Electronic Journal of Statistics
Vol. 3 (2009) 747–780
ISSN: 1935-7524
DOI: 10.1214/08-EJS298

Quantifying the cost of simultaneous non-parametric approximation of several samples

P.L. Davies*

University of Duisburg-Essen
Technical University Eindhoven
e-mail: laurie.davies@uni-due.de

A. Kovac*

University of Bristol
e-mail: A.Kovac@bristol.ac.uk

Abstract: We consider the standard non-parametric regression model with Gaussian errors but where the data consist of different samples. The question to be answered is whether the samples can be adequately represented by the same regression function. To do this we define for each sample a universal, honest and non-asymptotic confidence region for the regression function. Any subset of the samples can be represented by the same function if and only if the intersection of the corresponding confidence regions is non-empty. If the empirical supports of the samples are disjoint then the intersection of the confidence regions is always non–empty and a negative answer can only be obtained by placing shape or quantitative smoothness conditions on the joint approximation, or by making additional assumptions about the support points. Alternatively, a simplest joint approximation function can be calculated which gives a measure of the cost of the joint approximation, for example, the number of extra peaks required.

AMS 2000 subject classifications: Primary 62G08; secondary 62G15, 62P35, 82D25.
Keywords and phrases: Modality, non-parametric regression, penalization, regularization, total variation.

Received September 2008.

1. Introduction

1.1. The problem

We consider the following problem in non-parametric regression: given k samples

$$\boldsymbol{y}_{in_i} = \{(t_{ij}, y_{ij}) : j = 1, \dots, n_i\},\ i = 1, \dots, k, \tag{1}$$

*Research supported in part by Sonderforschungsbereich 475, Fakultät Statistik, Technische Universität Dortmund.

with empirical supports

$$S_{in_i} = \{t_{i1} < t_{i2} < \ldots < t_{in_i}\},\ i = 1, \ldots, k, \tag{2}$$

the question to be answered is whether they can be simultaneously represented by a common function f. The standard approach is to assume that the data were generated according to the model

$$Y_{in_i}(t) = f_i(t) + \sigma_i Z_i(t),\ \ i = 1, \ldots, k, \quad t \in [0, 1], \tag{3}$$

where the $Z_i, i = 1, \ldots, k$ are independent, standard Gaussian white noise processes and then to consider the null and alternative hypotheses

$$H_0 : f_1 = \ldots = f_k \quad H_1 : f_i \neq f_j \quad \text{for some } i, j. \tag{4}$$

Individual samples generated under (3) will be denoted by

$$\boldsymbol{Y}_{in_i} = \{(t_{ij}, Y_{ij}) : j = 1, \ldots, n_i\},\ i = 1, \ldots, k.$$

Here and in the following we use minuscule letters to denote general data sets and majuscule letters for data generated under (3). We shall mostly restrict attention to the case $k = 2$; the extension to more samples poses no problems.

Within this model it is possible to construct tests which are asymptotically consistent if $\lim n_i = \infty$, $i = 1, 2$, and which can detect alternatives converging to the null hypothesis at certain rates. This may be formalized by putting

$$f_1(t) - f_2(t) = f_{1,n_1}(t) - f_{2,n_2}(t) = \Delta_n(t), \quad n = \min(n_1, n_2) \tag{5}$$

where Δ_n is a difference function and measures the rate of convergence to the null hypothesis. An asymptotic approach requires some assumptions about the design points t_{ij}. Very often these are taken to be random variables T_{ij} with values in $[0, 1]$ and whose density has support $[0, 1]$. In this paper our use of the word ‘support’ refers always to the empirical support S_{in_i} and not to the support of some underlying density. With this in mind the best result seems to be that of Neumeyer and Dette (2003) who construct a test which can detect alternatives which converge to the null hypothesis at the optimal rate asymptotic rate $\Delta_n = \mathrm{O}(n^{-1/2})$. If the supports are equal, $S_{in_i} = \{t_1, \ldots, t_{n_i}\}$, $i = 1, 2$, then it is not difficult to construct such a test as the differences $Y_{1n_1}(t_j) - Y_{2n_2}(t_j)$ do not depend on f (see for example Delgado (1992) and Fan and Lin (1998)). The result of Neumeyer and Dette (2003) continues to hold even if the supports are disjoint, $S_{1n_1} \cap S_{2n_2} = \emptyset$. In this case, however, there are conceptual difficulties as we now show.

We consider firstly the case of exact data

$$y_{ij} = f_i(t_{ij}),\ \ t_{ij} \in S_{in_i}, \quad i = 1, 2,$$

with disjoint supports S_{1n_1} and S_{2n_2}. If we denote the supremum norm on $[0, 1]$ by $\|\cdot\|_\infty$ then the null and alternative hypotheses of (4) may be rewritten as

$$H_0 : \|f_1 - f_2\|_\infty = 0, \quad H_1 : \|f_1 - f_2\|_\infty > 0. \tag{6}$$

As the values of f_1 and f_2 are known only on the disjoint sets S_{1n_1} and S_{2n_2} respectively, it is not possible to decide between H_0 and H_1: there always exists a function f which agrees with f_1 on S_{1n_1} and with f_2 on S_{2n_2}. It does not help to impose qualitative smoothness assumptions such as infinite differentiability as it is always possible to interpolate the data points with such a function. In spite of this, all conditions imposed in the literature are of a qualitative form: Hall and Hart (1990), a bounded first derivative; Härdle and Marron (1990), Hölder continuity; King et al. (1990), at least uniform continuity; Kulasekera (1995), Kulasekera and Wang (1997), a continuous second derivative; Munk and Dette (1998), Hölder continuity of order $\beta > 1/2$; Dette and Neumeyer (2001), a continuous rth derivative: Lavergne (2001), a second derivative which is uniformly Lipschitz of order β, $0 \leq \beta < 1$; Neumeyer and Dette (2003), continuous derivatives of order $d \geq 2$. The problem cannot be solved by asymptotic considerations because the situation for $n = \infty$ can be completely different from that for any finite n. This is related to the following result in real analysis. It is possible to construct a sequence of infinitely differentiable functions f_n on $[0, 1]$ which converge pointwise to a function f_∞ which is discontinuous at every rational point: for finite n we have infinite differentiability, for $n = \infty$ we have a discontinuity at every rational point. One remedy is to place *quantitative* smoothness conditions on the functions f_n such as $\|f_n^{(1)}\|_\infty \leq 1$ for all n. In this case the limiting function f_∞ is at least continuous and it is now possible to distinguish between H_0 and H_1 for finite n. This continues to hold if noise is added. A possibly more acceptable alternative is to impose shape constraints such as monotonicity. If for example it is assumed that under H_0 the common function is monotone then it is possible to decide between H_0 and H_1 for finite n. There is a third possibility if it is reasonable to assume that the supports are interchangeable, for example that they are i.i.d random variables. This is discussed in Section 3.2.

In general there is an understandable reluctance to place a priori quantitative smoothness conditions unless these have some scientific justification. This applies also to shape constraints but to a lesser degree. We propose the following method. If the supports are disjoint then there will always be a common function consistent with the data. However if f_1 and f_2 differ then this common function will become increasingly complex as the sample sizes increase. Complexity can be measured by smoothness such as the value of $\|f^{(1)}\|_\infty$ or by shape by specifying the smallest number of local extreme values required to be consistent with the data. It is this increase in complexity which we call the cost of the simultaneous approximation.

1.2. An example

The top panel of Figure 1 shows two data sets of sizes $n_1 = n_2 = 500$ generated according to (3) with $f_1(t) = \exp(1.5t)$, $f_2(t) = \exp(1.5t) + 4$ and $\sigma_i = 0.25$, $i = 1, 2$. The support of the first sample is $j/500, j = 0, \ldots, 499$ and of the second sample $(2j + 1)/1000, j = 0, \ldots, 499$. The bottom panel shows a sample

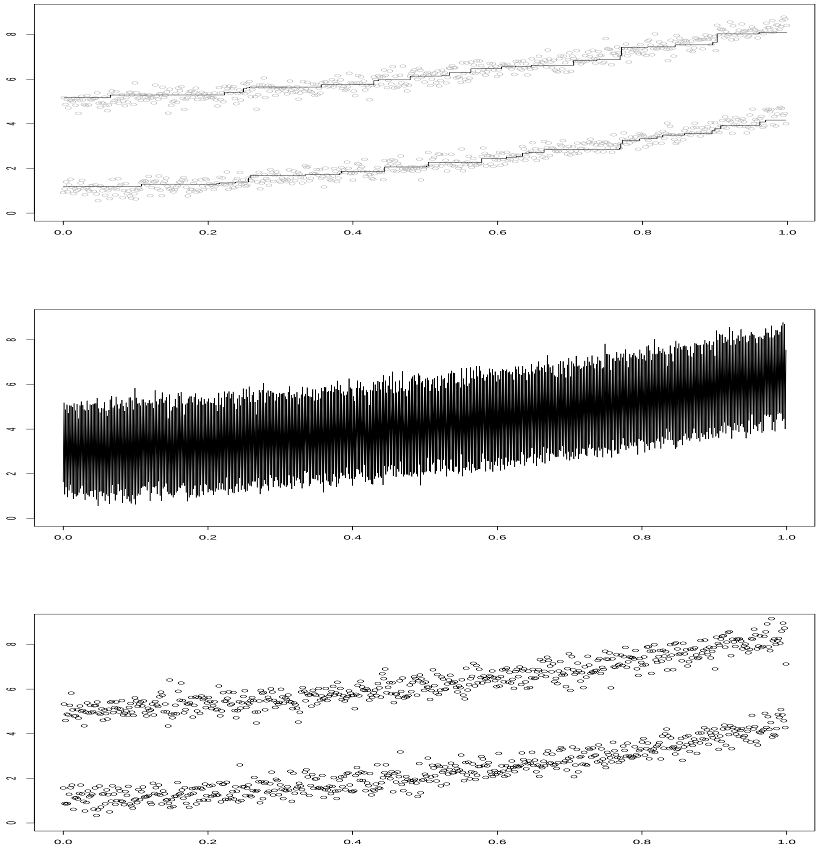

Fig 1. *The top panel shows two samples each of size 500 generated by* $Y_1(t) = \exp(1.5t) + 0.25Z(t)$ *and* $Y_2(t) = \exp(1.5t) + 4 + 0.25Z(t)$ *together with the approximating monotonic curves. The design points were* $j/500, j = 0, \ldots, 499$ *for the first sample and* $(2j+1)/1000, j = 0, \ldots, 499$ *for the second sample. The centre panel shows a joint approximating function with 998 local extreme values. The bottom panel shows a sample of size* $n = 1000$ *generated using the function of the centre panel.*

of size $n = 1000$ generated using the function of the centre panel. It is similar to the data generated under f_1 and f_2. The function shown in the centre panel is piecewise constant. However it can be made infinitely differentiable with very little change in its values by convoluting it with a Gaussian kernel with a very small bandwidth. Such a function would also produce the sample shown in the bottom panel, that is, it is consistent with both individual samples. Furthermore as an infinitely differentiable function it satisfies all the qualitative smoothness conditions to be found in the literature. If however one imposes the quantitative smoothness condition $\|f^{(1)}\|_\infty \leq 1$ then it is clear that no such function can

