TESTING FOR MONOTONICITY OF A REGRESSION MEAN WITHOUT SELECTING A BANDWIDTH[1]

Peter Hall[2] and Nancy E. Heckman[3]

Technical Report 178, Statistics Department, University of British Columbia

July 1998

ABSTRACT. A new approach to testing for monotonicity of a regression mean, not requiring computation of a curve estimator or a bandwidth, is suggested. It is based on the notion of 'running gradients' over short intervals, although from some viewpoints it may be regarded as an analogue for monotonicity testing of the dip/excess mass approach for testing modality hypotheses about densities. Like the latter methods, the new technique does not suffer difficulties caused by almost-flat parts of the target function. In fact, it is calibrated so as to work well for flat response curves, and as a result it has relatively good power properties in boundary cases where the curve exhibits shoulders. In this respect, as well as in its construction, the 'running gradients' approach differs from alternative techniques based on the notion of a critical bandwidth.

KEYWORDS. Bootstrap, calibration, curve estimation, Monte Carlo, response curve, running gradient.

SHORT TITLE. Testing for monotonicity.

---

[1] The manuscript is available directly from either of the authors, or from www.stat.ubc.ca/∼nancy.

[2] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia.

[3] Department of Statistics, University of British Columbia, 2021 West Mall, Vancouver BC V6T 1Z2, Canada. This author's research was supported by National Sciences and Engineering Research Council of Canada Grant A7969.

1. INTRODUCTION

The potential monotonicity of a response to the level of a stimulus is often of significant practical interest. For example, the size or condition of an animal population, in response to the level of a nutrient, can be an important indicator of the concentration at which the nutrient starts to become toxic. In radiocarbon dating problems, a monotone relationship between true and assessed age is critical to accuracy; and a monotone link between the levels of two medical symptoms is an important indicator of a common cause.

Particularly in cases where the null hypothesis of monotonicity fails, a response curve can be awkward to model parametrically, suggesting that nonparametric approaches to testing are of interest. Bowman, Jones and Gijbels (1998) developed a test based on Silverman's (1981) critical bandwidth method in nonparametric density estimation. They employed a local linear estimator $\hat{g} = \hat{g}_h$ of the response curve, depending on a bandwidth h, and calculated that value of h, h_{crit} say, which was as small as possible subject to $\hat{g}_h$ still being monotone. Proceeding by simulation they computed empirical approximations to critical points, and thereby implemented the test.

While the procedure of Bowman *et al* (1998) has many attractive features, it is clear that it also suffers difficulties. In particular, in cases where the true response function g is flat, or nearly flat, in places, the bandwidth approach can have low power. To appreciate why, consider the setting where g is nondecreasing within a compact interval $\mathcal{I}$ but there exists a subinterval $\mathcal{J} \subseteq \mathcal{I}$ where g is horizontal. Suppose too that the kernel employed in the local linear estimator is compactly supported, that the errors are Normal, and (for the sake of simplicity) that the n design points are regularly spaced through $\mathcal{I}$. Then, as n increases, h_{crit} converges to a positive constant, the value of which depends on the support of the kernel and the width of $\mathcal{J}$. However, the same is true if, at another location, the curve $y = g(x)$ has a sufficiently small dip. Consequently, the existence of a flat section in one part of the response curve can effectively mask a small dip at another place, resulting in zero power for detecting the dip. Similar problems occur in less extreme cases, for example with response curves that have almost-flat parts, and when using infinitely supported kernels such as the Gaussian.

These problems are to be expected, since they are inherited from difficulties that the bandwidth test has in the context of density estimation. There, places where the density is flat or almost flat — for example, in the tails of a distribution — can be almost impossible for the bandwidth test to distinguish from the sites of small modes. Indeed, theoretical regularity conditions under which the bandwidth approach produces a mode test with good power usually require the sampled density to be compactly supported and rise steeply in its tails, so much so that its derivative is bounded away from zero there. See for example Silverman (1983) and Mammon, Marron and Fisher (1992).

A remedy for the 'flatness' problem in the context of density estimation is to permit the bandwidth to have a degree of local variation, or to otherwise incorporate additional tuning parameters. See for example Fisher, Mammen and Marron (1994). Often the tuning is conducted in an *ad hoc* way, although a systematic approach, involving two tuning parameters in addition to the bandwidth, is suggested in a recent, unpublished manuscript of N.I. Fisher and J.S. Marron. One could incorporate such a method into bandwidth tests for monotonicity of response curves, in order to overcome the power problems noted earlier, but it would be complex to implement.

In both the density and regression cases these difficulties persist for very large samples; they do not occur just for small to moderate sample sizes. For example, in the context of testing a response curve for monotonicity it is straightforward to construct two similar-looking curves where (a) the first has a short flat section, and is monotone nondecreasing on $\mathcal{I}$; (b) the second is monotone except for a small dip instead of the flat section; and (c) the critical bandwidth computed from data generated by the j'th model (for $j = 1, 2$) converges in probability to a constant C_j, where $0 < C_2 < C_1 < \infty$. Therefore, from the viewpoint of the bandwidth test the second, non-monotone response curve looks more monotone than the first, monotone curve, even for very large samples.

We argue that a test for monotonicity should have reasonable power in marginal cases such as those discussed above, where it is difficult to distinguish a shoulder or a flat section in an increasing function from a small, downwards dip. Motivated by this idea, in the present paper we propose an alternative approach which eschews the notion of a bandwidth and, instead, focuses on 'running gradient' estimates over relatively short intervals. In a sense, our approach represents a version, in the context of testing for monotonicity in regression, of the dip test and excess mass approaches of Hartigan and Hartigan (1985) and Müller and Sawitzki (1991), respectively, for testing modality hypotheses. (These two tests are numerically equivalent.)

The dip/excess mass method has the advantage over the bandwidth test of being relatively immune to problems caused by flatness of the tails of the sampled density, and so it is to be expected that something similar would be true in the present setting. Indeed, our 'running gradients' method does not suffer problems caused by flat parts of the curve; our method is in fact calibrated for the case where the response curve is flat. Moreover, it is sensitive to any small dips in the curve, whether appearing in a flat section or elsewhere.

The latter property is demonstrated in our theoretical work in section 4; see in particular point (2) in section 4.1. Section 2 outlines our methodology and describes its implementation, and section 3 discusses numerical performance. There we avoid the obvious numerical demonstration of properties discussed above, since the greater power of our method for detecting dips in response curves that have almost-flat sections is self-evident. Instead, we focus on the class of functions introduced during the simulation study of Bowman *et al* (1998), so as to compare our approach with theirs in a setting where their method might be expected to perform well. Their class of response surfaces is governed by a single tuning parameter, a, and we show that our method is superior to theirs, in the sense of having greater power, if and only if a exceeds a certain threshold.

