Densities, Spectral Densities and Modality [1]

By P. L. Davies and A. Kovac

Universität Essen

[2] Given a data set $(x_1, \ldots, x_n)$ this paper considers the problem of specifying a simple approximating density function. Simplicity is measured by the number of local extremes but several different definitions of approximation are introduced. The taut string method is used to control the numbers of modes and to produce candidate approximating densities. Refinements are introduced that improve the local adaptivity of the procedures and the method is extended to spectral densities.

1. Contents In Section 2 we formulate the density problem in terms of obtaining the simplest density which is an adequate approximation for the given data. The taut string method of Davies and Kovac (2001) is adapted to the density problem and is used for producing candidate densities of increasing complexity. The difficulties of the density problem are discussed in Section 3. Section 4 contains a more detailed account of the application of the taut string method to the density problem. The asymptotics of the procedure on appropriate test beds are discussed in Section 5. A refinement based on cell occupancy frequencies which increases local sensitivity is described in Section 6. Section 7 compares the taut string method with kernel estimators in a small simulation study. Finally in Section 8 considers briefly the application of taut strings to the problem of spectral densities.

2. Introduction

2.1. *The density problem* Given a sample $\mathbf{x}_n = (x_1, \ldots, x_n)$ of size n we consider the problem of specifying a distribution F with the smallest number of modes such that the resulting model of i.i.d. random variables $\mathbf{X}_n^F = (X_1^F, \ldots, X_n^F)$ with common distribution F is an adequate approximation for the data $\mathbf{x}_n$.

We use different concepts of approximation one of which is the following. Let E_n,

$$E_n(x) = \frac{1}{n}\sum_{I=1}^{n}\{x_i \leq x\},$$

denote the empirical distribution of the data $\mathbf{x}_n$ and F_n the empirical distribution function of n i.i.d. random variables $\mathbf{X}_n^F$ with common distribution F. The

[2] *AMS 1991 subject classifications.* Primary 62G07; secondary 62M15.

Key words and phrases. Density estimation, time series analysis, spectral densities, modality, asymptotics, strings.

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

Kolmogoroff metric d_{ko} is defined by

$$d_{ko}(F,G) = \sup\{x : |F(x) - G(x)|\}.$$

The i.i.d. model with distribution F will be regarded as an adequate approximation to the data $\mathbf{x}_n$ if

$$d_{ko}(E_n, F) \le \mathrm{qu}(n, \alpha, d_{ko}). \tag{2.1}$$

where $\mathrm{qu}(n, \alpha, d_{ko})$ denotes the α-quantile of the random variable $d_{ko}(F_n, F)$. $\mathrm{qu}(n, \alpha, d_{ko})$ is independent of F for continuous F. This gives rise to the Kolmogoroff problem:

Problem 2.1 Kolmogoroff problem. *Determine the smallest integer k_n for which there exists a density f^n with k_n modes and whose distribution F^n satisfies*

$$d_{ko}(E_n, F^n) \le \mathrm{qu}(n, \alpha, d_{ko}). \tag{2.2}$$

We note that the problem is well posed: for any data set $\mathbf{x}_n$ it has a solution. We have posed the problem in terms of approximation so that no assumptions regarding the "true" data generating mechanism are required or made.

The problem (refkolprob) is formulated in terms of the smallest number of modes required for an adequate approximation. A detailed theoretical discussion of such one-sided problems is given by Donoho (1988): one of his examples is that of modality for nonparametric densities and spectral densities. His paper also raises interesting questions about statistical inference involving objects whose very existence cannot be shown, an example being the "underlying density" for the data. We avoid such problems by phrasing the paper in terms of approximation.

Hengartner and Stark (1995) also make use of the Kolmogoroff ball to determine nonparametric confidence bounds for densities subject to an upper bound for the number of modes. In the particular case of monotone or unimodal densities the width of their bounds on appropriate trest beds is $(\log n/n)^{1/3}$ which agrees with the results given in this paper. It seems that their bounds become difficult to calculate for more than one mode as the complexity is given as $\binom{n}{l}$ where l is the number of local extremes. The main differences to the work of Hengartner and Stark are as follows:

- we provide an explicit density but no bounds,
- neither the number of modes nor even an upper bound is specified in advance,
- the algorithmic complexity of our method is $\mathrm{O}(n)$ independently of the number of modes.

2.2. *The taut string methodology* The basic methodology we use for producing densities is the taut string methodology. Taut strings were first used in the context of monotonic regression: the greatest convex minorant of the integrated data is a taut string and its derivative is precisely the monotone increasing least squares approximation. This is described in Barlow, Bartholomew, Bremner and Brunk

(1972) who were the first to use the phrase "taut string". We refer also to Leurgans (1982). The first use of the taut string which goes beyond the monotone case and which explicitly deals with modality is in Hartigan and Hartigan (1985) where it is referred to as the "stretched string". Hartigan and Hartigan (1985) introduced their DIP test for unimodality which is based on the closest (in the Kolmogoroff metric) unimodal distribution to the empirical distribution function of the data. Based on the work of Hartigan and Hartigan (1985) Davies (1995) used the taut string method to produce candidate densities of low modality to approximate data. Mammen and van de Geer (1997) employed the taut string in the non-parametric regression problem. They considered a penalized least squares problem where the penalty is the total variation of the approximating function. The solution is the basic taut string confined to a tube centered at the integrated data. Mammen and van de Geer gave a detailed description of the taut string but did not mention the connection with modality. Hartigan (2000) recently proposed a generalization of the DIP test and examined for each antimode of a taut string approximation the supremum distance between the empirical distribution function and a monotone density on a "shoulder interval" including the antimode. Finally Davies and Kovac (2001) used the taut string methodology to control the number of local extremes of a nonparametric approximation to a data set. They also introduced the idea of local squeezing and residual driven tube widths which greatly increase the precision and flexibility of the taut string methodology.

2.3. *Smoothness* The taut string methodology produces densities which are piecewise constant and therefore not even continuous. Smoothness will not be a consideration in this paper but we point out that techniques for smoothing such functions have been developed. The idea is to obtain the smoothest density subject to shape and deviation constraints taken from the taut string. We refer to Metzner (1997), Löwendick and Davies (1998) and Majidi (2001).

2.4. *Previous work* Much work has been done on the problem of density estimation. One of the most popular methods is that of kernel smoothing. We refer to Nadaraya (1964), Watson (1964), Silverman (1986), Sheather and Jones (1991), Wand and Jones (1995), Sain and Scott (1996) and Simonoff (1996) and the references given there. The main problem here is the determination of appropriate global or local bandwidths. A further approach is based on wavelets. We refer to Donoho, Johnstone, Kerkyacharian and Picard (1996), Herrick, Nason and Silverman (2000) and to Chapter 7 of Vidakovic (1999). Mixtures of densities have been considered in the Bayesian framework by Richardson and Green (1997) and Roeder and Wasserman (1997). Other Bayesian methods are to be found in Verdinelli and Wasserman (1998).

None of the above approaches is directly concerned with modality. For example the non-Bayesian theory is generally based on integrated squared error or some similar loss function. In spite of this methods are often judged by their ability to identify peaks in the data as in Loader (1999) and Herrick et al (2000). Work directly concerned with modality has been done by Müller and Sawitzki (1991) using their concept of excess mass. Their ideas have been extended to multidimensional distributions by Polonik (1995a, 1995b, 1999). Hengartner and Stark (1995) use the Kolmogoroff ball centred at the empirical distribution function to obtain nonparametric confidence bounds for shape restricted densities. Another way of controlling modality is that of mode testing. We refer to Good and Gaskins (1980), Silverman (1986), Hartigan and Hartigan (1985) and Fisher, Mammen and Marron (1994).

3. The difficulties of the density problem Obtaining adequate approximate densities is a special case of nonparametric regression. Whereas nonparametric regression is usually concerned with the size of the dependent variable the density problem is concerned with measuring the degree of closeness of the design points. In spite of a formal similarity this is the more difficult problem and it may explain the modesty evident in the literature on densities. The difficulties may be illustrated by the following data set.

The upper panel of Figure 1 shows the so called claw density. It is one of ten introduced by Marron and Wand (1992) and is defined analytically by

$$\mathbb{Q} = 0.5 * \mathcal{N}(0,1) + 0.1 * \sum_{i=0}^{4} \mathcal{N}(i/2 - 1, 0.1).$$

Looking at the pronounced nature of the peaks one has the feeling, we put it no more strongly than this, that it should not be difficult to detect the peaks given a sample of size $n = 500$. The centre panel of Figure 1 shows a kernel estimate of the density for a sample of size 500. The result seems to be disappointing. The finger at -0.5 is missing, the remaining fingers are emasculated and there are two spurious modes at -2.2 and 3.0. Decreasing the bandwidth until the five peaks are clearly visible results in spurious peaks. This is shown in the bottom panel of Figure 1.

Figure 2 shows kernel and taut string approximations to the normal, uniform and claw distributions using samples of size 2048 each. In each case the bandwidth was chosen as small as possible while still retaining the correct modality. The kernel method performs quite well on the sample from the normal distribution. The approximation to the uniform density is rather poor and can only be improved using a smaller bandwidth and introducing additional modes. The claw distribution is approximated even worse. Only three peaks are identified correctly and two peaks in the tails near -2 and 3 have been introduced. The taut string method produces much better approximations, but still the difficult problem is left to find out how many modes have to be included in the approximation.

4. Taut strings, Kuiper metrics and densities

4.1. *Taut strings* We give a short description of the taut string method. A thorough analysis of properties of the taut string can be found in Hartigan (2000). Further details and an algorithm of complexity $O(n)$ are given by Davies and Kovac (2001).

Consider a sample $\mathbf{x}_n$ and form the ordered sample $\mathbf{x}_{(n)} = (x_{(1)}, \ldots, x_{(n)})$. For a given $\varepsilon > 0$ we consider the Kolmogoroff tube $T(E_n, \varepsilon)$ centred at the empirical

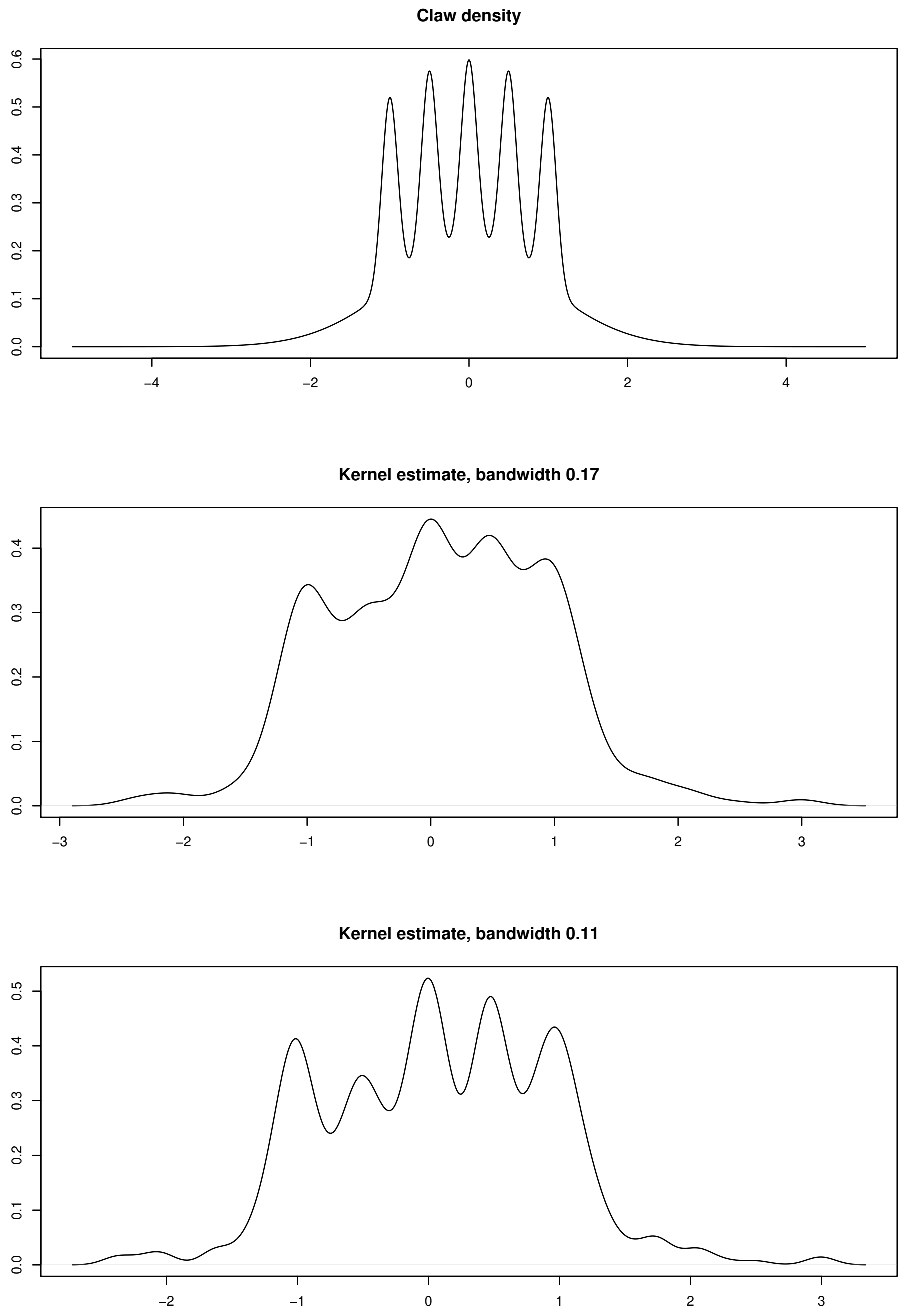

FIG. 1. *Claw density. The upper panel shows the claw density function, the centre panel a default kernel estimate based on a sample of size 500. The lower panel shows the effect of decreasing the bandwidth until the five peaks are clearly visible.*

