*The Annals of Statistics*
2000, Vol. 28, No. 1, 20–39

# TESTING FOR MONOTONICITY OF A REGRESSION MEAN BY CALIBRATING FOR LINEAR FUNCTIONS

By Peter Hall and Nancy E. Heckman[1]

*Australian National University and University of British Columbia*

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.

**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 $h$, $h_{\text{crit}}$, 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, Jones and Gijbels (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. There are even cases where, when the regression mean has a flat part and a nonmonotone dip in another portion

Received August 1998; revised October 1999.

[1]Supported in part by National Sciences and Engineering Research Council of Canada Grant A7969.

*AMS* 1991 *subject classifications*. 62G08, 62F40, G2F30.

*Key words and phrases*. Bootstrap, calibration, curve estimation, Monte Carlo, response curve, running gradient.

of the curve, the bandwidth test can fail asymptotically to detect the overall nonmonotonicity of the curve. Details will be given in Section 2.1.

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 body of the distribution, can be almost impossible for the bandwidth test to distinguish from the sites of small modes. Some of these issues are clear from the theoretical analysis of Silverman (1983) and Mammen, Marron and Fisher (1992).

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. Our approach has the flavor of the dip 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 sampled density, and so it is to be expected that something similar would be true in the present setting. Indeed, our “running gradients” approach does not suffer problems caused by flat parts of the curve; it is calibrated for the case where the response curve is flat. Moreover, it is sensitive to small dips in the curve.

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.

Schlee (1982) has also considered the problem of testing for monotonicity of a response surface, although not really from a practicable viewpoint. More recent work on the topic includes that of Ghosal, Sen and van der Vaart (1999) and Woodroofe and Sen (1999). Nonparametric regression subject to constraints of monotonicity has been treated by, for example, Ramsay (1988, 1998) 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. *Influence of flat or nearly-flat parts of the curve.*

We use the term “increasing” in the sense of “nondecreasing,” and qualify it by “strictly” if we are considering an increasing regression mean that is not flat on any nondegenerate interval. We argue, however, that in the nonparametric context of the present paper, there is not really statistical interest in distinguishing between the null hypothesis which prescribes that the regression mean $g$ be strictly increasing and the null which asks only that it be increasing.

Consider the analogous problem of testing a composite null hypothesis in a parametric context, for example testing $H_0$: $\theta \geq \theta_0$ against the alternative $H_1$: $\theta < \theta_0$. It would rarely be argued that the potential null hypothesis $H_0'$: $\theta > \theta_0$ be distinguished from $H_0$. The reason is one of continuity: the level of a test of $H_0$ is generally taken to be the supremum over $H_0$ of the probability of rejecting $H_0$. The supremum typically occurs at the "boundary," that is, at $\theta = \theta_0$, and the supremum is the limit as $\theta \downarrow \theta_0$ of probabilities calculated under $\theta$ in $H_0'$. The same argument applies when testing $H_0'$; the regression function is strictly increasing. If the regression error distribution is continuous and is considered fixed, then for most "plausible" tests, the supremum over $H_0'$ of the probability of rejecting $H_0$ will occur at the "boundary," that is, when the regression function is constant. In particular, this holds for our proposed method, as stated in result (1) in Section 4.1.

The test of Bowman, Jones and Gijbels (1998), referred to below as the bandwidth test, will generally perform well if the regression function does not have any flat or nearly flat parts, that is, if the regression function is far from the "boundary" of $H_0$. However, the test will have low power for detecting dips in regression functions that are flat or nearly flat in some parts. The bandwidth test can be so strongly influenced by flat or nearly flat parts of the curve that it can overlook places where the curve is not monotone.

We next give an example of functions which exhibit this type of behavior; we shall explore it numerically in Section 3.3. Suppose the true regression mean is defined on the interval $\mathscr{I} = [0, 1]$, is strictly increasing on $[0, \frac{1}{2})$, and is flat on $[\frac{1}{2}, 1]$. For the sake of definiteness, assume that the support of the symmetric kernel employed in the local linear estimator for the bandwidth test is the interval $[-1, 1]$, the errors are Normal, and the $n$ design points are regularly spaced through $\mathscr{I}$. In order to combat stochastic variability which will occur among values of the gradient of the local linear estimate on $[\frac{1}{2}, 1]$, $h_{\text{crit}}$ will tend to be relatively large. In particular, the probability that $h_{\text{crit}} \geq \frac{1}{6}$ will converge to a strictly positive number as $n \to \infty$. (To see where $\frac{1}{6}$ comes from, note that local linear estimators with bandwidth $\frac{1}{6}$, computed at $x = \frac{2}{3}$ and 1 respectively, are stochastically independent.) This means that, for large $n$, the local linear estimator computed at a point $x$ will with high probability involve averaging over all data whose design points lie within at least $\frac{1}{6}$ of $x$.

