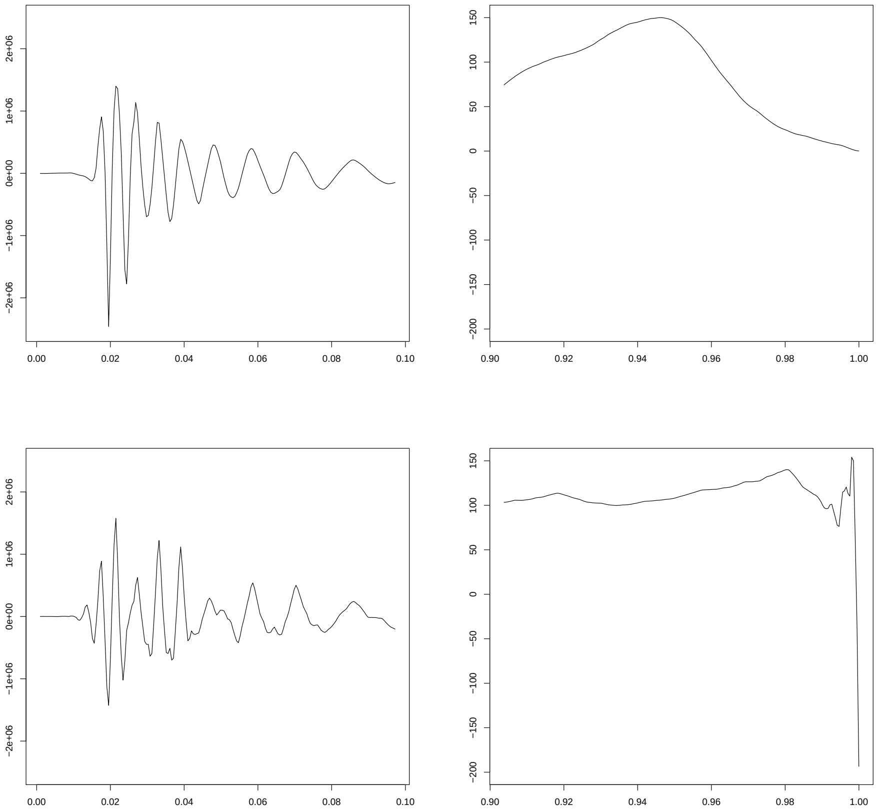

Figure 3: The upper row shows the second derivative of the smoothing spline regression function for the first and last 200 observations of the Doppler data. The bottom row shows the corresponding results for the wavelet reconstruction.

the λ_i (right panel). The top row of Figure 3 show the first and the last 200 values of the second derivative of the weighted smoothing spline reconstruction. The second row shows the corresponding values of the wavelet approximation.

The t_i of Figures 1 and 2 are of the form $t_i = i2^{-k}$ and were specifically chosen with wavelets in mind. In general the t_i will not form a grid and the number of data points will not be a power of two. In this case the performance of weighted smoothing splines is hardly impaired but the performance of the wavelet reconstruction deteriorates perceptibly. This is shown in Figure 4 where $n = 100$ and the t_i were chosen at random. For a description of the used irregular wavelet transform see Kovac and Silverman (2000).

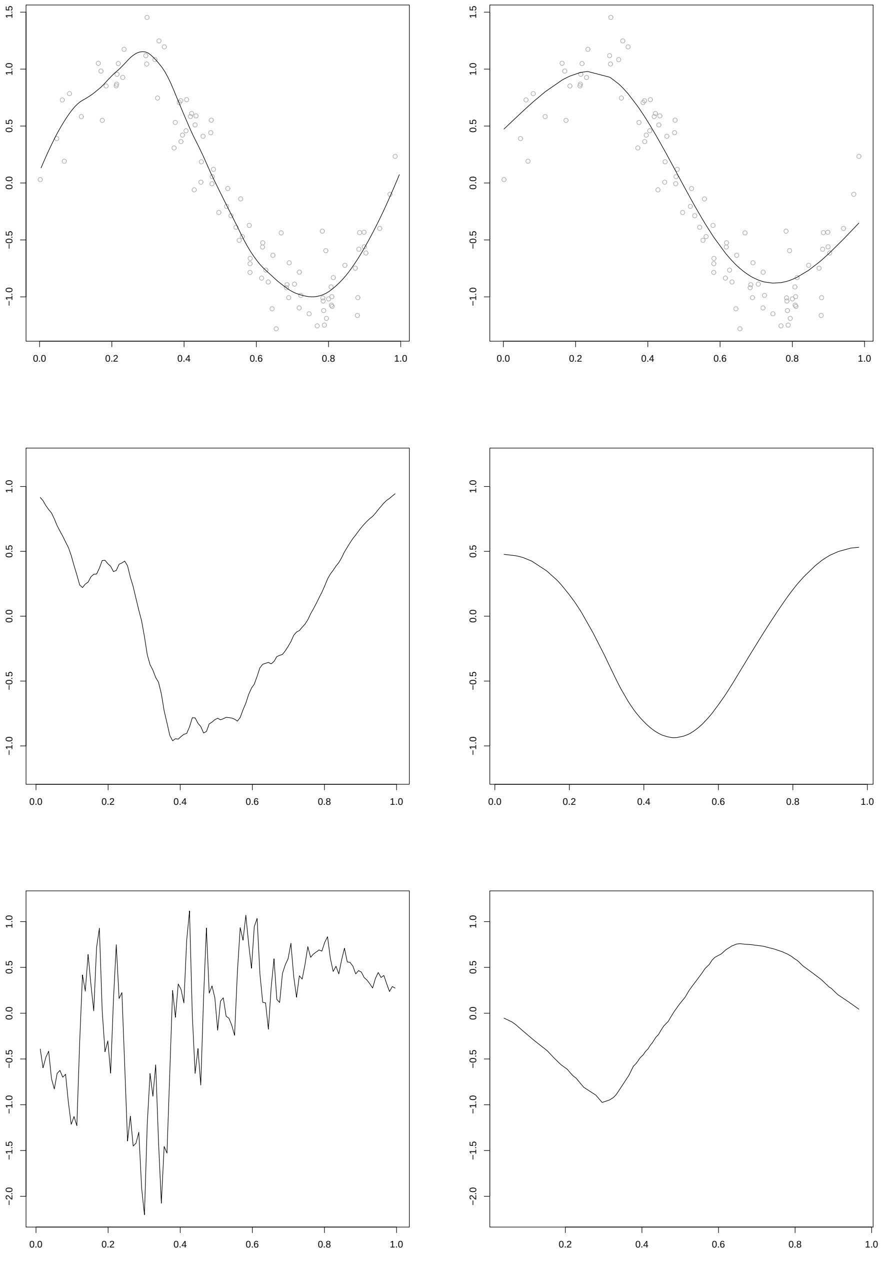

Figure 4: An irregular spaced noisy sine of sample size 100. The upper left panel shows the wavelet approximation and the upper right panel the weighted smoothing spline. Second and third row show the corresponding first and second derivatives.