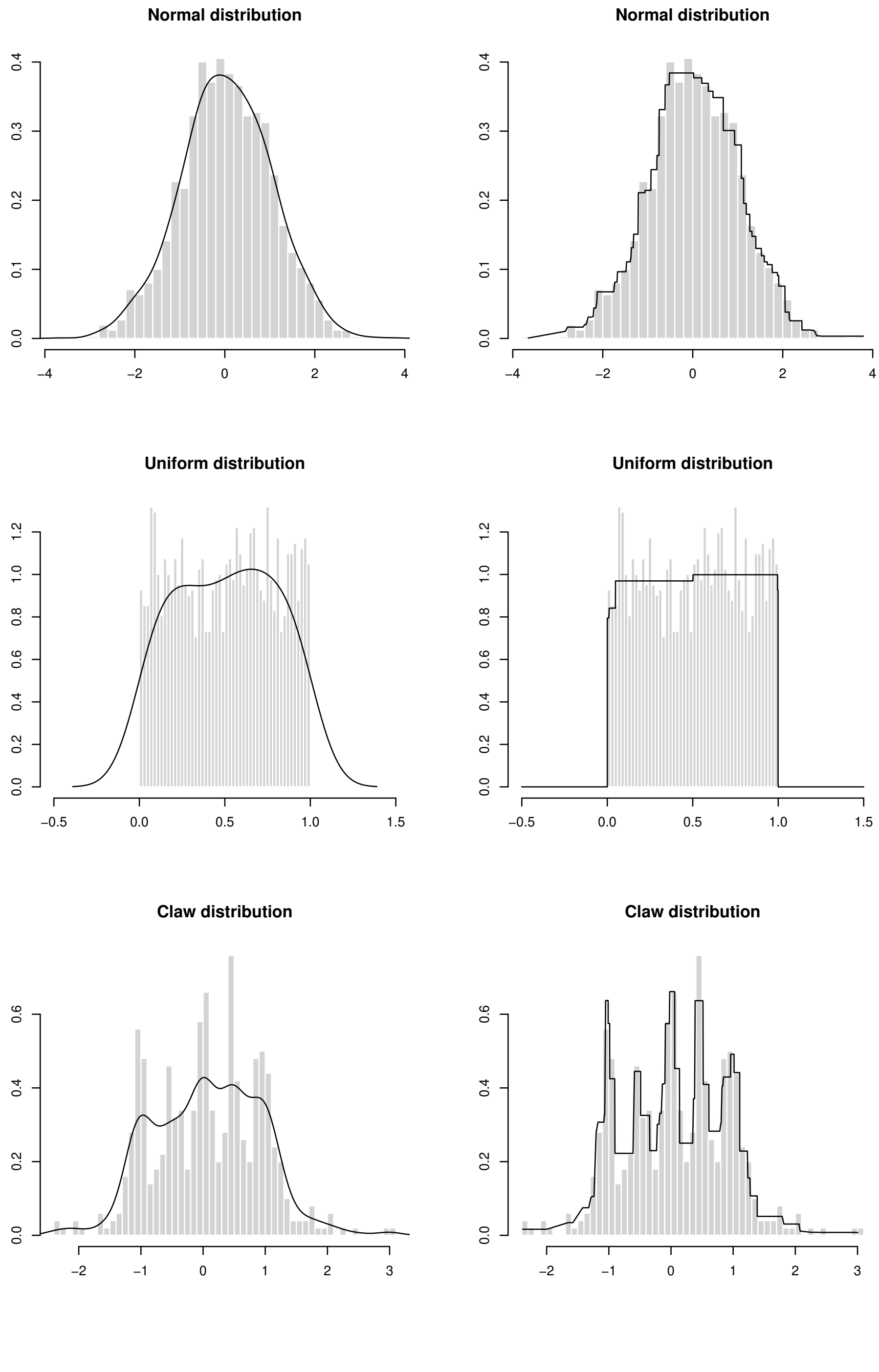

Fig. 2. *Normal, uniform and claw density. The panels show kernel and taut string approximations using the smallest bandwidth that retains the correct modality.*

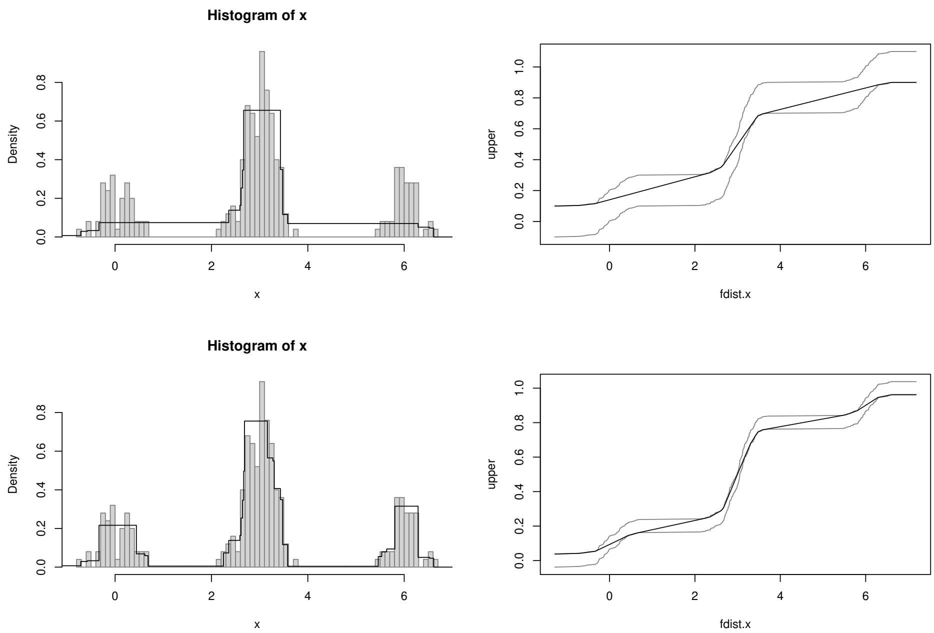

FIG. 3. *These figures illustrate the taut string method applied to a sample of mixture of normal distributions with two different bandwidths. The right column shows the tubes and the taut strings whilst the left column shows histograms of the data and the corresponding densities of the taut string.*

distribution E_n and of radius $\varepsilon > 0$

$$T(E_n, \varepsilon) = \{G : G \text{ monotone } \sup_x |G(x) - E_n(x)| \leq \varepsilon\}$$

Imagine now a taut string which is attached to the upper boundary at the point $x_{(1)}$ and the lower boundary at the point $x_{(n)}$ and is constrained to lie within the Kolmogoroff tube. Such a string is shown in the right panels of Figure 3 for two different values of ε. The taut string defines a function S_n on the interval $[x_{(1)}, x_{(n)}]$. Although S_n depends on E_n and ε we suppress this dependency to relieve the burden on the notation. We denote the density of S_n by s_n. It is defined as the left hand derivative of S_n except at the smallest data point $x_{(1)}$ where we use the right hand derivative. The left panels of Figure 3 show histograms of the data with the corresponding densities s_n superimposed.

The taut string is a spline with knots at the points at which it touches the lower or upper boundaries of the Kolmogoroff tube. It has the following properties (see Mammen and van de Geer, 1997; Davies and Kovac, 2001):

(a) S_n is monotonic increasing and linear between knots

(b) s_n is nonnegative and piecewise constant between knots

(c) s_n has the minimum modality of all functions whose integral over $[x_{(1)}, x_{(n)}]$ lies in $T(E_n, \varepsilon)$ and satisfies the end point conditions

(d) S_n switches from the upper boundary $E_n + \varepsilon$ to the lower boundary $E_n - \varepsilon$ at points where s_n has a local maximum

(e) S_n switches from the lower boundary $E_n - \varepsilon$ to the upper boundary $E_n + \varepsilon$ at points where s_n has a local minimum

(g) If ξ_j and ξ_{j+1} are consecutive knots on the same boundary then on the interval $(\xi_j, \xi_{j+1}]$

$$s_n(x) = \frac{|\{i : \xi_j < x_i \leq \xi_{j+1}\}|}{n(\xi_{j+1} - \xi_j)} \tag{4.3}$$

It is property (c) which is of importance and allows control of the number of modes. If consecutive knots ξ_j and ξ_{j+1} are on opposite boundaries then it follows from (d) and (e) above that (4.3) must be replaced by

$$s_n(x) = \frac{|\{i : \xi_j < x_i \leq \xi_{j+1}\}| \pm 2\varepsilon}{n(\xi_{j+1} - \xi_j)} \tag{4.4}$$

with a minus sign at local maxima and a plus sign at local minima. This means that the derivative underestimates local maxima and overestimates local minima. This can be remedied as follows (Davies and Kovac, 2001). We define the modified string $\tilde{S}_n$ by

$$\tilde{S}_n(\xi_j) = E_n(\xi_j) \text{ at all knots } \xi_j \tag{4.5}$$

and linear in between. The corresponding derivative $\tilde{s}_n$ satisfies

$$\tilde{s}_n(\xi_j) = \frac{|\{i : \xi_j < x_i \leq \xi_{j+1}\}|}{n(\xi_{j+1} - \xi_j)} \text{ between the knots } \xi_j \text{ and } \xi_{j+1}. \tag{4.6}$$

This modification has no effect on the modality. It is possible that the modified string $\tilde{S}_n$ does not lie in the Kolmogoroff tube but we ignore this.

4.2. *Data analysis* Even without further analysis the taut string can be used as a data analytical tool. If the radius of the Kolmogoroff tube is monotonically decreased then the number of modes of the derivative of the taut string increases monotonically. It is therefore possible to specify the number of modes of the approximate density. Figure 4 shows this for the same sample as used for Figure 1.

The densities of Figure 4 can also be interpreted as histograms with an automatic choice of the number of bins and the bin widths. We note that in this example the five peaks are correctly identified when the tube is squeezed until the string has five peaks. It should be compared with the bottom panel of Figure 1. Figure 5 shows the number of correctly identified peaks as a function of sample size when the tube is squeezed to have five peaks. A peak is defined as being correctly identified if the midpoint of the interval defining a peak differs by less than 0.15 from the position of the corresponding peak of the claw density. This shows that the taut string method is extremely good at finding peaks. For samples of size 200 the five peaks will be

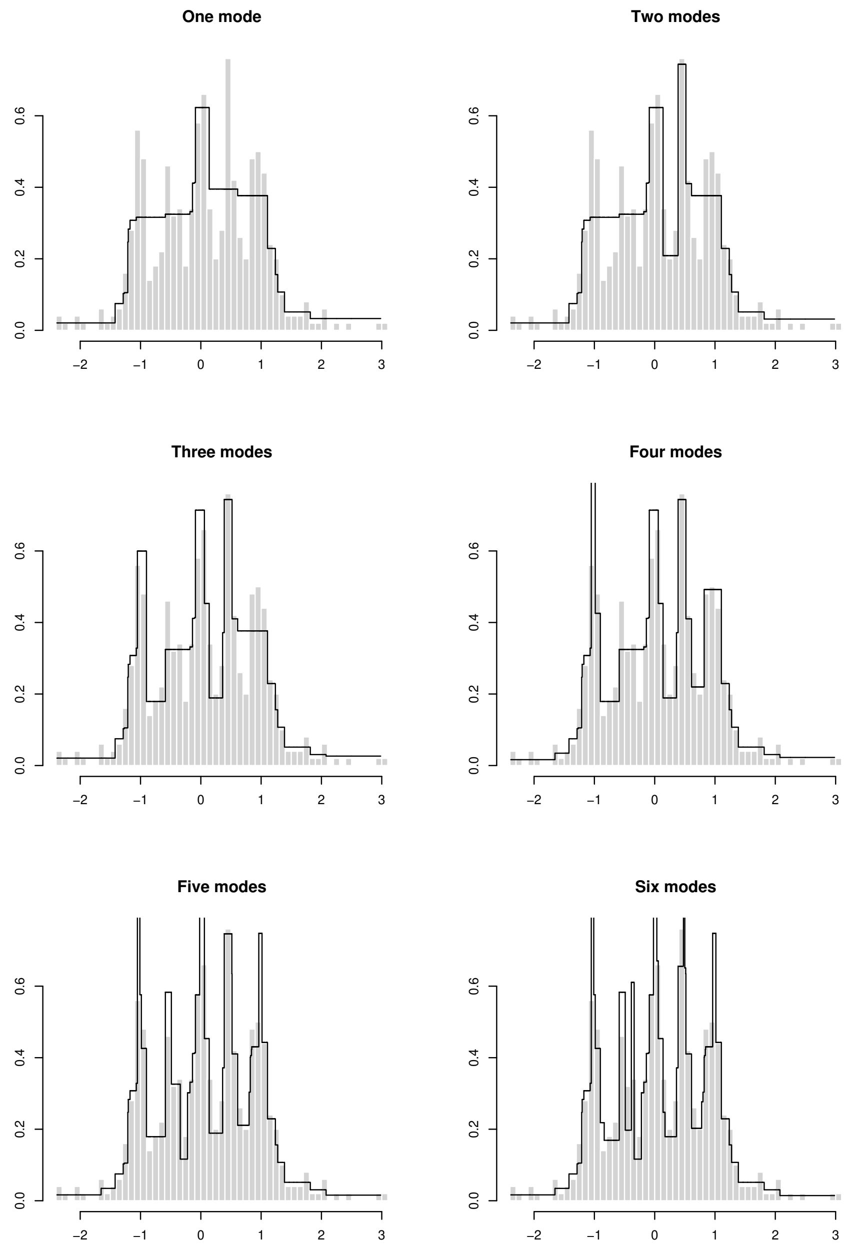

Fig. 4. *Six taut string estimates of a sample of the claw distribution with increasing number of modes.*

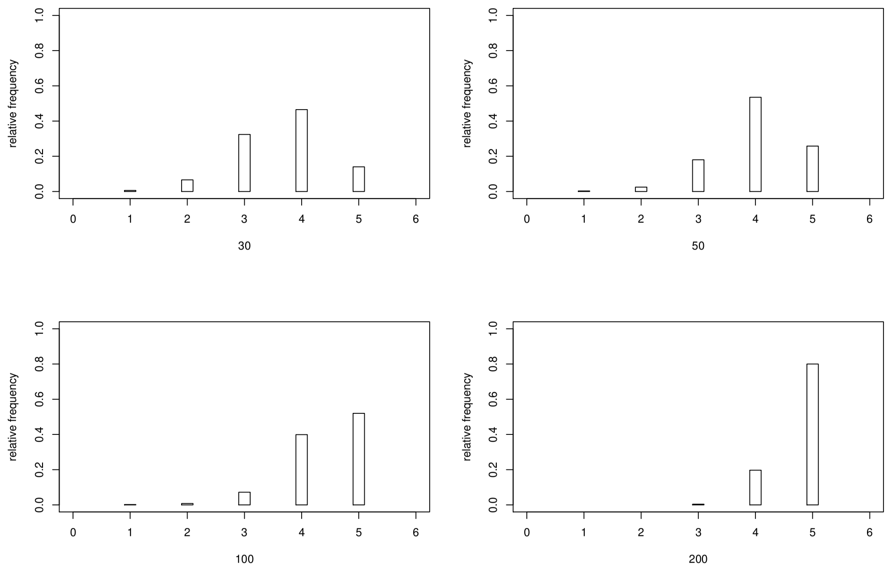

FIG. 5. *Five modal taut string: number of correctly identified peaks of the claw density as a function of sample size.*

correctly identified in over 80% of the cases. This in a sense confirms Loader (1999) who, on the basis of theoretical results of Marron and Wand (1992), claims that for samples of size $n = 193$ the claws should be detectable. The problem we now address is the difficult one of defining an automatic procedure to do this.

4.3. *An automatic procedure* The taut string method described above is the same as that used in Davies and Kovac (2001) in the nonparametric regression setting. The particular difficulty of the density problem is deriving an automatic procedure. In Davies and Kovac (2001) an autmatic procedure is obtained using a multiresolution analysis of the residuals which is similar to the hard thresholding method for wavelets. This is sensitive and works extrewmely well for many data sets. There is unfortunately no obvious equivalent for the density problem and this is the main difference between the two problems.

The following theorem is an immediate consequence of the properties of the taut string listed above.

THEOREM 4.1. *The density s_n of the taut string through the tube $T(E_n, \mathrm{qu}(n, \alpha, d_{ko}))$ is a solution of the Kolmogoroff density problem.*

For finite n the values of $\mathrm{qu}(n, \alpha, d_{ko})$ can be obtained by simulation. In the limit $\sqrt{n}\mathrm{qu}(n, \alpha, d_{ko})$ tends to the corresponding quantile of

$$\max_{0 \le t \le 1} B_0(t) - \min_{0 \le t \le 1} B_0(t)$$

where B_0 denotes a Brownian bridge and for which an explicit expression exists (Dudley, 1989).