Now imagine putting a dip in that part of the regression mean on $[0, \frac{1}{2})$. Let the dip have its center at $\frac{1}{4}$ and extend strictly less than $\frac{1}{6}$ on either side of $\frac{1}{4}$. Provided the dip is not too deep relative to the gradient of other parts of the curve on $[0, \frac{1}{2})$, the following will be true: (a) the gradient of any local linear smooth with bandwidth exceeding $\frac{1}{6}$ will, with probability tending to 1, have strictly positive gradient on $[0, \frac{1}{2}]$; and (b) the probability that $h_{\text{crit}} \geq \frac{1}{6}$ will continue to converge to a strictly positive number as $n \to \infty$. Therefore, the probability that the dip will be detectable by the bandwidth test will not converge to 1. The test proposed in Section 2.2 does not suffer from this deficiency.

2.2. *Test statistic.* Suppose data $\mathscr{X} = \{(x_i, Y_i),\ 1 \le i \le n\}$ are generated by the model $Y_i = g(x_i)+\varepsilon_i$, 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 $\mathscr{I}$ and the errors $\varepsilon_i$ are independent and identically distributed with zero mean and variance $\sigma^2$ for some $\sigma > 0$. We wish to test the null hypothesis $H_0$ that $g$ is nondecreasing on the interval $\mathscr{I}$.

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

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

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 $\sigma^2$ for each pair $(r, s)$. We incorporate this standardization into our test statistic, which is

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

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,

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

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

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 less attractive, and we do not consider them further.

2.3. *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

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

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 neighborhood 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 $\sigma^2$, using any of a variety of different methods; see, for example, 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), Seifert, Gasser and Wolf (1993) and Dette, Munk and Wagner (1998). The technique is generally not as important in the present context as it is in more general applications of nonparametric regression, since under the constraint of monotonicity the influence of bias, which is the main factor determining relative performance of different variance estimator types, is usually relatively small.

Having estimated variance, condition on $\hat{\sigma}^2$ and simulate values of $\varepsilon_i$ from the Normal $N(0, \hat{\sigma}^2)$ distribution. 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 $\hat{t}_{m,a}$ to the point $t_{m,a}$ such that $P_{0,\text{Norm}}(T_m > t_{m,a}) = \alpha$, where $P_{0,\text{Norm}}$ 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 $\hat{t}_{m,a}$.

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 $\hat{\varepsilon}_i = Y_i - \hat{g}(x_i)$. We might wish to center 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 $\sigma^2$. Centering 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 $\mathscr{I}$ might be considered unreliable because of excessive bias or variance, and not be used for that reason.

Let $\mathscr{E}$ be the set of residuals that we have computed, and let $\varepsilon_1^*, \dots, \varepsilon_n^*$ denote a resample drawn by sampling randomly, with replacement, from $\mathscr{E}$. Put

$$\begin{aligned} S^*(a, b|r, s) &= \sum_{i=r+1}^{s} \{\varepsilon_i^* - (a + bx_i)\}^2, \\ (\hat{a}^*, \hat{b}^*) &= \operatorname{argmin}_{(a,b)} S^*(a, b|r, s), \\ T_m^* &= \max\{-\hat{b}^*(r, s)\, Q(r, s) \colon 0 \le r \le s - m \le n - m\}. \end{aligned} \tag{2.4}$$

The nonparametric bootstrap estimator of the $\alpha$-level critical point of $T_m$ under the null hypothesis is $\tilde{t}_{m,\alpha}$, defined by $P(T_m^* > \tilde{t}_{m,\alpha}|\mathscr{X}) = \alpha$. In the case of approximately Normal errors, the calibration step suggested earlier can itself be calibrated using a bootstrap argument.

2.4. *Heteroscedasticity.* If the errors $\varepsilon_i$, 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 $\sigma(x_i)^2 = \operatorname{var}(\varepsilon_i|x_i)$ 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.

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 $Y_i = \varepsilon_i$, for $i = 1, \dots, n = 100$, where the $\varepsilon_i$'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: $\varepsilon_i$ Normally distributed with standard deviation 0.1 (the Normal calibration model), and $\varepsilon_i = \rho W$, where $W$ had Student's $t$ distribution with 5 degrees of freedom, denoted below by $t_5$. The value of $\rho$ was chosen so that the interquartile range of $\varepsilon_i$ 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),