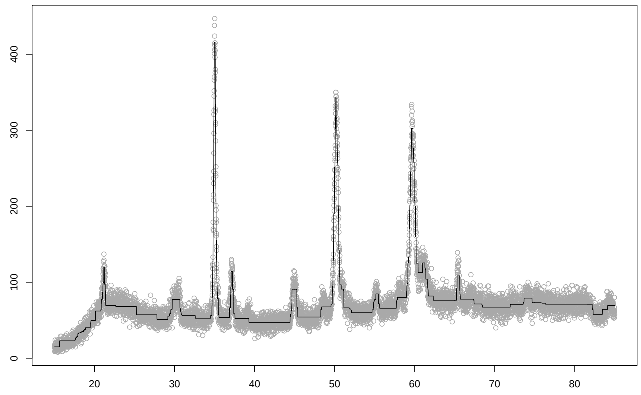

Figure 5: Taut string approximation to the thinfilm data.

3.3 An example from thin-film physics

Figure 5 shows some data from thin-film physics together with the default taut string approximation (Davies and Kovac (2001)). The data were kindly supplied by Prof. Dieter Mergel of the University Duisburg-Essen and show the photon counts of reflected X-rays as a function of the angle of deflection. Interest centres on the location and the power of the peaks. As can be seen from Figure 5 the taut string identifies the number and location of the peaks very well but the power is mote complicated as it is to be measured from the slowly varying baseline evident in Figure 5. We solve the problem by using the weighted spline approximation and determine the baseline by the values of the first derivative. These are shown in Figure 6. The upper panel shows the weighted spline approximation and the lower panels shows the first derivative truncated to values to lie between 0.15 and -0.15. Figure 7 shows the baseline.

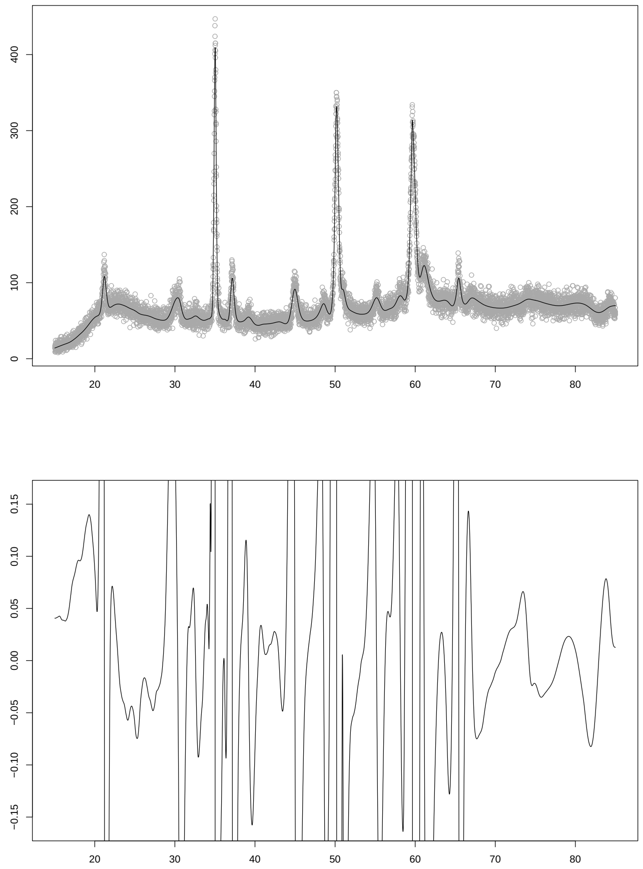

Figure 6: The weighted smoothing spline approximation to the thinfilm data and its first derivative.

4 Heteroscedasticity and robustness

4.1 Nonparametric scale approximations

The ideas developed in the previous section can also be used to obtain nonparametric approximations to data with varying scale. The model we use it

$$Y(t) = \sigma(t)Z(t), \quad 0 \le t \le 1, \quad Z(t) \quad \text{Gaussian white noise,} \tag{11}$$

and given data $(t_i, y(t_i), i = 1, \ldots, n$ we look for a representation $y(t_i) = \sigma_n(t_i)r(t_i)$ where σ_n is simple and the $r(t_i)$ "look like" standard Gaussian white noise. The concept of approximation we use is based on the sums

$$v(I, s_n) = \sum_{t_i \in I} r(t_i)^2 = \sum_{t_i \in I} y(t_i)^2/\sigma_n(t_i)^2 \tag{12}$$

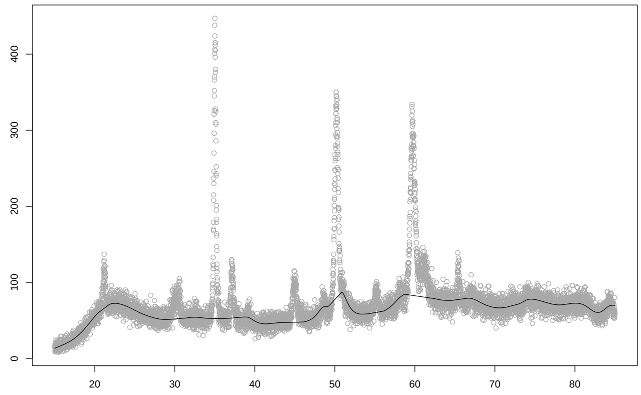

Figure 7: Baseline approximation.

Under the model (11) these should "look like" chi-squared random variables with $|I|$ degrees of freedom. This leads to the following set of inequalities

$$\mathrm{qu}((1-\gamma_n)/2, |I|) \le v(I, \sigma_n) \le \mathrm{qu}((1+\gamma_n)/2, |I|), \quad I \in \mathcal{I}, \tag{13}$$

where $\mathrm{qu}(\gamma, k)$ denote the $\gamma-$quantile of the chi-squared distribution with k degrees of freedom. The default value of γ_n we use is

$$\gamma_n = 1 - \exp(-1.15 \log(n)) = 1 - n^{-1.15} \tag{14}$$

which corresponds to the default choice of $\tau = 2.3$ in (9). As we are looking for a smooth approximating function s we consider the solution of the weighted smoothing spline problem

$$\text{minimize} \quad \sum_{i=1}^{n} \lambda_i(|y_i| - \sigma_n(t_i))^2 + \int_0^1 \sigma_n^{(2)}(t)^2\, dt. \tag{15}$$