The solution of the Kolmogoroff density problem defines a procedure which can be evaluated. If we do this on an i.i.d. test bed that is with data of the form $X_1(F), \ldots, X_n(F)$ where f has a k-modal density function f then it is clear that the taut string density s_n will have at most k modes with probability at least α. This follows on noting that F lies in the tube with probability α and that in this case s_n has at most as many modes as f. In particular if $k = 1$ we have

THEOREM 4.2. *Let $X_1(F), \ldots, X_n(F)$ be an i.i.d. sample with common unimodal distribution F and let s_n be the solution of the Kolmogoroff density problem (2.2). Then*

$$\mathbf{P}(s_n \textit{ unimodal}) \ge \alpha. \tag{4.7}$$

A simulation was performed to investigate the performance of the procedure with $\alpha = 0.9$ and the corresponding tube width $1.245/\sqrt{n}$ on test beds defined by the uniform distribution on $[0, 1]$, the standard normal distribution and the claw density for several sample sizes between 50 and 15000. It is clear that the modality is correctly estimated with probability at least 0.9 for the unimodal distributions in accordance with Theorem 4.2. Indeed the actual probability exceeds 0.9: all simulations resulted in exactly one peak. The results for the claw distribution are, in contrast, disappointing. They are shown in Table 1. Asymptotically the modality

Size	50	500	2000	5000	10000	15000
Uniform	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Normal	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Claw	1 (4)	1 (4)	1 (4)	1 (3.96)	5 (0.35)	5 (0)

TABLE 1

Medians of number of detected modes in samples of three densities using the 0.9*-quantile of the Kolmogoroff metric. In brackets average absolute deviation from the true modality.*

will be correctly estimated with probability at least 0.9 but the rate of convergence is very slow. We now try and obtain an improved procedure in two ways. Firstly we note that the choice of $\mathrm{qu}(n, \alpha, d_{ko})$ for the radius of the tube means that a probability of at least α is guaranteed for all unimodal test beds. If we provisionally accept that many unimodal distributions, for example the uniform distribution, are poor models for most data sets then we may accept a worse performance on such test beds in return for an enhanced performance on others. Silverman (1986) and Müller and Sawitzki (1991) argue in a similar vein. The second way of gaining an improved performance is to use a generalized Kuiper metric rather than the Kolmogoroff metric. Kuiper metrics consider the differences in probability over a fixed number of disjoint intervals and are therefore better at detecting modality.

4.4. *Calibrating unimodality* To implement the first way of improving performance let $\mathrm{qu}(n, \alpha, F, d_{ko})$ denote the $\alpha-$quantile of the Kolomogoroff distance of the closest unimodal distribution (given by the taut string) to the empirical distribution F_n of n i.i.d. random variables with common distribution F. We have the following theorem.

THEOREM 4.3. *Let $X_1(F), \ldots, X_n(F)$ be an i.i.d. sample with common unimodal distribution F and empirical distribution F_n. Let s_n be the derivative of the string S_n through the tube $T(F_n, \mathrm{qu}(n, \alpha, F, d_{ko}))$. Then*

$$\mathbf{P}(s_n \text{ unimodal}) = \alpha. \tag{4.8}$$

Clearly

$$\mathrm{qu}(n, \alpha, F, d_{ko}) \leq \mathrm{qu}(n, \alpha, d_{ko})$$

but it is not clear whether

$$\sup_{F \text{ unimodal}} \mathrm{qu}(n, \alpha, F, d_{ko}) = \mathrm{qu}(n, \alpha, d_{ko}).$$

We point out that the uniform distribution does not maximize $\mathrm{qu}(n, \alpha, F, 1, d_{ko})$ (Hartigan and Hartigan, 1985). We now take $F = U$ to be the uniform distribution on the basis that it is not an adequate approximation for most data sets and set $\alpha = 0.5$. This means that on uniform test beds the modality will be correctly determined with probability 0.5. The uniform distribution has the advantage that

Distribution	50	100	200	500	1000	2000
Uniform	2 (0.54)	2 (0.52)	1 (0.52)	1 (0.47)	1 (0.49)	1 (0.51)
Normal	1 (0.07)	1 (0.03)	1 (0.01)	1 (0)	1 (0)	1 (0)
Claw	1 (3.72)	1 (3.72)	1 (3.53)	3 (2.41)	4 (0.77)	5 (0)

TABLE 2

Medians of number of detected modes in samples of three densities using the 0.5*-quantile of the improved Kolmogoroff metric. In brackets average absolute deviation from the true modality.*

the asymptotics of the quantiles $\mathrm{qu}(n,\alpha,U,d_{ko})$ can be calculated. We have

$$\lim_{n\to\infty}\sqrt{n}\,\mathrm{qu}(n,\alpha,U,d_{ko}) = \mathrm{qu}(\alpha,B_0) \tag{4.9}$$

where $\mathrm{qu}(\alpha,B_0)$ denotes the α-quantile of the random variable

$$\min_H \sup_x |B_0(x) - H(x)| \tag{4.10}$$

where the function $H : [0,1] \to \mathbb{R}$ is convex on $[0,t_H]$ and concave on $[t_H,1]$ for some $t_H, 0 \le t_H \le 1$. Simulations show that the 0.5-quantile of (4.10) is 0.432. A correction for finite n gives

$$\mathrm{qu}(n,0.5,U,d_{ko}) = 0.43/\sqrt{n} - 0.64/n.$$

with a percentage error (based on simulations) of at most 0.0045. Table 2 shows the results. We see that the performance for the Gaussian test bed is hardly impaired. On the claw test bed we note that the performance for $n = 1000$ is now comparable to that of the simple Kolmogoroff quantile for $n = 10000$.

Instead of the uniform distribution one can use the Gaussian distribution Φ. Heuristic arguments indicate that

$$\lim_{n\to\infty}\sqrt{n}\,\mathrm{qu}(n,\alpha,\Phi,d_{ko}) = 0$$

but we have no exact asymptotic result. For sample sizes $50 \le n \le 10000$ the following approximation may be used

$$\mathrm{qu}(n,0.5,\Phi,d_{ko}) = 0.385/n^{0.566} - 0.578/n^{1.02} \tag{4.11}$$

again with a percentage error of at most 0.0047.

4.5. *Kuiper metrics* Suppose that the density s_n of the taut string is unimodal. Part of the description of the taut string S_n given in Section 4.1 is that it swaps from the upper bound to the lower bound at each maximum. Consider now the Kuiper metric d_{ku} defined by

$$d_{ku}(F,G) = \sup\{a < b : |(F(b) - F(a)) - (G(b) - G(a))|\} \tag{4.12}$$

It follows from the above that if $d_{ko}(E_n,S_n) = \varepsilon$ and s_n is unimodal then $d_{ku}(E_n,S_n) = 2\varepsilon$. The α-quantile $\mathrm{qu}(n,\alpha,d_{ku})$ of $d_{ku}(F_n,F)$ is independent of F for continuous F and is less than twice the α-quantile of $d_{ko}(F_n,F)$. This suggests that the Kuiper metric is more appropriate for unimodality than the Kolmogoroff metric. To demonstrate this we firstly define the Kuiper problem:

Problem 4.1 Kuiper density problem. *Determine the smallest integer k_n for which there exists a density f^n with k_n modes and whose distribution F^n satisfies*

$$d_{ku}(E_n, F^n) \le \mathrm{qu}(n, \alpha, d_{ku}).$$

Suppose now that F^n is a unimodal distribution which solves the Kuiper density problem. Let $\varepsilon_1 = \max\{x : F^n(x) - E_n(x)\}$ and $\varepsilon_2 = \max\{x : G(x) - F^n(x)\}$. As $d_{ku}(E_n, F^n) = \varepsilon_1 + \varepsilon_2 = \mathrm{qu}(n, \alpha, d_{ku})$ it follows by shifting F^n by an amount $\frac{1}{2}|\varepsilon_2 - \varepsilon_1|$ that the solution of the Kolmogoroff problem with $\varepsilon = \frac{1}{2}\mathrm{qu}(n, \alpha, d_{ku})$ is also unimodal. As $\frac{1}{2}\mathrm{qu}(n, \alpha, d_{ku}) < \mathrm{qu}(n, \alpha, d_{ko})$ this implies that if the solution of the Kuiper density problem for a given α is unimodal, so is the solution of the Kolmogoroff problem for the same α.

To cover the case of multimodality we define the Kuiper metric d_{ku}^{κ} of order κ by

$$d_{ku}^{\kappa}(F, G) = \max\{\sum_1^{\kappa} |(F(b_j) - F(a_j)) - (G(b_j) - G(a_j))|\} \tag{4.13}$$

where the maximum is taken over all a_j, b_j with

$$a_1 \le b_1 \le a_2 \le b_2 \cdots \le a_\kappa \le b_\kappa.$$

Again the distribution of $d_{ku}^{\kappa}(F_n, F)$ is independent of F for continuous F. If we denote the $\alpha-$quantile by $\mathrm{qu}(n, \alpha, d_{ku}^{\kappa})$ we can formulate the κ-Kuiper problem.

Problem 4.2 κ-Kuiper density problem. *Determine the smallest integer k_n for which there exists a density f^n with k_n modes and whose distribution F^n satisfies*

$$d_{ku}^{k}(E_n, F^n) \le \mathrm{qu}(n, \alpha, d_{ku}^{\kappa}).$$

If the density s_n of the taut string has k modes then for the Kuiper metric $d_{ku}^{2k-1}$of order $2k - 1$ we have

$$d_{ku}^{2k-1}(E_m, S_n,) = (2k - 1)\varepsilon.$$

This follows on noting that the strings swaps boundaries at each of the k local maxima of s_n and also at the $k - 1$ local minima. As

$$\mathrm{qu}(n, \alpha, d_{ku}^{2k-1}) < (2k - 1)\mathrm{qu}(n, \alpha, d_{ko})$$

this indicates that the Kuiper metric d_{ku}^{2k-1} is more efficacious when the data exhibit k modes. We have no simple algorithm for solving the κ-Kuiper problem so we use the strategy of Davies and Kovac (2001) and decrease the radius ε of the Kolmogoroff tube gradually until

$$d_{ku}^{2k-1}(E_n, S_n) \le \mathrm{qu}(n, \alpha, d_{ku}^{2k-1}).$$

For large n approximations to $\mathrm{qu}(n, \alpha, d_{ku}^{\kappa})$ are available using the weak convergence result

$$\sqrt{n} d_{ku}^{\kappa}(F_n, F) \Rightarrow \max\{\sum_1^{\kappa} |B_0(b_j) - B_0(a_j)|\}$$

Distribution	50	100	200	500	1000	2000
Uniform	2 (0.66)	1 (0.60)	1 (0.65)	1 (0.65)	1 (0.68)	1 (0.77)
Normal	1 (0.14)	1 (0.08)	1 (0.04)	1 (0)	1 (0)	1 (0)
Claw	1 (3.39)	2 (3.08)	3 (2.44)	4 (1.37)	4 (0.66)	5 (0.08)

TABLE 3

Medians of number of detected modes in samples of three densities using the quantile $\mathrm{qu}(n, 0.5, U, d_{ku}^9)$. *The values in brackets give the average absolute deviation from the true modality.*

Distribution	50	100	200	500	1000	2000
Uniform	3 (1.56)	3 (2.06)	4 (2.85)	5 (4.18)	7 (5.66)	8 (7.38)
Normal	1 (0.67)	1 (0.71)	2 (0.88)	2 (0.95)	1 (1.02)	2 (1.19)
Claw	2 (2.52)	3 (1.91)	4 (1.12)	5 (0.25)	5 (0.09)	5 (0.12)

TABLE 4

Medians of number of detected modes in samples of three densities using the quantile $\mathrm{qu}(n, 0.5, \Phi, d_{ku}^9)$ *The values in brackets give the average absolute deviation from the true modality.*

where B_0 denotes the standard Brownian bridge on $[0, 1]$ and

$$a_1 < b_1 < a_2 < b_2 \ldots < a_\kappa < b_\kappa.$$

The distribution of $\max\{|B_0(b) - B_0(a)|\}$ corresponding to the unimodal case $k = 1$ is known (for example Dudley (1989), Proposition 12.3.4.) Sufficiently accurate quantiles for finite n and for the other asymptotic cases may be obtained by simulations. Best results are obtained if κ is related to the modality k of the test bed by $\kappa = 2k - 1$. In practice a default value of κ is required and we use $\kappa = 9$. This performed well over a wide class of test beds and larger values of κ brought little improvement even for test beds with many peaks.

We combine the κ-Kuiper-metric with the ideas of Section 4.4. Let $\mathrm{qu}(n, \alpha, F, d_{ku}^\kappa)$ denote the α-quantile of the κ-Kuiper distance of the closest unimodal distribution to the empirical distribution F_n of n i.i.d. random variables with common distribution F. We use the modified string $\tilde{S}_n$ (see (4.5)) as the closest unimodal distribution. If F is the uniform distribution of $[0, 1]$, then we have again a $1/\sqrt{n}$ asymptotic. For example for $\kappa = 9$ and $\alpha = 0.5$ simulations showed that

$$\mathrm{qu}(n, 0.5, U, d_{ku}^9) \approx 5.46/\sqrt{n} - 13.7/n$$

is a good approximation. If instead of the uniform distribution we calibrate using the Gaussian distribution we have corresponding to (4.11)

$$\mathrm{qu}(n, 0.5, \Phi, d_{ku}^9) \approx 8.215/n^{0.584} - 14.97/n^{0.914}.$$

Table 3 shows the results using the quantile $\mathrm{qu}(n, 0.5, U, d_{ku}^9)$ and Table 4 the corresponding results for the quantile $\mathrm{qu}(n, 0.5, \Phi, d_{ku}^9)$. On comparing these with Table 2 we see that there has been an improvement for the claw density and some deterioration for the uniform and normal densities.

Distribution	50	100	200	500
Conservative Kolmogoroff	1 (9.00)	1 (9.00)	1 (8.99)	1 (8.61)
Improved Kolmogoroff	3 (7.32)	4 (6.04)	7 (2.83)	10 (0.02)
Kuiper (Uniform)	4 (5.79)	7 (3.14)	9 (0.81)	10 (0)
Kuiper (Normal)	5 (5.08)	8 (2.47)	10 (0.4)	10 (0)
9-Kuiper (Uniform)	5 (4.83)	8 (2.50)	9 (0.85)	10 (0.01)
9-Kuiper (Normal)	6 (3.64)	9 (1.33)	10 (0.14)	10 (0)

TABLE 5
Medians of number of detected modes in samples of a mixture of ten normal distributions. The values in brackets give the average absolute deviation from the true modality.

A sample of size $n = 100$ is generated from the the following mixture of ten normal distributions:

$$f(x) = \frac{1}{10}\sum_{i=1}^{10} \phi(x - (10i - 5)).$$

This mixture has recently been used by Loader (1999) to compare several bandwidth selectors for kernel estimators. The Kolmogoroff tube is squeezed, until in turn each of the following holds: $d^1_{ku}(s^\circ_n, \Delta_n) = \mathrm{qu}(n, 0.5, U, d^1_{ku})$, $d^1_{ku}(s^\circ_n, \Delta_n) = \mathrm{qu}(n, 0.5, \Phi, d^1_{ku})$, $d^9_{ku}(s^\circ_n, \Delta_n) = \mathrm{qu}(n, 0.5, U, d^9_{ku})$ and $d^9_{ku}(s^\circ_n, \Delta_n) = \mathrm{qu}(n, 0.5, \Phi, d^9_{ku})$. The resulting densities are shown in Figure 6.