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

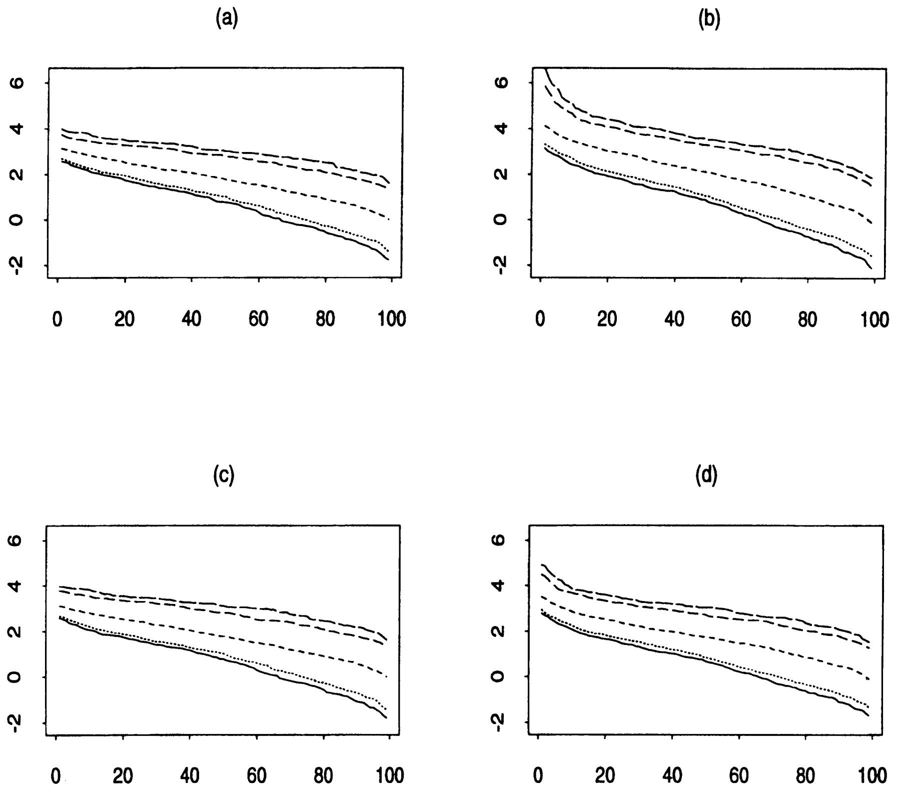

FIG. 1. *Distribution of $T_m$ when $g \equiv 0$. The 5th, 10th, 50th, 90th and 95th pointwise percentiles are graphed in each panel, for (a) the distribution of $T_m/0.1$ with Normal errors, (b) $T_m/0.1$ with $t_5$ errors, (c) $T_m/\hat{\sigma}$ with Normal errors, and (d) $T_m/\hat{\sigma}$ with $t_5$ errors.*

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

Figure 1 graphs the 5th, 10th, 50th, 90th and 95th 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, standardizing by dividing by $\hat{\sigma}$ tends to counteract much of that variability.

3.2. *Level accuracy.* 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 2.3. In the case of *Tp* the $\varepsilon_i^*$'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 Studentized 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, Jones and Gijbels (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.

To check level accuracy of *Tp*, *Tnp* and *Tst* we simulated data $Y_i = g(x_i) + \varepsilon_i$, for $i = 1, \dots, n = 100$, with $x_i = i/(n+1)$ and $g \equiv 0$. In this context, 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 horizontal axis gives values of $m$.) The actual level for the test *Tst* is of course exactly 0.05, and that for *Bow* is 0.06 in the same setting. It is seen that, in terms of level accuracy, *Tp* and *Bow* are comparable, although *Tnp* is slightly more liberal. The former result is true also at nominal level 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 for an $\alpha = 0.05$ level test. For this comparison we first calibrated the tests so that, in the case of Normal errors, they all had level 0.05, to within simulation error (which here is $\pm 0.01$, denoting plus or minus one standard deviation). In particular, we did not rely on the nominal levels. The same adjustment was made to *Bow*, to ensure that that test had level 0.05 when errors were Normal. Note that there is no need to adjust *Tst*.

We see from panels (c)–(e) that after this calibration, level accuracy of *Tp* and *Tnp* is reasonable for $m$ above 10 or 15. 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. The relative conservatism of *Bow* persists for nominal levels 0.10 and 0.25.

3.3. *Power.* To explore numerically the properties discussed in Section 2.1 we considered a function $g$ that was monotone except for a pronounced dip on an interval around $x = 1/4$. Specifically, we took $g(x) = g_1(x) - g_2(x)$, where

$$g_1(x) = \begin{cases} 15\left(x - \frac{1}{2}\right)^3 + M\left(x - \frac{1}{2}\right), & \text{on } \left[0, \frac{1}{2}\right], \\ M\left(x - \frac{1}{2}\right), & \text{on } \left(\frac{1}{2}, 1\right] \end{cases}$$

and $g_2(x) = \exp\{-250(x - \frac{1}{4})^2\}$. Thus, $g$ is increasing with a continuous second derivative, and $g_2$ introduces the dip. For each version of $g$ considered below we generated 500 datasets of the form $Y_i = g(i/101) + \varepsilon_i$, for $i = 1, \dots, 100$, where the $\varepsilon_i$'s were independent Normal $N(0, 0.1^2)$. Figure 3 illustrates both $g$ and a typical simulated dataset in the case $M = 0.3$.