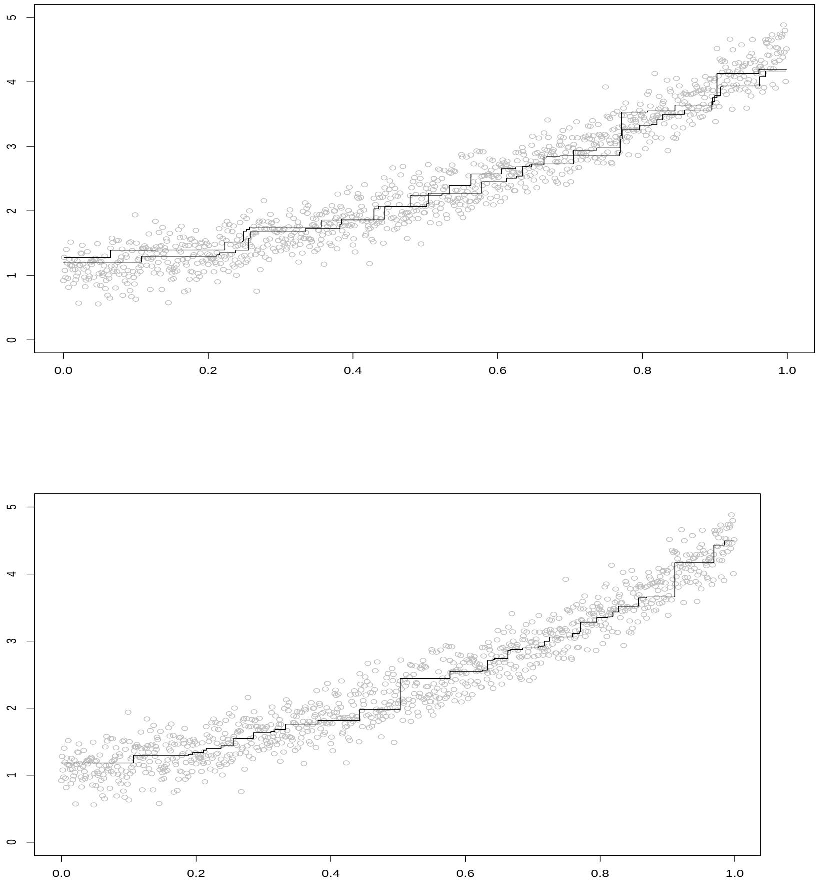

Fig 2. *The upper panel shows the data of Figure 1 but with the values of* Y_2 *now given by* $Y_2(t) = \exp(1.5t) + 0.103 + 0.25Z(t)$. *There is now a joint monotonic approximating function which is shown in the lower panel.*

approximate both data sets simultaneously. Another way of looking at the function is in terms of shape. The joint piecewise constant approximating function has 998 local extreme values and this is the minimum number which is consistent. As the individual curves are both monotone the extra 998 local extreme values can be seen as the cost of the joint approximation. Alternatively one can place a shape constraint such as monotonicity on the joint function. In this case there does not exists a joint approximating function which satisfies the shape constraint of monotonicity.

Using exactly the same noise we now move the two data sets closer together by putting $f_2(t) = \exp(1.5t) + 0.103$. This is shown in the top panel of Figure 2

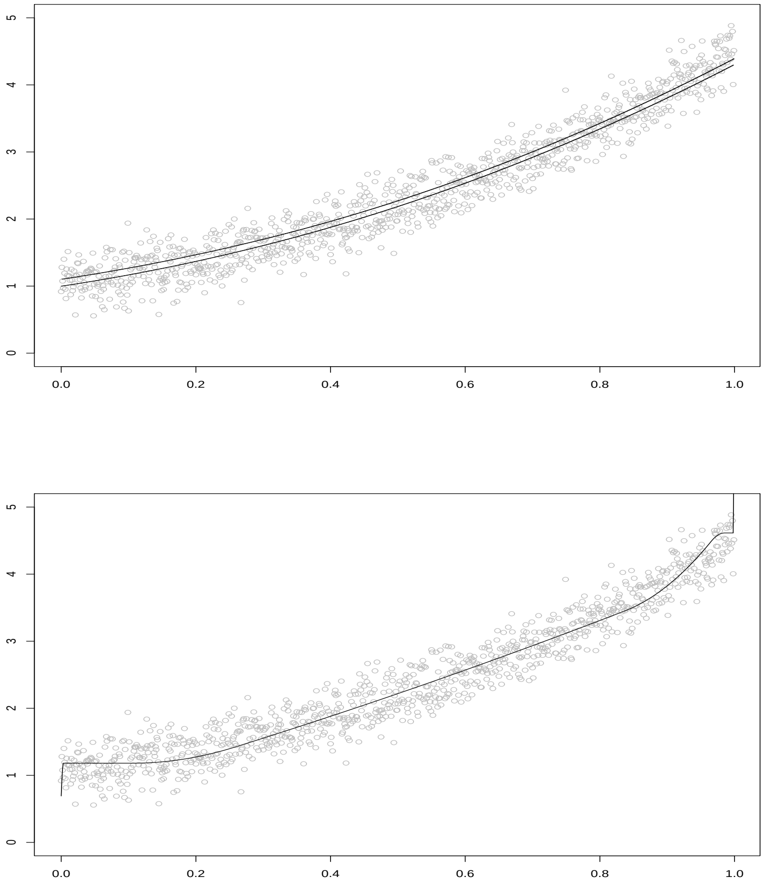

Fig 3. *The upper panel shows the results of minimizing the total variation of the second derivatives for the data of the upper panel of Figure 2 subject to monotonicity. The lower panel shows the corresponding result for the lower panel of Figure 2.*

together with the approximating functions. There now does exist a monotone joint approximation which is shown in the lower panel of Figure 2. The cost of the joint approximation in terms of the number of local extreme values is now zero. The cost in terms of smoothness as measured by the total variation of the second derivative $TV(f^{(2)})$ is however very high. The top panel of Figure 3 shows the result of minimizing the total variation of the second derivative subject to the monotonicity of the function for the two samples separately. The values of the total variation of the second derivative are 0.36 and 0.70 for the first and second samples respectively. These may be compared with a total variation of the second derivative of $\exp(1.5t)$ which is 7.83. The second derivatives are shown in the

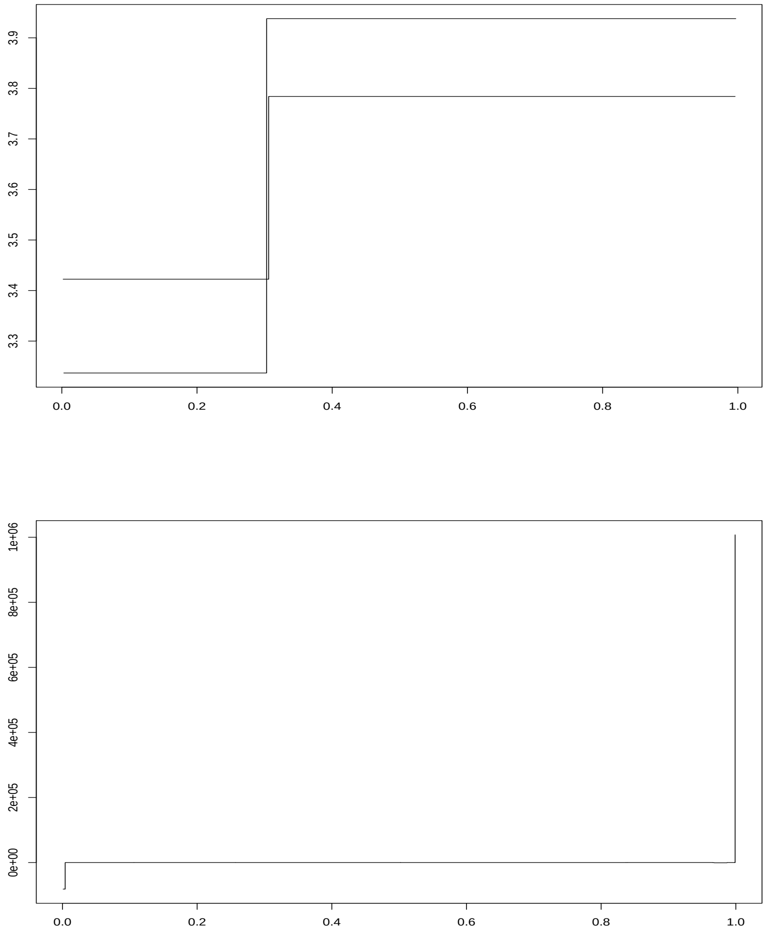

Fig 4. *The upper panel shows the second derivative of the functions in the upper panel of Figure 3: the lower panel shows the corresponding result for the lower panel of Figure 3.*

top panel of Figure 4. The smoothest joint approximation is shown shown in the lower panel of Figure 3 and its second derivative in the lower panel of Figure 4. It is clear that the cost in terms of smoothness is high and indeed the value of the total variation of the second derivative for the joint approximation is 1090530. The reason is that for these two data sets a monotone joint approximation is just possible. If we move the samples slightly closer close together by putting $f_2(t) = \exp(1.5t) + 0.1$ then the joint approximation becomes much smoother. It is shown in the upper panel of Fig 5 and its second derivative in the lower panel. The total variation of the second derivative is now 12.42. We note that

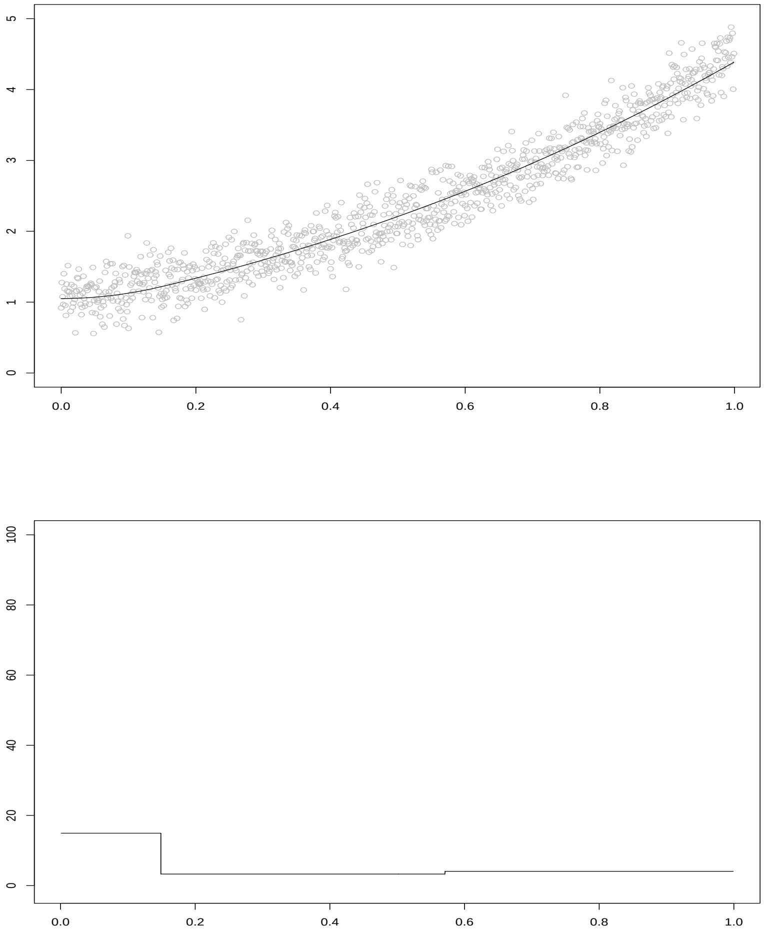

Fig 5. *The upper panel shows the smoothest joint approximating function when the two samples are moved closer together with* $f_2(t) = \exp(1.5t)+0.1$. *The bottom panel shows the second derivative of the approximating function.*

we place numerical values on all our measures of complexity, the number of local extreme values, the total variation of the second derivative.