Figure 6 shows in the top panel the result with the Kolmogorov metric. The second panel shows the result using the standard Kuiper metric and the third panel the result for the generalized d^9_{ku} metric.

In this example the differences between the usual Kuiper metric and the generalized Kuiper metrics are substantial. Simulations on other test beds show that the gains using this metric are real but not as large as the last example suggests.

4.6. *Discrete data* So far we have looked for an approximation to the data in the form of a Lebesgue density. However at little cost we can extend the methodology to integer-valued data which typically arise from counts. Suppose the data set $\mathbf{x}_n = (x_1, \ldots, x_n)$ contains only N different values $t_1 < t_2 < \cdots < t_N$. We look for an approximation in terms of N probabilities $p_j = \mathbb{P}(X = t_j), j = 1, \ldots N$ where the random variable X has support $t_1 < t_2 < \cdots < t_N$. Let $e_1, \ldots, e_N$ be the empirical frequencies of the t_j in the data and consider the cumulative sums

$$E_j = \sum_{i=1}^{j} e_i$$

and the tube constructed by linear interpolation of the points $(j/N, E_j), j = 0, \ldots, N$. Differentiating yields an approximation of $p_1, \ldots p_N$. This procedure corresponds to the taut string algorithm in the regression context (Davies and Kovac, 2001) with time points $t_1, \ldots t_n$ and with observations $e_1, \ldots e_n$. Our default procedure uses the Kolmogoroff tube of radius but other forms of approximation can be accommodated without much difficulty. An example is shown in Figure 7

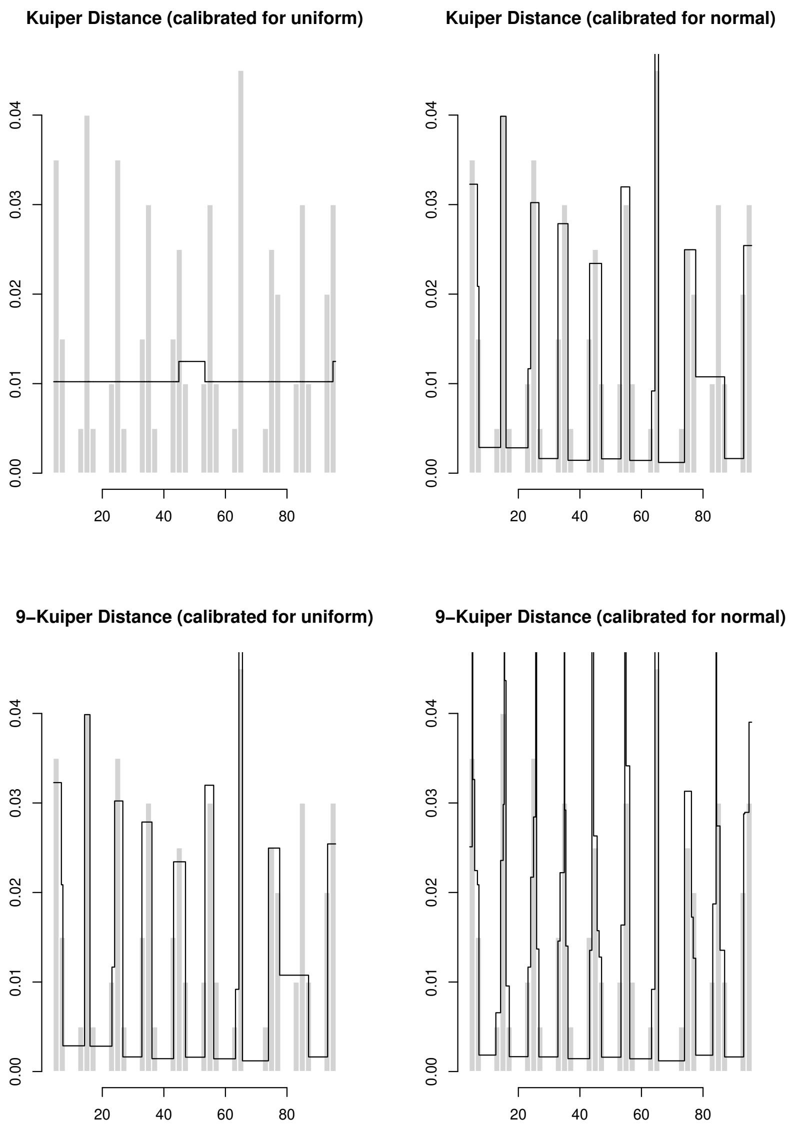

FIG. 6. *Mixture of ten normal distributions. The figures show histograms of a sample of the ten-modal density and taut string densities. The densities in the top row are based on the usual Kuiper distance, those in the bottom row are based on the generalized* 9*-Kuiper metric. In the left column the uniform distribution served as the calibrating distribution, in the right column the normal distribution was used instead.*

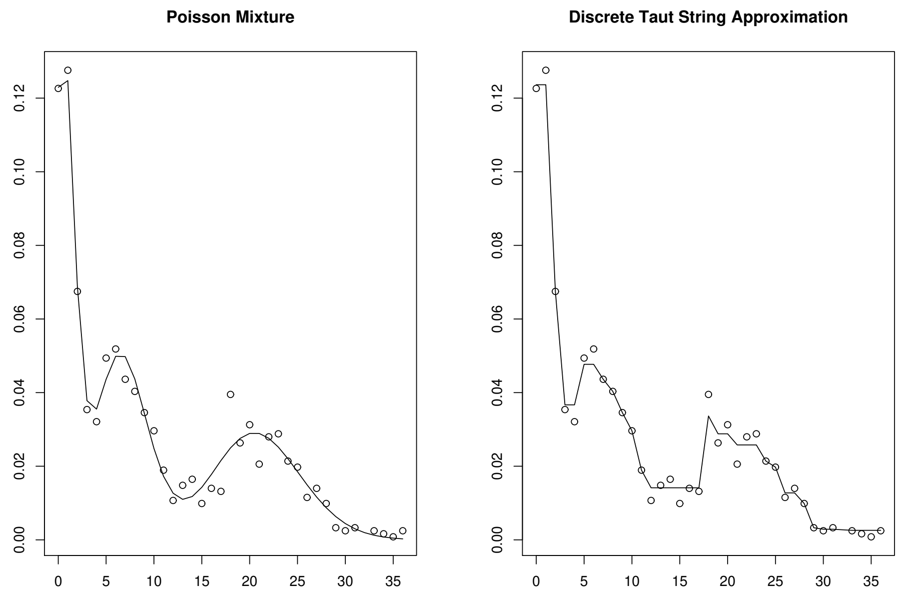

FIG. 7. *Discrete data. The left panel shows the density function of the mixture of three Poisson distributions and the frequencies of a sample of size 1200. The discrete taut string approximation is shown in the right panel.*

where the discrete taut string method was applied to 1200 observations from a mixture of three Poisson distributions

$$\mathbb{Q} = \frac{1}{3}(\mathfrak{P}(1) + \mathfrak{P}(7) + \mathfrak{P}(21)).$$

The other situation is where repeated values occur not because of the nature of the data (counting) but because of rounding. The rounding of data is very common and it can cause difficulties when looking for an approximation based on Lebesgue densities. To see the difficulties assume that some data point x is observed k times. Depending on the exact implementation of the taut string algorithm two problems may occur. If the tube is centred around the empirical distribution function and the tube width is smaller than $k/2n$, the derivative of the taut string at x will be ∞. If on the other hand the tube is constructed by linear interpolation of the empirical distribution function then the empirical mass at x of k/n is spread over the interval $[x_l, x]$ where x_l is the largest data point smaller than x. To overcome these problems we propose the following. Let ε be the precision or cut-off-error which we set to $\varepsilon = 10^{-3}\mathrm{MAD}(\mathbf{x}_n)$ where MAD denotes the Median Absolute Deviation. We construct a modified data set $\tilde{x}_1, \ldots \tilde{x}_n$ where the identical observations at x are equally spread over the interval $[x - \varepsilon/2, x + \varepsilon/2]$. To be precise we replace $x_{(j+1)} = x_{(j+2)} = \cdots = x_{(j+k)}$ by

$$\tilde{x}_{j+i} = x + \varepsilon \cdot (-\frac{1}{2} + \frac{1}{2k} + \frac{i-1}{k})$$

for $i = 1, \ldots, k$. The taut string method described above is then applied to $\tilde{x}$ instead of x.

5. Asymptotics on test beds The asymptotic behaviour of the taut string may be analysed on appropriate test beds. It turns out that asymptotically the modality is correctly estimated and that the optimal rate of convergence is attained except in small intervals containing the local extremes of the density f.

We denote the modality of the derivative of the taut string in the supremum tube $T(F_n, C/\sqrt{n})$ by k_n^C. The taut string based on the radius $C/\sqrt{n}$ will be denoted by S_n^C with derivative s_n^C. We write $I_i^e(n, C), 1 \le i \le k_n^C$, for the intervals where s_n^C attains its local extreme values and denote the midpoints of these intervals by $t_i^e(n, C), 1 \le i \le k_n^C$. The length of an interval I will be denoted by $|I|$.

THEOREM 5.1. *Let f be a k-modal density function on $\mathbb{R}$ such that*

$$\min_{g(k-1)-modal} |F(x) - G(x)| > 0.$$

Then we have for all $\delta > 0$

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

In the following A denotes a generic constant which depends only on f and whose value may differ from appearance to appearance.

THEOREM 5.2. *Assume that*

- *f has a compact support on* $[0,1]$
- *f has exactly k local extreme values at the points* $0 < t_1^e < \ldots < t_k^e < 1$
- *f has a bounded second derivative $f^{(2)}$ which is non-zero at the k local extremes.*
- $f^{(1)}(t) = 0$ *only for* $t \in \{t_1^e, \ldots, t_k^e\}$

Then the following statements hold.

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

(b) *For every* $C_1 < 6$ *and* $C_2 > 12$
$$\lim_{C\to\infty} \liminf_{n\to\infty} \mathbf{P}(|I_i^e(n,C)| \cdot \left(\frac{\sqrt{n}|f^{(1)}(t_i^e)|}{C}\right)^{1/3} \in [C_1^{1/3}, C_2^{1/3}], 1 \le i \le k) = 1.$$

(c) *Let $\xi_j^{n,C}$ be the knots of the taut string S_n^C and denote*
$$m(n,C) = \max\{\xi_{j+1}^{n,C} - \xi_j^{n,C} : \xi_j^{n,C}, \xi_{j+1}^{n,C} \in (0,1)\backslash \cup_1^k I_i^e(n,C)\}.$$
For some constant A only depending on f we have
$$\lim_{C\to\infty} \liminf_{n\to\infty} \mathbf{P}\left(m(n,C) \le \left(A|f^{(1)}(x_j)|^{-2/3}\left(\frac{\log n}{n}\right)^{1/3}\right)\right) = 1.$$

(d) *Denote*
$$M(n,C) = [A\left(\frac{\log n}{n}\right)^{1/3}, 1 - A\left(\frac{\log n}{n}\right)^{1/3}]\backslash \cup_i^n I_i^e(n,C).$$
Then for some constant A only depending on f we have
$$\lim_{C\to\infty} \liminf_{n\to\infty} \mathbf{P}\left(\max_{t\in M(n,C)} |f(t) - f_n^C(t)| \le \left(A|f^{(1)}(t)|^{1/3}\left(\frac{\log n}{n}\right)^{1/3}\right)\right) = 1.$$

(e) *For some constants A_1 and A_2 only depending on f we have*
$$\lim_{C\to\infty} \liminf_{n\to\infty} \mathbf{P}(\max_{t\in\cup_1^n I_i^e(n,C)} |f(t) - f_n^C(t)| \le AC^{2/3}n^{-1/3}) = 1.$$

Part (d) of the theorem shows that bounded away from the local extrema the taut string density attains the optimal rate of convergence up to a logarithmic factor. The proofs follow the lines of Davies and Kovac (2001) and we omit them.

6. Cell occupancy frequencies and local squeezing

6.1. *Cell occupancy frequencies* The multiresolution procedure of Davies and Kovac (2001) is based on comparing the residuals of some regression function with those of Gaussian white noise. The comparison is based on the means on intervals which form a multiresolution scheme. A similar idea can be applied to the density problem. A distribution F is an adequate model for the data $\mathbf{x}_n = (x_1, \ldots, x_n)$ of the transformed data

$$\mathbf{u}_n = F(\mathbf{x}_n) = (F(x_1), \ldots, F(x_n))$$

looks like an i.i.d. sample of size n from the uniform distribution on $[0,1]$. This is done by comparing the frequencies

$$w_{jk}^n = |\{l : k2^{-j} < u_i \le (k+1)2^{-j}\}|, \quad 0 \le k \le 2^j,\, 1 \le j \le m,$$

with those to be expected from i.i.d. uniform random variables. The maximum resolution level m is taken to be the smallest integer such that $n \le 2^m$. Suppose that $U_1, \ldots, U_n$ are independently and uniformly distributed on $[0,1]$. Then

$$W_{jk}^n = |\{l : k2^{-j} < U_i \le (k+1)2^{-j}\}|$$

is binomially distributed $b(n, 2^{-j})$. For given α we define the upper bounds $v_{j,k}^n(\alpha)$ by

$$v_j^n(\alpha) = \min\left\{l : \mathbf{P}(Z_j^n \ge l) \le \frac{1-\alpha}{2n}\right\} \tag{6.14}$$

where Z_j^n satisfies the binomial distribution $b(n, 2^{-j})$. It follows that

$$\mathbf{P}(W_{jk}^n < v_j^n(\alpha), 1 \le k \le 2^j, 1 \le j \le n) \ge \alpha.$$

This gives rise to the following problem:

PROBLEM 6.1 CELL OCCUPANCY PROBLEM. *Determine the smallest integer k_n for which there exists a density f^n with k_n modes and whose distribution F^n is such that the cell frequencies $w_{j,k}^n$ satisfy*

$$w_{j,k}^n = |\{l : k2^{-j} < F^n(x_i) \le (k+1)2^{-j}\}| \le v_j^n(\alpha) \tag{6.15}$$

where the $v_{j,k}^n(\alpha)$ are given by (6.14).

Although the cell occupancy problem is well defined there is no obvious connection between the modality of the density f^n and the set of inequalities (6.15). We therefore again adopt the strategy of producing test densities derived from the taut string and gradually increase the modality until the inequalities (6.15) hold. The knowledge of which inequalities fail to hold provides further information which we are able to exploit as described in the next section.