Schlee (1982) has also considered the problem of testing for monotonicity of a response surface, although not really from a practicable viewpoint. Nonparametric regression subject to constraints of monotonicity has been treated by, for example, Ramsay (1988) and Mammen (1991). There is an extensive literature on issues of modality, including testing, in the context of density estimation; it includes work of Cox (1966), Good and Gaskins (1980), Minnotte and Scott (1992), Roeder (1994) and Polonik (1995a,b).

2. METHODOLOGY

2.1. Test statistic. Suppose data $\mathcal{X} = \{(x_i, Y_i),\ 1 \le i \le n\}$ are generated by the model $Y_i = g(x_i) + \epsilon_i$, where g is a smooth function, the x_i's (which may be either conditioned values of random variables, or regularly-spaced points) are distributed in a compact interval $\mathcal{I}$, and the errors ϵ_i are independent and identically distributed with zero mean and variance σ^2 for some $\sigma > 0$. We wish to test the null hypothesis H_0 that g is nondecreasing on the interval $\mathcal{I}$.

Our test statistic is defined as follows. Let $0 \le r \le s-2 \le n-2$ be integers, let a, b be constants, and put

$$S(a, b|r, s) = \sum_{i=r+1}^{s} \{Y_i - (a + bx_i)\}^2 . \tag{2.1}$$

For each choice of (r, s), define $\hat{a} = \hat{a}(r, s)$ and $\hat{b} = \hat{b}(r, s)$ by

$$(\hat{a}, \hat{b}) = \mathrm{argmin}_{(a,b)}\, S(a, b|r, s) ,$$

and let

$$Q(r, s)^2 = \sum_{i=r+1}^{s} \left\{ x_i - (s-r)^{-1} \sum_{j=r+1}^{s} x_j \right\}^2 .$$

Then, $\hat{b}(r, s)\, Q(r, s)$ has variance σ^2 for each pair (r, s). We incorporate this standardisation into our test statistic, which is

$$T_m = \max \big\{ -\hat{b}(r, s)\, Q(r, s) : 0 \le r \le s-m \le n-m \big\} , \tag{2.2}$$

where m, satisfying $2 \le m \le n$, is an integer. The test consists of rejecting H_0 if the value of T_m is too large.

From some points of view m plays the role of a smoothing parameter, in that choosing m relatively large tends to 'smooth out' (and consequently reduce) the effects of outlying data values. Thus, larger values of m provide greater resistance against the effects of heavy-tailed error distributions, and this is the principal reason for taking m to be other than 2.

Alternative but related definitions of T_m have useful features that T_m lacks. They include, for example,

$$T_{1m}(u) = \max \big\{ s - r : \hat{b}(r, s)\, Q(r, s) \le -u \quad \text{and} \quad 0 \le r \le s-m \le n-m \big\} ,$$
$$T_{2m}(u) = \max \big\{ x_s - x_r : \hat{b}(r, s)\, Q(r, s) \le -u \quad \text{and} \quad 0 \le r \le s-m \le n-m \big\} ,$$

where u denotes a positive number that may be interpreted as the 'average' standard error of $\hat{b}(r, s)\, Q(r, s)$. (Here and below we take the left-hand side to be 0 if the set on the right-hand side is empty.) Unlike T_m, the statistics T_{1m} and T_{2m} focus specifically on the lengths of runs of design points where the average gradient is negative, length being measured in terms of number of design points and distance between design points, respectively. However, the need to select u as a precursor to tests based on T_{1m} and T_{2m} makes such statistics a little less attractive.

2.2. Calibration. Our approach to calibration is based on the fact that a constant function is the most difficult nondecreasing form of g for which to test. Thus, we develop approximations to the distribution of T_m when $S(a, b|r, s)$ is given by

$$S(a, b|r, s) = \sum_{i=r+1}^{s} \{\epsilon_i - (a + bx_i)\}^2 \,, \tag{2.3}$$

instead of by the definition at (2.1). We shall show in section 4 that, provided m increases sufficiently fast, asymptotically correct levels may be obtained by calibrating as though the errors were Normal. 'Sufficiently fast' can actually be quite slow. For example, if the sampling distribution has a moment generating function in the neighbourhood of the origin then m need only increase faster than $(\log n)^2$ in order for calibration based on approximation by the Normal-error case to be asymptotically valid.

Calibration for Normal errors would usually involve application of the parametric bootstrap, as follows. First, compute an estimator $\hat{\sigma}^2$ of σ^2, using any of a variety of different methods (see e.g. Rice 1984; Gasser, Sroka and Jennen-Steinmetz 1986; Buckley, Eagleson and Silverman 1988; Buckley and Eagleson 1989; Hall and Marron 1990; Hall, Kay and Titterington 1990; Carter and Eagleson 1992; and Seifert, Gasser and Wolf 1993). Then, conditional on $\hat{\sigma}^2$, simulate values of ϵ_i from the Normal $N(0, \hat{\sigma}^2)$ distribution, and thereby compute Monte Carlo approximations to the distribution of first $S(a, b|r, s)$ (defined on this occasion by (2.3) for Normal errors) and then to the distribution of T_m. In particular, compute an approximation $\hat{t}_{m,\alpha}$ to the point $t_{m,\alpha}$ such that $P_{0,\mathrm{Norm}}(T_m > t_{m,\alpha}) = \alpha$, where $P_{0,\mathrm{Norm}}$ denotes probability measure under the model where g is identically constant and the errors are Normal $N(0, \sigma^2)$. Reject the null hypothesis if the value of T_m computed from the data, this time using the definition of S at (2.1), exceeds $\hat{t}_{m,\alpha}$.

More generally, if the error distribution were known up to a vector of parameters then the step of simulating from the $N(0, \hat{\sigma}^2)$ distribution would be replaced by simulating from the model for the error distribution, with unknown parameter values replaced by estimates. Alternatively we may use nonparametric bootstrap methods, as follows. Let $\hat{g}$ be a consistent estimator of g, for example computed by local linear methods, and calculate the residuals $\hat{\epsilon}_i = Y_i - \hat{g}(x_i)$. We might wish to centre or rescale these quantities, for example so that their average value is zero and their (sample) variance is the same as a standard estimate of σ^2. Centring does not affect our estimate of the distribution of T_m under H_0, however. Note too that we do not need to compute $\hat{\epsilon}_i$ for all values of i. In particular, values corresponding to design points at the ends of $\mathcal{I}$ might be considered unreliable because of excessive bias or variance, and not be used for that reason.