We considered $M = 0$, 0.15 and 0.3, and present here results for the bandwidth and *Tst* tests. Results for the *Tp* and *Tnp* tests are similar to those for *Tst*. In the comparisons below the tests were calibrated so that their levels were all equal to 0.05. In particular, we did not rely on accuracy of nominal levels.

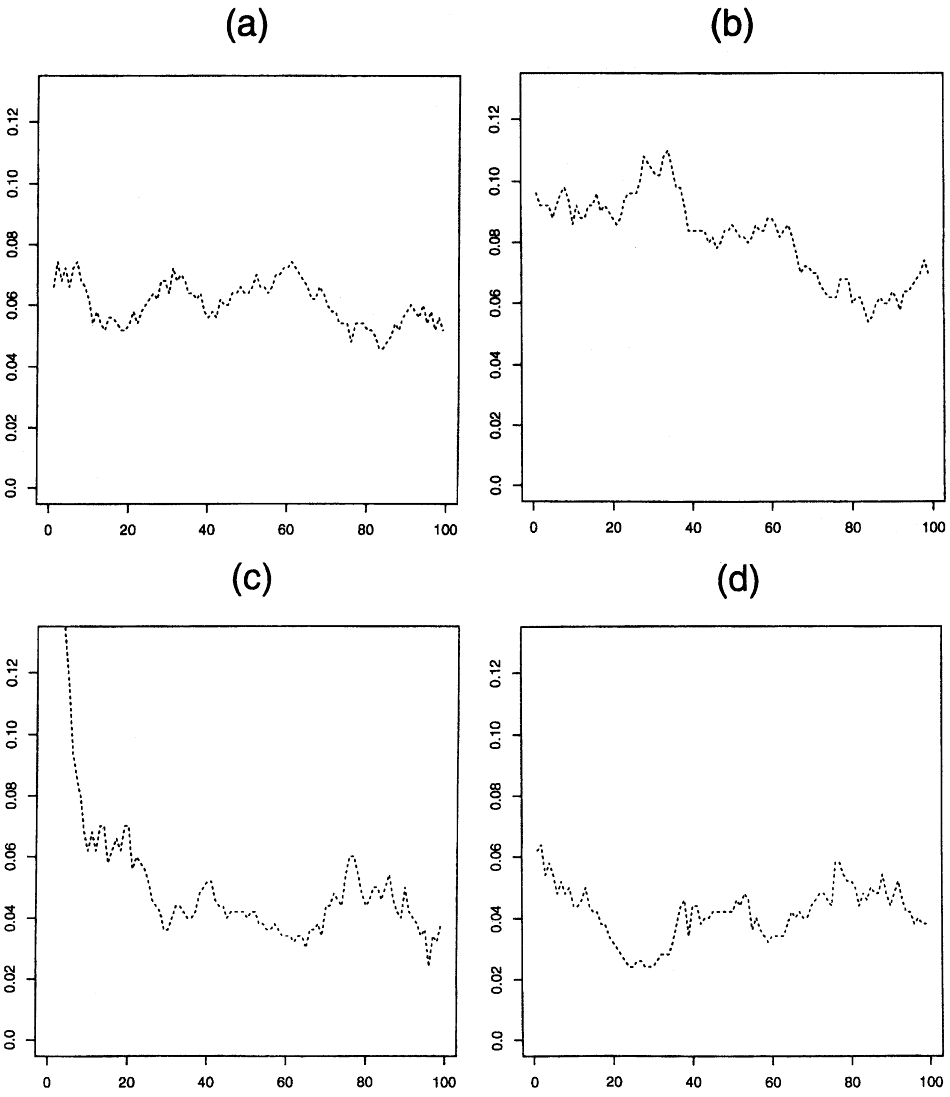

FIG. 2. *Level accuracy of tests Tp, Tnp and Tst. For* $\alpha = 0.05$, *panels give* (*a*) *Monte Carlo approximations to actual level of Tp for Normal errors,* (*b*) *level of Tnp for Normal errors,* (*c*) *level of Tp for* $t_5$ *errors, after calibration for Normal errors,* (*d*) *level of Tnp for* $t_5$ *errors, after calibration, and* (*e*) *level of Tst for* $t_5$ *errors, after calibration.*