1.3. Approximation and regularization

Our approach is as follows:

(1) Firstly, for each sample $\boldsymbol{y}_{in_i}$ we define a so called approximation region

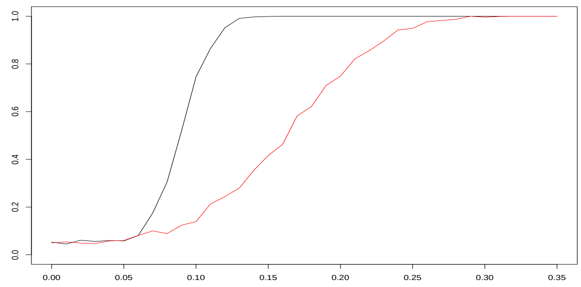

Fig 6. *The probability of rejecting the equality of functions as a function of δ for the shape constrained method (black) and the rank method (red).*

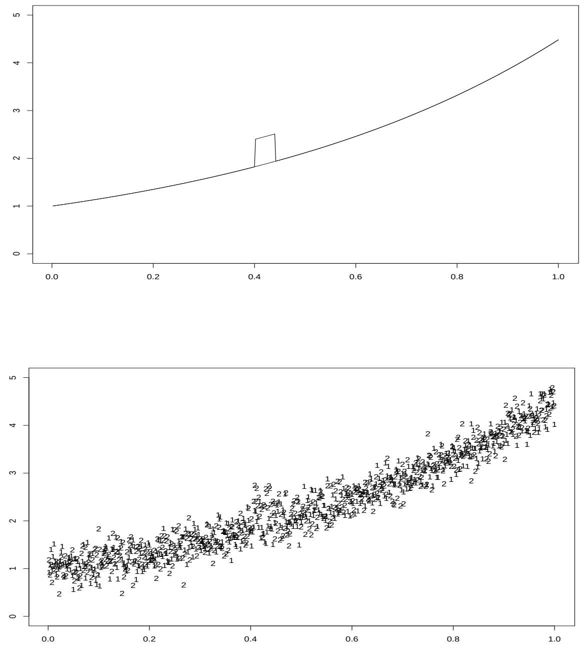

Fig 7. *The upper panel shows the function $f_1(t) = \exp(1.5t)$ and the function f_2 which is equal to f_1 apart from the interval $[0.402, 0.440]$ where $f_2(t) = f_1(t) + 0.575$. The lower panel shows the two data sets $Y_1(t_j) = f_1(t_j) + 0.25Z_1(t_j)$ and $Y_2(t_j) = f_2(t_j) + 0.25Z_2(t_j)$ for $j = 1, \ldots, 500$ and with $t_j = j/500$.*

$\mathcal{A}_{in_i}$ which specifies those functions f_i for which the model (3) is an adequate approximation for the sample. The intersection of the approximation regions $\mathcal{A}_{1,n_1} \cap \mathcal{A}_{2,n_2}$ contains all those functions which simultaneously approximate both samples. It is also the approximation region for the simultaneous approximation. A similar idea in the context of the one-way table in the analysis of variance is expounded in Davies (2004).

(2) Secondly, using some measure of complexity we regularize within each approximation region by choosing the simplest function which is consistent with the data. This is in the spirit of Donoho (1988) who pointed out that in non-parametric regression and density problems it is possible only to give lower bounds on certain quantities of interest such the number of local extremes.

We give an example which may clarify our purpose. The top panel of Figure 8 shows a sample of size $n = 1000$ generated by

$$Y(t) = \Phi(t) + 0.25Z(t)$$

where Φ denotes the standard normal distribution function. The grid used was $t_j = -49.95 + j/10, j = 1, \ldots, 1000$. The second panel shows a kernel estimator using a global bandwidth $h = 2$. The resulting estimate is smooth but lies so far away from the data that it does not belong to the approximation region. The third panel shows the estimator derived by choosing the largest bandwidth consistent with the estimator lying in the approximation region. The function is close to the data but exhibits superfluous wiggles. The bottom panel shows a function obtained by regularizing within the approximation region first by shape and then for smoothness as measured by the total variation of the second derivative. The function has the correct shape and is smooth.

Regularization plays an important and often unrecognized role in statistics. Even the simple location problem requires regularization. Suppose we are given a sample of independently and identically distributed random variables $X_1, \ldots, X_n$ with distribution function $F(\cdot - \mu)$. We are told that F is symmetric, continuous and has variance 1 but no more. We require an estimate for the unknown μ and a 95% confidence interval. Which F do we choose and why? Firstly we restrict the choice to all those F in

$$\mathcal{F}_n = \{F : d_{ko}(F, F_n) \leq 1.52/\sqrt{n}\,\}$$

where d_{ko} denotes the Kolmogorov metric and F_n the empirical distribution function of the data. This defines a 0.99-approximation region for the unknown F and corresponds to the $\mathcal{A}_n$ above. We regularize within $\mathcal{F}_n$ and choose the simplest F which is symmetric and has variance 1. On the basis of TINSTAAFL, there is no such thing as a free lunch, at least in statistics, (Tukey (1993)) the simplest F is the least favourable one, that is, it minimizes the Fisher information within $\mathcal{F}_n$. If the data are so close to normal that $\Phi \in \mathcal{F}_N$ then the result is Φ itself. In other words, the standard choice of the normal distribution for the location problem is an act of regularization. We refer to Davies (2008) for a more detailed discussion.

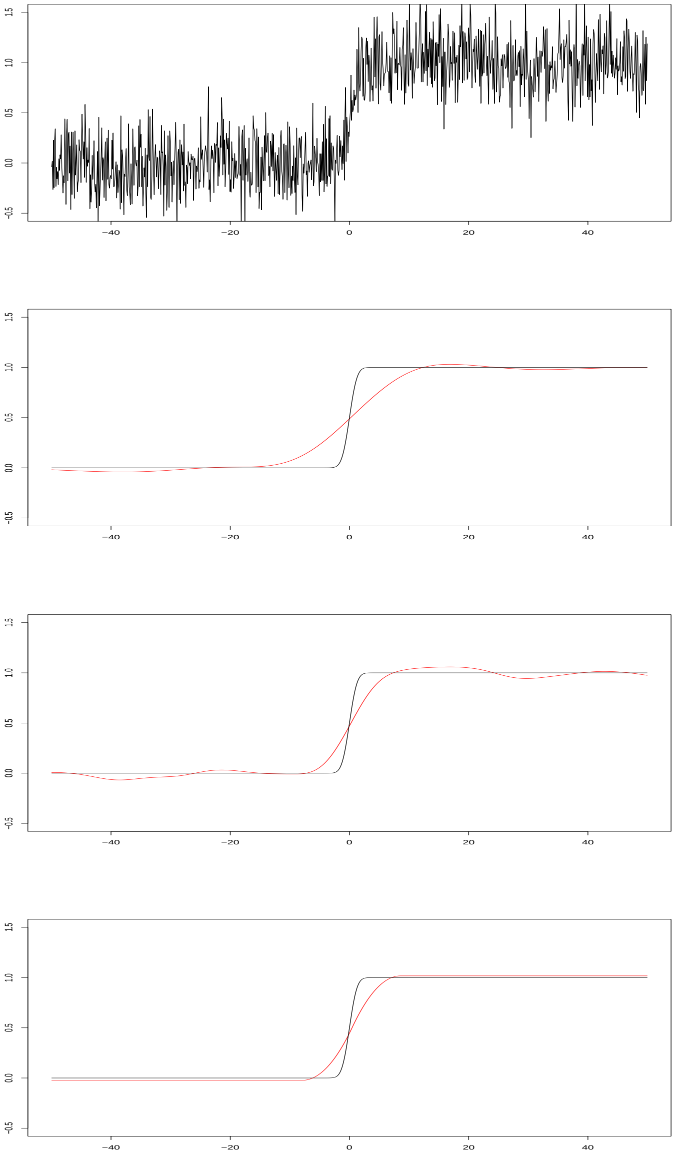

Fig 8. *Top panel: sample of size* $n = 1000$ *generated according to* $Y(t) = \Phi(t) + 0.25Z(t)$. *Second panel: a kernel estimate with a global bandwidth not in the approximation region. Third panel: a kernel estimate with the largest global bandwidth so that it is in the approximation region. Bottom panel: a function in the approximation region regularized for shape and smoothness.*

1.4. Contents

In Section 2 we define the approximation or confidence regions $\mathcal{A}_{in_i}$. Section 3 is devoted to the case of disjoint supports and Section 4 to the case of intersecting supports. A short comparison with other methods is given in Section 5 and the problem of heteroskedastic noise is discussed in Section 6. The main shape regularization we consider is that of minimizing the number of local extreme values. The algorithm we use is the taut string algorithm of Davies and Kovac (2001). In Section 7 we describe how it can be adapted to the case of several samples with differing noise levels. We end with a short discussion of the results in Section 8.

2. Approximation regions

2.1. Single samples

The following is based on Davies et al. (2009). We consider a single sample of data $\boldsymbol{Y}_n = (t_i, Y(t_i))_1^n$ generated under the model

$$Y(t) = f(t) + \sigma Z(t) \tag{7}$$

where we take the t_i to be ordered. Based on this model we consider two different approximation or confidence regions $\mathcal{A}_n$ and $\mathcal{A}_n^*$ defined as follows. For any function g and any interval $I \subset [0,\,1]$ we put

$$w(g, \boldsymbol{Y}_n, I) = \frac{1}{\sqrt{|I|}} \sum_{t_i \in I} (Y(t_i) - g(t_i)) \tag{8}$$

where $|I|$ denotes the number of points $t_i \in I$. The confidence region $\mathcal{A}_n$ is defined by

$$\mathcal{A}_n(\boldsymbol{Y}_n, \mathcal{I}_n, \sigma, \tau_n) = \{g : \max_{I \in \mathcal{I}_n} |w(g, \boldsymbol{Y}_n, I)| \; \leq \sigma\sqrt{\tau_n \log(n)}\;\}. \tag{9}$$

where $\mathcal{I}_n$ is a collection of intervals of $[0,\,1]$. We restrict attention to the cases where $\mathcal{I}_n$ is either the family of all intervals or a family of intervals of the form

$$\mathcal{I}_n(\lambda) = \big\{[t_{l(j,k)},\, t_{u(j,k)}] \,:\, l(j,k) = \lfloor (j-1)\lambda^k + 1 \rfloor, u(j,k) = \min(\lfloor j\lambda^k \rfloor, n), \\ j = 1, \ldots, \lceil n\lambda^{-k} \rceil, k = 1, \ldots, \lceil \log n / \log \lambda \rceil \big\} \tag{10}$$

for some $\lambda > 1$. Our default choice is the (wavelet) dyadic scheme $\mathcal{I}_n(2)$. For any given α and collection of intervals $\mathcal{I}_n$ we define $\tau_n(\alpha)$ by

$$\boldsymbol{P}\left(\max_{I \in \mathcal{I}_n} \frac{1}{\sqrt{|I|}} \Big| \sum_{i \in I} Z(t_i) \Big| \leq \sqrt{\tau_n(\alpha) \log n} \right) = \alpha. \tag{11}$$

One immediate consequence is

$$\boldsymbol{P}(f \in \mathcal{A}_n(\boldsymbol{Y}_n, \mathcal{I}_n, \sigma, \tau_n(\alpha))) = \alpha \tag{12}$$

so that $\mathcal{A}_n$ is a universal, exact and non-asymptotic confidence region for f of size α. The value of $\tau_n(\alpha)$ may be determined by simulations. For the dyadic scheme $\mathcal{I}_n = \mathcal{I}_n(2)$ these show that $\tau_n(0.95) \le 3$ for all $n \ge 500$. If $\mathcal{I}_n$ contains all singletons $\{t_i\}$, as will always be the case, it follows from Dümbgen and Spokoiny (2001) and Kabluchko (2007) that $\lim_{n\to\infty} \tau_n(\alpha) = 2$ for any α.

