TESTS FOR MONOTONICITY OF A REGRESSION MEAN WITH GUARANTEED LEVEL

Irène Gijbels,[1,2] Peter Hall,[2] M.C. Jones[2,3] and Inge Koch[2,4]

ABSTRACT. In this paper a nonparametric procedure for testing for monotonicity of a regression mean with guaranteed level is proposed. The procedure is based on signs of differences of observations from the response variable. The test is calibrated against the most difficult null hypothesis, when the regression function is constant, and produces an exact test in this context. In general, the test is conservative. The power of the test is good, and comparable with that of other nonparametric tests. It is shown that the testing procedure has asymptotic power 1 against certain local alternatives. The method is also robust against heavy-tailed error distributions, and even maintains good power when the errors are for example Cauchy distributed. A simulation study is provided to demonstrate finite-sample behaviour of the testing procedure.

KEYWORDS. Calibration, Counts, Exact test, Lengths of runs, Local alternatives, Monte Carlo simulation, Power, Signs of differences.

SHORT TITLE. Testing for monotonicity.

[1] Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium.

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

[3] Department of Statistics, Open University, Milton Keynes MK7 6AA, UK.

[4] Department of Statistics, University of Newcastle, Newcastle NSW 2308, Australia.

1. INTRODUCTION

Postulating monotonicity of a regression relationship between a response variable and a covariate is often reasonable on physical grounds. For example, the age of an object determined by a radiocarbon dating process is expected to be monotonically related to the true age of the object. The monotonicity is a crucial issue for accuracy of the radiocarbon dating process. If a monotonicity assumption is justified then using this additional information will enhance performance of an estimator of the regression relationship. See Friedman and Tibshirani (1984), Mukerjee (1988), Ramsay (1988, 1998) and Mammen (1991), among others, for discussions on the performance of estimators under a monotonicity constraint. See also Turlach (1997) for a review of estimation under constraints using splines.

Problems related to testing for monotonicity are testing for the number of bumps in a regression mean and testing for multimodality of a probability density. There is an extensive literature on testing procedures for the number of modes in a density. Nonparametric estimation of a density involves the choice of a smoothing parameter, such as for example a bandwidth in kernel density estimation. In testing for multimodality Silverman (1981) proposed using as a test statistic the smallest bandwidth which just forces the density estimate to have the modality of the null hypothesis. Theoretical properties of this approach have been studied by Silverman (1983) and Mammen, Marron and Fisher (1992). This 'critical bandwidth' test can exhibit low power when there are flat parts in the density, for example in the tails of the distribution. A possible remedy is to allow for a varying bandwidth; see for example Fisher, Mammen and Marron (1994). Minnotte and Scott (1993), Minnotte (1997) and Chaudhuri and Marron (1997) have developed alternative approaches for dealing with varying values of bandwidths. A second class of tests for multimodality of a density is provided via the dip test of Hartigan and Hartigan (1985) and the excess mass approach of Müller and Sawitzki (1991). See also Cheng and Hall (1997). These tests do not involve choice of a smoothing parameter and are more immune to the flatness problem.

The literature on testing for monotonicity of a regression mean is less extensive. Schlee (1982) proposed a test based on the greatest discrepancy of an estimate of the derivative of the regression mean from zero. Theoretical foundations of the method are provided but a discussion on the practical implementation is lacking. An analogue of the 'critical bandwidth', using kernel type estimators, has been developed by Bowman, Jones and Gijbels (1998). See Ramsay (1998) for a similar approach using spline smoothing. The Bowman et al. (1998) test suffers from a drawback similar to that for the Silverman (1981) test for multimodality of a density: its power can be low when there are flat parts in the regression function. This has been pointed out by Hall and Heckman (1998), who proposed a test statistic based on running gradients, an approach which has connections to the dip/excess mass method in the density context.