The local weights are data dependent and are chosen so that the solution s satisfies (13). The procedure we use is similar to that described in Section 3.1 but with some modifications. On intervals I where the inequality (13) is not satisfied we increase the weights by a factor of q but we do this firstly for single observations, that is interval of length one. When (13) is satisfied for all such intervals we consider intervals of length two. When again all the inequalities are satisfied we move on to the next longer intervals until finally all inequalities are satisfied. A similar procedure was used in Davies and Kovac (2004) in the context of approximating spectral densities. Figure 8 shows the result of the procedure applied to data generated according to the model

$$Y(t) = \sin(4\pi t)^2 Z(t). \tag{16}$$

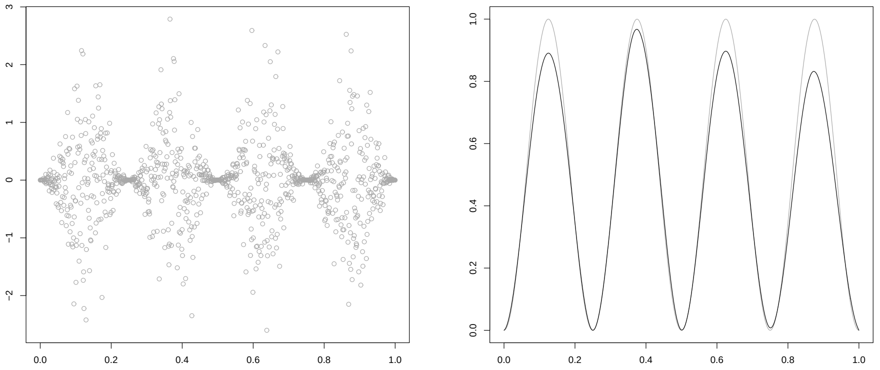

Figure 8: The left panel shows data generated according to (16). The right panel shows the generating curve $\sin(4\pi t)^2$ and the reconstruction using weighted splines.

The left panel shows the data and the right panel shows the generating curve and the reconstruction.

4.2 Heteroscedastic data

We can combine the procedures of Sections 3.1 and 4.1 to deal with heteroscedastic data. The model is

$$Y(t) = f(t) + \sigma(t)Z(t), \quad 0 \le t \le 1. \tag{17}$$

Given data $(t_i, y(t_i)), i = 1, \ldots, n$ we start by quantifying the noise. We put

$$v(t_i) = |y(t_{i+1}) - y(t_i)|, \quad i = 1, \ldots, n-1 \tag{18}$$

and then apply the procedure of Section 4.1 to the points $(t_i, v(t_i)), i = 1, \ldots, n-1$. This gives a non-parametric approximation $\sigma_n(t_i)$ to the noise level of the $v(t_i)$. We put

$$s_n(t_i) = \sigma_n(t_i)/\sqrt{2}, \quad 1 \le i \le n-1; \quad s_n(t_n) = \sigma_n(t_{n-1})/\sqrt{2} \tag{19}$$

to obtain an approximation to the noise level of the original data. We now replace the $w(I, f_n)$ of (6) by

$$w(I, f_n) = \frac{1}{\sqrt{|I|}} \sum_{t_i \in I} r(t_i, f_n)/s_n(t_i) \tag{20}$$

and then use (9) as the definition of approximation. The procedure of Section 3.1 is now applied to the data $(t_i, y(t_i)), i = 1, \ldots, n$.

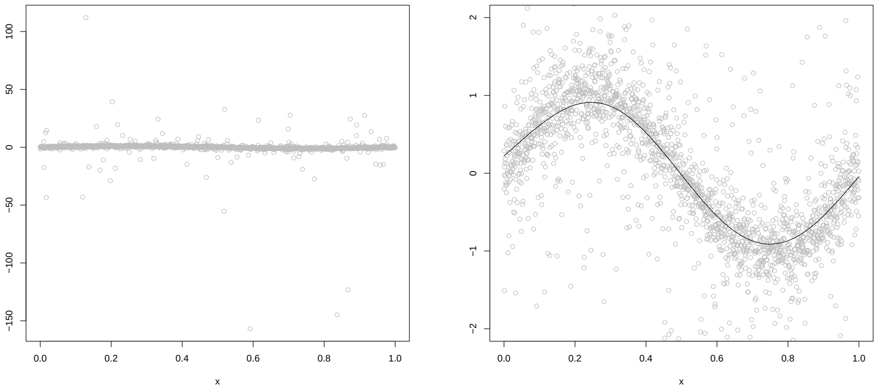

Figure 9: The robustified weighted spline procedure applied to a sine curve contaminated with cauchy noise.

4.3 Robust smoothing

A complete robustification of the procedure described in Section 3.1 would entail replacing (10) by, for example,

$$\text{minimize } S(f, \boldsymbol{\lambda}) := \sum_{i=1}^{n} \lambda_i |y(t_i) - f(t_i)| + \int_0^1 f^{(2)}(t)^2 \, dt, \tag{21}$$

the definition of approximation (6) by

$$\tilde{w}(I, f_n) = \frac{1}{\sqrt{|I|}} \sum_{t_i \in I} \operatorname{sgn}(r(t_i, f_n)) \tag{22}$$

and finally (9) by

$$\max_{I \in \mathcal{I}_n} |\tilde{w}(I, f_n)| \leq \sigma_n \sqrt{2 \log(n)} \tag{23}$$

(see Kovac (2002)). A much simpler but reasonably effective method is the following. The noise level σ_n is quantified by (8). A running median with a window width of say five observations is applied to the data

$$m_5(t_i) := \text{median}(y(t_{i-2}), y(t_{i-1}), y(t_i), y(t_{i+1}), y(t_{i+2})$$

and any data point $y(t_i)$ for which

$$|y(t_i) - m_5(t_i)| \geq 3.5\sigma_n,$$

is replaced by $m_5(t_i)$ (see Hampel (1985)). The weighted splines procedure is now applied to the cleaned data set. The procedure will work well as long as no group of five successive observations contains more than two outliers. Figure 9 shows the result of applying this robustified procedure to a sine curve contaminated with Cauchy noise.

5 Approximations using kernels and local polynomials

The techniques described above can also be applied to determining the local bandwidths h_i for kernel and local polynomial approximations. For simplicity we describe the method only for kernel approximations f_n^k of the form

$$f_n^k(t) = \frac{\sum_{i=1}^n y(t_i) K\left(\frac{t_i - t}{h(t)}\right)}{\sum_1^n K\left(\frac{t_i - t}{h(t)}\right)}. \tag{24}$$

Here $K : [0,\, 1] \rightarrow \mathbb{R}$ denotes a smooth symmetric kernel of the sort usually chosen in this situation. We commence with constant bandwidths

$$h(t_1) = \ldots = h(t_n) = h_0$$

for some large h_0, calculate the function $f_{n,1}^k$ according to (24) and then the associated residuals. If (9) holds the procedure terminates. Otherwise we reduce the size of the local bandwidths $h(t_i)$ at all points t_i which lie in intervals where (9) does not hold to $qh(t_i)$. The default value we use for q is $q = 0.8$. This process is repeated until (9) is satisfied for all intervals. Figure 10 shows the result using local polynomials of order 1 applied to the Doppler data. The artefact close to $t = 0.75$ in the upper left panel is due to the large discontinuity in the local bandwidths at this point (lower left panel). If the bandwidths are smoothed but whilst still maintaining (9) then the artefact disappears as may be seen from the upper right panel.