6.2. *Local squeezing* Local squeezing is described in Davies and Kovac (2001). We apply it to the density problem as follows. Suppose that one of the inequalities of (6.15) fails. We suppose that

$$w^n_{j,k} = |\{l : k2^{-j} < F^n(x_i) \le (k+1)2^{-j}\}| \ge v^n_{j,k}(\alpha)$$

Clearly there exists an interval $[x_{(i1)}, x_{(i2)}]$ such that $k2^{-j} < F^n(x_i) \le (k+1)2^{-j}$ for all points x_i in $[x_{(i1)}, x_{(i2)}]$. We now squeeze the tube locally on this interval and do this for all intervals where an inequality fails. The general procedure for doing this is as follows. Firstly, a suitable initial global tube radius γ_0 is chosen using the Kolmogorov or generalized Kuiper metrics and the taut string is calculated. If all the inequalities (6.15) hold the procedure terminates. If not we reduce the radius by a factor $\rho, 0 < \rho < 1$, on all intervals where an inequality fails. Typical choices for ρ are 0.9 or 0.95. The taut string through the modified tube is calculated and using this new test distribution it is checked whether the inequalities (6.15) hold. If so, the procedure terminates. Otherwise the tube radius is again decreased by the factor ρ on all intervals where an inequality fails. This is continued until all the inequalities are satisfied.

It is not easy to analyse the behaviour of the local squeezing procedure. In the case of nonparametric regression Davies and Kovac (2001) give a heuristic indicating that the procedure improves the behaviour at local extremes. A similar argument can be given for densities but as it remains heuristic we omit it.

The ability of the local squeezing method to detect low power peaks (see Davies and Kovac, 2001) is shown by the following example. The data consist of a sample of size 1000 drawn from a mixture of four normal distribution $\mathbb{P}$ given by

$$\mathbb{P} = 0.8 * \mathcal{N}(0,3) + 0.015 * \mathcal{N}(8, 0.02) + 0.015 * \mathcal{N}(9, 0.02) + 0.17 * \mathcal{N}(15, 0.2).$$

The density is shown in the upper left corner of Figure 8. It exhibits a main peak, a moderate peak on the right and in the centre two low power but very concentrated and very close peaks.

The upper right panel shows a kernel estimate which was calculated using a Gaussian kernel. The mode on the right-hand side was detected, but is considerably broader than the normal component of the original density function. The main component is well captured but there are three superfluous peaks. Finally, the two sharp peaks in the centre of the data result in one flat local maximum. The lower left panel shows the result with the taut string method and two global tube radii. The solid line is derives from the d^1_{ku} metric. There are no spurious local extremes but the small central peaks are not detected. The dashed line shows that further global squeezing would only lead to additional spurious modes on the left before the central peaks are detected. Finally, the lower right panel shows the result of local squeezing. The number and locations of the local extrema are estimated correctly and the difference to the original density function is very small.

7. Simulations The performance of the taut string method on ten densities was analysed by a simulation study. Figure 9 shows the densities that were used. They are the uniform distribution on $[0, 1]$, the Gaussian distribution and eight

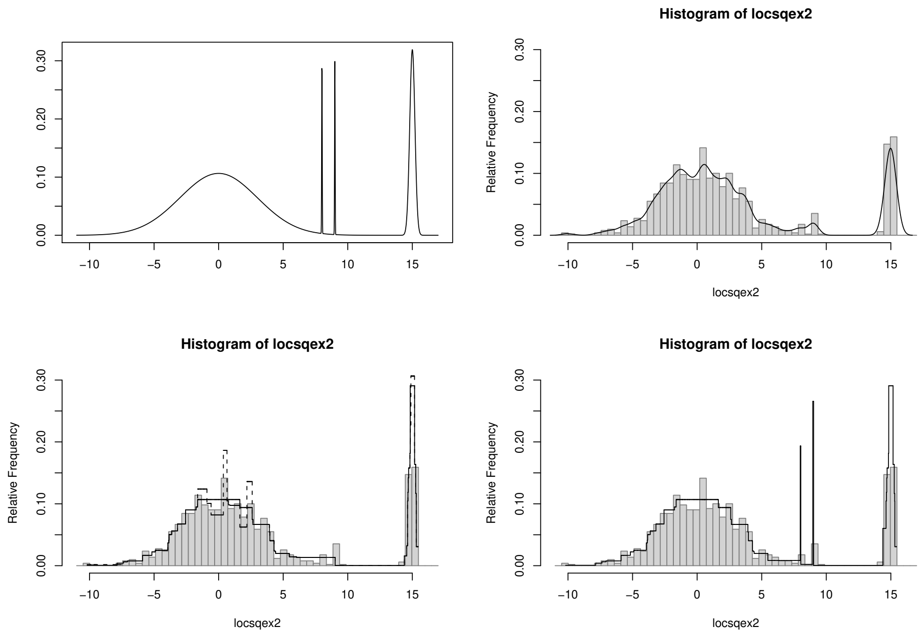

FIG. 8. *Local squeezing: The upper left panel shows the density of a mixture of four normal distributions. A kernel estimate is drawn in the upper right panel. The lower left panel illustrates global squeezing first with a solid line using the Kolmogorov bounds and then with a dashed line the taut string density with four modes. The local squeezing estimate is depicted in the lower right panel.*

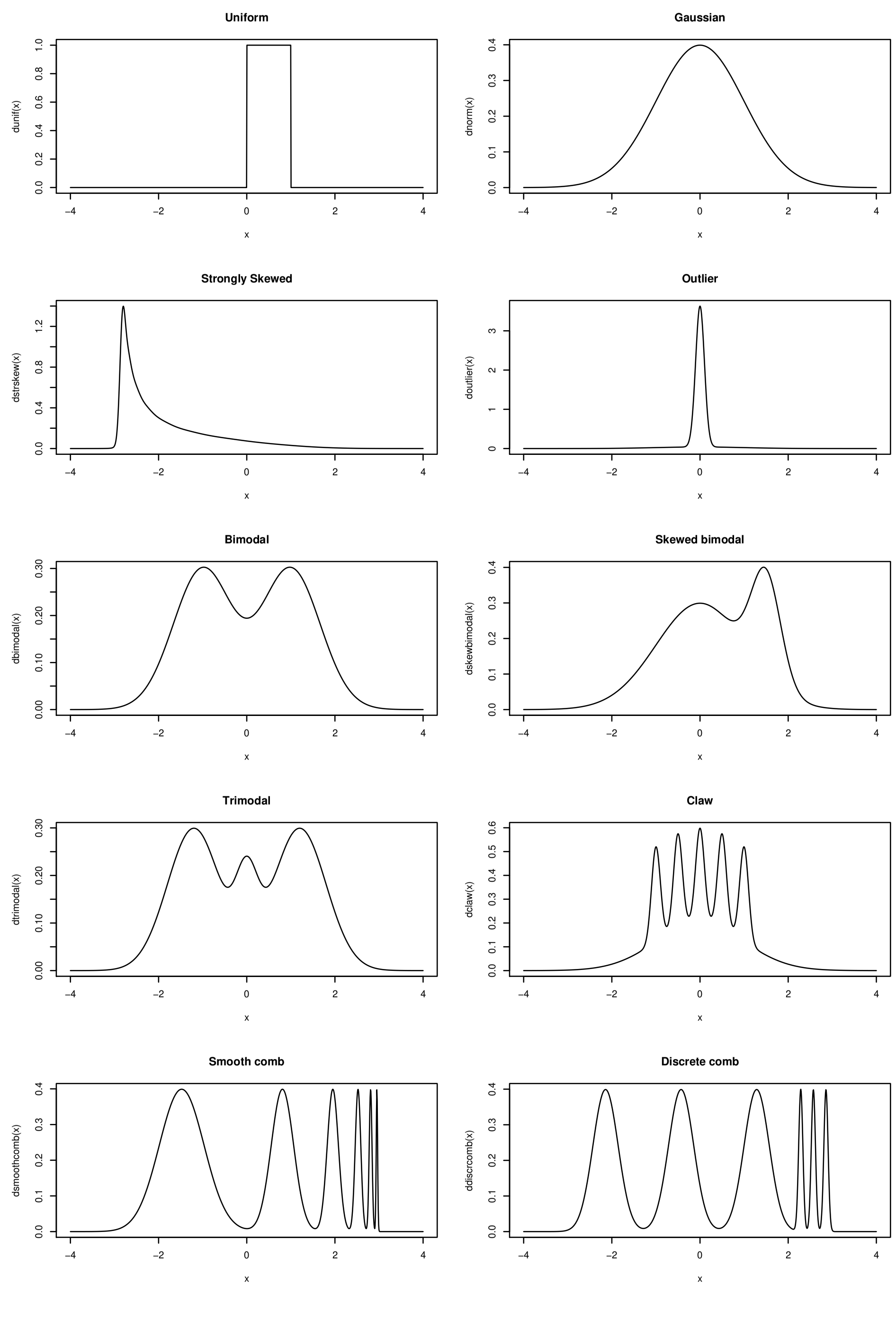

FIG. 9. *Ten densities that were used in a simulation study.*

Density	size	KERNCV	KERNSJ	STRINGKOL50	STRINGKUIP50LOCAL
Uniform	100	1	16	99	50
	500	0	1	100	44
	2000	0	0	99	55
Gaussian	100	77	79	100	95
	500	79	78	100	99
	2000	74	59	100	99
Strongly skewed	100	4	0	99	96
	500	1	0	100	100
	2000	0	0	98	100
Outlier	100	15	0	100	96
	500	0	0	99	100
	2000	0	0	93	99
Bimodal	100	71	81	0	33
	500	75	84	1	68
	2000	75	73	54	95
Skewed bimodal	100	32	46	0	26
	500	45	37	0	20
	2000	34	12	0	31
Trimodal	100	29	12	0	1
	500	57	67	0	4
	2000	81	82	0	10
Claw	100	1	0	0	0
	500	2	2	0	5
	2000	0	0	0	92
Smooth comb	100	18	0	0	0
	500	5	0	0	0
	2000	1	1	0	8
Discrete comb	100	12	0	0	0
	500	2	0	0	1
	2000	0	82	0	88

Table 6
Correctly detected modes in samples of various densities and for several automatic methods.

normal mixtures from Marron and Wand (1992). Four automatic methods were compared:

- KERNCV: A kernel estimator using likelihood cross-validation for the choice of the bandwidth
- KERNSJ: A kernel estimator using Sheather-Jones plugin bandwidths
- STRINGKOL50: Taut string estimates based on the conservative Kolmogorov metric with $\alpha = 0.5$
- STRINGKUIP50LOCAL: Taut string estimates based on the 9-Kuiper metric with $\alpha = 0.5$ and local squeezing

Each method was applied to 100 samples of each of the densities and three different sample sizes (100,500,2000). For each estimate it was checked if the correct number of modes was found and if the positions of the modes corresponded to those of the densities. Table 6 shows how often the modes were determined correctly for the various densities and methods. Some comments are in order. Firstly all the densities are mixtures of a small number of Gaussian distributions with the exception of the uniform density. It is precisely this density where the kernel methods based on a Gaussian kernel fail. The trimodal distribution is the one where the kernel methods perform clearly better than the taut string method. If however the central Gaussian distribution is replaced by a uniform distribution then the kernel methods again fail. This may indicate that the comparison is weighted in favour of the kernel methods as both they and the densities are based on the Gaussian kernel. Finally we note that the performance of the kernel methods seems to deteriorate with increasing sample size.

8. Hidden periodicities, spectral densities and taut strings

8.1. *Hidden periodicities* The second problem we consider is that of detecting hidden periodicities in a data set $\mathbf{x}_n$. One method of formulating the problem is the following: calculate an appropriate spectral density function f^n and identify the hidden periodicities in the data with the peaks of f^n (Brillinger, 1981; Priestley, 1981; Brockwell and Davis, 1987 and the references given there).

Existing methods by and large belong to one of two different categories of procedures. The first is nonparametric and uses some form of smoothing of the periodogram. This may take the form of kernel estimators or splines or wavelets or averages of periodograms obtained by splitting the data into blocks (see Chapter 5 of Brillinger (1981), Neumann (1996) and the references given there). The second possibility is to model the data by an autoregressive process whose order is determined using some criterion such as AIC (Akaike, 1977), BIC (Akaike, 1978) or HQ (Hannan and Quinn, 1979). The spectral density associated with the autoregressive process is then used to determine the hidden periodicities. None of these methods controls the number of peaks directly although the problem of hidden peaks is one of modality.

Before proceeding we assume that the data have been normalized to have sample mean zero and variance 1. To ease the notation the transformed data will also be denoted by $\mathbf{x}_n$. In the context of time series e_n will denote the empirical spectral density or the periodogram defined by

$$e_n(\omega) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} x_t \exp(i\omega t)\right|^2, \quad 0 \le \omega \le 2\pi. \tag{8.16}$$

The corresponding empirical spectral distribution function E_n given by

$$E_n(\omega) = \int_0^{\omega} e_n(\lambda)d\lambda. \tag{8.17}$$

The candidate spectral densities we use are based on the taut strings S_n through the Kolmogoroff tubes centred at E_n. We assume that the taut string is constrained to go through $(0, E_n(0)) = (0,0)$ and $(2\pi, E_n(2\pi)) = (2\pi, 1)$. One difference to the i.i.d. model is the fact that the empirical spectral distribution function is defined

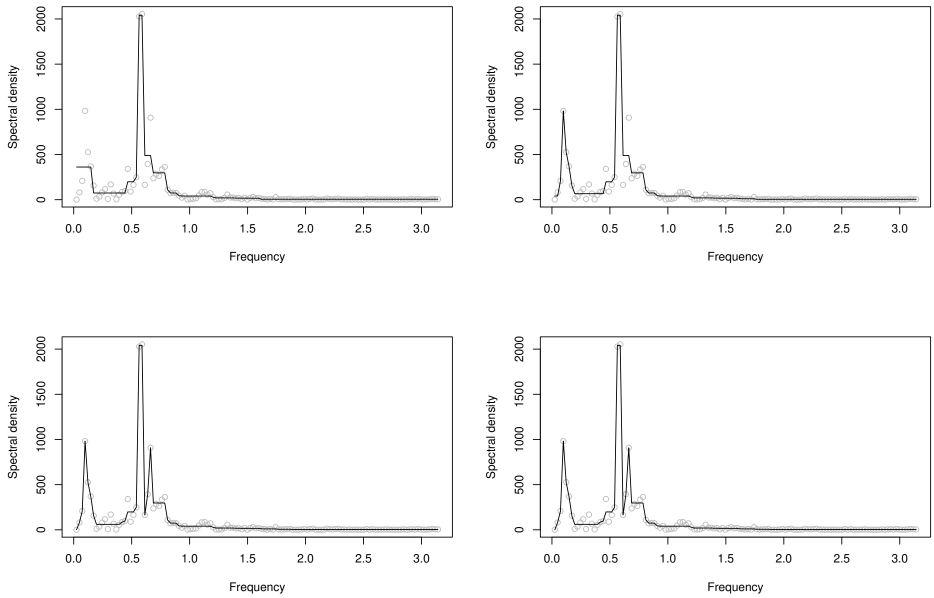

FIG. 10. *Sunspot data with number of peaks increasing from 1 to 4.*

for all ω. In practice a grid must be chosen which, when analysing the asymptotic behaviour on test beds, becomes increasingly fine. This has no theoretical effect. We use the Fourier frequencies $\frac{2\pi j}{n}, j = 0, \ldots n-1$, where the data have been augmented by zeros to produce a power of two. Choosing a finer grid has had no effect on the data sets we have analysed so far.

8.2. *Data analysis* Just as in Section 4.2 it is possible to use the taut string as a data analytical tool. The radius of the Kolmogoroff tube is gradually decreased and the resulting densities give information about the power and positions of the peaks. We give two examples. Figure 10 shows the first four peaks for the sunspot data (Anderson 1971).