The confidence region (9) treats all intervals equally. The second confidence region $\mathcal{A}_n^*$ downweights the importance of small intervals and is constructed as follows. Dümbgen and Spokoiny (2001) extended Lèvy's uniform modulus of continuity of the Brownian motion and showed that

$$\sup_{0<s<t<1} \frac{\frac{(B(t)-B(s))^2}{t-s} - 2\log(1/(t-s))}{\log(\log(e^e/(t-s)))} < \infty \quad \text{a.s.} \tag{13}$$

If we embed the partial sums $\sum_{i\in I}^{j} Z(t_i)/\sqrt{|I|}$, $I \in \mathcal{I}_n$, in a standard Brownian motion it follows that

$$\sup_{I\in\mathcal{I}_n} \frac{(\sum_{t_j\in I} Z(t_j))^2/|I| - 2\log(n/|I|)}{\log(\log(e^e n/|I|)))} = \Gamma_n < \infty \quad \text{a.s.} \tag{14}$$

where Γ_n is bounded in probability as n tends to infinity. This implies that for any α we can find a $\gamma_n = \gamma_n(\alpha)$ such that

$$\begin{aligned}\mathcal{A}_n^*(\boldsymbol{Y}_n, \mathcal{I}_n, \sigma, \gamma_n(\alpha)) = \big\{ g : |w(g, \boldsymbol{Y}_n, I)| \le \\ \sigma\sqrt{2\log(n/|I|) + \gamma_n(\alpha)\log(\log(e^e n/|I|))} \text{ for all } I \in \mathcal{I}_n) \big\}.\end{aligned} \tag{15}$$

is a universal, exact and non-asymptotic α-confidence for f. The values of γ_n may be determined by simulation. For $\alpha = 0.95$ and with $\mathcal{I}_n = \mathcal{I}_n(2)$ a good approximation for $\gamma_n(\alpha)$ for $n \ge 100$ is given by

$$\gamma_n(0.95) \approx 5.77 - \exp(2.89 - 0.6\log(n)). \tag{16}$$

The confidence regions $\mathcal{A}_n(\boldsymbol{Y}_n, \mathcal{I}_n, \sigma, \tau_n)$ and $\mathcal{A}_n^*(\boldsymbol{Y}_n, \mathcal{I}_n, \sigma, \gamma_n)$ both require the true value of σ. We show how it is possible to obtain an estimate $\hat\sigma_n$ such that on replacing σ by $\hat\sigma_n$ both regions become honest in the sense of Li (1989) rather than exact. The following argument corrects the somewhat casual remarks on the problem made in Davies et al. (2009). Consider n independently distributed $N(0,\sigma^2)$ random variables. The median of the $|N(0,1)|$ distribution is $\Phi^{-1}(3/4) = 0.6745$ and hence $\mathbb{P}(1.4826|W_i| \ge \sigma) = 0.5$. If we denote the jth order statistic of $|W_1|, \ldots, |W_n|$ by $|W|_{(j)}$ then the normal approximation for the binomial $(n, 1/2)$ distribution implies the following approximation

$$\boldsymbol{P}\big(1.4826|W|_{(\lceil n/2 + z_\beta\sqrt{n}/2\rceil)} \ge \sigma\big) \approx \beta \tag{17}$$

where z_β denotes the β-quantile of the standard normal distribution. It follows from Anderson (1955) that if we replace the W_i by $W_i' = W_i - c_i$ then whatever the c_i there is an upward bias in the estimation of σ and we have

$$\boldsymbol{P}\big(1.4826|W'|_{(\lceil n/2 + z_\beta\sqrt{n}/2\rceil)} \ge \sigma\big) \ge \beta.$$

If, in particular, we put $\beta = 0.995$ we obtain

$$\boldsymbol{P}\big(1.4826|W'|_{(\lceil n/2+1.288\sqrt{n}\rceil)} \geq \sigma\big) \geq 0.995.$$

We can apply this to $W'_i = Y(t_{2i}) - Y(t_{2i-1}), i = 1, \ldots, \lfloor n+1 \rfloor/2 =: m$ and obtain

$$\boldsymbol{P}\big(1.4826|W'|_{(\lceil m/2+1.288\sqrt{m}\rceil)} \geq \sqrt{2}\sigma\big) \geq 0.995$$

which holds for *any* function f. On putting

$$\hat{\sigma}_n = \frac{1.4826}{\sqrt{2}}|W'|_{(\lceil m/2+1.288\sqrt{m}\rceil)} = 1.0484|W'|_{(\lceil m/2+1.288\sqrt{m}\rceil)}$$

we see that $\mathbb{P}(\hat{\sigma}_n \geq \sigma) \geq 0.995$ and hence

$$\boldsymbol{P}\big(f \in \mathcal{A}_n(\boldsymbol{Y}_n, \mathcal{I}_n, \hat{\sigma}_n, \tau_n(\alpha))\big) \geq \alpha - 0.005 \tag{18}$$

is an honest confidence region with $\alpha - 0.005$ in place of α. The corresponding inequality for $\mathcal{A}_n^*$ also holds. In spite of this the default value for $\hat{\sigma}_n$ we shall use in this paper is

$$\hat{\sigma}_n = 1.0484\text{median}\,(|Y(t_2) - Y(t_1)|, \ldots, |Y(t_n) - Y(t_{n-1})|). \tag{19}$$

It is simpler, the difference is in general small, it was used in Davies and Kovac (2001), Davies et al. (2008c), Davies et al. (2008a) and it also corresponds to using the first order Haar wavelets to estimate σ.

In Davies (1995) implicit use is made of an confidence region based on the lengths of runs of the signs of the residuals. Explicit universal, honest and non-asymptotic confidence regions which based on the signs of the residuals are to be found in Dümbgen (1998, 2003, 2006, 2007) and Dümbgen and Johns (2004).

2.2. A one-way table for regression functions

This section extends the approach given in Davies (2004) for the one-way table to the case of regression functions. We consider k samples $\boldsymbol{Y}_{in_i} = (t_{ij}, Y_i(t_{ij}))_{j=1}^{n_i}$ generated under (3). As a first step we replace the α in (11) and (15) by $\alpha_k = \alpha^{1/k}$ or the more general Bonferroni bound $\alpha_k = 1 - (1-\alpha)/k$ where k is the number of samples. This adjusts the size of each confidence region to take into account the number of samples. The confidence region for the ith sample is given by

$$\begin{aligned}\mathcal{A}_{in_i} = \mathcal{A}_{in_i}(\boldsymbol{Y}_{in_i}, \mathcal{I}_{in_i}, \hat{\sigma}_{in_i}, \tau_{in_i}(\alpha_k)) = \\ \Big\{g : \max_{I \in \mathcal{I}_{in_i}} |w(g, \boldsymbol{Y}_{in_i}, I)| \leq \hat{\sigma}_{in_i}\sqrt{\tau_{in_i}(\alpha_k)\log(n_i)}\Big\}.\end{aligned} \tag{20}$$

We denote by $\boldsymbol{P}_{\boldsymbol{f}}$ with $\boldsymbol{f} = (f_1, \ldots, f_k)$ the probability model where all the samples $\boldsymbol{Y}_{in_i}$, $i = 1, \ldots, k$, are independently distributed and $\boldsymbol{Y}_{in_i}$ was generated under (3) with $f = f_i$, $i = 1, \ldots, k$. It follows from the choice $\alpha_k = \alpha^{1/k}$ that

$$\boldsymbol{P}_{\boldsymbol{f}}\,(f_i \in \mathcal{A}_{in_i}(\boldsymbol{Y}_{in_i}, \mathcal{I}_{in_i}, \hat{\sigma}_{in_i}, \tau_{in_i}(\alpha_k)),\ i = 1, \ldots k) \geq \alpha \quad \text{for all} \quad \boldsymbol{f}. \tag{21}$$

All questions concerning the relationships between the functions f_i can now be answered by using the confidence regions $\mathcal{A}_{in_i}$. For example, the question as to whether the f_i are all equal translates into the question as to whether

$$\mathcal{A}_{\boldsymbol{n}_k} = \cap_{i=1}^{k} \mathcal{A}_{in_i} = \cap_{i=1}^{k} \mathcal{A}_{in_i}(\boldsymbol{Y}_{in_i}, \mathcal{I}_{in_i}, \hat{\sigma}_{in_i}, \tau_{in_i}(\alpha_k)), \quad \boldsymbol{n}_k = (n_1, \ldots, n_k) \tag{22}$$

is empty or not. If the supports S_{in_i} of the samples are not disjoint then it is possible that the linear inequalities which define the confidence regions are inconsistent. In this case $\mathcal{A}_{\boldsymbol{n}_k} = \emptyset$ and there is no joint approximating function. If the supports S_{in_i} of the samples are pairwise disjoint then $\mathcal{A}_{\boldsymbol{n}_k}$ is non–empty and so there always is a joint approximation function. Without further restrictions on the joint approximating function nothing more can be said. If however the joint approximating function is required to satisfy, for example, a shape constraint such as monotonicity, then it may be the case that there is no joint approximating function. Figure 1 shows just such a case where there are monotone approximations for each sample individually but no monotone joint approximation. To answer questions of this nature we must regularize within $\mathcal{A}_{\boldsymbol{n}_k}$ and this is the topic of the next section.

3. Disjoint supports

3.1. Regularization

We consider firstly the case when the supports $S_{in_i}, i = 1, \ldots, k$, are pairwise disjoint. In this case the joint approximation region $\mathcal{A}_{\boldsymbol{n}_k}$ is non-empty and will in general include many functions which would not be regarded as being acceptable. Indeed, it may be that $\mathcal{A}_{\boldsymbol{n}_k}$ does not contain any acceptable function.

3.2. Exchangeable supports

The joint approximating function shown in the centre panel of Figure 1 may be regarded as unacceptable. It is however possible to think of situations where the joint approximating function is of this complexity. The function is characterized by many very thin peaks so that a random choice of design points will often miss them all (needles in a haystack). On the basis of some additional information (metal detectors) the design points of the second sample are chosen intentionally and the peaks (needles) are discovered. The joint approximating function of Figure 1 is clearly incompatible with independently distributed design points. If we choose a random sample uniformly distributed on $[0, 1]$ and generate the corresponding y-values using the joint approximating function the resulting sample will, with overwhelming probability, not be approximable by a monotone function, even though each of the individual samples was approximable by a monotone function. In the following we assume that the experimental conditions are such that the supports of the two samples may be regarded as exchangeable. We give two different ways of making use of this information.

The first method is to accept a joint approximating function if and only if the approximating functions of the samples and the joint approximating function all have the same shape. The definition of shape we use for the purpose of demonstrating the method is the number of local extreme values. On the basis of this we would conclude for the data exhibited in Figure 1 that the two data sets cannot be adequately approximated by the same function.

The second method does not require the calculation of approximating functions but it restricted to the case where the noise levels of the two samples are the same. If we generated two samples $\boldsymbol{Y}_{1n_1}$ and $\boldsymbol{Y}_{2n_2}$ under the model (3) where the support points of the two samples are interchangeable, then the resulting samples will be interchangeable: what is labelled as being from sample 1 could equally have been labelled as being from sample 2. In particular, if we permutate all observations at random and assign the first n_1 to sample 1 and the remaining n_2 to sample 2, then the new data will have exactly the same distribution as the original data. This continues to hold if we replace the Y_i-values by their ranks R_i in the joint sample. To compare the two permutated samples we multiply the ranks of the second sample by -1. This gives us n_1+n_2 observations of the form $(T_i, \pm R_i), i=1,\ldots,n$ with $T_i < T_{i+1}, i=1,\ldots,n-1$ and with $+$ if (T_i,R_i) has been allocated to the first sample and $-$ if it was allocated to the second sample. Given a collection of intervals $\mathcal{I}_n$ we consider the sums

$$S_j = \sum_{T_i \in I_j} \pm R_i, \quad I_j \in \mathcal{I}_n.$$

The means m_j and standard deviations s_j of the S_j conditional on the original data (T_i,R_i) can be calculated on the basis of random permutations and corresponding allocations of the signs $\pm$. In a second simulation the say 0.95-quantile $\lambda(0.95)$ of

$$\Lambda = \max_{I_j \in \mathcal{I}_n} \frac{|S_j - m_j|}{s_j}$$

can be calculated. If the value $\tilde{\Lambda}$ of Λ for the original sample exceeds $\lambda(0.95)$ then we conclude that the two samples cannot be described by the same function and exchangeable support points.

Figure shows the result of a small simulation study to demonstrate the two methods. Data were generated according to (3) with

$$n_1 = n_2 = 500, \sigma_1 = \sigma_2 = 0.25, f_1(t) = \exp(1.5t), f_2(t) = \exp(1.5t) + \delta$$

and with support points independently and uniformly distributed over $[0,1]$ for both samples. The family $\mathcal{I}_n$ of subsets of $[0,1]$ was taken to be $\mathcal{I}_n(2)$ and to make the two procedures comparable we set $\tau_n = 2.26$. The figure shows the probability of rejecting a joint approximation as a function of δ.

4. Intersecting supports

4.1. Quantifying detectable differences

As mentioned in Section 1 the Neumeyer and Dette (2003) procedure can detect differences of the order of $n^{-1/2}$. We now consider the size of detectable differences for our procedure in the case of equal supports. For simplicity we consider only the case $k = 2$ and assume that the supports $S_{1n_1} = S_n$ and $S_{2n_2} = S_n$ are the same and that the data $\boldsymbol{Y}_{1n}$ and $\boldsymbol{Y}_{1n}$ are generated by (3) where we allow for differing σ_1 and σ_2. We take $\mathcal{I}_n$ to be the set of all intervals but we indicate below the adjustments required if $\mathcal{I}_n = \mathcal{I}_n(\lambda)$ as in (10). We state the results using σ_1 and σ_2 rather than the estimates (19) and write $\tau_n = \tau_n(\alpha^{1/2})$.