6 Image analysis and weighted thin plate splines

6.1 Weighted thin plate splines

We consider data $(\boldsymbol{t}_i, y(\boldsymbol{t}_i))$, $i = 1, \ldots, n^2$ with the $\boldsymbol{t}_i$ of the form

$$\boldsymbol{t}_i = (j_i/n, k_i/n), \quad j_i, k_i = 0, \ldots, n-1.$$

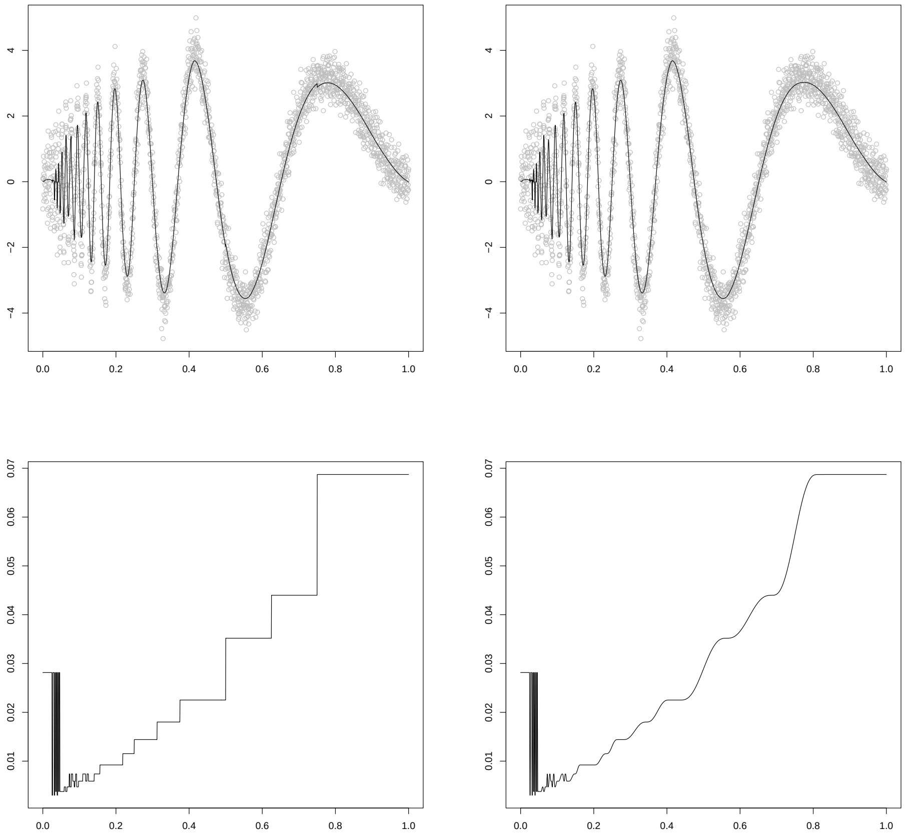

Figure 10: Local polynomial (p=1) approximation with piecewise constant and smoothed local bandwidth.

Corresponding to (10) we consider minimizing

$$\begin{aligned} S(f,\boldsymbol{\lambda}) := & \sum_{i=1}^{n^2} \lambda(\boldsymbol{t}_i)(y(\boldsymbol{t}_i) - f(\boldsymbol{t}_i))^2 \qquad (25) \\ + & \int_0^1 \int_0^1 \left(\left(\frac{\partial^2 f(s,t)}{\partial^2 s} \right)^2 + \left(\frac{\partial^2 f(s,t)}{\partial s \partial t} \right)^2 + \left(\frac{\partial^2 f(s,t)}{\partial^2 t} \right)^2 \right) dsdt. \end{aligned}$$

It can be shown that the solution is a natural thin plate spline. We refer to Green and Silverman (1994).

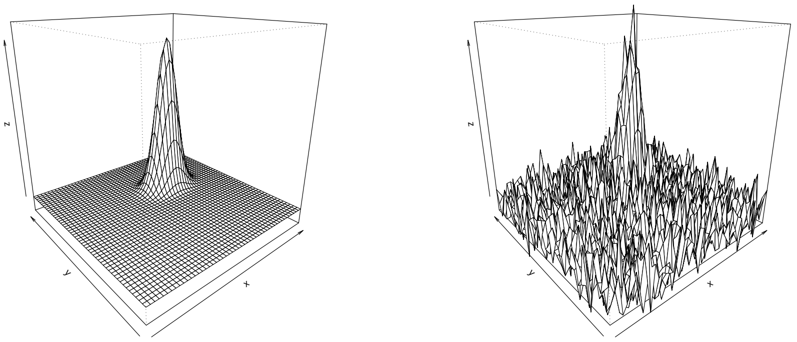

Figure 11: The original function (left) and the noisy data (right).

6.2 Approximation in two dimensions

The given a function f_n the corresponding residuals are given by

$$r(\boldsymbol{t}_i, f_n) = y(\boldsymbol{t}_i) - f_n(\boldsymbol{t}_i), \quad \boldsymbol{t}_i = (j_i/n, k_i/n), \quad j_i, k_i = 0, \ldots, n-1. \tag{26}$$

For a given family $\mathcal{C}_n$ of subsets C of $[0,\ 1]^2$ we define their normalized sums by

$$w(C, f_n) = \frac{1}{\sqrt{|C|}} \sum_{\boldsymbol{t}_i \in C} r(\boldsymbol{t}_i, f_n). \tag{27}$$

which leads to the following definition of approximation

$$\max_{C \in \mathcal{C}_n} |w(C, f_n)| \le \sigma_n \sqrt{4.6 \log(n)} \tag{28}$$

where the factor 4.6 replaces the factor 2.3 in (9) as we now have n^2 observations. The noise level σ_n is defined

$$\sigma_n = \frac{1.48}{2} \text{median}(|y(\tfrac{j_i+1}{n}, \tfrac{k_i+1}{n}) - y((\tfrac{j_i+1}{n}, \tfrac{k_i}{n}) - y(\tfrac{j_i}{n}, \tfrac{k_i+1}{n}) + y(\tfrac{j_i}{n}, \tfrac{k_i}{n})|,\ i = 1, \ldots, n^2) \tag{29}$$

The quality of the results depends on the choice of $\mathcal{C}_n$. If $\mathcal{C}_n$ contains too few sets then the concept of approximation is too crude. We therefore require $\mathcal{C}_n$ to allow fines divisions of $[0,\ 1]^2$ and also to be such that the residuals (26) can be efficiently calculated. Work in this direction has been done and we refer to Friedrich (2005). Such a segmentation can for example be the subdivision of $[0, 1]^2$ into all possible squares, containing at least one point $\boldsymbol{t}_i$. This is the one we use here. Others are possible and might also provide partitions bounded by line or arc segments.