The second example is an artificial data set generated according to a scheme of Gardner (1988). Gardner does not explicitly specify the spectral density except that it has Gaussian shape with centre frequency $2\pi\lambda$ with $\lambda = 0.35$. The density f of (8.18) approximates the graph shown in Gardner's Figure 9.4 (a)

$$f(\omega) = \frac{1}{3} e^{-300\left(\frac{\omega}{2\pi} - 0.35\right)^2}. \tag{8.18}$$

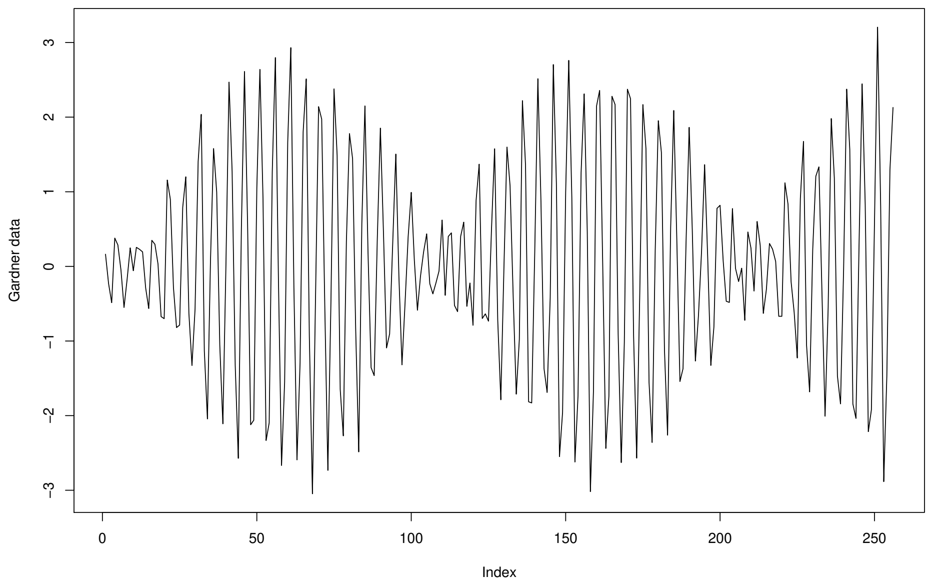

FIG. 11. *The Gardner data*

A realization of length 2048 was generated by filtering in the frequency domain. The following pure sine terms were added

$$\sqrt{2}\sin\left(2\pi(0.2\times t-106/360)\right), \sqrt{2}\sin\left(2\pi(0.21\times t-45.1/360)\right),$$
$$\sqrt{2}/10\sin\left(2\pi(0.1\times t-32.6/360)\right).$$

A segment of length 256 starting at $t = 1023$ was taken as the simulated sample. It is shown in Figure 11.

A similar data set was analysed by Gardner (Chapter 9.E, Experimental Study, Gardner, 1988) in an experimental study of the performance of different spectral estimates. Figure 12 shows the first four peaks (in a log scale) for the data set of Figure 11. Finally Figure 13 shows the four peak density together with the periodogram.

8.3. *Two concepts of approximation* The concepts of approximation used in the i.i.d. case had the advantage that the distributions involved were independent of the approximating model. This is no longer the case for stationary models. Furthermore, specifying the spectral distribution function F does not specify the joint distribution of the stationary sequence. If however one is prepared to accept a Gaussian model

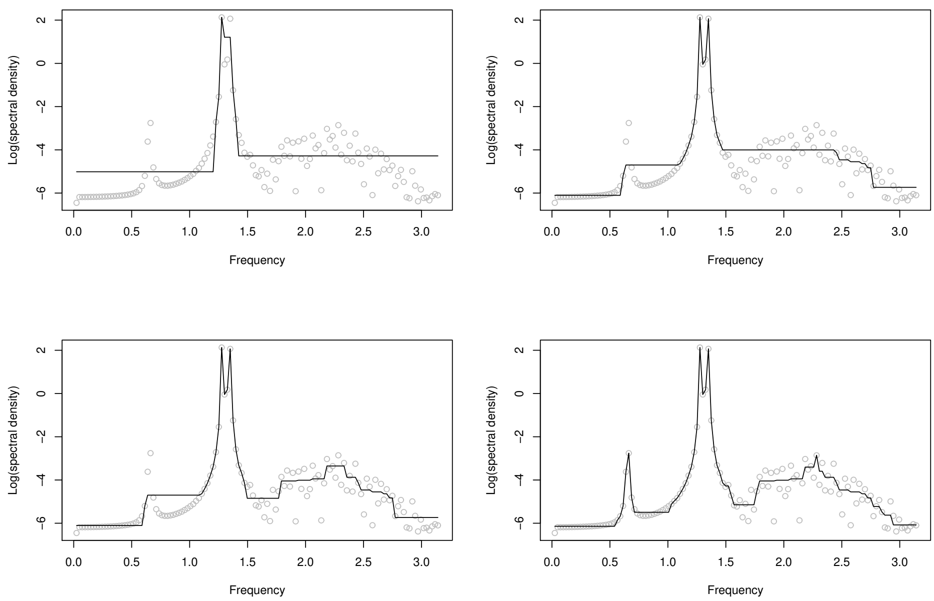

FIG. 12. *Gardner data with number of peaks increasing from 1 to 4*

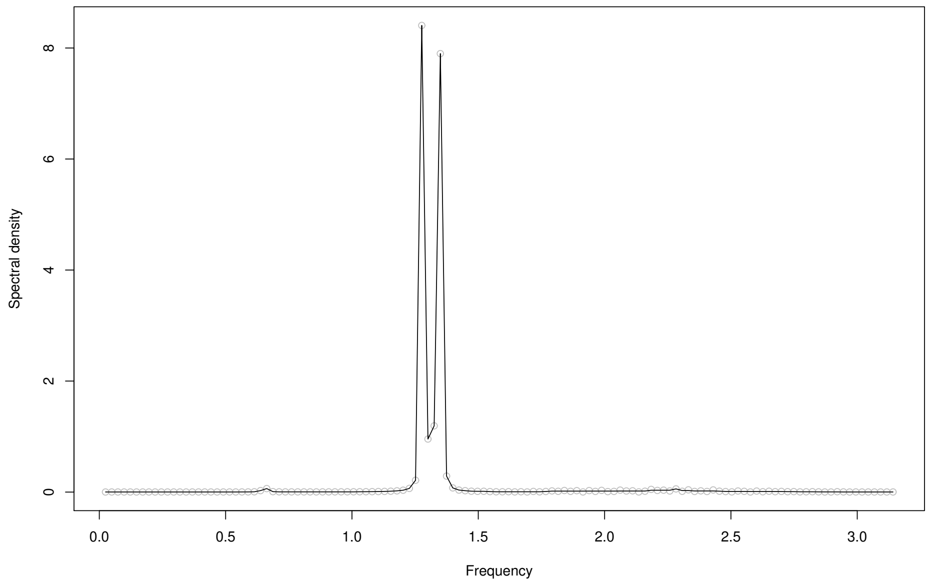

FIG. 13. *Gardner data with four peaks and the periodogram*

then the distribution $\mathbb{P}_F$ of the sequence is determined by F. In analogy with the i.i.d. case we have

PROBLEM 8.1 KUIPER SPECTRAL DENSITY PROBLEM. *Determine the smallest integer k_n for which there exists a spectral density f^n with k_n modes and whose distribution F^n satisfies*

$$d_{ku}(E_n, F^n) \le qu(n, \alpha, \mathbb{P}_{F^n}, d_{ku}) \tag{8.19}$$

where $\mathbb{P}_{F^n}$ denotes the distribution of the observations under the model.

There are two disadvantages with the procedure based on this concept of approximation. One is that the quantile in (8.19) depends on F^n. It would be possible to overcome this by at each stage using the taut string S_n and then simulating the quantile $qu(n, \alpha, \mathbb{P}_{S^n}, d_{ku})$. This is clearly very time consuming. The second disadvantage is the following. Under appropriate conditions (Dahlhaus, 1988) we have the weak convergence result

$$\sqrt{n}(F_n - F) \Rightarrow Z$$

where F_n denotes the empirical spectral distribution function of the model with spectral distribution function F and density f and Z denotes a continuous zero mean Gaussian process defined by

$$\mathbb{E}(Z(\lambda_1)Z(\lambda_2)) = \int_0^{\min(\lambda_1,\lambda_2)} f(\omega)^2 d\omega \tag{8.20}$$

It follows from (8.20) that any large peaks will swamp smaller peaks which may be present and so prevent their detection. The one advantage of (8.19) is that it allows an asymptotic evaluation.

A more sensitive procedure is based on some kind of multiresolution analysis. Suppose for the moment that the sample size n is a power of two $n = 2^m$. Given a spectral density function f we define

$$g_n(f, \omega) = \frac{e_n(\omega)}{f(\omega)}. \tag{8.21}$$

and consider the multiresolution scheme

$$w_{jk}(f) = \sum_{l=(j-1)2^k+1}^{j2^k} g_n(f, \omega_{l,n}), j = 1, \ldots, 2^{m-k-1}, k = 0, \ldots, m-1, \tag{8.22}$$

where the $\omega_{l,n} = 2\pi l/n$ are the Fourier frequencies. The class of stationary processes with spectral density function f is too large to provide a meaningful definition of approximation so we now restrict attention to Gaussian processes. Corresponding to level dependent thresholds for wavelets we specify lower and upper bounds $l_{k,n}$ and $u_{k,n}$ respectively for the multiresolution coefficients (8.22). These now define the

Problem 8.2 Multiresolution spectral density problem. *Determine the smallest integer k_n for which there exists a spectral density f^n with k_n modes such that*

$$l_{k,n} \le w_{jk}(f^n) \le u_{k,n}, j = 1, \ldots, 2^{m-k-1}, k = 0, \ldots, m-1. \tag{8.23}$$

The default bounds we use are $l_{k,n} = \mathrm{qu}(\alpha_{1n}, 2^k)$ and $u_{k,n} = \mathrm{qu}(\alpha_{2n}, 2^k)$ where $\mathrm{qu}(\beta, \nu)$ denotes the β-quantile of the Gamma distribution with ν degrees of freedom, $\alpha_{1n} = (1-\alpha)/2n$ and $\alpha_{2n} = 1 - \alpha_{1n}$ with $\alpha = 0.9$. The bounds are based on the Gaussian model and the asymptotic results for such processes as given for example by Theorem 5.2.6 of Brillinger (1981). If the asymptotic results held precisely for finite n then the bounds are chosen such that for a stationary Gaussian process with spectral density function f the inequalities (8.23) hold with probability at least 0.9 for $f^n = f$. As the individual $g_n(f, \omega)$ of (8.21) for $\omega = \frac{2\pi j}{n}$ are asymptotically independent the bounds will be approximately of the correct order, again for Gaussian processes with a spectral density function. The usefulness of the bounds for real data sets is an empirical matter. In particular they will be be too slack if the spectral distribution function contains point masses. This is the case for the Gardner data given above and may be seen in Figure 13. The absolute continuous part of the spectrum shows a degree of noise whereas the remainder of the spectrum is noise free. The default bounds we propose will detect the first peak but they are not sufficiently tight to split the two main peaks. On the other hand if the bounds are sufficiently tight to separate the two peaks then superfluous peaks will be produced in the absolutely continuous part of the spectrum. There would seem to be no easy solution which will work equally well for continuous as well as for discrete spectra.

We have no algorithm to solve the problem as it stands so again we use the local squeezing variant of the taut string method. The string is squeezed locally on the intervals where (8.23) fail and this is continued until all the inequalities are satisfied. When doing this however care must be taken regarding the order in which the inequalities are treated. From the form of $g_n(f, \omega)$ in (8.21) it is clear that a particular $g_n(f, \omega)$ can be very large and influence all interval containing this particular frequency and this although the corresponding $e_n(\omega)$ is very small. Squeezing locally over all intervals effected by this frequency will often produce many superfluous peaks. To avoid this we consider the intervals in order of size commencing with intervals of size one. When all the inequalities are satisfied we then move on to intervals of size two and continue in this manner until all the inequalities are satisfied. This is the default version of the algorithm. If global squeezing is used then the peaks will be introduced according to their power and may be introduced on intervals where the inequalities (8.23) are satisfied. This is the case for the Gardner data. If the default version with local squeezing is used the main peak is not split. If however global squeezing is used then it is split.

8.4. *Asymptotics on test beds* We indicate briefly the results of an asymptotic analysis using the Kuiper concept of approximation. The test bed we consider is that of a stationary process $X_n(F), 1 \le n < \infty$, with a spectral distribution function F and spectral density function f as follows.

TEST BED 8.1.

- *F has exactly k local extreme values on the interval* $(0, \pi)$.
-
$$inf_{G \in \mathcal{F}(k-1)} \sup_{\omega \in [0,\, \pi]} |F(\omega) - G(\omega)| > 0$$
where $\mathcal{F}(k-1)$ denotes the set of distributions with at most $k-1$ local extreme values.

To investigate the behaviour of the taut string on the test bed (8.1) we consider a tube of width $2C/\sqrt{n}$ and denote the taut string through this tube by $S_n(C)$ with derivative $s_n(C)$ and modality k_n^C. The intervals on which $s_n(C)$ takes on its local extreme values will be denoted by $I_i^e(n, C), i = 1, \dots, k_n^C$ with midpoints $\omega_i^e(n, C)$. The first theorem shows that on test bed (8.1) the number and locations of the local extreme values are determined in a consistent manner.

THEOREM 8.1. *Consider the test bed (8.1). Then for all $\delta > 0$*

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

To obtain rates of convergence on appropriate test beds we must impose further conditions.

TEST BED 8.2.

- *all spectral densities f^j of order j exist and* $\sup_\omega |f^j(\omega)| \le B^j$ *for some constant B*
- *the spectral density function $f = f^2$ has a continuous second derivative $f^{(2)}$*
- *f has exactly k local extreme values, $0 < \omega_1, \dots, \omega_k < 2\pi$, and $f^{(1)}(\omega) \neq 0$ for $\omega \in [0, 2\pi] \backslash \{\omega_1, \dots, \omega_k\}$*
- $f^{(2)}(\omega_j) \neq 0, j = 1, \dots, k$
- *the fourth order spectral density is continuous.*

The above conditions correspond to (i) of Assumption 2.1 of Dahlhaus (1988).

Rates of convergence require a modulus of continuity for the process $Z_n = \sqrt{n}\,(F_n - F)$ where F_n denotes the empirical spectral distribution function of the sample $(X_1(F), \dots, X_n(F))$. Under the conditions of Theorem 2.4 of Dahlhaus (1988) it follows that

$$\sup_{0\le\omega_1<\omega_2\le 2\pi, \omega_2-\omega_1<\delta} |Z_n(\omega_2) - Z_n(\omega_1)| \le C\sqrt{\omega_2 - \omega_1}\,|\log(\omega_2 - \omega_1)| \qquad (8.24)$$

with probability tending to one as δ tends to zero. From this the it can be shown that the rate of uniform convergence away from the local extremes is $O\left(\left(\frac{(\log n)^2}{n}\right)^{1/3}\right)$. This differs from the rate of convergence for the test beds considered in Davies and Kovac (2001) by an extra $\log n$ term. This is explained by the different modulus of continuity. On the test beds of Davies and Kovac (2001) it is $\sqrt{\delta|\log\delta|}$ whereas above it is $\sqrt{\delta}\,|\log\delta|$.