Let $\mathcal{E}$ be the set of residuals that we have computed, and let $\epsilon_1^*, \ldots, \epsilon_n^*$ denote a resample drawn by sampling randomly, with replacement, from $\mathcal{E}$. Put

$$S^*(a,b|r,s) = \sum_{i=r+1}^{s} \{\epsilon_i^* - (a + bx_i)\}^2 , \quad (\hat{a}^*, \hat{b}^*) = \mathrm{argmin}_{(a,b)} S^*(a,b|r,s) ,$$

$$T_m^* = \max\big\{ -\hat{b}^*(r,s)\, Q(r,s) : 0 \le r \le s - m \le n - m \big\} . \tag{2.4}$$

The nonparametric bootstrap estimator of the α-level critical point of T_m under the null hypothesis is $\tilde{t}_{m,\alpha}$, defined by

$$P(T_m^* > \tilde{t}_{m,\alpha} | \mathcal{X}) = \alpha .$$

In a test at the α level we would reject H_0 if $T_m > \tilde{t}_{m,\alpha}$, where T_m now denotes the value of the test statistic computed from the data.

2.3. Heteroscedasticity. If the errors ϵ_i, conditional on the x_i's, may be assumed identically distributed except for a scale factor that depends on x_i, then we should estimate scale. We may parametrically model the function $\sigma(\cdot)$ defined by $\sigma(x_i)^2 = \mathrm{var}(\epsilon_i | x_i)$, and estimate the parameters of the model; or we might use nonparametric methods to estimate conditional variance. Either way, we obtain an estimator $\hat{\sigma}(x_i)^2$ of $\sigma(x_i)^2$, which should be constrained to be bounded away from zero. If the weight $\hat{\sigma}(x_i)^{-2}$ is incorporated into the series at (2.1) and (2.3) then our method may proceed as before. For the sake of simplicity and brevity, however, we shall confine attention to the homoscedastic case when describing properties of the test. Concise results in the general case depend on the accuracy with which we may estimate $\sigma(\cdot)$.

3. NUMERICAL PROPERTIES AND EXAMPLE

3.1. Distribution of T_m when g is constant. Without loss of generality, $g \equiv 0$. Define T_m as at (2.2). In each of 500 simulations we generated data $Y_i = \epsilon_i$, for

$i = 1, \ldots, n = 100$, where the ϵ_i's were independent and identically distributed with mean zero. Our test statistics were based on $S(a, b|r, s)$, defined as at (2.3) with $x_i = i/(n+1)$ (i.e. equally spaced x_i's).

Two error distributions were considered: ϵ_i Normally distributed with standard deviation 0.1 (the Normal calibration model), and $\epsilon_i = \rho W$, where W had Student's t distribution with 5 degrees of freedom, denoted below by t_5. The value of ρ was chosen so that the interquartile range of ϵ_i was the same as that for the Normal distribution with standard deviation 0.1. Note that t_5 is the lowest-index Student's t distribution for which the fourth moment is finite. We addressed the distributions of both $T_m/0.1$ and $T_m/\hat{\sigma}$, where $\hat{\sigma}^2$ was the variance estimator proposed by Rice (1984):

$$\hat{\sigma}^2 = \frac{1}{2\,(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2 \,. \tag{3.1}$$

When g is constant the distribution of $T_m/\hat{\sigma}$ does not depend on σ.

Figure 1 graphs the 5'th, 10'th, 50'th, 90'th and 95'th pointwise percentiles of the distribution of $T_m/0.1$ (panels (a) and (b)) and $T_m/\hat{\sigma}$ (panels (c) and (d)) against m. As expected, the quantiles are monotone decreasing functions of m; and the statistics are more variable and stochastically larger when the errors are distributed as t_5 than they are for Normal errors. However, standardising by dividing by $\hat{\sigma}$ tends to counteract much of that variability.

[*Place figure 1 near here, please* .]

3.2. Testing procedures. We considered four tests: *Tp*, *Tnp*, *Tst* and *Bow*. Tests *Tp* and *Tnp* employed the statistic T_m with p-values derived using the parametric and nonparametric bootstrap, respectively, described in section 2.2. In the case of *Tp* the ϵ_i^*'s were Normal with mean 0 and variance $\hat{\sigma}^2$, defined at (3.1). For *Tnp*, residuals were computed from a local linear fit with automatic bandwidth choice, using the Splus routine `KernSmooth` written by M.P. Wand and described by Wand and Jones (1995). The test *Tst* employed the Studentised statistic $T_m/\hat{\sigma}$, with p-value calculated using the simulated distribution of $T_m/\hat{\sigma}$, with Normal errors, from section 3.1. The test *Bow* is described by Bowman *et al* (1998), and is based on the bandwidth of a local linear estimator of g. The code for *Bow* was kindly supplied by A.W. Bowman. The number of bootstrap resamples for methods *Tp*, *Tnp* and *Bow* was 300.

We simulated data $Y_i = g(x_i) + \epsilon_i$, for $i = 1, \ldots, n = 100$, with $x_i = i/(n+1)$ and either $g \equiv 0$ or $g(x) = g_a(x) \equiv 1 + x + a \exp\{-50\,(x - .5)^2\}$, the latter being the function employed by Bowman *et al* (1998) in their numerical work. As in that study we took $a = 0,\ 0.15,\ 0.25,\ 0.45$. Note that g is strictly increasing for $a = 0$ and 0.15, but is not monotone for $a = 0.25$ or $a = 0.45$. We used the two error distributions introduced in section 3.1.

Panels (a) and (b) of figure 2 graph Monte Carlo approximations to the actual level when the nominal level is 0.05, in the case of tests *Tp* and *Tnp*, respectively, and for Normal errors. The actual level for the test *Tst* is of course exactly 0.05, and that for *Bow* is 0.060 in the same setting. Thus, *Tp* and *Bow* are comparable. This is true also at level nominal 0.10, where both have an actual level of about 0.11, but at nominal level 0.25 the actual level of *Bow* is about 0.34 (compared with about 0.27 for *Tp*).

Panels (c)–(e) of figure 2 give Monte Carlo approximations to exact levels of *Tp*, *Tnp* and *Tst*, for t_5 errors and nominal level $\alpha = 0.05$, after calibration as though the errors were Normal. (See section 2.2 for discussion. Graphs for Normal errors, analogous to those in panels (c)–(e), would be horizontal lines at height 0.05.) Level accuracy for $m \geq 20$ is good, and as noted in section 2, power can be expected to decrease as m increases. Therefore we take $m = 20$ in the remainder of this section. The *Bow* test tends to be more conservative than *Tp*, *Tnp* and *Tst*, its actual level being 0.030 in the case of t_5 errors after the same calibration. The relative conservatism of *Bow* persists for nominal levels 0.10 and 0.25.

[*Place figure 2 near here, please* .]