All the above tests have the problem that there is no guarantee for the actual level of the test. In this paper we propose a procedure with guaranteed level, maintaining at the same time good power properties. It does not involve choice of a smoothing parameter. The test is based on signs of differences of observations from the response variable Y, and is calibrated for the most difficult null hypothesis, when the regression function is constant. Under this hypothesis, differences of the Y observations reduce to differences of observations of the error variable. The distribution of the signs of these quantities does not depend on the error distribution, and this simple fact gives easy access to the distribution of the test statistic under the most difficult null hypothesis. A simple Monte Carlo simulation from an arbitrary continuous distribution does the job, and leads to exact tests. Under the more general null hypothesis of monotonicity, the resulting test procedure is guaranteed to be conservative. If the test is applied to only that portion of the data where monotonicity of the regression function is in doubt, the extent of conservatism is particularly small. Strictly speaking the exactness of the test is obtained under the condition of homoscedasticity, but the test has approximately correct level in case of heteroscedastic errors. Since calibration does not directly involve the error distribution, our method turns out to be relatively robust against heavy-tailed error distributions, such as for example a Student t distribution with a small number of degrees of freedom, or a Cauchy distribution. In contrast to, for example, the Hall and Heckman (1998) approach the calibration does not involve estimation of the error variance, and indeed does not require finiteness of any moments.

Our approach relies on signs of differences, and various test statistics based on these signs can be proposed. We focus on two statistics, based on counts of equal signs and on runs of equal signs, respectively. For these statistics we study asymptotic power against local alternatives, where the regression mean departs linearly from the null hypothesis at a certain rate. It is shown that both tests have asymptotic power one. The class of local alternatives for which this holds is larger for the counts test than for the runs test, however. Hence, the counts test is, at least asymptotically, superior to the runs test. On the other hand, the runs test performs as well as or better than the counts test in many finite sample situations and is extremely fast computationally, almost ruling out the counts test when computational speed is an issue.

The paper is further organised as follows. In Section 2 we introduce our approach to the testing problem, and discuss test statistics. Further details on practical implementation, as well as a simulation study, are provided in Section 3. In Section 4 we investigate the asymptotic power of test statistics based on counts and runs. Outlines of the proofs of the theoretical results are given in the Appendix.

2. METHODOLOGY

Assume data pairs (x_i, Y_i), $1 \leq i \leq n$, are generated by the model $Y_i = g(x_i) + \epsilon_i$, where the errors ϵ_i are independent random variables and the x_i's form an increasing sequence of non-random quantities, possibly representing conditioned values of random variables. Our methods are based on signs of the differences $Y_{i+k} - Y_i$, and produce tests with exactly known level, if the ϵ_i's have an identical, continuous distribution. In this context we do not require any moment assumptions about the error distribution; our test is valid for Cauchy-distributed errors, for example. Although the function g is not identifiable unless we make an assumption about the centre of the distribution of ϵ_i, for example that ϵ_i has median 0, hypotheses about monotonicity of g are well defined without such assertions. Our test has approximately correct level if we assume that $\epsilon_i = \sigma(x_i)\, \eta_i$, where the η_i's are independent and identically distributed random variables with zero mean and finite variance, σ is a smooth function, and we confine attention to differences $Y_{i+k} - Y_i$ where $x_{i+k} - x_i$ is small for each i.

Suppose we wish to test the null hypothesis H_0 that g is nondecreasing on the interval $\mathcal{I} = (x_1, x_n)$, against the alternative H_1 that g is decreasing on a subinterval $\mathcal{J}$ of $\mathcal{I}$. In effect, H_0 is a composite hypothesis, and an equivalent test is of the simple hypothesis H_0': g is constant on $\mathcal{I}$, against the alternative H_1.

Our test statistic T is based on the signs $S_{i,k} \equiv \mathrm{sgn}\,(Y_{i+k} - Y_i)$ for $1 \le i \le n-k$, which under H_0 or H_0' have no particular tendency towards negativity but under H_1 are more likely to be negative than positive on $\mathcal{J}$. There is of course a wide variety of approaches to defining T. Their relative merits depend, for example, on the scale of the error distribution relative to the derivative of g on $\mathcal{J}$. We consider two approaches here, based respectively on counts and runs of the signs $S_{i,k}$.