8.5. *Examples* The default version we use is the procedure deriving from the multiresolution problem with $\alpha = 1 - 0.1/n$ and a squeezing factor of 0.9. For the sunspot data the result is the one peak density shown in the top left panel of Figure 10. For the Gardner data the result is the four peak density shown in the bottom right panel of Figure 12. Finally we consider data generated according to a scheme of Neumann (1996) which is as follows:

$$X_n = Y_n + c_0 Z_n \tag{8.25}$$

where

$$Y_n + a_1 Y_{n-1} + a_2 Y_{n-2} = b_0\varepsilon_n + b_1\varepsilon_{n-1} + b_2\varepsilon_{n-2}$$

and $\{\varepsilon_n\}, \{Z_n\}$ are independent Gaussian white noise processes with variance 1. Neumann chose the coefficient values as follows: $a_1 = 0.2$, $a_2 = 0.9$, $b_0 = 1$, $b_1 = 0$, $b_2 = 1$ and $c_0 = 0.5$. A sample of size 1024 was generated according to this scheme. Figure 14 shows the logarithm of the spectral density of the sequence $\{X_n\}$ together with the logarithm obtained from the default version of the taut string method. The two peaks are correctly identified. The wavelet method used by Neumann results in 6 peaks ((b) of Figure 2 of Neumann 1996) for the data set he considered.

REFERENCES

H. Akaike. On entropy maximization principle. In P.R. Krishnaiah, editor, *Applications of Statistics*, pages 27–41. North Holland, Amsterdam, 1977.

H. Akaike. A new look at the bayes procedure. *Biometrika*, 65:53–59, 1978.

T.W. Anderson. *The Statistical Analysis of Time Series.* Wiley, New York, 1971.

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

P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1968.

D.R. Brillinger. *Time Series: Data Analysis and Theory.* Holden–Day, San Francisco, 1981.

P.J. Brockwell and R.A. Davis. *Time Series:Theory and Methods.* Springer, New York, 1987.

R. Dahlhaus. Empirical spectral processes and their applications to time series analysis. *Stochastic Processes and their Applications*, 30:69–83, 1988.

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

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

D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

R.M. Dudley. *Real Analysis and Probability.* Wadsworth and Brooks/Cole, California, 1989.

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

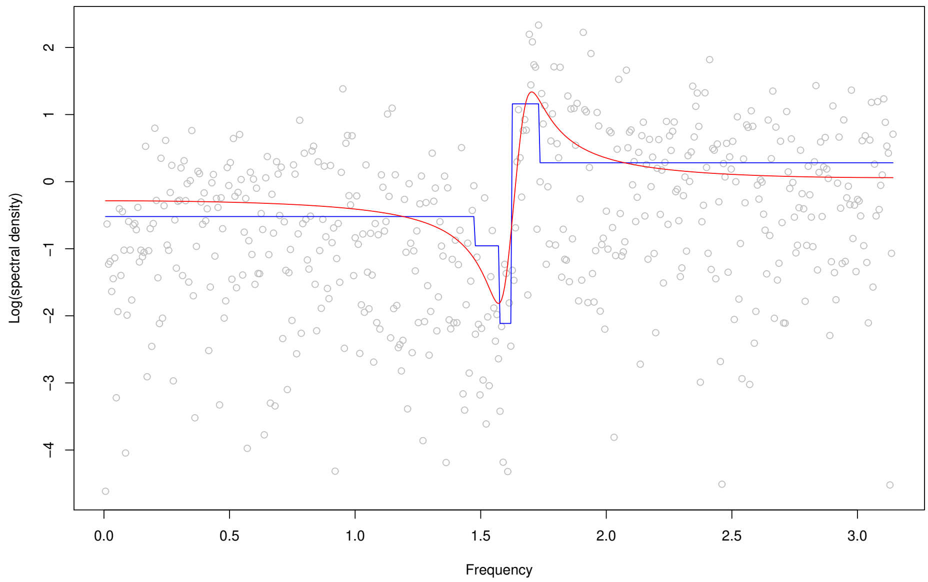

FIG. 14. *Log spectral densities of a sample of size 1024 generated by the scheme (8.25).*

W.A. Gardner. *Statistical Spectral Analysis: A Nonprobabilistic Theory.* Prentice Hall, New Jersey, 1988.

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

E.J. Hannen and B.G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Satistical Society, Series B*, 41:190–195, 1979.

J.A. Hartigan. Testing for antimodes. In W. Gaul, O. Optiz, and M. Schader, editors, *Data Analysis, Scientific Modeling and Practical Applications*, pages 169–181, New York, 2000. Springer.

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

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

D.R.M. Herrick, G.P. Nason, and B.W. Silverman. Some new methods for wavelet density estimation. Technical report, Department of Mathematics, University of Bristol, United Kingdom, 200.

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

C.R. Loader. Bandwidth selection: classical or plug-in? *Annals of Statistics*, 27:415–438, 1999.

M. Löwendick and P.L. Davies. Smoothing functions under shape constraints. Department of Mathematics and Computer Science, University of Essen,Essen, 1998.

A Majidi. Nicht-parametrische regression unter modalitätskontrolle. Master's thesis, Fachbereich Mathemttik und Informatik, Universität Essen, 2000.

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

J.S. Marron and M.P. Wand. Exact mean integrated squared error. *Annals of Statistics*, 20:712–736, 1992.

D.M. Mason, G.R. Shorack, and J.A. Wellner. Strong limit theorems for oscillation moduli of the uniform empirical process. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65:83–97, 1983.

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

D. W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.*, 86(415):738–746, 1991.

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

M.H. Neumann. Spectral density estimation via nonlinear wavelet methods for stationary non-gaussian time series. *Journal of Time series Analysis*, 17(6):601–633, 1996.

W. Polonik. Density estimation under qualative assumptions in higher dimensions. *Journal of Mulitivariate Analysis*, 55:61–81, 1995.

W. Polonik. Measuring mass concentrations and estimating density contour clusters–an excess mass approach. *Annals of Statistics*, 23:855–881, 1995.

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

M.B. Priestly. *Spectral Analysis and Times Series*, volume 1: Univariate Series. Academic Press, London, 1981.

S. Richardson and P.J. Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society*, 59:731–792, 1997.

K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.*, 92(439):894–902, 1997.

S.R. Sain and D.W. Scott. On locally adaptive density estimation. *Journal of the American Statistical Association*, 91:1525–1534, 1996.

S.J. Sheather and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, 53:683–690, 1991.
B. W. Silverman. *Density estimation for statistics amd data analysis.* Chapman and Hall, London, 1986.
J.S. Simonoff. *Smoothing methods in statistics.* Springer-Verlag, New York, 1996.
I. Verdinelli and L. Wasserman. Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.*, 26(4):1215–1241, 1998.
B Vidakovic. *Statistical modeling based on wavelets.* Wiley, New York, 1999.
M. P. Wand and M. C. Jones. *Kernel smoothing.* Chapman and Hall Ltd., London, 1995.
G. S. Watson. Smooth regression analysis. *Sankhyā*, 26:101–116, 1964.

PROOFS

9. Proof of Theorem 5.1 Let B denote the Brownian Bridge on $[0,1]$ and $E_n = \sqrt{n}(F_n - F)$ the empirical process. Then

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{t\in\mathbb{R}} |E_n(t)| < x\right) = \mathbb{P}\left(\max_{t\in\mathbb{R}} |B(F(t))| < x\right) \geq 1 - 2\exp(-2x^2)$$

using Proposition 12.3.3 of Dudley (1989). It follows that

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{t\in\mathbb{R}} |F_n(t) - F(t)| \leq \frac{C}{\sqrt{n}}\right) \geq 1 - \exp(-2C^2).$$

As the taut string minimizes the modality in $T(F_n, \frac{C}{\sqrt{n}})$ we see that

$$\lim_{C\to\infty} \lim_{n\to\infty} \mathbb{P}\left(k_n^C \leq k\right) = 1.$$

On the other hand, if we denote by M all functions on $\mathbb{R}$ of modality at most $k-1$, then

$$\inf_{g\in M} \sup_{t\in\mathbb{R}} |F(t) - G(t)| > 0. \tag{9.26}$$

Using the Glivenko-Cantelli theorem this implies that for for arbitrary $C > 0$ and for all n large enough the modality of the taut string is at least k. We conclude that

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

The other claims are proved similarly. For example, assume that the lengths of the intervals $I_i^e(n, C)$ do not converge in probability to zero. Then, for some subsequence n_j and constant $\delta > 0$

$$\inf_j \mathbb{P}(\max_i |I_i^e(n_j, C)| > \delta) > 0.$$

Denote by M the set of all functions on $\mathbb{R}$ with finitely many discontinuities and such that they attain a local extreme values on an interval of size at least δ. Then we are again in the situation of (9.26) and have a contradiction to the Glivenko-Cantelli theorem. A similar argument applies to the convergence of the location of the local extreme points. $\square$

10. Proof of Theorem 5.2 *Proof of (a):*
We require

LEMMA 10.1. *Assume that F is a convex function on $[0, x_0]$ and has three continuous derivatives with $F''(x_0) = 0$ and $F^{(3)}(x_0) < 0$. Consider a sequence $a_n \in [0, x_0]$ converging to x_0. Then*

$$G(h) = \frac{F(x) - F(a_n)}{x - a_n}$$

is strictly increasing on $[a_n, x_0 + \frac{1}{2}(x_0 - a_n) + o(x_0 - a_n)]$.

We now turn to the proof of (a). Since the empirical process

$$E_n = \sqrt{n}(F_n - F)$$

is tight, we conclude (Billingsley, 1968, p. 106).

$$\lim_{C\to\infty} \lim_{n\to\infty} \mathbb{P}\left(\sup_{s \le t \le s+2\tau_n} |E_n(s) - E_n(t)| \le \frac{1}{C} \right) = 1$$

where

$$\tau_n = \max(t_j^e - t_j^l),$$

t_j^e denotes the point where f takes its j-th local extreme value and t_j^l the left endpoint of the j-th local extreme interval of f_n^C respectively.

From Theorem 5.1 we deduce that for C and n sufficiently large f_n^C has the correct modality and

$$\sup_{s \le t \le 2\tau_n} |E_n(s) - E_n(t)| \le \frac{1}{C} \tag{10.27}$$

with arbitrarily high probability,

We consider the case of the first extremum only since the general case is proved analogously, but with some more technical notation. We can also assume without any restriction that $t_1^l < t_1^e$ as if $t_1^r > t_1^e$ a similar argument holds. Suppose F_n^C is initially convex. Then F_n^C is the largest convex minorant of $F_n + C/\sqrt{n}$ (Barlow et al, 1972) until it reaches the left endpoint $t_1^l(n, C)$ of $I_1^e(n, C) = [t_1^l(n, C), t_1^r(n, C)]$.

For some constant $\delta > 0$ such that for each C and sufficiently large n

$$t_1^r - t_1^e = \text{argmax}_{0 \le h \le \delta} H(h)$$

where

$$H(h) = \frac{F_n(t_1^l + h) - F_n(t_1^l) - \frac{2C}{\sqrt{n}}}{h}. \tag{10.28}$$

As F is convex on $[t_1^l, t_1^e]$. Lemma 10.1 implies that

$$G(h) = \frac{F(t_1^e + h) - F(t_1^e)}{h} \tag{10.29}$$

defines a strictly increasing function on $[0, \frac{4}{3}\mu]$ where $\mu = t_1^e - t_1^l$. Furthermore, for all $\tau < \mu$

$$
\begin{aligned}
H\left(\frac{4}{3}\mu\right) - H(\tau) &\geq G\left(\frac{4}{3}\mu\right) - G(\tau) + \frac{2C}{\sqrt{n}\tau} - \frac{2C}{\sqrt{n}\frac{4}{3}\mu} - \frac{2}{C\sqrt{n}\tau} \\
&\geq \frac{2C}{\sqrt{n}} \cdot \left(\frac{1}{\tau} - \frac{3}{4\mu} - \frac{1}{C^2\tau}\right) \qquad (10.30) \\
&> 0.
\end{aligned}
$$

This shows that H cannot attain its maximum on $[0, \mu]$ and consequently $t_1^r > t_1^e$.

Proof of (b):
We suppose that S_n has a local maximum on

$$I_1^e(n, C) = [t_1^l(n, C), t_1^r(n, C)],$$

that $t_1^e \in I_1^e$ and that (10.27) is satisfied. Define G by

$$H(h) = \frac{F(t_1^l + h) - F(t_1^l) - \frac{2C}{\sqrt{n}}}{h}.$$

and consider

$$h_0 = \text{argmax}_{0 \leq h \leq \delta} G(h).$$

Then $G'(h_0) = 0$ implies

$$f(t_1^l + h_0)h_0 = F(t_1^l + h_0) - F(t_1^l) - \frac{2C}{\sqrt{n}}.$$

Using Taylor expansions in t_1^e and the fact that $f'(t_1^e) = 0$ we obtain

$$3h_0(t_1^l + h_0 - t_1^e)^2 = (t_1^l + h_0 - t_1^e)^3 + (t_1^e - t_1^l)^3 - \frac{12C}{\sqrt{n}f''(t_1^e)} + o(h_0^3).$$

Thus

$$h_0(3(t_1^l - t_1^e)h_0 + 2h_0^2) = -\frac{-12C}{\sqrt{n}f''(t_1^e)} + o(h_0^3)$$

and finally

$$h_0^3 \geq -\frac{6C}{\sqrt{n}f''(t_1^e)} + o(h_0^3).$$

In the other direction we consider

$$h_1 = \text{argmax}_{0 \leq h \leq \delta} \frac{F(t_1^e + h) - F(t_1^e) - \frac{2C}{\sqrt{n}}}{h} \qquad (10.31)$$

and

$$h_2 = \text{argmin}_{0 \leq h \leq \delta} \frac{F(t_1^e - h) - F(t_1^e) - \frac{2C}{\sqrt{n}}}{h}$$

It is not difficult to see that $h_0 \leq h_1 + h_2$. Setting the derivative of the right-hand side of (10.31) to zero and using a Taylor expansion in t_1^e yields

$$\frac{1}{2}h_1^3 f''(t_1^e) = \frac{1}{6}h_1^3 f''(t_1^e) - \frac{2C}{\sqrt{n}} + o(h_1^3).$$

Thus

$$h_1^3 = -\frac{6C}{\sqrt{n} f''(t_1^e)} + o(h_1^3).$$

The same argument holds for h_2 as well and both together show that

$$h_0^3 \leq -\frac{12C}{\sqrt{n} f''(t_1^e)} + o(h_0^3).$$

Define H as in (10.28) and consider

$$\tilde{G}(h) = G(h) - \frac{2}{\sqrt{C \cdot n h}}$$

and

$$\tilde{h}_0 = \text{argmax} \tilde{G}(h).$$

The considerations above show that

$$\left(-\frac{6(C + \frac{1}{\sqrt{C}})}{\sqrt{n} f''(t_1^e)} \right)^{\frac{1}{3}} \leq \tilde{h}_0 (1 + o(1)) \leq \left(-\frac{12(C + \frac{1}{\sqrt{C}})}{\sqrt{n} f''(t_1^e)} \right)^{\frac{1}{3}}.$$

Furthermore for all $h > (1 + \frac{1}{\sqrt{C}})\tilde{h}_0$

$$\begin{aligned} H(\tilde{h}_0) - H(h) &\geq G(\tilde{h}_0) - G(h) - \frac{2}{C\sqrt{n h}} \\ &= \tilde{G}(\tilde{h}_0) - \tilde{G}(h) + \frac{2}{\sqrt{C \cdot n \tilde{h}_0}} - \frac{2}{\sqrt{C \cdot n h}} - \frac{2}{C\sqrt{n h}} \\ &> 0. \end{aligned} \tag{10.32}$$

Consequently, H cannot attain its maximum in $h > \tilde{h}_0(1 + \frac{1}{\sqrt{C}})$ and hence

$$\text{argmax}_{0<h<\delta} H(h) < (1 + \frac{1}{\sqrt{C}}) \cdot \left(-\frac{12(C + \frac{1}{\sqrt{C}})}{\sqrt{n} f''(t_1^e)} \right)^{\frac{1}{3}}.$$

Similarly it can be shown that

$$\text{argmax}_{0<h<\delta} H(h) < (1 - \frac{1}{1 + \sqrt{C}}) \cdot \left(-\frac{6(C - \frac{1}{\sqrt{C}})}{\sqrt{n} f''(t_1^e)} \right)^{\frac{1}{3}}.$$

Proof of (c):
The proof relies on the modulus of continuity of the empirical process E_n.