Theorem 4.1. *Let $\mathcal{I}_n$ be the set of all intervals, suppose $f_1(t) > f_2(t) + \eta_n$ on an interval I_n containing $|I_n|$ support points $t_i \in S_n$ and set*

$$\zeta_n = \left(\max\left\{0, \eta_n\sqrt{|I_n|} - (\sigma_1+\sigma_2)\sqrt{\tau_n\log(n)}\right\}\right)/\sqrt{\sigma_1^2+\sigma_2^2}\,. \tag{23}$$

Then with probability at least $2\Phi(\zeta_n) - 1$ there will be no joint approximation for the data sets $\boldsymbol{Y}_{1n}$ and $\boldsymbol{Y}_{2n}$.

Proof. Suppose there exists a joint approximation $\tilde{f}_n$. Then

$$\frac{1}{\sqrt{|I_n|}}\left|\sum_{t_i\in I_n}(Y_j(t_i) - \tilde{f}_n(t_i))\right| \le \sigma_j\sqrt{\tau_n\log(n)}\,,\ j = 1,\,2,$$

or equivalently

$$\frac{1}{\sqrt{|I_n|}}\left|\sum_{t_i\in I_n}(f_1(t_i) + \sigma_j Z_j(t_i) - \tilde{f}_n(t_i))\right| \le \sigma_j\sqrt{\tau_n\log(n)}\,,\ j = 1,\,2.$$

which implies

$$\begin{aligned}\frac{1}{\sqrt{|I_n|}}\left|\sum_{t_i\in I_n}(f_1(t_i) - f_2(t_i))\right| &\le (\sigma_1+\sigma_2)\sqrt{\tau_n\log n}\\ &\quad+\frac{1}{\sqrt{|I_n|}}\left|\sum_{t_i\in I_n}(\sigma_1 Z_1(t_i) - \sigma_2 Z_2(t_i))\right|\\ &= (\sigma_1+\sigma_2)\sqrt{\tau_n\log n} + \sqrt{\sigma_1^2+\sigma_2^2}\,|N(0,1)|.\end{aligned}$$

As

$$\frac{1}{\sqrt{|I_n|}}\left|\sum_{t_i\in I_n}(f_1(t_i) - f_2(t_i))\right| \ge \sqrt{|I_n|}\,\eta_n$$

it follows

$$|N(0,1)| \ge \left(\eta_n\sqrt{|I_n|} - (\sigma_1+\sigma_2)\sqrt{\tau_n\log n}\right)/\sqrt{\sigma_1^2+\sigma_2^2}$$

from which the theorem follows. □

If the supports points in S_n are equidistant and I_n has length δ_n then $|I_n| \approx n\delta_n$ and we can replace ζ_n of (23) by

$$\zeta_n = \left(\max\left\{0, \eta_n\sqrt{n\delta_n} - (\sigma_1+\sigma_2)\sqrt{\tau_n \log(n)}\right\}\right)/\sqrt{\sigma_1^2+\sigma_2^2} \tag{24}$$

The theorem is based on the assumption that $\mathcal{I}_n$ is the set of all intervals. If $\mathcal{I}_n = \mathcal{I}_n(\lambda)$ as in (10) it follows that there exists an interval $I_n' \subset I$ in $\mathcal{I}_n(\lambda)$ and containing $|I_n'| \geq |I_n|/\lambda = \delta_n/\lambda$ points t_i of S_n for which $f_1(t_i) - f_2(t_i) > \eta_n$. This requires replacing (23) by

$$\zeta_n = \left(\max\left\{0, \eta_n\sqrt{|I_n/\lambda|} - (\sigma_1+\sigma_2)\sqrt{\tau_n \log(n)}\right\}\right)/\sqrt{\sigma_1^2+\sigma_2^2} \tag{25}$$

and (24) by

$$\zeta_n = \left(\max\left\{0, \eta_n\sqrt{n\delta_n/\lambda} - (\sigma_1+\sigma_2)\sqrt{\tau_n \log(n)}\right\}\right)/\sqrt{\sigma_1^2+\sigma_2^2}\,. \tag{26}$$

If $I_n = [0,1]$ then (26) reduces to

$$\zeta_n = \left(\max\left\{0, \eta_n\sqrt{n/\lambda} - (\sigma_1+\sigma_2)\sqrt{\tau_n \log(n)}\right\}\right)/\sqrt{\sigma_1^2+\sigma_2^2}\,. \tag{27}$$

so that discrepancies on the whole interval of $O(\sqrt{(\log n)/n}\,)$ will be detected. This results in an extra $\log n$ term when compared with the result of Neumeyer and Dette (2003).

The same analysis can be carried through using the approximation region $\mathcal{A}_n^*$. In the theorem we simply replace γ_n by

$$\zeta_n^* = \frac{\max\left\{0, \eta_n\sqrt{|I_n|} - (\sigma_1+\sigma_2)\sqrt{2\log(n/|I_n|) + \gamma_n(\alpha)\log\log(e^e n/|I_n|)}\right\}}{\sqrt{\sigma_1^2+\sigma_2^2}} \tag{28}$$

and corresponding to (26) we obtain

$$\zeta_n^* = \frac{\max\left\{0, \eta_n\sqrt{n\delta_n/\lambda} - (\sigma_1+\sigma_2)\sqrt{2\log(\lambda/\delta_n) + \gamma_n(\alpha)\log\log(e^e \lambda/\delta_n)}\right\}}{\sqrt{\sigma_1^2+\sigma_2^2}}. \tag{29}$$

In particular, if $\delta_n = 1$, then deviations of order $O(1/\sqrt{n})$ can be detected.

4.2. An example

We consider a situation similar to that of Figure 1 as is shown in Figure 7. The sample sizes are $n = 500$ with common supports $t_j = j/n$ and we take α to be 0.95 so that $\alpha_k = 0.95^{1/2} = 0.9747$. For this choice of α and with $\mathcal{I}_n = \mathcal{I}_n(2)$ simulations give $\tau_n = 2.973$. We set $f_1(t) = \exp(1.5t)$ and put

$f_2(t) = f_1(t)$ except for $t \in [0.402, 0.44]$ where $f_2(t) = f_1(t) + \eta_n$. For this interval $\delta_n = 20/500$. To be able to detect deviations η_n with a probability of at least 0.95 it is sufficient to require $\zeta_n \geq 1.96$. This leads to

$$\eta_n \geq (1.96\sqrt{1/2} + 0.5\sqrt{2.973 \log 500}\,)/\sqrt{10} = 1.12.$$

For the data shown in Figure 7 the difference is detected with $\eta_n = 0.575$ but not with $\eta_n = 0.574$. Simulations show that for this particular example a deviation of 0.65 on the interval $[0.402, 0.44]$ is detected with probability 0.95.

5. Comparison with other procedures

5.1. Analysis and simulations

When the supports are disjoint the approach developed in this paper is not comparable with others. There are two reasons for this. Firstly, other approaches postulate equality of the regression functions and either accept or reject this postulate. In contrast our approach results in a joint approximating function which is either accepted or rejected, the decision being taken outside of statistics. Secondly, in the case of disjoint supports the other approaches require estimates of the two functions. In all cases known to us these estimates depend on a smoothing parameter for which there is no default choice: there are only suggestions based on asymptotics which may be arbitrarily bad for any given n. We therefore restrict attention to the case of equal supports. For simplicity we take $k = 2$. For such data Delgado (1992) proposed the test statistic

$$T_n = \sqrt{n} \max_{1 \leq j \leq n} |R(j)|/s_n^* = \max_{1 \leq j \leq n} \left| \sum_{i=1}^{j} (Y_1(t_i) - Y_2(t_i)) \right| /(\sigma_n \sqrt{n}) \qquad (30)$$

where σ_n is some quantifier of the noise. Under the null hypothesis $f_1 = f_2 = f$ the distribution of T_n does not depend on f. In this special case the test statistic of Neumeyer and Dette also reduces to (30). If the data were generated under (3) then under H_0 the distribution of T_n converges to that of $\max_{0 \leq t \leq 1} |B(t)|$ where B is a standard Brownian motion. The 0.95-quantile is approximately 2.24 which leads to rejection of H_0 if

$$T_n \geq 2.24. \qquad (31)$$

Suppose now that the data are generated as in (3) with $f_1(t) = f_2(t)$ apart from t in an interval I of length δ_n where $f_1(t) - f_2(t) \geq \eta_n$. It follows from (31) that H_0 will be rejected with high probability if

$$\delta_n \eta_n \geq 4.48\sigma/\sqrt{n} \qquad (32)$$

where $\sigma^2 = \sigma_1^2 + \sigma_2^2$. If $\delta_n = 1$ deviations of the order of $O(\sigma/\sqrt{n})$ can be picked up which contrasts with the $O(\sigma\sqrt{\log(n)/n})$ of (26) for the method based on

$\mathcal{A}_n$. The method based on $\mathcal{A}_n^*$ however will also pick up deviations of the order of $O(\sigma/\sqrt{n})$ as can be seen from (28). For $\delta_n = 1/\sqrt{n}$ it follows from (32) that the test statistic T_n will pick up deviations of the order of σ. In this situation we see from (26) and (28) that the methods based on $\mathcal{A}_n$ and $\mathcal{A}_n^*$ will both pick up deviations of the order of $O(\sigma\sqrt{\log(n)/\sqrt{n}})$.

Another test which is applicable in this situation is due to Fan and Lin (1998). If we denote the Fourier transform of the data sets by $\tilde{Y}_1(i)$ and $\tilde{Y}_2(i), i = 1, \ldots, n$, and order them as described in Fan and Lin (1998), their test statistic reduces to

$$T_n^* = \max_{1 \le m \le n} \left| \frac{1}{\sqrt{m}} \sum_{i=1}^{m} ((\tilde{Y}_2(i) - \tilde{Y}_1(i))^2/\tilde{\sigma}_n^2 - 1) \right| \tag{33}$$

where $\tilde{\sigma}_n$ is some estimate of the standard deviation of the $\tilde{Y}_2(i) - \tilde{Y}_1(i)$. For data generated under the model (3) the critical value of T_n^* can be obtained by simulations. It is not as simple to determine the size of the deviations which can be detected by the test (33) as the test statistic is a function of the Fourier transforms and the differences in the functions must be translated into differences in the Fourier transforms. The first member of the sum in (33) is the difference of the means and this is given the largest weight. We do not pursue this further but give the results of a small simulation study.

We put $n = 500$ and consider two samples of the form

$$Y_1(i/n) \quad = \quad Z_1(i/n), \quad i = 1, \ldots, n = 500 \tag{34}$$

$$Y_2(i/n) \quad = \quad g(i/n) + Z_2(i/n), \quad i = 1, \ldots, n = 500 \tag{35}$$

were generated where the $Z_j(i/n)$ are i.i.d $N(0,1)$ random variables with g given by one of

$$\begin{aligned}
g_1(t) &= \eta,\ 0 \le t \le 1 \\
g_2(t) &= \begin{cases} \eta, & 0 \le t \le 1/2, \\ -\eta, & 1/2 < t \le 1, \end{cases} \\
g_3(t) &= \begin{cases} 0, & 0 \le t \le U, \\ \eta, & U < t \le U + 1/4, \\ 0, & U + 1/4 < t \le 1, \end{cases} \\
g_4(t) &= \begin{cases} 0, & 0 \le t \le U, \\ \eta, & U < t \le U + 1/8, \\ -\eta, & U + 1/8 < t \le U + 1/4, \\ 0, & U + 1/4 < t \le 1, \end{cases}
\end{aligned}$$

where U is uniformly distributed on $[0, 3/4]$ and independent of the $Z_i, i = 1, 2$. The four procedures, Delgado–Neumeyer–Dette, Fan–Lin and those based on $\mathcal{A}_n$ and $\mathcal{A}_n^*$ were all calibrated to give tests of size 0.05 for testing $g \equiv 0$. The critical values for Delgado–Neumeyer–Dette and Fan–Lin tests are 2.22 and 6.97 respectively. The value of τ_n for the test based on $\mathcal{A}_n$ is 1.46 and the corresponding value of γ_n for that based on $\mathcal{A}_n^*$ is 0.66. Figure 9 shows the power of the tests for different values of η. The upper panels are the results for $g = g_1$