First we define the 'count' statistic. Let a 'string' of signs, $\mathcal{S}_r = \mathcal{S}_r(j,k)$, denote the sequence $S_{j+1,k}, \ldots, S_{j+r,k}$. Given the string, let $N(\mathcal{S}_r)$ equal the number of $+1$'s it contains, and observe that under H_0, $N(\mathcal{S}_r) \ge \frac{1}{2}r + O(r^{1/2})$ with high probability, whereas $N(\mathcal{S}_r)$ tends to take smaller values under H_1. Therefore, we define

$$T_{\mathrm{count}} = T_{\mathrm{count}}(k, r_1, r_2) = \max_{r_1 \le r \le r_2} \max_{0 \le j \le n-r-k} r^{-1/2} \big[\tfrac{1}{2}r - N\{\mathcal{S}_r(j,k)\}\big]\,, \qquad (2.1)$$

which tends to be larger under H_1 than under H_0. It is possible to maximise over a range of values of k as well as over r and j. Taking k larger gives the test greater power against functions g that have more gently decreasing parts. For a given choice of r_1 and r_2, the test will have relatively good power against alternatives where the interval $\mathcal{J}$ is of approximately the same length as the distance between x_i and $x_{i+k+r_2-r_1}$, assuming this does not vary greatly with i. On the other hand, we could take $r_2 = n-k$ and r_1 to be a relatively small integer. See Section 4 for theory in this case.

Next we define the 'run' statistic. Let $R(j,k)$ denote the supremum of integers r such that $S_{j+1,k}, \ldots, S_{j+r,k}$ is a sequence solely of -1's, and put

$$T_{\mathrm{run}} = T_{\mathrm{run}}(k) = \max_{0 \le j \le n-k} R(j,k)\,. \qquad (2.2)$$

Again we may choose to maximise over k. The value of T_{run} tends to be larger under H_1 than it is under H_0.

Let T denote either T_{count} or T_{run}. We test H_0 against H_1 by rejecting H_0 if T is too large. The exact level of the test under H_0' is easily obtained. If $\epsilon_1, \ldots, \epsilon_n$ are independent random variables with common continuous distribution F, and if $S_{1,k}, \ldots, S_{n-k,k}$ is a sequence of ± 1's, then $q(S_{1,k}, \ldots, S_{n-k,k}) \equiv P\{\mathrm{sgn}\,(\epsilon_{i+k} - \epsilon_i) =$

$S_{i,k}, 1 \leq i \leq n-k\}$ does not depend on F. To appreciate why, note that $\epsilon_{i+k} - \epsilon_i > 0$ if and only if $U_{i+k} - U_i > 0$, where $U_i = F(\epsilon_i)$, $1 \leq i \leq n-k$, is a sequence of independent Uniform random variables. Therefore, $q(S_{1,k}, \ldots, S_{n-k,k})$ is identical to the value it would take in the Uniform case.

To determine a critical point we can use the following simple Monte Carlo algorithm. Write $\mathcal{Y}$ for the set $\{Y_1, \ldots, Y_n\}$, let $\mathcal{Y}^* = \{Y_1^*, \ldots, Y_n^*\}$ be a set of independent random variables with an arbitrary continuous distribution, and let T^* be the version of T that arises when $\mathcal{Y}$ is replaced by $\mathcal{Y}^*$. Suppose we wish the test to have level α, a rational number between 0 and 1. Conduct B simulations, where B is chosen so that, for an integer ν satisfying $1 \leq \nu < B$, $\nu/(B+1) = 1-\alpha$. Reject H_0, at level α, if T is strictly greater than the ν'th largest value of T^*.