Next we treat the case $g \equiv g_a$, using throughout the Normal calibration method mentioned in the previous paragraph, taking $\alpha = 0.05$ and $m = 20$, and employing either Normal or t_5 errors. When $a = 0.45$ and errors are Normally distributed, Monte Carlo estimates of the probability of rejecting the null hypothesis are 0.932, 0.964 and 0.936 for the tests *Tp*, *Tnp* and *Tst*, respectively, but only 0.758 in the case of *Bow*. For $a = 0.45$ and t_5 errors the rejection rates are 0.744, 0.784 and 0.742 for *Tp*, *Tnp* and *Tst*, but only 0.536 for *Bow*. On the other hand, *Bow* enjoys greater power in the marginal case $a = 0.25$. There it rejects the null with probability 0.202 for Normal errors, and 0.122 for t_5 errors, whereas the rejection rates of *Tp*, *Tnp* and *Tst* are below the nominal significance level.

This reflects a feature that is apparent from the construction of our tests: in cases where g is monotone for all, or almost all, of $\mathcal{I}$, the values of our test statistics tend to be swamped by the influence of regions where g is monotone, and so tend to under-reject. In particular, when $a = 0$ or 0.15, in which case the null hypothesis of monotonicity holds, our tests reject the null at rates substantially below the significance level α, whereas the rejection rate of *Bow* is only a little below α. More extensive simulation studies show that (when $g \equiv g_a$, $\alpha = 0.05$ and $m = 20$) the break-even point for the performance of our tests relative to *Bow* is approximately $a_0 \equiv 0.415$, in the sense that our tests tend to have greater power than *Bow* if $a > a_0$ but not otherwise. When $a = a_0$ and errors are Normal, the rejection rate of all four tests (i.e. *Tp*, *Tnp*, *Tst* and *Bow*) is approximately 0.8.

3.3 Analysis of radio carbon dating data. Bowman *et al* (1998) analysed a data set of size 49, extracted from Pearson and Qua (1993) and representing the known calendar age, and the age estimated by radiocarbon dating, for 49 specimens. The purpose of the analysis was to determine if the relationship between the two ages was monotone, since monotonicity could be used to more accurately calibrate the radiocarbon dating process. In this example, all three of our tests gave relatively high p-values for all values of m. The p-values were smallest when $m = 4$, indicating that a least-squares linear fit over four points had a negative slope. However, even for $m = 4$ the slope was significant only at $\alpha = 0.75$ (for *Tp*) and above, suggesting that there was insufficient evidence to reject the monotonicity hypothesis. This is in contrast to conclusions of Bowman *et al* (1998).

4. THEORETICAL PROPERTIES

4.1. Properties of distribution of T_m when g is not constant. Result (1) below confirms that, by taking the null hypothesis to assert that g is identically constant, we obtain a test that is conservative against other nondecreasing functions g. Result (2) shows that if g is strictly decreasing over some portion of the interval supporting the design, then the test statistic T_m assumes values at least as large as $n^{1/2}$. In section 4.2 we shall demonstrate that, for tests calibrated using either the Normal approximation or bootstrap methods, critical points are an order of magnitude smaller than $n^{1/2}$; they are in fact of size n^δ for all $\delta > 0$. Together, these results confirm that (i) large values of the test statistic T_m indicate that g is not nondecreasing, (ii) our method of calibration tends to be conservative, and (iii) our calibration method produces tests for which power increases to 1 with increasing sample size.

As a prelude to stating our results, write P_0 for the probability measure P when g is taken to be identically constant but the error distribution is the same as that for the data. Assume that the error distribution has finite variance, and that either (a) the design points x_i are conditioned values of a sequence of independent and identically distributed random variables from a continuous distribution on $\mathcal{I} = [0,1]$, with density bounded away from zero and infinity, or (b) the design points are regularly spaced on $\mathcal{I} = [0,1]$. In case (a), probability statements should be interpreted as holding with respect to classes of sequences $\{x_i\}$ that arise with probability 1. The following results hold.

(1) For any nondecreasing function g, for all $2 \leq m \leq n$ and for all $0 \leq x < \infty$, $P(T_m \leq x) \leq P_0(T_m \leq x)$.

(2) If there exists a nondegenerate interval $\mathcal{J} \subseteq \mathcal{I}$ on which the derivative of g exists and is bounded below 0, then T_m is of size at least $n^{1/2}$, uniformly in values m satisfying $m = o(n)$, in the following sense. There exists $C > 0$ such that, for each sequence $m_0 = m_0(n) \to \infty$ with $m_0 = o(n)$,

$$\inf_{2 \leq m \leq m_0(n)} P\big(T_m > C\, n^{1/2}\big) \to 1$$

as $n \to \infty$.

4.2. Properties of null distribution. We take the null distribution to be that which arises under the assumption that g is constant, and do not insist that the errors be Normally distributed. Result (1) below shows that under this null hypothesis, and provided the error distribution is reasonably light-tailed, T_m is of smaller order than n^{δ} for any positive δ. This property should be contrasted with (2) in section 4.1, where we showed that if at least some portion of g is decreasing then T_m is of size at least $n^{1/2}$. Therefore, T_m can distinguish nondecreasing functions g from functions that do not have this property. However, result (2) below shows that heavy-tailedness of the error distribution does tend to increase the size of T_m under the null hypothesis, and so can be expected to reduce power. This property motivated us to incorporate the tuning parameter m in our definition of T_m.

Result (3) below demonstrates that, provided m is chosen large enough (depending on the tailweight of the error distribution), the null distribution of T_m

is asymptotically equivalent to its form in the case of Normal errors. This motivated our suggestion that the Normal-error approximation be considered as one approach to calibration. Result (4) shows that in the case of Normal errors the size of T_m grows very slowly indeed, at rate $(\log n)^{1/2}$. Finally, (5) establishes that our nonparametric bootstrap approach to calibration produces statistically consistent results.

Let $P_{0,\mathrm{Norm}}$ denote P_0-measure when the errors ϵ_i are Normally distributed with variance σ^2. Assume g is identically constant, and that the design sequence satisfies either (a) or (b) of section 4.1. Then the following results hold.

(1) If the error distribution has all moments finite then T_m is of smaller order than n^δ for all $\delta > 0$, in the sense that

$$\text{for all } \delta > 0\,, \quad \sup_{2\leq m\leq n} P_0\big(T_m > n^\delta\big) \to 0$$

as $n \to \infty$.

(2) If the error distribution has the property that $P(|\epsilon| > x) \geq \text{const.}\, x^{-c}$ for all sufficiently large x, then for each fixed $m \geq 2$,

$$\liminf_{n\to\infty} P_0\big(T_m > Cn^{1/c}\big) \to 1$$

as $C \to 0$.