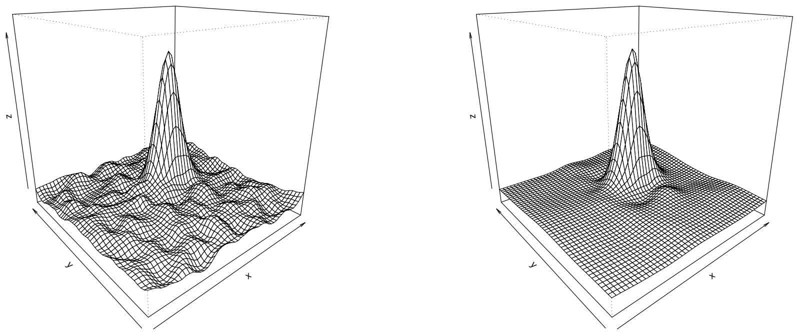

Figure 12: A normal thin plate approximation using GCV (left) and the automatically weighted version.

6.3 An example

As a simple example we consider the function $F : \mathbb{R}^2 \to \mathbb{R}$

$$F(x, y) = 10 \exp(-x^2 - 2y^2) \tag{30}$$

on a 50×50 grid on $[-7, 4]^2$ with added normal noise, $\varepsilon_i \sim N(0, 1)$ (Figure 11). The weighted thin plate approximation outperforms the procedure with a global penalizing parameter chosen by generalized cross validation as is seen from Figure 12. The collection $\mathcal{C}_n$ of subsets used in this example was the set of all possible squares containing at least one point $\boldsymbol{t}_i$. The main drawback of weighted thin plate splines is the numerical difficulty of calculating them for larger grids. More work in this direction is required.

7 Asymptotics

We consider the one- and the two-dimensional case, $d \in \{1, 2\}$, which can be written in the form

$$\text{minimize} \quad S_\lambda(f) := \sum_{i=1}^{n} \lambda(y_i - f_n(i))^2 + f_n^t \Omega_n f_n \tag{31}$$

where Ω_n is an $n \times n$-non-negative definite matrix with eigenfunctions g_{ni} and corresponding eigenvalues $\gamma_{ni}, 1 \le i \le n$ with $\gamma_{n1} = \gamma_{n2} = 0$. The remaining eigenvalues satisfy the inequalities

$$c_1 \frac{i^{4/d}}{n} \le \gamma_{ni} \le c_2 \frac{i^{4/d}}{n}, \quad 3 \le i \le n \tag{32}$$

with the constants c_1 and c_2 being independent of n. We denote the corresponding normalized eigenvectors by g_{ni}. For an interval I and squares respectively we denote by θ_I the vector whose elements θ_i are $1/\sqrt{|I|}$ for $i \in I$ and 0 otherwise. We see that $\|\theta_I\| = 1$ and for the solution $\tilde{f}_n$ of (31) the $w(I, \tilde{f}_n)$ of (6) are given by

$$w(I, \tilde{f}_n) = \theta_I^t(y_n - \tilde{f}_n), \quad I \in \mathcal{I}. \tag{33}$$

We have

Theorem 7.1

(a) $\tilde{f}_n^t(\lambda)\Omega_n\tilde{f}_n(\lambda)$ *is an increasing function of* λ.

(b) $\mathbb{E}\left(\tilde{f}_n^t(\lambda)\Omega_n\tilde{f}_n(\lambda)\right) \le cn^{d/4}\lambda^{d/4+1}$ *for some constant* c.

(c) For all $\lambda > 0$, *for* $\mathcal{I}_n$ *with* $|\mathcal{I}_n| \le qn$ *for some fixed* q *and for all* $\tau > 2$ *we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{I\in\mathcal{I}_n} |w(I, \tilde{f}_n(\lambda))| \le \sigma\sqrt{\tau \log(n)}\right) = 1.$$

Proof. (a) The solution $\tilde{f}_n(\lambda)$ of (31) is given by

$$\tilde{f}_n(\lambda) = \lambda(\lambda I_n + \Omega_n)^{-1}y_n$$

and on writing $y_n = \sum_{i=1}^n \eta_{ni}g_{ni}$ we obtain

$$\tilde{f}_n^t(\lambda)\Omega_n\tilde{f}_n(\lambda) = \lambda^2 \sum_{i=3}^n \frac{\eta_{ni}^2\gamma_{ni}}{(\lambda + \gamma_{ni})^2}$$

from which the claim follows on noting that $\gamma_{ni} > 0$ for $i \ge 3$.

(b) We note that

$$\tilde{f}_n^t(\lambda)\Omega_n(\lambda)\tilde{f}_n = \lambda^2 y_n^t(\lambda I_n + \Omega_n)^{-1}\Omega_n(\lambda I_n + \Omega_n)^{-1}y_n$$

and hence

$$\begin{aligned}\mathbb{E}\left(\tilde{f}_n^t(\lambda)\Omega_n\tilde{f}_n(\lambda)\right) &= \lambda^2 f_n^t(\lambda I_n + \Omega_n)^{-1}\Omega_n(\lambda I_n + \Omega_n)^{-1}f_n \\ &\quad + \mathbb{E}\left(\lambda^2\epsilon_n^t(\lambda I_n + \Omega_n)^{-1}\Omega_n(\lambda I_n + \Omega_n)^{-1}\epsilon_n\right).\end{aligned}$$

Arguing as above we obtain

$$\begin{aligned}\lambda^2 f_n^t(\lambda I_n + \Omega_n)^{-1}\Omega_n(\lambda I_n + \Omega_n)^{-1}f_n &= \lambda^2 \sum_3^n \alpha_{ni}^2 \frac{\gamma_{ni}}{(\lambda + \gamma_{ni})^2} \\ &\le \sum_3^n \alpha_{ni}^2\gamma_{ni} = f_n^t\Omega f_n\end{aligned}$$

and

$$\mathbb{E}(\lambda^2\epsilon_n^t(\lambda I_n+\Omega_n)^{-1}\Omega_n(\lambda I_n+\Omega_n)^{-1}\epsilon_n)=\lambda^2\sum_3^n\frac{\gamma_{ni}}{(\lambda+\gamma_{ni})^2}.$$

On splitting the last summation into two parts, from $i=3$ to $i=n^{d/4}\lambda^{d/4}$ and from $i=n^{d/4}\lambda^{d/4}$ to $i=n$ and on using (32) it follows that

$$\mathbb{E}(\lambda^2\epsilon_n^t(\lambda I_n+\Omega_n)^{-1}\Omega_n(\lambda I_n+\Omega_n)^{-1}\epsilon_n)\le cn^{d/4}\lambda^{d/4+1}$$

for some constant c which completes the proof of the theorem.
(c) We have

$$y_n-\tilde{f}_n(\lambda)=(\lambda I_n+\Omega_n)^{-1}\Omega_n y_n.$$

and on writing $y_n=f_n+\epsilon_n$ we obtain

$$y_n-\tilde{f}_n(\lambda)=h_n+\delta_n \tag{34}$$

with

$$h_n=(\lambda I_n+\Omega_n)^{-1}\Omega_n f_n,\quad \delta_n=(\lambda I_n+\Omega_n)^{-1}\Omega_n\epsilon_n. \tag{35}$$