Owing to discreteness of the distribution of T under either H_0 or H_0', it is generally not possible to compute a critical point such that the level of the test is exactly equal to α. However, provided the errors ϵ_i are independent and have identical, continuous distributions, the Monte Carlo form of our test has, for each choice of B, a level under H_0' that does not exceed α; if the distribution of T were continuous then the level would equal α. This is easily seen by noting that, under H_0', the values of T and of the B realisations of T^* comprise $B+1$ independent random variables all distributed as T. Furthermore, as $B \to \infty$ for fixed n, the critical point of the Monte Carlo test converges to a quantity t_α, equal to the infimum of values t such that $\pi(t) \equiv P(T \leq t | H_0')$ satisfies $\pi(t) \geq 1-\alpha$.

An alternative would be to calculate the p-value directly, by calculating the proportion of T^* values, among the B simulated values, which are greater than or equal to the observed value T, and reject H_0, at level α, if this proportion was less than or equal to α. To deal with discreteness of the distribution, we only count half of the T^*'s which are exactly equal to the observed T when calculating the proportion. In our practical study we derived the p-value directly, instead of determining the critical value.

3. SOME PRACTICALITIES AND SIMULATIONS

In this section we explore practical aspects of our testing procedures via simulation. We focus mainly on power evaluations of the tests, obtained via calculating the number of times the null hypothesis was rejected among the number of simulations carried out. If the null hypothesis were true, this proportion should be small, and if the null hypothesis were false this proportion should be close to one.

First let us concentrate on use of the run statistic T_{run}. Options here include choosing a particular value k_0 of k for use in $T_{\mathrm{run}}(k)$, (2.2), and maximising over $1 \leq k \leq k_{\max}$. For samples of size $n = 100$ from each of a couple of situations considered in Table 1 of Bowman et al. (1998) and described below, powers were calculated for many values of k_0 and of $k_{\max}$ between 1 and 99. The cases chosen were such that values of k around 20 were particularly good for capturing non–monotonicity. Maximising over k displayed a consistent and worthwhile improvement in power over utilising specific values of k. A further alternative, that of normalising $T_{\mathrm{run}}(k)$ by dividing by $n - k$ before maximising over k, was also considered. This gave similar results to the unnormalised case, except when $k_{\max}$ was taken very large: as k increased towards $n-1$, the power of the normalised version reduced dramatically, while that of the unnormalised version retained essentially its maximum power. We prefer the version of $T_{\mathrm{run}}(k)$ maximised over $1 \leq k \leq k_{\max}$ in its unnormalised form, because of its robustness to choice of $k_{\max}$.

Likewise, for T_{count}, maximisation over $1 \leq k \leq k_{\max}$ is the preferred option. For a number of experiments in which $n = 100$, the power performance of T_{count} was consistently fairly similar to, but marginally less effective than, that of T_{run}. Computation times for T_{count} were, however, considerably greater than those for T_{run}, the latter being especially trivial computationally.

However, as suggested by the theory in Section 4, there are situations where the count statistic can be expected to be superior to the runs statistic. Consider for example the gently downward sloping regression mean $g(x) = -0.075x$, $0 < x < 1$, with $\sigma = 0.025$. Thus, the value of g at $x = 1$ is three standard deviations below the value 0 at which it started. Unless stated otherwise, estimated powers are based on 1000 values from the null distribution and 2000 replications. When $n = 100$, the power of T_{run} with $k_{\max} = 20$ was 0.511, while that of T_{count} with $k_{\max} = 12$ was 0.501. However, when n was increased to 250 the power of T_{run} with $k_{\max} = 50$ was 0.704, while that of T_{count} with $k_{\max} = 30$ had increased to 0.819. This illustrates the theoretically better performance of T_{count} for large n. Our choices of $k_{\max}$ were somewhat arbitrary, but larger values gave only minor improvements, and for the counts test larger values would also increase computation times considerably.