(3) If either (i) $E|\epsilon|^c < \infty$ and $m = m(n)$ satisfies $n^{2/c}(\log n)^{1+\epsilon} = O(m)$ and $m = O(n^\rho)$, for some $c > 3$, $\epsilon > 0$ and $\rho \in (2/c, 1)$; or (ii) the distribution of ϵ has a finite moment generating function in some neighbourhood of the origin, and $m = m(n)$ satisfies $(\log n)^{3+\epsilon} = O(m)$ and $m = O(n^\rho)$ for some $\epsilon > 0$ and $\rho \in (0, 1)$; then the distribution of T_m is asymptotically equivalent to that in the case of Normal errors (with the same variance σ^2), in the sense that

$$\sup_{0\leq x<\infty} \big|P(T_m \leq x) - P_{0,\mathrm{Norm}}(T_m \leq x)\big| \to 0$$

as $n \to \infty$. Moreover, this result remains true if, when effecting the Normal-error approximation, we employ a variance σ_n^2 that is within $O(n^{-\delta})$ of the true variance σ^2, for some $\delta > 0$. (This property justifies our empirical calibration by Normal-error approximation, using an estimate of σ^2.)

(4) If the error distribution is Gaussian then, uniformly in $2 \leq m \leq n^\rho$ for each $0 < \rho < 1$, T_m is of exact size $(\log n)^{1/2}$, in the sense that for each ρ there exist

constants $0 < t_1 < t_2 < \infty$ such that

$$\sup_{2 \le m \le n^\rho} \left| P_{0,\mathrm{Norm}}\left\{ t_1 \le T_m/(\log n)^{1/2} \le t_2 \right\} - 1 \right| \to 0$$

as $n \to \infty$.

(5) Suppose the error distribution is continuous with density f, let F denote the corresponding distribution function, and put $\bar{F} = \min(F, 1-F)$. If the random function $\hat{g}$ used to compute the residuals $\hat{\epsilon}_i = Y_i - \hat{g}(x_i)$ satisfies $\sup_{y_1 \le x \le y_2} |\hat{g}(x) - g(x)| = O_p(n^{-\eta})$ for some $\eta > 0$ and $0 \le y_1 < y_2 \le 1$; if we compute all those residuals ϵ_i that correspond to design points x_i satisfying $y_1 \le x_i \le y_2$; if $f \ge C\,\bar{F}\,|\log \bar{F}|^\gamma$ for constants $C, \gamma > 0$; and if $m = m(n)$ satisfies $(\log n)^{3+\epsilon} = O(m)$ and $m = O(n^\rho)$ for some $\epsilon > 0$ and $0 < \rho < 1$; then the bootstrap distribution of T_m^* is consistent for the unconditional null distribution of T_m, in the sense that

$$\sup_{0 \le x < \infty} |P(T_m^* \le x | \mathcal{X}) - P_0(T_m \le x)| \to 0$$

in probability as $n \to \infty$.

The assumption on f in (5) is just a little stronger than the requirement that the error distribution have finite moment generating function, and in this sense is a version of condition (ii) in result (3) above. Likewise, we may formulate and prove result (5) under a slightly stronger form of condition (i) in (3).

APPENDIX

A.1. Outline proofs of results in section 4.1. (1) Let T_{0m} have the distribution of T_m at (2.2), but in the case where $S(r,s)$ is defined by (2.3) instead of (2.1). (Thus, P-measure for T_{0m} is equivalent to P_0-measure for T_m.) Put $\bar{x}_{rs} = (s - r)^{-1} \sum_{r+1 \le j \le s} x_j$, $w_i = w_i(r,s) = (x_i - \bar{x}_{rs})/Q(r,s)^2$ if $r+1 \le i \le s$, and $w_i = 0$ otherwise. Then $T_m = T_{0m} + \Delta_m$ where $\Delta_m = \sum_i w_i\, g(x_i)$. If g is nondecreasing then $\Delta_m \ge 0$, and so

$$P_0(T_m \le x) = P(T_{0m} \le x) = P(T_m \le x + \Delta_m) \ge P(T_m \le x)\,.$$

(2) Let $\mathcal{J}$ denote an interval from c_1 to $c_2 > c_1$, and suppose the derivative of g is bounded below $-C_1$ on $\mathcal{I}$, where $C_1 > 0$. Let $c_1 < c_3 < c_4 < c_2$, and put

$$\beta(r,s) = E\{\hat{b}(r,s)\} = \sum_i w_i\, g(x_i)\,, \quad \Delta(r,s) = \hat{b}(r,s) - \beta(r,s)\,,$$

$$\mathcal{S}_m = \Big\{ (r,s) : 0 \le r \le s - m \le n - m\,,\ c_3 \le x_{r+1} \le x_s \le c_4 \quad \text{and} \quad x_s - x_{r+1} \ge \tfrac{1}{2}\,(c_4 - c_3) \Big\}\,.$$

Then there exist $C_2, C_3 > 0$ such that

$$\sup_{(r,s)\in\mathcal{S}_m} \beta(r,s) \le -C_2 \quad \text{and} \quad \inf_{(r,s)\in\mathcal{S}_m} Q(r,s)^2 \ge C_3^2\, n$$

for all sufficiently large n; and for all $\delta > 0$,

$$\sup_{(r,s)\in\mathcal{S}_m} P\big\{\big|\hat{b}(r,s) - \beta(r,s)\big| > \delta\big\} \to 0$$

as $n \to \infty$. Therefore, for all $\delta > 0$,

$$\sup_{(r,s)\in\mathcal{S}_m} P\big\{ - \hat{b}(r,s)\, Q(r,s) \ge (1-\delta)\, C_2\, C_3\, n^{1/2}\big\} \to 0\,,$$

which implies the desired result.

A.2. Outline proofs of results in section 4.2. (1) Since g is constant, $\hat{b}(r,s) = \sum_i w_i \epsilon_i$. Hence, by Rosenthal's inequality (Hall and Heyde 1980, p. 23) we have for all $k \ge 1$,

$$E_0\big\{\hat{b}(r,s)^{2k}\big\} \le B(k)\left\{Q(r,s)^{-2k} + E\big(|\epsilon|^{2k}\big) \sum_{i=r+1}^{s} w_i^{2k}\right\},$$

where $B(k)$ depends only on k and E_0 denotes expectation in P_0-measure. Now, $|w_i| \le (x_s - x_r)/Q(r,s)^2$ for $r+1 \le i \le s$, and, uniformly in

$$(r,s) \in \mathcal{R} = \big\{(r,s) : 0 \le r \le s-m \le n-m \quad \text{and} \quad 2 \le m \le n\big\}\,,$$