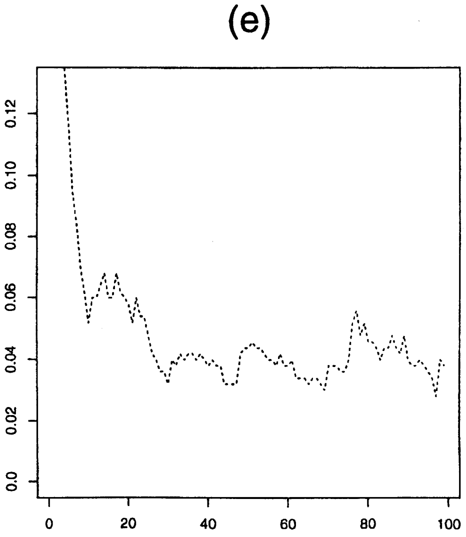

FIG. 2. *(Continued)*

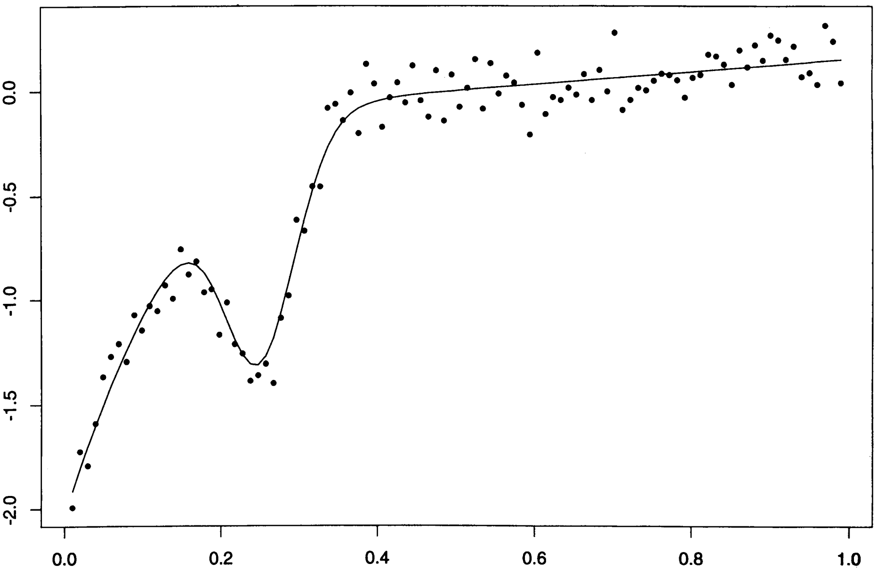

FIG. 3. *Simulated dataset for the function g when $M = 0.3$. Definition of g is given in Section* 3.3*; errors are Normal* $N(0, 0.1^2)$.

Our choice of $g$ is of particular interest since, although the corresponding curve is partly flat only when $M = 0$, the dip is more pronounced for smaller values of $M$. This means that the bandwidth test could actually have more trouble finding dips in functions $g$ without flat sections, than in the case where they were flat over half the design interval. We found this to be the case.

For example, when $M = 0.3$ the bandwidth test rejected $H_0$ 5.0% of the time, that is, its power was equal to its significance level. On the other hand, its power for $M = 0.15$ and $M = 0$ was 19% and 59%, respectively. The test *Tst* performed substantially better in all these cases. For example, no matter what the value taken for $m$ in the range $1 \le m \le 15$, when $M = 0.3$, 0.15 and 0 the power of *Tst* was never less than 98%, 99% and 99%, respectively.

Power of the *Tst* test decreased monotonically with increasing $m$. This was to be expected, since the dip extends over a relatively short interval and was not seen as a problem. Indeed, a plot of the $p$-value as a function of $m$ provided valuable insight into the shape of $g$, suggesting the length of the abnormality that was leading to nonmonotonicity.

Bowman, Jones and Gijbels (1998) studied the test function $g(x) = g_a(x) \equiv 1 + x + a \exp\{-50\,(x - \frac{1}{2})^2\}$, where $a > 0$. Note that $g$ is strictly increasing for $0 < a < a_1 \equiv 0.1e^{1/2} \approx 0.165$, but neither increasing nor decreasing for $a > a_1$. In a simulation study we used the two error distributions introduced in Section 3.1. We found that when $g \equiv g_a$, $\alpha = 0.05$ and $m = 20$, the break-even point for performance of our tests relative to *Bow* was $a_2 \equiv 0.415$, in the sense that our tests tended to have greater power than *Bow* if $a > a_2$ but not otherwise. When $a = a_2$ and errors were Normal, the rejection rate of all four tests (i.e., *Tp*, *Tnp*, *Tst* and *Bow*) was approximately 0.8. (As always, we have adjusted all tests so that they are indeed 5% level tests and thus comparable.)