For further comparisons we continue with T_{run}. In Table 1 we give power values computed for two modelling situations, each with normal errors and $n = 100$. Model I, as above, has $g(x) = -ax$, $0 < x < 1$, for $a = 0$ and two nonzero values of a, using $\sigma = 0.025$. Model II, adapted from Bowman et al. (1998), is a flat regression function with a normal shaped dip: $g(x) = -a\exp\{-50(x-0.5)^2\}$, for two further nonzero values of a. Models I and II are both identically zero for $a = 0$. Table 2 gives power values for Model III, the linearly increasing function with dip $g(x) = 1 + x - a\exp\{-50(x-0.5)^2\}$, for three values of a and three values of σ. Model III was used by Bowman et al. (1998); see Figure 1 of that paper. The values for the method of Bowman et al. in Table 2 are taken from their Table 1.

The first entry in the second row of Table 1 confirms the exact nature of the runs test when $a = 0$, which is the limiting case of monotonicity. The table also reveals good power for the method under each of the alternatives considered. In Table 2, the strongly increasing nature of g when $a = 0$ leads to very small proportions of rejection both for the runs test, which is guaranteed to be conservative, and for Bowman et al.'s test, which is not.

While both Bowman et al.'s method and the runs test clearly identify the non–monotonicity of Model III when $a = 0.45$, at least provided σ is not too large, when $a = 0.25$ and σ is small, the Bowman et al. method still identifies non–monotonicity which the runs test does not. It is only in this type of situation, where non–monotonicity is particularly small and, most importantly, the monotonicity of the regression function away from the interval $\mathcal{J}$ is strong, that the runs test is unable to match the Bowman et al. test in terms of power. Note how the same dip is picked up for much smaller a in Table 1 where $1 + x$ is not present in g.

In practice, however, strong increases will be clear and attention can be focused on regions in which the monotonicity is in more doubt, in which case the runs test performs essentially as well as the Bowman et al. test. And it does so very much faster, computationally. Moreover, the level of the new tests is guaranteed, and that of the Bowman et al. test is not. In more limited experiments with $g(x) = 0$, we found that, when a cut–off value in Professor Bowman's program was taken into account, the null hypothesis was rejected 53% of the time. Thus, given the guarantees about the level of the new tests, we feel that they do very well to have good power, comparable to that of the Bowman et al. test, in most situations.

Finally we turn to another scenario where the new tests outperform the Bowman et al. test. As already emphasised, the new tests are quite robust to outliers in the error distribution. Experiments involving Student t and Cauchy errors indicate that for sufficiently heavy-tailed departures from normality, the Bowman et al. test is unable to identify non–monotonicity, in suitable versions of Model III, while the runs test still can. An example, based on just 50 simulation replications, is that of Model III with $a = 0.6$ and errors that are standard Cauchy distributed, scaled by a factor of 0.05; the runs test maintained power of 0.96, the Bowman et al. test managed power of only 0.06.

4. THEORETICAL PROPERTIES

In this section we investigate the power of the tests under local alternatives. Assume that the errors ϵ_i are independent with a common distribution having a bounded density, and for the sake of definiteness that the design points are given by $x_i = i/n$. Note that we make no assumption about the shape or tail-weight of the error distribution. Take δ_n to be a sequence of positive numbers. Let $0 < u_1 < u_2 < 1$ be functions of n, for which there exists a constant $\xi_1 \in (0, 1)$ such that $u_2 - u_1 > n^{-\xi_1}$ for all sufficiently large n; let g_0 be a function which is nondecreasing on $[0, u_1]$, constant on $\mathcal{J} = [u_1, u_2]$, and nondecreasing on $[u_2, 1]$; and let g_n be the function that is identical to g_0 on $[0, u_1]$, equal to $g_0(u_1) - (x - u_1)\,\delta_n$ on (u_1, u_2), and equal to $g_0(x) - (u_2 - u_1)\,\delta_n$ on $[u_2, 1]$. In particular, g_0 satisfies H_0 but g_n does not, and g_n differs from g_0 in being linearly decreasing, rather than constant, on $\mathcal{J}$. Generally, both g_0 and g_n vary with n.