On writing

$$f_n=\sum_1^n\alpha_{ni}g_{ni}$$

we obtain

$$h_n=\sum_3^n\alpha_{ni}\frac{\gamma_{ni}}{(\lambda+\gamma_{ni})}g_{ni}$$

and hence

$$\|h_n\|^2=\sum_3^n\alpha_{ni}^2\frac{\gamma_{ni}^2}{(\lambda+\gamma_{ni})^2}=\frac{1}{\lambda}\sum_3^n\alpha_{ni}^2\frac{\gamma_{ni}^2/\lambda}{(1+\gamma_{ni}/\lambda)^2}\le\frac{1}{\lambda}\sum_3^n\alpha_{ni}^2\gamma_{ni}.$$

As $f_n^t\Omega_n f_n=\sum_3^n\alpha_{ni}^2\gamma_{ni}$ we see that at least asymptotically

$$\|h_n\|^2\le\frac{1}{\lambda}f_n^t\Omega_n^{(d)}f_n. \tag{36}$$

We turn to δ_n. We write

$$\epsilon_n=\sum_1^n Z_{ni}g_{ni}$$

where, because of the transformation is orthonormal, the Z_{ni} are i.i.d. Gaussian random variables with zero mean and variance σ^2. It follows

$$\delta_n=\sum_3^n Z_{ni}\frac{\gamma_{ni}}{(\lambda+\gamma_{ni})}g_{ni}$$

and on writing

$$\theta_I = \sum_1^n \theta_{ni} g_{ni}$$

we obtain

$$\mathbb{E}((\theta_I^t \delta_n)^2) = \sigma^2 \sum_3^n \theta_{ni}^2 \left(\frac{\gamma_{ni}}{\lambda + \gamma_{ni}} \right)^2 \le \sigma^2.$$

The claim of the theorem follows from the usual upper bound for the tail of a Gaussian distribution. □

We consider the following modified procedure. We consider the solutions $\tilde{f}_n(\lambda)$ of (31) and determine the smallest value of λ for which the multiresolution conditions (7) are fulfilled with $\tilde{f}_n = \tilde{f}_n(\lambda)$. It follows from (c) of Theorem 7.1 this smallest value is asymptotically with arbitrarily large probability smaller than any given λ_0. If we denote this solution by $\tilde{f}_n(\lambda_n^*)$ then it follows from (a) and (b) of Theorem 7.1 that

$$\lim_{n\to\infty} \mathbb{P}\left(\tilde{f}_n^t(\lambda_n^*) \Omega_n \tilde{f}_n(\lambda_n^*) \le c n^{d/4} \lambda_0^{d/4+1} \right) = 1. \tag{37}$$

Let $\hat{f}_n$ be the solution obtained from the weighted splines procedure described in Section 3 and 6 respectively. If

$$\hat{f}_n^t \Omega_n^{(d)} \hat{f}_n \le \tilde{f}_n^t(\lambda_n^*) \Omega_n \tilde{f}_n(\lambda_n^*)$$

then we accept $\hat{f}_n$ and otherwise we accept $\tilde{f}_n(\lambda_n^*)$ and denote the solution by f_n^*.

7.1 The one-dimensional case

For the one-dimensional case we have

Theorem 7.2 *If f has a continuous second derivative then*

$$\lim_{c\to\infty} \lim_{n\to\infty} \mathbb{P}\left(\|f_n^* - f\|_{n,\infty} \le c n^{-1/3} \log(n)^{1/3} \right) = 1. \tag{38}$$

Proof. Consider a point i_0/n and an interval I which is such that i_0 lies in the central half of I. We firstly consider the case where $f_n^*(i/n)$ is either monotone increasing or monotone decreasing for $i \in I$. We suppose that $f_n^*(i_0/n) \ge f(i_0/n)$ and that $f_n^*(i/n)$ is monotone increasing. The other three cases are deal analogously. We have

$$|f_n^*(i_0/n) - f(i_0/n)| = f_n^*(i_0/n) - f(i_0/n) \le \frac{1}{4|I|} \sum_{i \ge i_0, i \in I} (f_n^*(i/n) - f(i_0/n))$$

and hence

$$\begin{aligned}|f_n^*(i_0/n)-f(i_0/n)| &\le \frac{1}{4|I|}\Bigg(\sum_{i\ge i_0, i\in I}(f_n^*(i/n)-f(i/n)-\epsilon(i/n)) \\ &\quad + \sum_{i\ge i_0, i\in I}\frac{i}{n}f^{(1)}(\theta_i i/n)+\sum_{i\ge i_0, i\in I}\epsilon(i/n)\Bigg)\end{aligned}$$

with $0<\theta_i<1$. The first and last terms on the right-hand side are $\mathrm{O}(\sqrt{|I|\log(n)})$ because of (7) and $\lim_{n\to\infty}\sigma_n=\sigma$ almost surely. The middle term is of order $|I|^2/n$ as $f^{(1)}$ is bounded and we obtain

$$|f_n^*(i_0/n)-f(i_0/n)|\le C\left(\sqrt{\frac{\log(n)}{|I|}}+\frac{|I|}{n}\right).$$

On choosing I so that $|I|=n^{2/3}(\log(n))^{1/3}$ we obtain the rate $(\log(n)/n)^{1/3}$. Suppose now that$f_n^*(i/n)$ is not monotone. In this case the first derivative of the spline has a zero, say at t_0, and we have

$$\begin{aligned}|f_n^{*(1)}(t)| &\le \int_{t_0}^{t}|f_n^{*(2)}(u)|\,du \\ &\le \sqrt{|t-t_0|}\sqrt{\int_0^1 f_n^{*(2)}(u)^2\,du} \\ &\le c\sqrt{|t-t_0|}\;n^{1/8}\end{aligned}$$

for sufficiently large n with high probability where we have used (37). From this we obtain

$$f_n^*(i/n)=f_n^*(i_0/n)+\mathrm{O}\left(\frac{|i-i_0|}{n}\sqrt{|I|}n^{-3/8}\right)$$

and it follows arguing as before

$$|f_n^*(i_0/n)-f(i_0/n)|\le C\left(\sqrt{\frac{\log(n)}{|I|}}+\frac{|I|^{3/2}}{n^{11/8}}+\frac{|I|}{n}\right).$$