## 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 equal $O(n^\delta)$ 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 $\mathscr{I} = [0,1]$, with density bounded away from zero and infinity, or (b) the design points are regularly spaced on $\mathscr{I} = [0,1]$. 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 $2 \le m \le n$ and for all $0 \le x < \infty$, $P(T_m \ge x) \le P_0(T_m \ge x)$. That is, the probability of Type I error is no more for a general nondecreasing $g$ than it is for an identically constant $g$.
2. If there exists a nondegenerate interval $\mathscr{J} \subseteq \mathscr{I}$ 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 $m_0 = o(n)$,
$$\inf_{2 \le m \le m_0(n)} P\big(T_m > C\, n^{1/2}\big) \to 1$$
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, $T_m = O(n^\delta)$ for all $\delta > 0$. This property should be contrasted with result 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, result 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 $\varepsilon_i$ are Normally distributed with variance $\sigma^2$. 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 $n^\delta$ for all $\delta > 0$, in the sense that
$$\text{for all } \delta > 0, \quad \sup_{2 \le m \le n} P_0\big(T_m > n^\delta\big) \to 0$$
as $n \to \infty$.

2. If the error distribution has the property that $P(|\varepsilon| > x) \geq \text{const.}\, x^{-c}$ 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) $E|\varepsilon|^c < \infty$ and $m = m(n)$ satisfies $n^{2/c}(\log n)^{1+\varepsilon} = O(m)$ and $m = O(n^\rho)$, for some $c > 3$, $\varepsilon > 0$ and $\rho \in (2/c, 1)$ or (ii) the distribution of $\varepsilon$ has a finite moment generating function in some neighborhood of the origin, and $m = m(n)$ satisfies $(\log n)^{3+\varepsilon} = O(m)$ and $m = O(n^\rho)$ for some $\varepsilon > 0$ 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 $\sigma^2$), in the sense that

$$\sup_{0\leq x<\infty} \big|P_0(T_m \leq x) - P_{0,\,\text{Norm}}(T_m \leq x)\big| \to 0$$

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

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 $\rho$ there exist constants $0 < t_1 < t_2 < \infty$ such that

$$\sup_{2\leq m\leq n^\rho} \big|P_{0,\,\text{Norm}}\{t_1 \leq T_m/(\log n)^{1/2} \leq t_2\} - 1\big| \to 0$$

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 $\hat{\varepsilon}_i = Y_i - \hat{g}(x_i)$ satisfies $\sup_{y_1\leq x\leq y_2} |\hat{g}(x) - g(x)| = O_p(n^{-\eta})$ for some $\eta > 0$ and $0 \leq y_1 < y_2 \leq 1$; if we compute all those residuals $\hat{\varepsilon}_i$ that correspond to design points $x_i$ satisfying $y_1 \leq x_i \leq y_2$; if $f \geq C\,\bar{F}\,|\log \bar{F}|^\gamma$ for constants $C, \gamma > 0$; and if $m = m(n)$ satisfies $(\log n)^{3+\varepsilon} = O(m)$ and $m = O(n^\rho)$ for some $\varepsilon > 0$ 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

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

in probability as $n \to \infty$.

The assumption on $f$ in result 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 result 3.

## APPENDIX

**A.1. Outline proofs of results in Section 4.1.** (i) Define

$$S_1(x)=\sum_{i=r+1}^{s}(x_i-\bar{x}_{rs})^2,\quad \Delta_1(r,s)=\sum_{i=r+1}^{s}g(x_i)(x_i-\bar{x}_{rs}),$$

$$\hat{b}(r,s)=S_1(x)^{-1}\sum_{i=r+1}^{s}Y_i\,(x_i-\bar{x}_{rs}),\quad \hat{b}_0(r,s)=S_1(x)^{-1}\sum_{i=r+1}^{s}\varepsilon_i\,(x_i-\bar{x}_{rs})$$

and $\Delta_2(r,s)=S_1(x)^{-1}\Delta_1(r,s)$, where $\bar{x}_{rs}=(s-r)^{-1}\sum_{r+1\le i\le s}x_i$. Trivially, $\hat{b}(r,s)=\hat{b}_0(r,s)+\Delta_2(r,s)$. We claim that $\Delta_1(r,s)$, and hence $\Delta_2(r,s)$, is nonnegative for each $r<s$, provided $g$ is increasing. To appreciate why, put $\delta_i=g(x_i)-g(x_{i-1})$ for $r+1\le i\le s$, and observe that by Abel's method of summation,

$$\Delta_1(r,s)=\sum_{i=r+1}^{s}(x_i-\bar{x}_{rs})\sum_{j=r+1}^{i}\delta_j=\sum_{j=r+1}^{s}\delta_j\sum_{i=j}^{s}(x_i-\bar{x}_{rs}).\tag{A.1}$$

Since $x_{r+1}\le\cdots\le x_s$ then $\sum_{j\le i\le s}(x_i-\bar{x}_{rs})\ge 0$ for $r+1\le j\le s$. Also, the monotone increasing property of $g$ implies that each $\delta_i>0$. Hence, by (A.1), $\Delta_1(r,s)\ge 0$ for each $r,s$.