Note that $\sup_{\mathcal{I}} |g_n - g_0| \asymp \delta_n(u_2 - u_1)$ as $n \to \infty$, where for positive constants a_n and b_n the notation '$a_n \asymp b_n$' means that the ratio a_n/b_n is bounded away from 0 and ∞ as $n \to \infty$. Thus, g_n is a local alternative to g_0, distant approximately $\delta_n(u_2 - u_1)$ from g_0. We are interested to know how small δ_n may be and yet g_n still be distinguishable from g_0 using our tests.

First we treat the 'count' test. There, we claim, g_n is distinguishable from g_0 if and only if δ_n exceeds a sufficiently large constant multiple of $n^{1/2}k^{-1}(\log n)^{1/2}$.

Theorem. *Assume the above model for the functions g_0 and g_n, and the data (x_i, Y_i). If r_1, r_2 satisfy $r_2 = n - k$ and $r_1 \asymp n^{\xi_2}$, where $0 < \xi_2 < 1 - \xi_1$; if $1 \le k \le O(n^{\xi_2 - \eta})$ for some $\eta > 0$; if*

$$\delta_n = \delta_n(c) = n^{1/2} k^{-1} (\log n)^{1/2} c \tag{4.1}$$

for a positive constant c; and if $p(n, c)$ denotes the probability that, for $B = \infty$, our 'count' test based on $T_{\mathrm{count}}(k, r_1, r_2)$ rejects H_0 when $g = g_n$; then

$$\lim_{c \to \infty} \liminf_{n \to \infty} p(n, c) = 1 \, , \quad \limsup_{c \to 0} \limsup_{n \to \infty} p(n, c) \le \alpha \, .$$

If g_0 is constant on $\mathcal{I}$ then $\limsup_{n\to\infty} p(n, c) \to \alpha$ as $c \to 0$.

An outline proof is given in Appendix (a).

When maximising over $1 \le k \le k_{\max}$, the statement in the theorem remains valid, with k replaced by $k_{\max}$ in (4.1). Thus, maximising over k allows for a larger class of local alternatives, and hence increases the power of the test.

When g is known up to a finite number of parameters, and u_1, u_2 do not vary with n, conventional hypothesis tests are generally just able to detect alternatives distant order $\delta_n \asymp n^{-1/2}$ from the null. In the present nonparametric setting, taking $k \asymp n^{1-\zeta}$ for an arbitrary small but positive ζ, we see from the theorem that δ_n can be as small as $n^{-(1/2)+\zeta}$ for any given ζ. Minor modifications to our proof show that this is true for a wide range of alternatives g_n to g_0, not just those which decrease linearly on $\mathcal{J}$. Therefore, the price paid for addressing an infinite-dimensional class of departures from the null hypothesis is only a multiplicative penalty of size n^{ζ} for arbitrarily small ζ.

A version of the theorem may be proved for the tests proposed by Bowman et al. (1998) and Hall and Heckman (1998). However, the regularity conditions there require the error distribution to be particularly light-tailed. For example, calibration of the Bowman-Jones-Gijbels and Hall-Heckman tests typically involves estimation of error variance, and in particular, variance must be finite. It is clear from theoretical results of Hall and Heckman (1998) that for distributions with sufficiently light tails we may take δ_n as small as $n^{-(1/2)+\zeta}$; by way of contrast, note that our theorem does not require assumptions about tail weight.

The penalty for choosing k large is that we can miss relatively small intervals $\mathcal{J}$ where the null hypothesis is violated. This problem manifests itself through the regularity conditions $0 < \xi_2 < 1 - \xi_1$ and $1 \le k \le O(n^{\xi_2 - \eta})$. In order to have k large, i.e. $k \asymp n^{1-\zeta}$ where ζ is small, we must have ξ_2 close to 1, hence ξ_1 small, and hence the width of $\mathcal{J}$ large. Similar difficulties arise in parametric settings, however; if the null hypothesis is violated on a small interval than we have relatively little information about the violation, and so our chance of detecting it is correspondingly reduced.