Lemma 10.2. *Let $Y(n,C)$ denote random variables such that for all $\varepsilon > 0$*

$$\lim_{C\to\infty}\lim_{n\to\infty}\mathbb{P}\left(|Y(n,C)| < \varepsilon\right) = 1.$$

Consider

$$\alpha_n = \frac{1}{n^\gamma}$$

for some $\gamma < 1$ and

$$\beta_n^C = \max\left\{\frac{1}{\log(n)}, Y(n,C)\}\right\}$$

Then for all $B > 2$ we have

$$\lim_{C\to\infty}\lim_{n\to\infty}\mathbb{P}(\max_{\alpha_n<|s-t|<\beta_n^C}\frac{|E_n(s)-E_n(t)|}{\sqrt{|t-s|\cdot\log(\frac{1}{|t-s|})}} > B) = 0.$$

Proof: Define random integer-valued variables K_n by

$$K_n = \lfloor\log_2\left(\frac{\beta_n^C}{\alpha_n}\right)\rfloor$$

Simple considerations show that

$$\begin{aligned}
&\mathbb{P}\Big(\max_{\alpha_n<|s-t|<\beta_n^C}\frac{|E_n(s)-E_n(t)|}{\sqrt{|t-s|\cdot\log(\frac{1}{|t-s|})}} > B\Big)\\
&\quad\le\sum_{k=0}^{\infty}\mathbb{P}\Big(|E_n(s)-E_n(t)| > B\cdot\sqrt{|t-s|\cdot\log(\frac{1}{|t-s|})}\text{ for some } s,t \text{ with}\\
&\qquad\qquad 2^k\alpha_n < |s-t| < 2^{k+1}\alpha_n\ \Big|\ k\le K_n\Big)\\
&\quad\le\sum_{k=0}^{\infty}\mathbb{P}\Big(|\omega_n(2^{k+1}\alpha_n) > B\cdot\sqrt{2^k\alpha_n\cdot\log(\frac{1}{2^k\alpha_n})}\ \Big|\ k\le K_n\Big)
\end{aligned}$$

where $\omega_n(a) = \sup_{|s-t|<a}|E_n(s)-E_n(t)|$. Denote $2^{k+1}\alpha_n$ by a_k and

$$\lambda_k = B\cdot\sqrt{\frac{\log(\frac{1}{\alpha_n})}{2}}.$$

Using a result of Mason, Shorack and Wellner (1983) we conclude that

$$\begin{aligned}
&\mathbb{P}\Big(\omega_n(a_k) > \sqrt{a_k}\cdot\lambda_k\ \Big|\ k\le K_n\Big)\\
&\qquad\le\frac{20}{a_k\cdot(\beta_n^C)^3}\cdot\exp\left(-(1-\beta_n^C)^4\frac{\lambda_k^2}{a}\psi\left(\frac{\lambda_k}{\sqrt{na_k}}\right)\right)
\end{aligned}\tag{10.33}$$

where

$$\psi(x) = 2\cdot\frac{(1+x)(\log(1+x)-1)+1}{x^2}.$$

provided

$$\beta_n^C < \frac{1}{2}. \tag{10.34}$$

As

$$\frac{\lambda_k}{\sqrt{na_k}} = \frac{B \cdot \sqrt{\log(\frac{1}{\alpha_n})}}{\sqrt{2 \cdot n2^{k+1} \cdot \alpha_n}} \leq \frac{B \cdot \sqrt{\log(\frac{1}{\alpha_n})}}{\sqrt{n \cdot \alpha_n}} = \tilde{C} \cdot \frac{\sqrt{\log(n)}}{n^{1/3}}$$

it follows that $\psi(\frac{\lambda_k}{\sqrt{na_k}}) \to 1$.

As ψ is strictly decreasing for $x \geq 1$,

$$\lim_{C,n\to\infty} \mathbb{P}((1-\beta_n^C)^4 \cdot \psi\left(\frac{\lambda_k}{\sqrt{na_k}}\right) > \frac{2}{B}) = 1 \tag{10.35}$$

This implies

$$\begin{aligned}
\mathbb{P}\Big(\omega_n(a_k) > \sqrt{a_k} \cdot \lambda_k \;\Big|\; k \leq K_n\Big) &\leq \frac{20}{a_k(\beta_n^C)^3} \exp\left(-\frac{2}{B} \cdot \frac{B^2 \cdot \log(\frac{1}{\alpha_n})}{4}\right) \\
&= \frac{20}{2^{k+1}\alpha_n(\beta_n^C)^3}\alpha_n^{B/2} \\
&= \frac{10\log(n)^3}{2^k n^{\gamma(B/2-1)}}
\end{aligned}$$

provided that the inequalities (10.34) and (10.35) are also satisfied.

Putting this together we deduce that

$$\begin{aligned}
&\mathbb{P}\Big(|E_n(s) - E_n(t)| > B \cdot \sqrt{|t-s| \cdot \log(\frac{1}{|t-s|})} \text{ for some } s,t \text{ with } \alpha_n < |s-t| < \beta_n^C\Big) \\
&\qquad \leq \sum_{k=0}^{\infty} \frac{1}{2^k} \cdot \frac{10\log(n)^3}{2^k n^{\gamma(B/2-1)}} \\
&\qquad < \frac{20\log(n)^3}{2^k n^{\gamma(B/2-1)}}
\end{aligned}$$

This completes the proof of the Lemma. □

We proceed now with the proof of (c). Since is f is twice continuously differentiable, there is some constant $D > 0$ such that

$$|F(x+h) - F(x) - hf(x) - \frac{1}{2}h^2 f'(x)| \leq Dh^3$$

for all x and h.

Let B be an arbitrary constant greater than 2 and

$$d(n,C) = \min\{|f'(x)| \; | x \in [0,1] \setminus \cup_i I_i^e(n,C)\}.$$

Define a random sequence $h(n,C)$ by

$$h(n,C) = \frac{(8B)^{2/3}\log(d(n,C)^2 n)^{1/3}}{(3n)^{1/3} d(n,C)^{2/3}}.$$

We consider the situation where

- f_n^C attains the correct modality
- $t_i^e \in I_i^e(n, C)$ for all i.
- The empirical process satisfies

$$\sup_{|s-t|<Y(n,C)} |E_n(t) - E_n(s)| < B \cdot \sqrt{|s-t| \cdot \log(1/|s-t|)}$$

  where $Y(n, C)$ is defined by

$$Y(n,C) = \max\{x_{j+1} - x_j \mid x_j, x_{j+1}\text{knots, } [x_j, x_{j+1}] \neq I_i^e(n,C) \text{ for all } i\}.$$

- For all $x \in [0,1] \setminus \cup_i I_i^e(n, C)$

$$h_n \leq \frac{f'(x)}{32D}$$

  holds.
- For each extreme interval $I_i^e(n, C)$, the distances of each endpoint to t_1^e are both smaller than $4h_n$.

The preceding lemmas and parts of this theorem show that the probability that all these assumptions are satisfied simultaneously converges to 1 as n and C tend to ∞. For example, (10) follows from (b) which provides a constant $A > 0$ such that $|f'(x)| \geq A \cdot n^{-1/6}$.

Consider now an arbitrary point $t_1 \in [0,1] \setminus \cup_i I_i^e(n, C)$ where $f'(t_1) > 0$. Then

$$\frac{F_n(t_1 + h_n) - F_n(t_1)}{h_n} \leq f(t_1) + \frac{1}{2} h_n f'(t_1) + D h_n^2 + \frac{B\sqrt{\log(1/h_n)}}{\sqrt{n h_n}}.$$

Plugging in the expression for h_n and using the assumptions made above we see that

$$\frac{F_n(t_1 + h_n) - F_n(t_1)}{h_n} \leq f(t_1) + \frac{1}{2} h_n f'(t_1)(1 + \frac{1}{4} + \frac{1}{4}).$$

Similarly, we conclude that for all $h \in [4h_n, t_j^e]$

$$\frac{F_n(t_1 + h) - F_n(t_1)}{h} \geq f(t_1) + \frac{1}{2} h f'(t_1)(1 - \frac{1}{4} - \frac{1}{4})$$

where t_j^e is the smallest local extreme value greater than t_1.

Suppose that there are knots x_j and x_{j+1} that do not embrace a local extreme interval such that $h_0 = x_{j+1} - x_j > 4h_n$ and such that f is increasing on $[x_j, x_{j+1}]$. The width $\tilde{h}$ is the local argmin

$$\tilde{h} = \text{argmin}_{0<h<\delta} \frac{F_n(x_1 + h) - F_n(x_1)}{h}.$$

On the other hand the considerations above show that

$$\frac{F_n(x_1 + h_n) - F_n(x_1)}{h_n} < \frac{F(x_1 + h) - F_n(x_1)}{h}.$$

Therefore, the distance between two knots that do not embrace an extreme interval is bounded by $4h_n$.

Proof of (d):

We assume that all the assumptions made in the proof of (c) are again satisfied and that each two extreme intervals I_i^e and I_{i+1}^e are separated by at least two additional knots x_j and x_{j+1}:

$$\max I_i^e < x_j < x_{j+1} < \min I_{i+1}^e.$$

Define h_n as in (10.31). Consider a knot x_i which does not delimit a local extreme interval I_i^e. We take f to be increasing in x_i. Then the proof of (c) shows that

$$f_n^C(x_i) \le \frac{F_n(x_i + h_n - F_n(x_i)}{h_n} \le f(x_i) + C_1|f'(x_i)|^{1/3}\left(\frac{\log(n)}{n}\right)^{1/3}.$$

Similar arguments show that

$$f_n^C(x_i) \ge \frac{F_n(x_i) - F_n(x_i - h_n}{h_n} \ge f(x_i) - C_1|f'(x_i)|^{1/3}\left(\frac{\log(n)}{n}\right)^{1/3}.$$

Analogous inequalities can be derived in the case where f is decreasing in x_i.

Suppose now that t is an arbitrary point in

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

Let x_i be the nearest knot which does not delimit a local extreme interval. Then

$$|f(t) - f_n^C(t)| \le |f(t) - f(x_i)| + |f(x_i) - f_n^C(x_i)| + |f_n^C(x_i) - f_n^C(x)|. \tag{10.36}$$

The inequalities above show that the second term is bounded by

$$C_2|f'(x_i)|^{1/3}\left(\frac{\log(n)}{n}\right)^{1/3}.$$

The first term is bounded by

$$C_3 \cdot |t - x_i| \cdot |f'(x_i)| \le C_3 \cdot |f'(x_i)|^{1/3}\left(\frac{\log(n)}{n}\right)^{1/3}.$$

This follows from (b).

Depending on the exact definition of $f_n^C(x)$ at knot points the third term is either 0 or bounded by $2C_1|f'(x_i)|^{1/3}\left(\frac{\log(n)}{n}\right)^{1/3}$.

This completes the proof of (d).

Proof of (e):

As in the other cases we assume that f_n^C attains the correct modality and that $t_i^e \in I_i^e(n, C)$ for each extreme point t_i^e. We also assume that for each extreme interval I_i^e

$$(1 - \frac{1}{1+\sqrt{C}}) \cdot \left(-\frac{6(C - \frac{1}{\sqrt{C}}}{\sqrt{n}f''(t_1^e)}\right)^{\frac{1}{3}} \le |I_i^e(n, C)| \le (1 + \frac{1}{\sqrt{C}}) \cdot \left(-\frac{12(C + \frac{1}{\sqrt{C}}}{\sqrt{n}f''(t_1^e)}\right)^{\frac{1}{3}}.$$

The regression function f_n^C takes in t_i^e the slope of the taut string in the extreme interval $I_i^e = [x_1, x_2]$. Taylor expansions in t_i^e using $f'(t_i^e) = 0$ show that

$$\begin{aligned}
&\frac{F(x_2)-F(x_1)}{x_2-x_1} - f(t_i^e) \\
&\qquad\qquad \leq \frac{1}{6} f''(t_i^e) \frac{(x_2-t_i^e)^3-(x_1-t_i^e)^3}{x_2-x_1} + O\left(\frac{(x_2-t_i^e)^4-(x_1-t_i^e)^4}{x_2-x_1}\right) \\
&\qquad\qquad \leq \frac{1}{6} f''(t_i^e)(x_2-x_1)^2 + O\left((x_2-x_1)^3\right)
\end{aligned}$$

and, similarly,

$$\frac{F(x_2)-F(x_1)}{x_2-x_1} - f(t_i^e) \geq \frac{1}{24} f''(t_i^e)(x_2-x_1)^2 + O\left((x_2-x_1)^3\right). \tag{10.37}$$

Another application of the modulus of continuity for the empirical process E_n as formulated in Lemma 10.2 yields

$$\begin{aligned}
f_n^C(t_i^e) &= \frac{F(x_2)-F(x_1)}{x_2-x_1} + \frac{E_n(x_2)-E_n(x_1)}{\sqrt{n}(x_2-x_1)} \\
&\leq f(t_i^e) + \frac{1}{6} f''(t_i^e)(x_2-x_1)^2 \cdot (1+o(1)) + B\frac{\sqrt{\log(\frac{1}{x_2-x_1})}}{\sqrt{n\cdot(x_2-x_1)}} \\
&\leq f(t_i^e) + C_1 \cdot \frac{f''(t_i^e)^{1/3}}{n^{1/3}}(1+o(1))
\end{aligned}$$

Applying the same arguments to (10.37) we conclude that

$$|f_n^C(t_i^e) - f(t_i^e)| \leq D_1 \cdot (1+o(1)) \cdot \frac{f''(t_i^e)^{1/3}}{n^{1/3}}.$$

The proof is now completed by extending the bound to arbitrary points in extreme intervals I_i^e. This is done in the usual way as in as in (10.36) using a Taylor expansion in t_i^e and shows that

$$|f(t) - f(t_i^e)| \leq D_2 |I_i^e|^2 f''(t_i^e).$$