On putting $|I|=n^{2/3}(\log(n))^{1/3}$ we again obtain the $(\log(n)/n)^{1/3}$ rate of convergence. □

7.2 The two-dimensional case

For the two-dimensional case we have with the corresponding definition of f_n^*

Theorem 7.3 *If f has continuous partial derivatives of order 2 then*

$$\lim_{c\to\infty}\lim_{n\to\infty}\mathbb{P}\left(\|f_n^* - f\|_{n,\infty} \le cn^{-1/6}\log(n)^{1/3}\right) = 1. \tag{39}$$

Proof. We have data $Y(s,t)$, $(s,t) \in [0,\, 1]^2$ generated by the model

$$Y(s,t) = f(s,t) + Z(s,t), \quad (s,t) \in [0,\, 1]^2 \tag{40}$$

with $Z(s,t)$ standard Gaussian white noise and evaluated on the grid $\{(i/n, j/n)\}$, $i,j = 0,\dots,n-1$. Corresponding to (4) we have for any function g

$$\begin{aligned} S_\lambda(g) : \;&=\; \sum_{i,j=0}^{n-1} \lambda\left(Y\left(i/n, j/n\right) - g\left(i/n, j/n\right)\right)^2 + \\ & \int_0^1\int_0^1 \left(\frac{\partial^2 g(s,t)}{\partial s^2} + \frac{\partial^2 g(s,t)}{\partial s \partial t} + \frac{\partial^2 g(s,t)}{\partial t^2}\right)^2 ds\, dt \end{aligned} \tag{41}$$

which we write in the discrete form

$$S(\boldsymbol{g}, \lambda) := (\boldsymbol{y} - \boldsymbol{g})^t \Lambda_n (\boldsymbol{y} - \boldsymbol{g})^t + \boldsymbol{g}^t \Omega_n \boldsymbol{g} \tag{42}$$

where

$$\begin{aligned} \boldsymbol{y} &= (y(0/n, 0/n), \dots, y((n-1)/n, (n-1)/n))^t \\ \boldsymbol{g} &= (g(0/n, 0/n), \dots, g((n-1)/n, (n-1)/n))^t \\ \Lambda_n &= \operatorname{diag}(\lambda(0/n, 0/n), \dots, \lambda((n-1)/n, (n-1)/n) \end{aligned}$$

and the symmetric, non-negative definite $(n+1)^2 \times (n+1)^2-$matrix Ω_n is defined by

$$\begin{aligned} &\boldsymbol{g}^t \Omega_n \boldsymbol{g} = \\ &\qquad \sum_{i,j=2,n} \left(\Delta_{11}\left(g\left(i/n, j/n\right)\right)^2 + \Delta_{12}\left(g\left(i/n, j/n\right)\right)^2 + \Delta_{22}\left(g\left(i/n, j/n\right)\right)^2\right) \end{aligned}$$

with

$$\begin{aligned} \Delta_{11}(g(i/n, j/n) &= g(i/n, j/n) + g((i-2)/n, j/n) - 2g((i-1)/n, j/n) && (43) \\ \Delta_{12}(g(i/n, j/n) &= g(i/n, j/n) + g((i-1)/n, (j-1)/n) - g((i-1)/n, j/n) \\ & \qquad - g(i/n, (j-1)/n) && (44) \\ \Delta_{22}(g(i/n, j/n) &= g(i/n, j/n) + g(i/n, (j-2)/n) - 2g(i/n, (j-1)/n). && (45) \end{aligned}$$

On writing

$$\begin{aligned} \Delta_1(g(i/n, j/n) &= g(i/n, j/n) - g((i-1)/n, j/n) \\ \Delta_2(g(i/n, j/n) &= g(i/n, j/n) - g(i/n, (j-1)/n) \end{aligned}$$

it follows with some manipulation that

$$\begin{aligned} g(i/n,j/n) &= g(0,0)+i\Delta_1(g(1/n,j/n))+j\Delta_2(g(0,1/n)) \\ &\quad +\sum_{\nu=2}^{i}(i-\nu+1)\Delta_{11}(g(\nu/n,j/n)) \\ &\quad +\sum_{\mu=2}^{j}(j-\mu+1)\Delta_{22}(g(0,\mu/n)). \end{aligned} \tag{46}$$

This implies

$$\begin{aligned} |g(i/n,j/n)-g(0,0)| &\le i|\Delta_1(g(1/n,j/n))|+j|\Delta_2(g(0,1/n))| \\ &\quad +i\sum_{\nu=2}^{i}|\Delta_{11}(g(\nu/n,j/n))| \\ &\quad +j\sum_{\mu=2}^{j}|\Delta_{22}(g(0,\mu/n))| \\ &\le i|\Delta_1(g(1/n,j/n))|+j|\Delta_2(g(0,1/n))| \\ &\quad +(i^{3/2}+j^{3/2})\left(\sum_{\mu,\nu=2}^{n}\big(\Delta_{11}(g(\nu/n,\mu/n))^2\right. \\ &\quad \left.\Delta_{22}(g(\nu/n,\mu/n))^2\big)\right)^{1/2} \end{aligned} \tag{47}$$

Let $\tilde{f}_n$ be the thin-plate spine for the data with smoothing parameter λ. Then for given $\epsilon > 0$ we can choose γ below so that

$$\mathbb{P}\left(\int_0^1\int_0^1\left(\left(\frac{\partial^2\tilde{f}_n(s,t)}{\partial s^2}\right)^2+\left(\frac{\partial^2\tilde{f}_n(s,t)}{\partial s\partial t}\right)^2+\left(\frac{\partial^2\tilde{f}_n(s,t)}{\partial t^2}\right)^2\right)ds\,dt\le\gamma n\right)\ge 1-\epsilon \tag{48}$$

and hence

$$\sum_{i,j=2}^{n}\Big(\Delta_{11}(\tilde{f}_n(i/n,j/n))^2+\Delta_{22}(\tilde{f}_n(i/n,j/n))^2\Big)\le\gamma/n \tag{49}$$

with high probability. On substituting this into (47) we obtain

$$\begin{aligned} |\tilde{f}_n(i/n,j/n)-\tilde{f}_n(0,0)| &\le i|\Delta_1(\tilde{f}_n(1/n,j/n))|+j|\Delta_2(\tilde{f}_n(0,1/n))| \\ &\quad +\sqrt{\gamma}(i^{3/2}+j^{3/2})/\sqrt{n}. \end{aligned} \tag{50}$$

A similar argument shows

$$\begin{aligned} i|\Delta_1(\tilde{f}_n(1/n,j/n))| &\le |\tilde{f}_n(i/n,j/n)-\tilde{f}_n(0,j/n)|+\sqrt{\gamma}i^{3/2}/\sqrt{n} \\ &\le 2\sigma_n\sqrt{\log(n)}+\sqrt{\gamma}\,i^{3/2}/\sqrt{n} \qquad (51) \end{aligned}$$