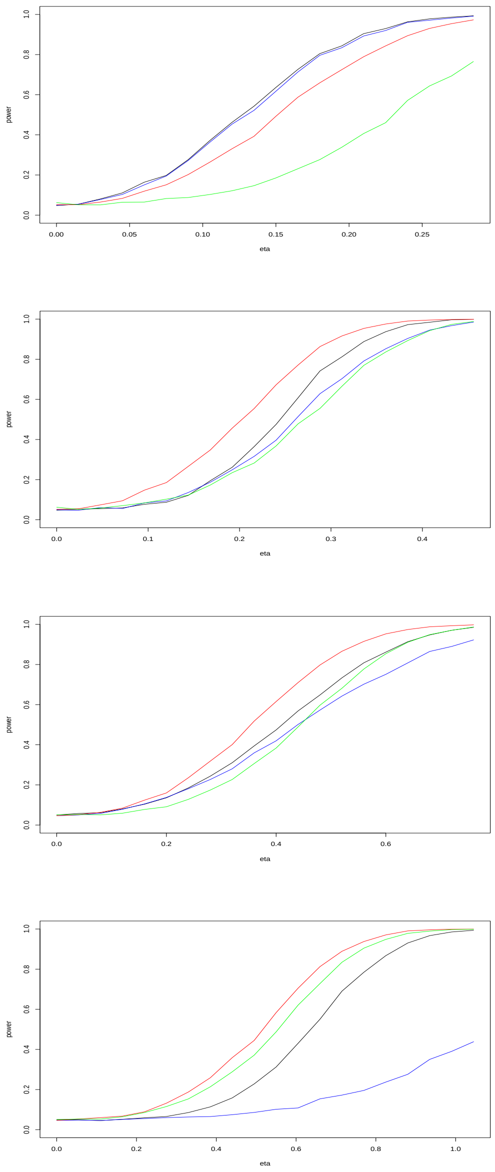

FIG 9. *The panels from top to bottom show the power functions of the four tests with $g = g_1$ to $g = g_4$ in order. The Delgado–Neumeyer–Dette is shown in blue, the Fan–Lin test in black, the test based on $\mathcal{A}_n$ in green and that based on $\mathcal{A}_n^*$ in red.*

and for $g = g_2$ and the lower panels give the results for $g = g_3$ and $g = g_4$. The colour scheme is as follows: Delgado–Neumeyer–Dette blue, the Fan–Lin

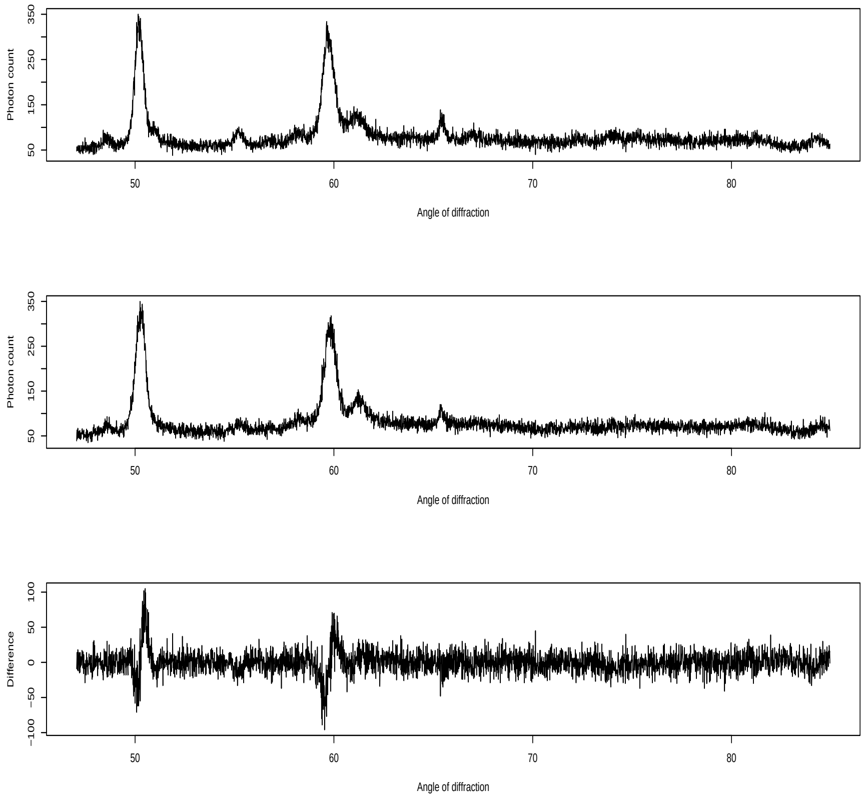

Fig 10. *The top and centre panels show two data sets each of 4806 observations with the same design points. The bottom panel shows the differences of the two samples.*

black, $\mathcal{A}_n$ green and $\mathcal{A}_n^*$ red. The results confirm the analysis given above. The Delgado–Neumeyer–Dette and Fan–Lin tests are better with g given by (1) but if the mean difference is zero (2), or the interval is small (3) or both (4) then they are outperformed by the procedure based on $\mathcal{A}_n^*$ and, in case 4, also by that based on $\mathcal{A}_n$.

5.2. An application

We give an example with some real data from the area of thin-film physics. They give the number of photons of refracted X-rays as a function of the angle of refraction and were kindly supplied by Professor Dieter Mergel of the University of Duisburg-Essen. Two such data sets are shown in the top panel of Figure 10; the differences $y_1(t_i) - y_2(t_i)$ are shown in the bottom panel. Each data set is composed of 4806 measurements and the design points are the same. The samples differ in the manner in which the thin film was prepared. One of the questions to be answered is whether the results of the two methods are substantially different.

The noise levels for the data sets are the same, namely 8.317, which is explainable by the fact that the data are integer valued. The differences between the two data sets are concentrated on intervals each containing about 40 observations. The estimate (32) suggests that the differences will have to be of the order of 92 to be detected with a degree of certainty by the Delgado–Neumeyer–Dette test. The actual differences are of about this order as can be seen from the bottom panel of Figure 10. In fact the test just fails to reject the null hypothesis at the 0.1 level. The realized value of the test statistic is 1.734 as against the critical value of 1.90 given in (31). The Fan-Lin test (33) rejects the null hypothesis at the 0.01 level. The realized value of the test statistic is 111.66 as against the critical value of 12.44 for a test of size $\alpha = 0.01$. Finally the tests based on $\mathcal{A}_n$ and $\mathcal{A}_n^*$ both reject the null hypothesis at the 0.01 level. The realized value of τ_n is 43.15 as against the critical value of 1.50. The realized value of γ_n is 53.27 as against the critical value of 0.733.

6. Heteroskedastic noise

The results so far were developed for data with homogeneous noise. If the noise is heteroskedastistic then we can use the results of Höhenrieder (2008) (see Davies (2005) for a preliminary report) to quantify the local noise level. The original motivation was to provide volatility estimates for financial data using the model

$$R(t) = \Sigma(t)Z(t) \tag{36}$$

where R denotes the daily returns, Σ is the volatility and Z is standard white noise.

$$\sum_{t_i \in I} \frac{R(t_i)^2}{\Sigma(t_i)^2} = \sum_{t_i \in I} Z(t_i)^2 \overset{D}{=} \chi^2(|I|)$$

for any interval I and where $\chi^2(k)$ denotes the chi-squared distribution with k degrees of freedom and $|I|$ denotes the number of points $t_i \in I$. We put

$$\mathcal{S}_n = \Big\{\sigma : \sigma : \{1, \ldots, n\} \to (0, \infty)$$
$$\chi^2_{|I|, \frac{1-\alpha_n}{2}} \le \sum_{t \in I} \frac{R(t)^2}{\sigma(t)^2} \le \chi^2_{|I|, \frac{1+\alpha_n}{2}}\ ,\ \forall I \subset \{1, \ldots, n\}\Big\}. \tag{37}$$

where $\chi^2_{k,\beta}$ denotes the β–quantile of the χ^2–distribution with k degrees of freedom and α_n is given by

$$P\left(\chi^2_{|I|, \frac{1-\alpha_n}{2}} \le \sum_{t \in I} Z(t)^2 \le \chi^2_{|I|, \frac{1+\alpha_n}{2}}\ ,\ \forall I \subset \{1, \ldots, n\}\right) = \alpha. \tag{38}$$

A simple approximation for α_n with $\alpha = 0.9$ is given by

$$\alpha_n = 1 - 0.0343\exp(-0.286\log\log(n))/n$$

which is valid at least for $100 \le n \le 20000$. It follows that $\mathcal{S}_n$ is a universal, exact and non-asymptotic confidence region for Σ of size α, that is

$$\mathbb{P}(\Sigma \in \mathcal{S}_n) = \alpha$$

whatever Σ. For real data $r(t_i)$ we replace the $R(t_i)$ in (37) by the $r(t_i)$. The form of regularization we choose is one based on sparsity. We take σ_n to be piecewise constant with the minimum number of intervals of all such functions in $\mathcal{S}_n$. The algorithmic complexity of this problem makes it practically unsolvable so we replace $\mathcal{S}_n$ by

$$\mathcal{S}_n^* = \Big\{\sigma : \sigma : \{1,\ldots,n\} \to (0,\infty)\chi^2_{|I|,\frac{1-\alpha_n}{2}} \le \sum_{t\in I}\frac{R(t)^2}{\sigma(t)^2} \le \chi^2_{|I|,\frac{1+\alpha_n}{2}}$$
$$\forall I \subset I_\nu,\ I_\nu \text{ a constancy interval of } \sigma\Big\}. \tag{39}$$

As $\mathcal{S}_n \subset \mathcal{S}_n^*$ we see that $\mathcal{S}_n^*$ is honest (Li (1989)) rather than exact. The sparsity problem can now be solved quite simply. For financial data it makes for greater interpretability if

$$\sigma_n(t_j)^2 = \frac{1}{|I|}\sum_{t_i\in I} r(t_i)^2, \quad t_j \in I,$$

on intervals I of constancy of σ_n. Amongst all such functions we then minimize the sum of the quadratic deviations $\sum_i (r(t_i)^2 - \sigma_n(t_i)^2)^2$. This problem can be solved using dynamic programming. We refer to Höhenrieder (2008): an R-package is available from

http://www.stat-math.uni-essen.de/cho/research.php.

The upper panel of Figure 11 shows the absolute daily returns of the standard and Poor's index over about 19260 days together with the piecewise constant volatility in red. The lower panel shows the first 5000 observations.