we have $(s-r)\,(x_s - x_r)^2 = O\{Q(r,s)^2\}$. (Here and below, in the case where $\{x_1, \ldots, x_n\}$ represents an ordered sequence of independent and identically distributed random variables, 'order' statements should be interpreted as holding with probability 1 with respect to such sequences.) Therefore,

$$\begin{aligned} Q(r,s)^{2k} \sum_{i=r+1}^{s} w_i^{2k} &\le Q(r,s)^{2k+2} \max_{r+1\le i\le s} w_i^{2(k-1)} \\ &\le Q(r,s)^{-2(k-3)} (x_s - x_r)^{2(k-1)} = O(1) \end{aligned}$$

uniformly in $(r,s) \in \mathcal{R}$. Hence,

$$\sup_{(r,s)\in\mathcal{R}} E_0\big\{|\hat{b}(r,s)\, Q(r,s)|^{2k}\big\} = O(1)\,,$$

and so by Markov's inequality, for all $k \ge 1/\delta$,

$$\begin{aligned} \sup_{2\le m\le n} P\big(T_m > n^{\delta}\big) &\le \sup_{2\le m\le n} \sum\sum_{(r,s):0\le r\le s-m\le n-m} P\big\{|\hat{b}(r,s)\, Q(r,s)| > n^{\delta}\big\} \\ &= O\big(n^2 .\, n^{-2k\delta}\big) \to 0\,. \end{aligned}$$

(2) Let ν denote the integer part of $(n-m)/m$, and for $0 \le j \le \nu$ put

$$R_j = -\hat{b}\{mj, m(j+1)\}\, Q\{mj, m(j+1)\}\,.$$

Then the R_j's are stochastically independent random variables, and so

$$P(T_m > t) \ge P\Big(\max_{0\le j\le \nu} R_j > t\Big) = 1 - \prod_{j=1}^{\nu} P(R_j \le t)\,. \tag{A.1}$$

Since $P(|\epsilon| > x) \ge \text{const.}\, x^{-c}$ then either or both $P(\epsilon^+ > x) > \text{const.}\, x^{-c}$ or $P(\epsilon^- > x) > \text{const.}\, x^{-c}$. Without loss of generality the former is true. Given constants $0 < B_1 < B_2 < \infty$, let $\mathcal{J}(B_1, B_2)$ denote the set of all $j \in [1, \nu]$ such that (a) $w_i \le -B_1$ for some $i \in [mj+1, m(j+1)]$, and (b) $|w_i| \le B_2$ for all $i \in [mj+1, m(j+1)]$. In view of the origin of the x_i's, for any $0 < u < 1$ we may choose B_1, B_2 such that the number of elements of $\mathcal{J}(B_1, B_2)$ exceeds νu for all sufficiently large n. Take $u = \frac{1}{2}$, and select B_1, B_2 accordingly.

Let $C_1, C_2, \ldots$ denote positive constants. Since $P(\epsilon^+ > x) > C_1 x^{-c}$ for large x then if $j \in \mathcal{J}(B_1, B_2)$,

$$P(R_j > x) \ge P\bigg(B_1\, \epsilon_1 - B_2 \sum_{j=2}^{m} |\epsilon_j| > x\bigg) \ge C_2\, x^{-c}$$

for large x, where C_2 depends only on m, B_1, B_2 and the error distribution. Therefore, by (A.1),

$$\begin{aligned} P\big(T_m > C\, n^{1/c}\big) &\ge 1 - \big(1 - C_2 C^{-c} n^{-1}\big)^{\nu} \\ &\ge 1 - \exp\big(-C_2 C^{-c} \nu n^{-1}\big) \ge 1 - \exp\big(-C_3 C^{-c}\big)\,, \end{aligned}$$

where $C_3 > 0$ does not depend on C. The desired result follows from this formula.

(3) Without loss of generality, $\sigma = 1$. Let $N_1, N_2, \ldots$ denote independent standard Normal random variables, and define $U_i = \sum_{1\le j\le i} \epsilon_j$ and $V_i = \sum_{1\le j\le i} N_j$. Put $W_i = w_{i+1} - w_i$. If g is identically constant then

$$\hat{b}(r,s) = \sum_{i=1}^{n} w_i\, (U_i - U_{i+1}) = \sum_{i=1}^{n} W_i\, U_i\,. \tag{A.2}$$

When the error distribution has finite moment generating function in a neighbourhood of the origin, Theorem 1 of Komlós, Major and Tusnády (1976) implies that the N_j's may be chosen so that $|U_i - V_i| = O_p(\log n)$, uniformly in $1 \le i \le n$, as $n \to \infty$. Hence, by (A.2),

$$\hat{b}(r,s) = \sum_{i=1}^{n} W_i\, V_i + O_p\bigg\{(\log n) \sum_{i=1}^{n} |W_i|\bigg\}\,, \tag{A.3}$$

uniformly in $(r,s) \in \mathcal{R}_m = \{(r,s) : 0 \le r \le s - m \le n - m\}$. In the cases of random design and regularly-spaced design,

$$Q(r,s) \sum_{i=1}^{n} |W_i| = O\{(x_s - x_r)/Q(r,s)\} = O\big\{(s-r)^{-1/2}\big\} \tag{A.4}$$

uniformly in $(r,s) \in \mathcal{R}_m$. Combining (A.3) and (A.4) we see that if $(\log n)^{3+2\epsilon} = O(m)$ for some $\epsilon > 0$ then

$$\hat{b}(r,s)\, Q(r,s) = Q(r,s) \sum_{i=1}^{n} W_i\, V_i + O_p\big\{(\log n)^{-(1/2)-\epsilon}\big\}\,, \tag{A.5}$$

uniformly in $(r,s) \in \mathcal{R}$, as $n \to \infty$. If $E|\epsilon|^c < \infty$ for some $c > 3$ then a similar argument, using Theorem 2 rather than Theorem 1 of Komlós, Major and Tusnády (1976), shows that (A.3) holds with $\log n$ replaced by $n^{1/c}$ and 'big oh' replaced by 'little oh.' This again leads to (A.5), provided $n^{2/c}(\log n)^{1+2\epsilon} = O(m)$.

The first term on the right-hand side of (A.5) has the distribution that T_m would enjoy if the errors were Normally distributed with the same variance as the original errors. Therefore, the claimed result will follow from (A.5) if we show that in the Normal-error case, stochastic perturbations of smaller order than $\delta_n = (\log n)^{-1/2}(\log\log n)^{-1}$ have asymptotically negligible effect on the distribution of T_m. In establishing this property we shall consider only the case of regularly-spaced design, where $x_i = i/n$.