On putting $i=\lfloor (4\sigma_n^2\,n\log(n))^{1/3}\rfloor$ we obtain

$$|\Delta_1(\tilde{f}_n(1/n,j/n))| = \mathrm{O}_{\mathbb{P}}\left((\log(n))^{1/6}/n^{1/3}\right) \qquad (52)$$

with a corresponding estimate for $|\Delta_2(\tilde{f}_n(0,1/n))|$. On using these in (50) we obtain

$$|\tilde{f}_n(i/n,j/n)-\tilde{f}_n(0,0)| \le \mathrm{O}_{\mathbb{P}}\left(k\,(\log(n))^{1/6}/n^{1/3}+k^{3/2}/\sqrt{n}\right) \qquad (53)$$

with $k=i+j$. For the function f of the theorem we have

$$|f(i/n,j/n)-f(0,0)| = \mathrm{O}\left((i+j)/n\right) \qquad (54)$$

We have

$$\begin{aligned} |\tilde{f}_n(0,0)-f(0,0)| &= \frac{1}{(k+1)^2}\left|\sum_{i,j=0}^{k}(\tilde{f}_n(0,0)-f(0,0))\right| \\ &\le \frac{1}{(k+1)^2}\left|\sum_{i,j=0}^{k}(\tilde{f}_n(i/n,j/n)-f(i/n,j/n))\right| \\ &\quad +\mathrm{O}_{\mathbb{P}}\left(k\,(\log(n))^{1/6}/n^{1/3}+k^{3/2}/\sqrt{n}\right)\Big) \\ &= \mathrm{O}_{\mathbb{P}}\Big(\sqrt{\log(n)}/k+k\,(\log(n))^{1/6}/n^{1/3} \\ &\qquad\qquad +k^{3/2}/\sqrt{n}\Big) \\ &= \mathrm{O}_{\mathbb{P}}\Big(n^{-1/6}(\log(n))^{1/3}\Big) \qquad (55) \end{aligned}$$

on putting $k=\mathrm{O}\big((n\log(n))^{1/6}\big)$. This applies for any point $(i/n,j/n)$ and consequently we have

$$\sup_{i,j}|\tilde{f}_n(i/n,j/n)-f(i/n,j/n)| = \mathrm{O}_{\mathbb{P}}\Big(n^{-1/6}(\log(n))^{1/3}\Big). \qquad (56)$$

Finally as f_n^* is as least as smooth as $\tilde{f}_n$ it also satisfies (56) and this completes the proof of the theorem. □

References

[1] M. Brockmann, T. Gasser, and E. Herrmann. Locally adaptive bandwidth choice for kernel regression estimators. *J. Amer. Statist. Assoc.*, 88:1302–1309, 1993.

[2] T. T. Cai and M. G. Low. Nonparametric estimation over shrinking neighborhoods: supperefficiency and adaption. *Ann. Statist.*, 33(1):184–231, 2005.

[3] P. Craven and G. Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31(4):377–403, 1978/79.

[4] P. L. Davies and A. Kovac. Densities, spectral densities and modality. *Ann. Statist.*, 32(3):1093–1136, 2004.

[5] P.L. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *Ann. Statist.*, 29:1–65, 2001.

[6] C. de Boor. *A practical guide to splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1978.

[7] C. de Boor. Calculation of the smoothing spline with weighted roughness measure. *Mathematical Models and Methods in Applied Sciences*, 11:33–41, 2001.

[8] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224, 1995.

[9] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57:301–369, 1995.

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

[11] R.L. Eubank. *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York, second edition edition, 1988.

[12] R.L. Eubank. *Nonparametric Regression and Spline Smoothing.* Marcel Dekker, New York, 1999.

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

[14] F. Friedrich. *Complexity Penalized Segmentations in 2D.* PhD thesis, Zentrum Mathematik, Technische Universität München, Munich, Germany, 2005.

[15] T. Gasser, A. Kneip, and W. Köhler. A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.*, 86:643–652, 1991.

[16] P.J. Green and B.W. Silverman. *Nonparametric Regression and Generalized Linear Models.* Chapman & Hall, London e.a., 1994.

[17] F. R. Hampel. The breakdown points of the mean combined with some rejection rules. *Technometrics*, 27:95–107, 1985.

[18] W. Härdle. *Applied nonparametric regression.* Cambridge University Press, New York u.a., 1990.

[19] W. Härdle, P. Hall, and J. S. Marron. How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.*, 83(401):86–101, 1988. With comments by David W. Scott and Iain Johnstone and a reply by the authors.

[20] W. Härdle, P. Hall, and J. S. Marron. Regression smoothing parameters that are not far from their optimum. *J. Amer. Statist. Assoc.*, 87(417):227–233, 1992.

[21] W. Härdle and J. S. Marron. Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, 13(4):1465–1481, 1985.

[22] E. Herrmann. Local bandwidth choice in kernel regression estimation. *J. Amer. Statist. Assoc.*, 6:35–54, 1997.

[23] C. M. Hurvich, J. S. Simonoff, and C-L. Tsai. Smoothing parameter selection in nonparametric regression using an improved aic criterion. *J. Roy. Statist. Soc. Series B*, 60:271–293, 1998.

[24] A. Kovac. Robust nonparametric regression and modality. In R. Dutter, P. Filzmoser, U. Gather, and P. Rousseeuw, editors, *Developments in Robust Statistics*, pages 218–227, Heidelberg, Germany, 2002. Physica.

[25] A. Kovac and B.W. Silverman. Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Amer. Statist. Assoc.*, 95:172–183, 2000.

[26] A. Majidi. *Glatte nichtparametrische Regression unter formerhaltenden Bedingungen.* PhD thesis, Universität Essen, 2003.

[27] M. Meise. *Residual Based Selection of Smoothing Parameters.* PhD thesis, Department of Mathematics, University Duisburg-Essen, Essen, Germany, 2004.

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

[29] J. Polzehl and V.G. Spokoiny. Adaptive weights smoothing with applications to image restoration. *J. R. Stat. Soc. B*, 62:335–354, 2000.

[30] J. Rissanen. Mdl-denoising. *IEEE Trans. Inform. Theory*, 46:2537–2543, 2000.

[31] D. Ruppert and M.P. Wand. Multivariate locally weighted least squares regression. *Ann. Statist.*, 22:1346–1370, 1994.

[32] L. L. Schumaker. *Spline functions: basic theory.* John Wiley & Sons Inc., New York, 1981.

[33] C. J. Stone. Optimal rates of convergence for nonparametric regression. *Ann. Statist.*, 10:1040–1053, 1982.

[34] G. Wahba. A survey of some smoothing problems and the method of generalized cross-validation for solving them. *Applications of Statistics*, pages 507–523, 1977.

[35] G. Wahba. *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, 1990.

[36] M.P. Wand and M.C. Jones. *Kernel Smoothing.* Chapman & Hall, London e.a., 1995.