Let $T_{0m}$ have the definition of $T_m$ at (2.2), but in the case where $S(r,s)$ is given by (2.3) instead of (2.1). That is, $T_{0m}$ is defined by (2.2) using $\hat{b}_0(r,s)$ instead of $\hat{b}(r,s)$. Since $\Delta_2(r,s)\ge 0$ and $\hat{b}(r,s)=\hat{b}_0(r,s)+\Delta_2(r,s)$, then $\hat{b}(r,s)\ge\hat{b}_0(r,s)$, and so by (2.2), $T_m\le T_{0m}$ with probability 1. Now, $P$-measure for $T_{0m}$ is equivalent to $P_0$-measure for $T_m$, and so for each $x$, $P(T_m\le x)\ge P(T_{0m}\le x)=P_0(T_m\le x)$, as had to be shown.

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

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

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

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

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

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

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

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

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

which implies the desired result.

**A.2. Outline proofs of results in Section 4.2.** (i) Since $g$ is constant, $\hat{b}(r,s) = \sum_i w_i \varepsilon_i$. Hence, by Rosenthal's inequality [Hall and Heyde (1980), page 23] we have for all $k \geq 1$,

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

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

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

we have $(s-r)\,(x_s - x_r)^2 = O\{Q(r,s)^2\}$. (Here and below, in the case where $x_1 < \cdots < x_n$ 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,

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

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

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

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

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

(ii) Let $\nu$ denote the integer part of $(n-m)/m$, and for $0 \leq j \leq \nu$ 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

$$P(T_m > t) \geq P\Big(\max_{0\leq j\leq\nu} R_j > t\Big) = 1 - \prod_{j=1}^{\nu} P(R_j \leq t). \tag{A.2}$$

Since $P(|\varepsilon| > x) \geq \text{const.}\,x^{-c}$ then either or both $P(\varepsilon^+ > x) > \text{const.}\,x^{-c}$ or $P(\varepsilon^- > x) > \text{const.}\,x^{-c}$. Without loss of generality the former is true. Given constants $0 < B_1 < B_2 < \infty$, let $\mathscr{J}(B_1, B_2)$ 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 $\mathscr{J}(B_1, B_2)$ exceeds $\nu u$ for all sufficiently large $n$. Take $u = \frac{1}{2}$, and select $B_1, B_2$ accordingly.

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

$$P(R_j > x) \ge P\left(B_1\,\varepsilon_1 - B_2 \sum_{j=2}^{m} |\varepsilon_j| > x\right) \ge C_2\, x^{-c}$$

for large $x$, where $C_2$ depends only on $m, B_1, B_2$ and the error distribution. Therefore, by (A.2),

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

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

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

$$\hat{b}(r, s) = \sum_{i=1}^{n} w_i\,(U_i - U_{i-1}) = \sum_{i=1}^{n} W_i\, U_i. \tag{A.3}$$

When the error distribution has finite moment generating function in a neighborhood 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.3),

$$\hat{b}(r, s) = \sum_{i=1}^{n} W_i\, V_i + O_p\left\{(\log n) \sum_{i=1}^{n} |W_i|\right\}, \tag{A.4}$$

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

$$Q(r, s) \sum_{i=1}^{n} |W_i| = O\{(x_s - x_r)/Q(r, s)\} = O\big\{(s - r)^{-1/2}\big\} \tag{A.5}$$

uniformly in $(r, s) \in \mathscr{R}_m$. Combining (A.4) and (A.5) we see that if $(\log n)^{3+2\varepsilon} = O(m)$ for some $\varepsilon > 0$ then

$$\hat{b}(r, s)\, Q(r, s) = Q(r, s) \sum_{i=1}^{n} W_i\, V_i + O_p\big\{(\log n)^{-(1/2)-\varepsilon}\big\}, \tag{A.6}$$

uniformly in $(r, s) \in \mathscr{R}$, as $n \to \infty$. If $E|\varepsilon|^c < \infty$ for some $c > 3$ then a similar argument, using Theorem 2 rather than Theorem 1 of Komlós, Major and

Tusnády (1976), shows that (A.4) holds with $\log n$ replaced by $n^{1/c}$ and "big oh" replaced by "little oh." This again leads to (A.6), provided $n^{2/c}(\log n)^{1+2\varepsilon} = O(m)$.

The first term on the right-hand side of (A.6) 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.6) 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

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

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

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

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

as $\varepsilon \downarrow 0$, 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 $\delta_n$.

(iv) 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

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

Therefore,

$$\text{(A.7)} \qquad \inf_{2\le m\le n} P_{0,\,\mathrm{Norm}}\left\{T_m/(\log n)^{1/2} \le C_1\right\} \to 1\,.$$

Let $m_0$ equal the integer part of $n^\rho$, let $\nu$ 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.7).