Finally we treat the 'run' test. We claim that in this case the analogue of δ_n, defined at (4.1), is $\delta_n(c) = nk^{-1}(\log n)^{-1}c$, with k replaced by $k_{\max}$ when maximisation over $1 \le k \le k_{\max}$ is considered. Thus, δ_n must be an order of magnitude larger, by a factor of $n^{1/2}(\log n)^{-3/2}$, than is necessary for the 'count' test, before the test is capable of detecting departures from the null hypothesis. We call this claim (C). Appendix (b) will give an outline proof of the claim, which implies inferior power properties for the run test in cases that are very difficult to distinguish from the null hypothesis.

This result obtains because, when used against local alternatives, the run test never has an opportunity to consolidate evidence against the null. There is too much tendency for run counts to stop early, and so provide insufficient information, owing to fluctuations in $\mathrm{sgn}\,(Y_{i+k} - Y_i)$ that are almost as common under local alternatives as they are under the null. By way of contrast, the count test focuses on relatively fixed amounts of evidence against the null, since it counts the number of -1's in a string of pre-specified length.

This result explains the inferior numerical performance of the run test observed in Section 3, when applied to detect very marginal departures from the null hypothesis. It has no relevance to numerical work on power under fixed alternatives, however, and in fact the simulation study in Section 3 suggests that the run test is often superior to the count test in that setting.

Appendix (a): *Outline proof of theorem*

Let $r = r(n) \in [r_1, r_2]$ be an integer, and put $\delta = (n/k)\,(r^{-1}\log n)^{1/2}c$ and $\gamma = (r^{-1}\log n)^{1/2}d$, where c, d are positive numbers depending on n and satisfying $c \to c_0$ and $d \to d_0$, with $0 \le c_0, d_0 < \infty$ and $d_0 \ne 0$. Define $D_i = \epsilon_{i+k} - \epsilon_i - (k\delta/n)$

and $S_i = I(D_i > 0)$, and let N denote the number out of $S_1, \ldots, S_r$ that equal $+1$. Then, $E(N)/r = \frac{1}{2} - C(k\delta/n) + o(k\delta/n)$, where the constant $C > 0$ depends on the error density, and $\mathrm{var}\,(N) = (r/12) + O(rn^{-\eta})$ for some $\eta > 0$. Moreover, using results on moderate-deviation probabilities for k-dependent random variables, see e.g. Heinrich (1985), we may prove that

$$\begin{aligned} P\big\{N \leq \tfrac{1}{2}\,(1-2\gamma)\,r\big\} &= P\big[(N-EN)/(\mathrm{var}\,N)^{1/2} \\ &\qquad \leq -\{1+o(1)\}\,12^{1/2}\,(d-cC)\,(\log n)^{1/2}\big] \\ &= \exp\big[-6\,\{1+o(1)\}\,(d_0-c_0C)^2\log n\big]\,, \end{aligned} \tag{A.1}$$

provided $d_0 > c_0 C$. Strictly speaking, Heinrich's result applies only to the case of fixed k and fixed $(k/n)\,\delta$, but our setting where $k = k(n)\to\infty$ and $(k/n)\,\delta\to 0$ is easily treated using the same arguments. Note too that several non-exponential terms that are usually present in moderate-deviation probabilities are in effect incorporated into the '$o(1)$' remainder on the right-hand side of (A.1).

When H_0' obtains, $c = 0$. Hence, by (A.1),

$$\begin{aligned} P\big\{T_{\mathrm{count}} \geq (\log n)^{1/2} d \,\big|\, H_0'\big\} &\leq \sum_{r=r_1}^{r_2}\sum_{j=0}^{n-r} P\Big(r^{-1/2}\big[N\{\mathcal{S}_r(j)\} - \tfrac{1}{2}r\big] \\ &\qquad\qquad \leq -(\log n)^{1/2}\,d\,\Big|\,H_0'\Big) \\ &= O\big(n^{2-6d_0^2+\eta}\big) \end{aligned}$$

for all $\eta > 0$. Therefore, the critical point for an α-level test under H_0' exceeds $-(\log n)^{1/2}\{(1/3)^{1/2}+\eta\}$ for all sufficiently large n, for each $\eta > 0$.