To apply the above methodology to non-parametric regression we replace the model (3) by

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

and consider the differences of the Y_i

$$Y_{i+1} - Y_i = f(t_{i+1}) - f(t_i) + \sigma(t_{i+1})Z(t_{i+1}) - \sigma(t_i)Z(t_i).$$

If the regression function f and the noise level σ are sufficiently smooth we have

$$Y_{i+1} - Y_i \approx \sigma(t_i)(Z(t_{i+1}) - Z(t_i))$$

which is close to (36). The main difference is that successive $Z(t_{i+1}) - Z(t_i)$ and $Z(t_{i+2}) - Z(t_{i+1})$ are correlated. This can either be ignored (our option)

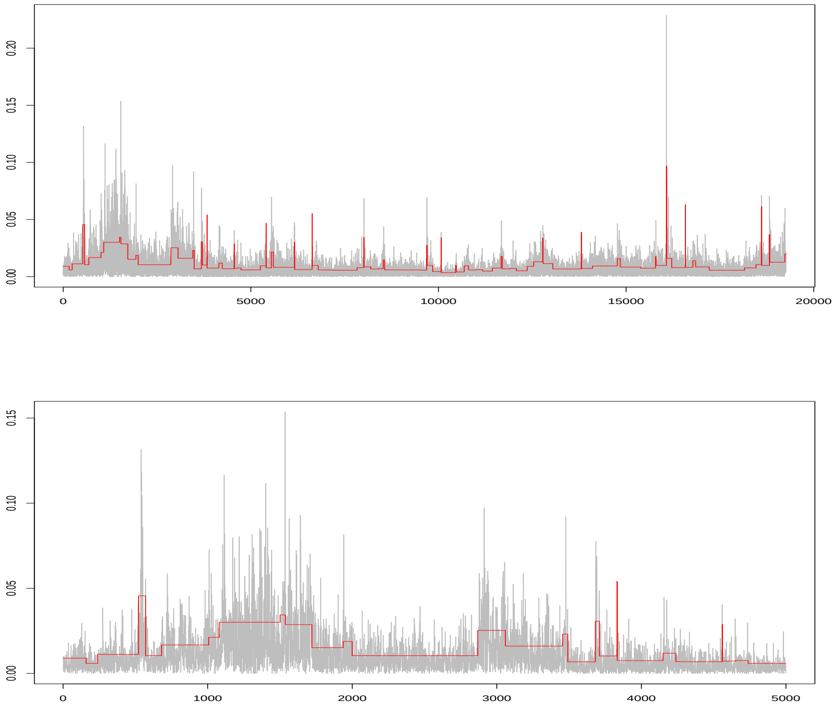

Fig 11. *Upper panel: absolute daily returns of the Standard and Poor's index over 19260 days together with the estimated piecewise volatility in red. The lower panel shows the first 5000 days.*

or every second difference can be eliminated. We can now calculate a piecewise constant noise level as for the finance data. The approximation region (9) is replaced by putting

$$w(g, \boldsymbol{Y}_n, I) = \frac{1}{\sqrt{|I|}} \sum_{t_i \in I} \frac{Y(t_i) - g(t_i)}{\sigma_n(t_i)} \tag{40}$$

and setting

$$\mathcal{A}_n(\boldsymbol{Y}_n, \mathcal{I}_n, \sigma, \tau_n) = \{g : \max_{I \in \mathcal{I}_n} |w(g, \boldsymbol{Y}_n, I)| \ \leq \sqrt{\tau_n \log(n)}\ \} \tag{41}$$

where σ is now a function. In applications we replace σ by an estimate of the noise level described above. This definition of the approximation region can also be found in Davies et al. (2008a) where high noise levels were essentially Poisson with parameter $f(t)$. The top panel of Figure 12 shows a data set of size $n = 500$ generated according to

$$Y(t) = \exp(1 - 5t) + 0.25 \exp(1.5t) Z(t) \tag{42}$$

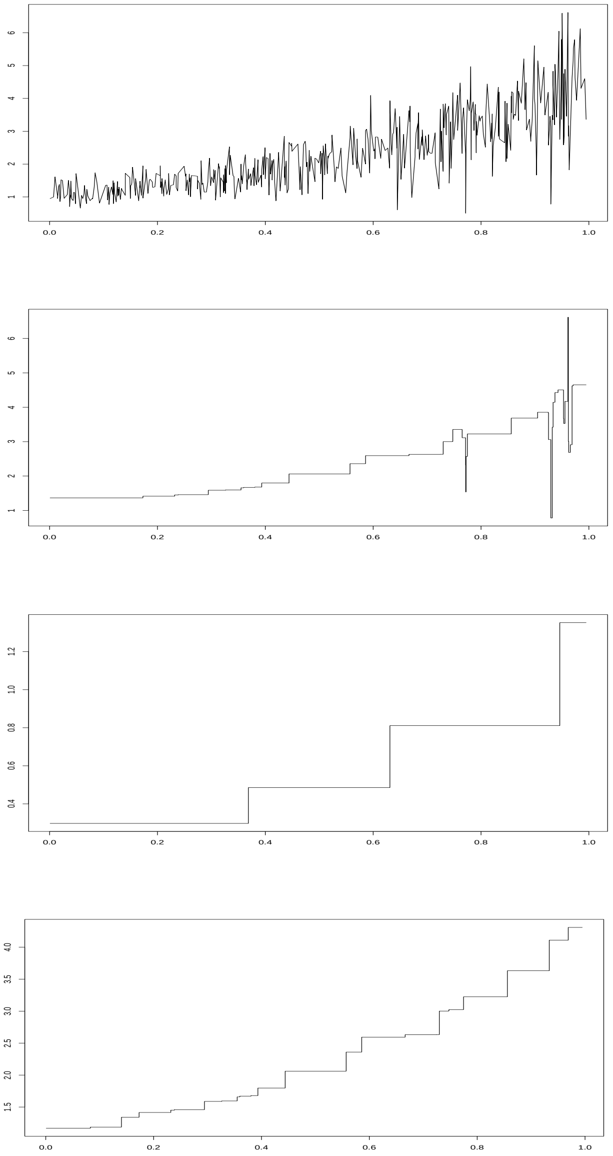

Fig 12. *Top panel: a sample of size* $n = 500$ *generated according to (42). Second panel: the taut string estimate based on homogeneous noise. Third panel: a piecewise constant estimate of the noise level. Bottom: the taut string estimate based on the piecewise constant noise estimate.*

and with the support being uniformly distributed on $[0, 1]$. The second panel shows the taut string estimate based on homogeneous noise. The third panel shows the piecewise constant noise estimate and the bottom panel the taut string estimate using this noise estimate. The extension to the problem of several samples is straightforward.

7. Adapting the taut string algorithm

The taut string algorithm of Davies and Kovac (2001) has proved to be very effective in determining the number of local extremes of a function contaminated by noise (see Davies et al. (2008b)). It has the following property. Two sets of points (t_i, u_i) and $(t_i, l_i), i = 0, \ldots, n$ with $t_i < t_{i+1}$ and $l_i \le u_i, l_0 = u_0, l_n = u_n$ describe a tube on $[t_0, t_n]$. The taut string algorithm calculates a function ts, the taut string, which lies within the tube, $l_i \le ts(t_i) \le u_i, i = 0, \ldots, n$ and has the smallest number of local extreme values of all functions which lie within the tube. An algorithm of complexity $O(n)$ for calculating the taut string is given in the appendix of Davies and Kovac (2001). When applied to a data set $(t_i, y_i), i = 1, \ldots, n$ the tube is centred at the partial sums $(t_i, \sum_{j=1}^{i} y_j)$ of the y_i whereby the tube is chosen automatically using local squeezing as described in Section 3.6 of Davies and Kovac (2001). The approximating function is taken to be the left-hand derivative of the piecewise linear taut string except for the first point where the right-hand derivative is taken. The method can be adapted to the case of k samples by weighting the observations according to the inverse noise level of the sample from which the observation comes as we now explain.

The first question to be decided is whether a common approximation exists or not. This will always be the case unless the intersection of the empirical supports is non-empty in which case it is possible that the linear inequalities defining the approximation regions are inconsistent. In principle this can be decided using linear programming but in practice it may not be possible for large data sets because of the corresponding large number of inequalities. A simple and effective way of overcoming the problem is as follows. Let $n \le \sum_{i=1}^{k} n_i$ denote the number of different t_{ij} values which we order as $0 \le t_1 < \ldots < t_m \le 1$. For each sample we calculate the values of $\hat{\sigma}_{in_i}$ given by (19) and put

$$y_\ell = \frac{\sum_{t_{ij}=t_m} y_{ij}/\hat{\sigma}^2_{in_i}}{\sum_{t_{ij}=t_\ell} 1/\hat{\sigma}^2_{in_i}}\,, \quad i = 1, \ldots, m. \tag{43}$$

We now simply check whether any function $\tilde{f}_m$ which interpolates the points $(t_\ell, y_\ell), \ell = 1, \ldots, m$ lies in each of the approximation regions. If this is the case then clearly an approximating function exists. If it is not the case then we conclude that no such function exists although this has not actually been shown. However if an interpolating function does not lie in all approximation regions then any function which does will be even more complex and so probably not acceptable.

If a joint approximation exists then the taut string algorithm can be used to find a joint approximation with the smallest, or close to the smallest, number of local extreme values. Applying the taut string algorithm directly to the points (t_ℓ, y_ℓ), $\ell = 1, \ldots, m$ may result in too many local extreme values, especially

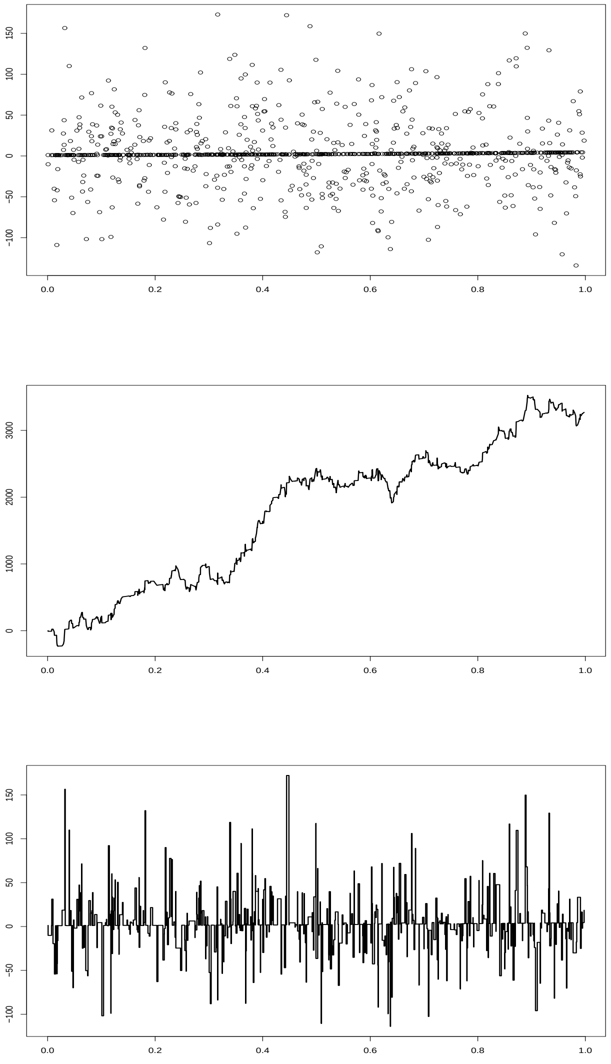

FIG 13. *Top: two data sets with* $f(t) = \exp(1.5t)$ *and noise levels* $\sigma_1 = 0.1, \sigma_2 = 50$. *Centre: plot of the unweighted partial sums. Bottom: the resulting taut string approximation.*

if the noise levels of the samples are very different. This is because the partial sums $\sum_{j=1}^{i} y_j$ of the y_i, on which the taut string algorithm relies, are dominated by those samples with the largest noise level. This is shown in Figure 13 where we generated two data sets each of size $n_i = 500$ according to (3) with

$$f(t) = \exp(1.5t), \quad \sigma_1 = 0.1, \quad \sigma_2 = 50$$

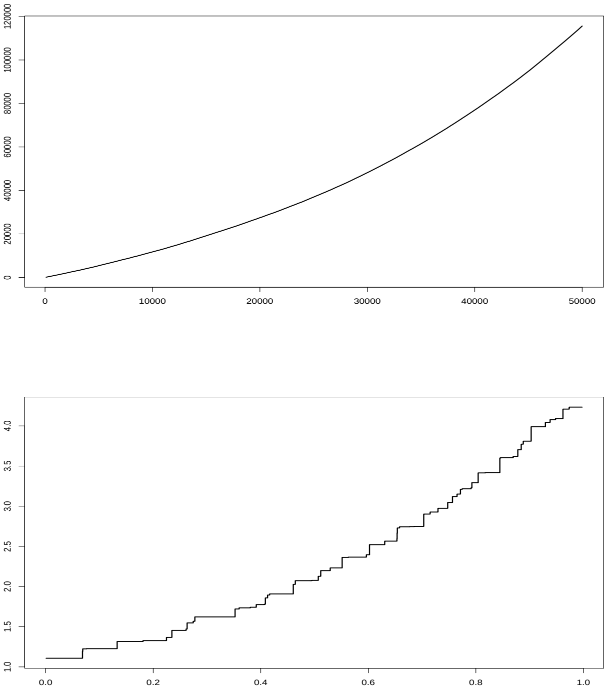

Fig 14. *Upper panel: plot of the weighted partial sums for the data of Figure 13. Lower panel: the resulting taut string approximation.*

and where the $t_{ij}, j = 1, \ldots, 500, i = 1, 2$ are taken to be independently and uniformly distributed on $[0, 1]$. The top panel of Figure 13 shows the two data sets, the centre panel shows the plot (t_ℓ, y^*_ℓ), $\ell = 1, \ldots, 1000$ of the partial sums $y^*_\ell = \sum_{j=1}^{\ell} y_j$ of the y_ℓ. The bottom panel shows the results of the taut string procedure when the tube is centred at the points (t_ℓ, y^*_ℓ). The partial sums are dominated by the second sample with $\sigma_2 = 50$ and this is reflected in the result of the taut string procedure. The problem may be overcome by weighting the observations according to the noise level. We put

$$\alpha_\ell = \sum_{t_{ij}=t_\ell} 1/\hat{\sigma}^2_{in_i}, \quad y^\circ_\ell = \sum_1^\ell \alpha_j y_j, \quad A_\ell = \sum_1^\ell \alpha_j, \quad y^\circ_0 = A_0 = 0$$

The plot of the partial sums $(A_\ell, y^\circ_\ell), \ell = 1, \ldots, 1000$ is shown in the upper panel of Figure 14. This is now dominated by the low noise sample. The lower panel shows the result of applying the the taut string procedure to these points.