(v) 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 $(-\log \bar{F})^{1-\gamma}/(1-\gamma)$ exceeds $C_1$, and hence that

$$\bar{F}(\pm x) \le \exp\big(-C_2\,|x|^{1/(1-\gamma)}\big). \tag{A.8}$$

It follows from (A.8) that the distribution $F$ has a bounded moment generating function in a neighborhood of the origin.

Write $\hat{\varepsilon}_i = \varepsilon_i + \tilde{\varepsilon}_i$, where $\tilde{\varepsilon}_i = g(x_i) - \hat{g}(x_i)$. Conditional on $\varepsilon_i^* = \hat{\varepsilon}_j$, put $\varepsilon_{1i}^* = \varepsilon_j$ and $\varepsilon_{2i}^* = \tilde{\varepsilon}_j$. Then, $\hat{b}^*(r,s) = \hat{b}_1^*(r,s) + \hat{b}_2^*(r,s)$ where $\hat{b}_j^*(r,s) = \sum_i w_i \varepsilon_{ji}^*$. 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 $\lambda > 0$,

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

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

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

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

Put $\xi_j = F^{-1}(j/n)$ for $1 \le j \le n-1$, $\xi_0 = -\infty$ and $\xi_n = +\infty$. Conditional on $\varepsilon$ (denoting a generic $\varepsilon_i$) lying in $\mathscr{K}_j = (\xi_j, \xi_{j+1})$, take $\varepsilon'$ to have the distribution of $\varepsilon$ given that $\varepsilon \in \mathscr{K}_j$. Given that $\varepsilon_{1i}^* \in \mathscr{K}_j$, let $\varepsilon_i^\#$ have the distribution of $\varepsilon$ given that $\varepsilon \in \mathscr{K}_j$ and be such that the pairs $(\varepsilon_i^*, \varepsilon_i^\#)$ for $1 \le i \le n$ are independent. Define $\hat{b}^\#(r,s) = \sum_i w_i\,\varepsilon_i^\#$ and $\Delta_j = \sum_{i\le j}(\varepsilon_{1i}^* - \varepsilon_i^\#)$. 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,

$$\hat{b}^*(r,s) - \hat{b}^\#(r,s) = \sum_{i=1}^n W_i\,\Delta_i. \tag{A.9}$$

We shall prove that

$$E(\varepsilon - \varepsilon')^2 \le C_3\,n^{-1}\,(\log n)^{-2(1+\gamma)}, \tag{A.10}$$

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

$$E\Big(\sup_{1\le i\le n}\Delta_i^2\Big)\le C_4\,(\log n)^{-2(1+\gamma)}. \tag{A.11}$$

Combining (A.5), (A.9) and (A.11) we deduce that

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

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.10). 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,

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

Result (A.10) follows from these bounds.

**Acknowledgments.** We are grateful to two reviewers and an Associate Editor for helpful comments on the paper.

## REFERENCES

BOWMAN, A. W., JONES, M. C. and GIJBELS, I. (1998). Testing monotonicity of regression. *J. Comput. Graph. Statist.* **7** 489–500.

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. *British J. Math. Statist. Psych.* **19** 39–47.

DETTE, H., MUNK, A. and WAGNER, T. (1998). Estimating the variance in nonparametric regression—what is a reasonable choice? *J. Roy. Statist. Soc. Ser. B* **60** 751–764.

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

GHOSAL, S., SEN, A. and VAN DER VAART, A. W. (1999). Testing monotonicity of regression. Technical Report WS–514, Faculteit der Exacte Wetenschappen, Vrije Univ., Amsterdam.

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 Press, 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 r.v.'s, and the sample df. II. *Z. Wahrsch. Verw. Gebiete* **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 Related Fields* **91** 115–132.

MINNOTTE, M. C. and SCOTT, D. W. (1992). The mode tree: a tool for visualization of nonparametric density features. *J. Comput. 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.

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. Multivariate Anal.* **55** 61–81.

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

RAMSAY, J. O. (1998). Estimating smooth monotone functions. *J. Roy. Statist. Soc. Ser. B* **60** 365–375.

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* **2** (B. V. Gnedenko, M. L. Puri and I. Vincze, eds.) 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* (J. F. C. Kingman and G. E. H. Reuter, eds.) 248–259. Cambridge Univ. Press.

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

WOODROOFE, M. and SUN, J. (1999). Testing uniformity versus a monotone density. *Ann. Statist.* **27** 338–360.

CENTRE FOR MATHEMATICS
AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA ACT 0200
AUSTRALIA
E-MAIL: halpstat@pretty.anu.edu.au

DEPARTMENT OF STATISTICS
UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER BC V6T 1Z2
CANADA