Let $T_{m,\mathrm{Norm}}$ denote the version of T_m in the case of standard Normal errors (and $g \equiv$ const.), let $W(\cdot)$ be a standard Wiener process on the positive half-line, and for $0 < s \le 1$ put

$$T_1(s) = \sup\left\{ \big(12/v^3\big)^{1/2} \int_t^{t+v} \big(u - t - \tfrac{1}{2}v\big)\, dW(u) : 0 \le t \le 1 - v\,,\ s \le v \le 1 \right\},$$

$$T_2(s) = \sup\left\{ \big(3/v^3\big)^{1/2} \int_t^{t+v} \{W(t+v) + W(t) - 2\,W(u)\}\, du : \right.$$
$$\left. 0 \le t < s^{-1} - v\,,\ 1 \le v < s^{-1} \right\}.$$

It can be proved that in the case of regularly-spaced design, and for sufficiently small $\delta > 0$, the process W may be chosen (depending on n) so that $T_{m,\mathrm{Norm}} = T_1(m/n) + O_p(n^{-\delta})$. Furthermore, it may be proved that

$$\limsup_{s \downarrow 0}\ (\log|\log s|) \sup_{-\infty < x < \infty} P\big\{T_2(s) \in \big(x, x + \epsilon\,|\log s|^{-1/2}\,(\log|\log s|)^{-1}\big)\big\} \to 0$$

as $\epsilon \downarrow 0$; and $T_1(s)$ and $T_2(s)$ have identical distributions. (We may obtain $T_1(s)$ from $T_2(s)$ by replacing $W(\cdot)$ in the definition of the latter by $W_s(\cdot)$, where $W_s(t) =$

$s^{-1/2}W(st)$, and integrating by parts in the formula for the integral.) The conditions imposed on m in result (3) in section 4.2 imply that, with $s = m/n$, we have $|\log s| \asymp \log n$. This establishes the desired immunity of T_m to perturbations of smaller order than δ_n.

(4) Without loss of generality, $\sigma = 1$. Under the assumption that the error distribution is Normal $N(0,1)$, each $-\hat{b}(r,s)\,Q(r,s)$ is Normal $N(0,1)$. Since T_m equals the maximum of at most n^2 such variables then with $t = C_1\,(\log n)^{1/2}$ and $C_1 > 2$ we have

$$P_{0,\mathrm{Norm}}(T_m > t) \le n^2\, P\{N(0,1) > t\} \le n^2\, \exp\big(-\tfrac{1}{2}\, C_1^2\, \log n\big) \to 0\,.$$

Therefore,

$$\inf_{2 \le m \le n} P_{0,\mathrm{Norm}}\big\{T_m/(\log n)^{1/2} \le C_1\big\} \to 1\,. \tag{A.6}$$

Let m_0 equal the integer part of n^ρ, let ν equal the integer part of $(n-m_0)/m_0$, and put $N_j = -\hat{b}\{m_0 j, m_0(j+1)\}\, Q\{m_0 j, m_0(j+1)\}$. Then the N_j's are independent standard Normal random variables, and so

$$P_{0,\mathrm{Norm}}(T_m \le t) \le P_{0,\mathrm{Norm}}\Big(\max_{0 \le j \le \nu} N_j \le t\Big) = P\{N(0,1) \le t\}^\nu\,.$$

Now, $\nu > n^{(1-\rho)/2}$ for all sufficiently large n, and $P\{N(0,1) > t\} \ge \exp(-t^2)$ for all sufficiently large t. Hence, if $t = C_2\,(\log n)^{1/2}$ and $C_2 < \{(1-\rho)/2\}^{1/2}$ then

$$P_{0,\mathrm{Norm}}(T_m \le t) \le \big\{1 - \exp\big(-C_2^2\, \log n\big)\big\}^\nu \to 0\,.$$

The desired result follows from this formula and (A.6).

(5) Let $C_1, C_2, \ldots$ denote positive constants. We may assume without loss of generality that $0 \le \gamma < 1$, in which case the inequality $f \ge C_1\, \bar{F}\, |\log \bar{F}|^\gamma$ implies that the derivative of $(-\log \bar{F})^{1-\gamma}$ exceeds C_1, and thence that

$$\bar{F}(x) \le \exp\big(-C_2\, |x|^{1/(1-\gamma)}\big)\,. \tag{A.7}$$

It follows that the distribution F has a bounded moment generating function in a neighbourhood of the origin.

Write $\hat{\epsilon}_i = \epsilon_i + \tilde{\epsilon}_i$, where $\tilde{\epsilon}_i = g(x_i) - \hat{g}(x_i)$. Conditional on $\epsilon_i^* = \hat{\epsilon}_j$, put $\epsilon_{1i}^* = \epsilon_j$ and $\epsilon_{2i}^* = \tilde{\epsilon}_j$. Then, $\hat{b}^*(r,s) = \hat{b}_1^*(r,s) + \hat{b}_2^*(r,s)$ where $\hat{b}_j^*(r,s) = \sum_i w_i \epsilon_{ji}^*$. Using Rosenthal's and Markov's inequalities, and the fact that $\hat{g} - g = O_p(n^{-\eta})$, we may prove that for all $0 < \delta < 1$ and $\lambda > 0$,

$$P\big\{|\hat{b}_2^*(r,s)\, Q(r,s)| > n^{-(1-\delta)\eta} \big| \mathcal{X}\big\} = O_p\big(n^{-\lambda}\big)\,.$$

Hence we may write $T_m^* = T_{1m}^* + R_{1m}^*$, where

$$T_{1m}^* = \max\big\{-\hat{b}_1^*(r,s)\, Q(r,s) : 0 \le r \le s - m \le n - m\big\}$$

and $P(R^*_{1m} > n^{-(1-\delta)\eta} | \mathcal{X}) = O_p(n^{-\lambda})$.

Put $\xi_j = F^{-1}(j/n)$ for $1 \le j \le n-1$, $\xi_0 = -\infty$ and $\xi_n = +\infty$. Conditional on ϵ (denoting a generic ϵ_i) lying in $\mathcal{K}_j = (\xi_j, \xi_{j+1})$, take ϵ' to have the distribution of ϵ given that $\epsilon \in \mathcal{K}_j$. Given that $\epsilon^*_{1i} \in \mathcal{I}_j$, let $\epsilon^\#_i$ have the distribution of ϵ given that $\epsilon \in \mathcal{K}_j$, and be such that the pairs $(\epsilon^*_i, \epsilon^\#_i)$ for $1 \le i \le n$ are independent. Define $\hat{b}^\#(r,s) = \sum_i w_i \, \epsilon^\#_i$ and $\Delta_j = \sum_{i \le j} (\epsilon^*_{1i} - \epsilon^\#_i)$. Let $T^\#_m$ be the version of T^*_m that arises if, in the definition at (2.4), we replace $\hat{b}^*$ by $\hat{b}^\#$. Then,

$$\hat{b}^*(r,s) - \hat{b}^\#(r,s) = \sum_{i=1}^{n} W_i \, \Delta_i \,. \tag{A.8}$$