The local squeezing procedure described in Davies and Kovac (2001) can be adopted to the case of k samples as follows. For a given tube, the taut string through the tube and constrained to pass through $(0, 0)$ and (y_n°, A_n) is calculated. The value of the estimate $\tilde{f}_n(t_m)$ at the point t_m is taken to be the left-hand derivative of the taut string except for the first point where the right-hand derivative is taken. For each data set individually it is now checked whether $\tilde{f}_n \in \mathcal{A}_{in_i}, i = 1, \ldots, k$. If this is the case the procedure terminates. Otherwise those intervals for which the inequalities defining the $\mathcal{A}_{in_i}$ do not hold are noted and the tube is squeezed at all points t_{j-1} and t_j for which t_j lies in such an interval. This is continued until a function $\tilde{f}_n \in \mathcal{A}_{\boldsymbol{n}_k}$ is found.

8. Discussion

Although the comparison of non-parametric regression functions is neither a fundamental theoretical nor practical problem it does pose some interesting conceptual challenges for statisticians, particularly in the case where the supports are disjoint. As pointed out in Section 1.1, if the supports of the samples are disjoint then there will always be a single function which could have generated the data and which is consistent with any qualitative smoothness condition. There can be no qualitative smoothness objections to the function in the centre panel of Figure 1 but there can be quantitative smoothness objections, for example that the absolute value of the first derivative exceeds a given bound. We note that this bound must be made explicit, simply assuming $\|f^{(1)}\|_\infty \le C$ without specifying C does not help. On a more general level the problem is one of lack of uniformity in the asymptotics. The set of functions which satisfy a qualitative smoothness condition is too large for uniform asymptotics over the class: if a sequence of functions f_n in this class converges pointwise to a function f, all that can be said about f is that its points of discontinuity have measure zero. A weak form of quantitative smoothness over the class $\|f^{(1)}\|_\infty \le C$ at least guarantees that the limiting function will be continuous. Another case of non-uniform asymptotics is the choice of a smoothing parameter. Articles which rely on kernel estimators with a global bandwidth h_n often recommend a choice $h_n \asymp cn^{-1/5}$ because it is asymptotically optimal. For any finite n it can be very poor and consequently $h_n \asymp cn^{-1/5}$ is of no help in any concrete application. Typically examples and simulations will be restricted to functions which are very smooth when measured in terms of the supremum norms of the first and second derivatives, $\|f^{(1)}\|_\infty \le 10$ for example. Here the asymptotics may start working for relatively small samples but none of this is captured by the theorems. Finally we note that lack of uniformity can cause problems in much simpler situations. An example is Hodges' super efficient estimator of the parameter μ of a $N(\mu, 1)$ distribution (see for example Lehmann (1983), page 405). For the connection between the Hodges' estimator and the effects of non-uniformity in other areas of statistics we refer to Pötscher and Leeb (2008).

The goal is to provide procedures which work well for the data at hand and, in particular, for the sample size n at hand. One consequence of this is that we are prepared to give specific numerical values such as 'at least 998 local minima' and 'the total variation of the second derivative is at least 0.7' to describe the data.

We do not embed the procedure in a sequence of procedures and then show that they will work asymptotically. Asymptotics are used only as a concrete approximation to calculate a probability, just as one may approximate the binomial $b(n, 1/2)$ probabilities using the normal approximation. One example is the determination of the values of the $\tau_n(\alpha)$. From the asymptotics $\lim_{n\to\infty} \tau_n(\alpha) = 2$ and simulations for finite n it is possible to derive a simple formula which will provide good approximations for all n. Another example is the use of the normal approximation in (17). The claim that the approximation regions are honest was made to emphasize the applicability to the data set at hand. If the noise is heteroskedastic then the regions may no longer be honest, but the intention is that they are still relevant for the data at hand. The main data constraints when defining the approximation region are the reasonably accurate determination of the noise level and the requirement that the noise can well be described by the Gaussian model. The work of Dümbgen shows that this latter restriction can be overcome to a large extent by defining the approximation region in terms of the signs or some other function of the residuals (see Dümbgen (1998, 2003, 2006, 2007); Dümbgen and Johns (2004); Dümbgen and Kovac (2009)). In Dümbgen and Kovac (2009) it is shown how the taut string algorithm can be adapted to these more general situations.

The approximation regions contain many functions which will in general not be acceptable. For example, any function which interpolates the data and whose residuals are consequently all zero will lie in the approximation region. If the statistician is interested in local extreme values then a function of interest would be one which minimizes this number subject to its belonging to the approximation region. This would in a sense answer the question as to which local extreme values ‘are really there’. Similarly, if the statistician is interested in smoothness then a function of interest would be one which maximizes this subject to its belonging to the approximation region. The approximation regions considered here are convex. It is therefore in principle possible to minimize any convex measure of complexity. In practice there may be numerical problems because of the large number of linear constraints which define the approximation regions. The number of local extreme values is not convex a convex function and we are lucky that the taut string algorithm performs not only very well in terms of its results, but also it is also very fast. By minimizing the complexity it is possible not only to exhibit a simple function which is consistent with the data but also to claim that this is the (or a) simplest function. In particular, if this simplest function has at least one local extreme value, then the data cannot be adequately approximated by a monotone function.

When comparing two data sets the analysis depends on whether the supports are disjoint or not. If there are not disjoint then it may well be the case that there is no common approximating function. If they are disjoint, then there always will be a joint approximating which can only be eliminated by complexity considerations. These need not be formulated in advance: it may well be that the experimenter cannot formulate these in advance, but, having seen the simplest function consistent with both data sets, is not prepared to accept it.

As already mentioned in Section 1.3 regularization plays an essential role in much of statistics. This is not generally recognized but well known exceptions are Hampel (Hampel et al. (1986)), Huber (Huber (1981)) and Tukey (Tukey (1993)). It is perhaps not a coincidence that all three were instrumental in laying down the foundations of robust statistics where the consideration of perturbations of models is part of the theory. The necessity of regularization derives from the fact that many statistical problems, and particularly those which involve the use of likelihood, are ill-posed. This is simply because the linear differential operator is not bounded and hence not continuous. This is the classical situation where regularization is required. For a more detailed discussion of the role of regularization in statistics and other related topics we refer to Davies (2008).

Acknowledgement

The authors thank an anonymous and sympathetic referee who brought the odyssey of this article to a happy conclusion and whose comments lead to an increase in clarity and precision.

References

Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2):170–176. MR0069229

Davies, P. L. (1995). Data features. *Statistica Neerlandica*, 49:185–245. MR1345378

Davies, P. L. (2004). The one-way table: In honour of John Tukey 1915-2000. *Journal of Statistical Planning and Inference*, 122:3–13. MR2057910

Davies, P. L. (2005). Universal principles, approximation and model choice. Invited talk, European Meeting of Statisticians, Oslo.

Davies, P. L. (2008). Approximating data (with discussion). *Journal of the Korean Statistical Society*, 37:191–240. MR2445029

Davies, P. L., Gather, U., Meise, M., Mergel, D., and Mildenberger, T. (2008a). Residual based localization and quantification of peaks in x-ray diffractograms. *Annals of Applied Statistics*, 2(3):861–886.

Davies, P. L., Gather, U., Nordman, D. J., and Weinert, H. (2008b). A comparison of automatic histogram constructions. *EIMS: Probability and Statistics*. to appear.

Davies, P. L., Gather, U., and Weinert, H. (2008c). Nonparametric regression as an example of model choice. *Communications in Statistics - Simulation and Computation*, 37:274 – 289. MR2422886

Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution (with discussion). *Annals of Statistics*, 29(1):1–65. MR1833958

Davies, P. L., Kovac, A., and Meise, M. (2009). Nonparametric regression, confidence regions and regularization. *Annals of Statistics*. To appear.

Delgado, M. A. (1992). Testing the equality of nonparametric regression curves. *Statistics and Probability Letters*, 17:199–204. MR1229937

Dette, H. and Neumeyer, N. (2001). Nonparametric analysis of covariance. *Annals of Statistics*, 29:1361–1400. MR1873335

Donoho, D. L. (1988). One-sided inference about functionals of a density. *Annals of Statistics*, 16:1390–1420. MR0964930

Dümbgen, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *Annals of Statistics*, 26:288–314. MR1611768

Dümbgen, L. (2003). Optimal confidence bands for shape-restricted curves. *Bernoulli*, 9(3):423–449. MR1997491

Dümbgen, L. (2006). Confidence bands for convex median curves using sign-tests. MR2459932

Dümbgen, L. (2007). Confidence bands for convex median curves using sign-tests. In Cator, E., Jongbloed, G., Kraaikamp, C., Lopuhaä, R., and Wellner, J., editors, *Asymptotics: Particles, Processes and Inverse Problems*, volume 55 of *IMS Lecture Notes - Monograph Series 55*, pages 85–100. IMS, Hayward, USA. MR2459932

Dümbgen, L. and Johns, R. (2004). Confidence bands for isotonic median curves using sign-tests. *J. Comput. Graph. Statist.*, 13(2):519–533. MR2063998

Dümbgen, L. and Kovac, A. (2009). Extensions of smoothing via taut strings. *Electronic Journal of Statistics*, 3:41–75. MR2471586

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics*, 29(1):124–152. MR1833961

Fan, J. and Lin, S. K. (1998). Test of significance when data are curves. *Journal of American Statistical Association*, 93:1007–1021. MR1649196

Hall, P. and Hart, D. H. (1990). Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association*, 85(412):1039–1049. MR1134500

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York. MR0829458

Härdle, W. and Marron, J. S. (1990). Semiparametric comparison of regression curves. *Annals of Statistics*, 18(1):63–89. MR1041386

Höhenrieder, C. (2008). *Nichtparametrische Volatilitäts- und Trendapproximation von Finanzdaten*. PhD thesis, Department of Mathematics, University Duisburg-Essen, Germany.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York. MR0606374

Kabluchko, Z. (2007). Extreme-value analysis of standardized Gaussian increments. arXiv:0706.1849.

King, E., Hart, J. D., and Wehrly, T. E. (1990). Testing the equality of two regression curves using linear smoothers. *Statistics and Probability Letters*, 12:239–247. MR1130364

Kulasekera, K. B. (1995). Comparison of regression curves using quasi-residuals. *Journal of the American Statistical Association*, 90(431):1085–1093. MR1354025

Kulasekera, K. B. and Wang, J. (1997). Smoothing parameter selection for power optimality in testing of regression curves. *Journal of the American Statistical Association*, 92(438):500–511. MR1467844

Lavergne, P. (2001). An equality test across nonparametric regressions. *Journal of Econometrics*, 103:307–344. MR1838202

Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley. MR0702834

Li, K.-C. (1989). Honset confidence regions for nonparametric regression. *Annals of Satistics*, 17:1001–1008. MR1015135

Munk, A. and Dette, H. (1998). Nonparametric comparison of several regression functions: Exact and asymptotic theory. *Annals of Statistics*, 26(6):2339–2368. MR1700235

Neumeyer, N. and Dette, H. (2003). Nonparametric comparison of regression curves - an empirical process approach. *Annals of Statistics*, 31:880–920. MR1994734

Pötscher, B. M. and Leeb, H. (2008). Sparse estimators and the oracle property, or the return of Hodges. *Journal of Econometrics*, 142:201–211. MR2394290

Tukey, J. W. (1993). Issues relevant to an honest account of data-based inference, partially in the light of Laurie Davies's paper. Princeton University, Princeton.