We shall prove that

$$E(\epsilon - \epsilon')^2 \le C_3 \, n^{-1} \, (\log n)^{-2(1+\gamma)} \,, \tag{A.9}$$

whence it follows via a square-function inequality for sums of independent random variables (e.g. Hall and Heyde 1980, p. 23) and properties of conditional expectations that

$$E\Big(\sup_{1 \le i \le n} \Delta_i^2 \Big) \le C_4 \, (\log n)^{-2(1+\gamma)} \,. \tag{A.10}$$

Combining (A.4), (A.8) and (A.10) we deduce that

$$E\bigg[\sup_{(r,s) \in \mathcal{R}_m} Q(r,s)^2 \left\{ \hat{b}^*(r,s) - \hat{b}^\#(r,s) \right\}^2 \bigg] = O_p\Big[\big\{ m \, (\log n)^{2(1+\gamma)} \big\}^{-1} \Big] \,.$$

Therefore, $T^*_{1m} = T^\#_m + R^*_{2m}$, where $E(R^*_{2m})^2 = o\{(\log n)^{-2(2+\gamma)}\}$. The distribution of $T^\#_m$, conditional on the data, is exactly the P_0-distribution of T_m, and so the desired result (5) in Section 4.2 follows on noting result (3) there.

It remains to prove (A.9). If $1 \le i \le \frac{1}{2} n$ then for some $\theta_i \in [0,1]$ we have, by Taylor expansion,

$$\begin{aligned} 0 \le F^{-1}\{(i+1)/n\} - F^{-1}(i/n) &= n^{-1} \Big(f\big[F^{-1}\{(i+\theta_i)/n\}\big] \Big)^{-1} \\ &\le C_1^{-1} \, (i+\theta_i)^{-1} \, |\log\{(i+\theta_i)/n\}|^{-\gamma} \,, \end{aligned}$$

with a similar bound holding for $\frac{1}{2} n < i \le n-2$. Hence, for any $C_5 > 0$,

$$\sum_{C_5 \log n \le i \le n - C_5 \log n} \big[F^{-1}\{(i+1)/n\} - F^{-1}(i/n) \big]^2 \le C_6 \, (\log n)^{-2(1+\gamma)} \,.$$

In view of (A.7), if $C_7 > 0$ is given then for C_5 sufficiently large, $\bar{F}(\pm \, n^{-1} \, C_5 \log n) \le C_8 \, n^{-C_7}$. Result (A.9) follows from these bounds.

REFERENCES

BOWMAN, A.W., JONES, M.C., AND GIJBELS, I. (1998). Testing monotonicity of regression. *J. Comput. Graph. Statist.*, to appear.

BUCKLEY, M.J. AND EAGLESON, G.K. (1989). A graphical method for estimating the residual variance in nonparametric regression. *Biometrika* **76**, 203–210.

BUCKLEY, M.J., EAGLESON, G.K. AND SILVERMAN, B.W. (1988). A graphical method for estimating residual variance in nonparametric regression. *Biometrika* **75**, 189–199.

CARTER, C.K. AND EAGLESON, G.K. (1992). A comparison of variance estimators in nonparametric regression. *J. Roy. Statist. Soc. Ser. B* **54**, 773–780.

COX, D.R. (1966). Notes on the analysis of mixed frequency distributions. *Brit. J. Math. Statist. Psychol.* **19**, 39–47.

FISHER, N.I., MAMMEN, E. AND MARRON, J.S. (1994). Testing for multimodality. *Computat. Statist. Data Anal.* **18**, 499–512.

GASSER, T., SROKA, L. AND JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika 73*, 625–633.

GOOD, I.J. AND GASKINS, R.A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75**, 42–73.

HALL, P. AND HEYDE, C.C. (1980). *Martingale Limit Theory and its Application.* Academic, New York.

HALL, P., KAY, J.W. AND TITTERINGTON, D.M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521–528.

HALL, P. AND MARRON, J.S. (1990). On variance estimation in nonparametric regression. *Biometrika* **77**, 415–419.

HARTIGAN, J.A. AND HARTIGAN, P.M. (1985). The DIP test of unimodality. *Ann. Statist.* **13**, 70–84.

KOMLÓS, J., MAJOR, P. AND TUSNÁDY, G. (1976). An approximation of partial sums of independent rv's, and the sample df. II. *Z. Wahrscheinlichkeitstheorie verv. Geb.* **34**, 33–58.

MAMMEN, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.* **19**, 724–740.

MAMMEN, E., MARRON, J.S. AND FISHER, N.I. (1992). Some asymptotics for multimodality tests based on kernel density estimates. *Probab. Theory Rel. Fields* **91**, 115–132.

MINNOTTE, M.C. AND SCOTT, D.W. (1992). The mode tree: a tool for visualization of nonparametric density features. *J. Computat. Graph. Statist.* **2**, 51–68.

MÜLLER, D.W AND SAWITZKI, G. (1991a). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86**, 738–746.

PEARSON, G.W. AND QUA, F. (1993). High precision 14^C measurement of Irish oaks to show the natural 14^C variations from AD 1840-5000 BC: a correction. *Radiocarbon* **35**, 105-123.

POLONIK, W. (1995a). Measuring mass concentrations and estimating density contour clusters — an excess mass approach. *Ann. Statist.* **23**, 855–881.

POLONIK, W. (1995b). Density estimation under qualitative assumptions in higher dimensions. *J. Multivar. Anal.* **55**, 61–81.

RAMSAY, J.O. (1988). Monotone regression splines in action (with discussion). *Statist. Sci.* **3**, 425–461.

RICE, J. (1984). Bandwidth choice for nonparametric kernel regression. *Ann. Statist.* **12**, 1215–1230.

ROEDER, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.* **89**, 487–495.

SCHLEE, W. (1982). Nonparametric tests of the monotony and convexity of regression. In: *Nonparametric Statistical Inference, vol. II*, Eds. B.V. Gnedenko, M.L. Puri and I. Vincze, pp. 823–836. North-Holland, Amsterdam.

SEIFERT, B., GASSER, T. AND WOLF, A. (1993). Nonparametric estimation of residual variance revisited. *Biometrika* **80**, 373–383.

SILVERMAN, B.W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43**, 97–99.

SILVERMAN, B.W. (1983). Some properties of a test for multimodality based on kernel density estimates. In: *Probability, Statistics and Analysis*, Eds J.F.C. Kingman and G.E.H. Reuter, pp. 248–259. Cambridge University Press, Cambridge, UK.

WAND, M. P. AND JONES, M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London.