Next, still in the case $c = 0$, take $r = \lfloor n^{\xi_3}\rfloor$, where $1-\xi_1 < \xi_3 < 1$ and $\lfloor\cdot\rfloor$ denotes the integer part function. Put $j_i = 1 + 2i\lfloor n^{\xi_3}\rfloor$ for $0 \leq i \leq i_0 = \lfloor\frac{1}{2}n^{1-\xi_3}\rfloor$. The variables $\mathcal{S}_r(j_i)$ for $0 \leq i \leq i_0$ are independent, and so, by (A.1),

$$\begin{aligned} P\big\{T_{\mathrm{count}} \geq (\log n)^{1/2} d \,\big|\, H_0'\big\} &\geq P\Big(r^{-1/2}\big[N\{\mathcal{S}_r(j_i)\} - \tfrac{1}{2}r\big] \leq -(\log n)^{1/2}\,d \\ &\qquad\qquad \text{for some } 0 \leq i \leq i_0\,\Big|\,H_0'\Big) \\ &= 1 - \Big(1 - \exp\big[-6\,\{1+o(1)\}\,d_0^2\,\log n\big]\Big)^{i_0}\,. \end{aligned}$$

The right-hand side converges to 1 if $d_0 < \{(1-\xi_3)/6\}^{1/2}$. Hence, the critical point for an α-level test under H_0' does not exceed $-(\log n)^{1/2}[\{(1-\xi_3)/6\}^{1/2} - \eta]$, for all sufficiently large n, for each $\eta > 0$. Combining this result with that obtained at the end of the previous paragraph we deduce that the α-level critical point t_α, determined by our Monte Carlo algorithm with $B = \infty$ and having the property that $1 - \alpha \le P(T_{\text{count}} \le t_\alpha | H_0') \to 1 - \alpha$ as $n \to \infty$, satisfies $t_\alpha \asymp (\log n)^{1/2}$. The theorem may be proved from this property and (A.1), taking $c_0 \neq 0$ in the latter.

Appendix (b): *Outline proof of claim (C)*

For the sake of brevity we shall verify (C) in a simplified setting where, rather than being based on possibly overlapping runs of k-dependent signs, the test uses independent runs of independent signs. After a little consideration it is clear that, in first-order asymptotic terms, these two approaches are virtually equivalent. To further simplify our discussion we assume u_1, u_2 are fixed and g_0 is identically constant.

Suppose there are m independent runs of -1, with respective lengths R_1, $\ldots, R_m$, where each R_i has a geometric distribution, i.e. $P(R_i = r) = \pi^r (1 - \pi)$ for $r \ge 0$. Here, π denotes the probability that $Y_{i+k} - Y_i$ is negative; it equals $\frac{1}{2}$ when $g = g_0$ and equals $\pi_1 \equiv \frac{1}{2} + C(k\delta_n/n) + o(k\delta_n/n)$ when $g = g_n$, where the constant C is as in Appendix (a). We ask of $m = m(n)$ only that $\log m \asymp \log n$.

The probability that the longest run of -1's does not exceed $r - 1$ equals $(1 - \pi^r)^m$. Our Monte Carlo method for calibration under H_0' is equivalent to choosing r_0 to be the smallest integer such that $\{1 - (\frac{1}{2})^{r_0}\}^m \ge 1 - \alpha$, and rejecting H_0' if the length of the longest of the m observed runs of -1 is greater than or equal to r_0. That is, r_0 is the smallest integer that is not less than

$$\nu_m \equiv -\frac{\log\{1 - (1-\alpha)^{1/m}\}}{\log 2} = \frac{\log m}{\log 2} - \frac{\log\log(1-\alpha)^{-1}}{\log 2} + o(1)\,,$$

and so we may write $r_0 = \nu_m + \eta_m$ where $0 \le \eta_m < 1$. If $c > 1$ and

$$\delta_n = \frac{n\,(\log 2)\,(\log c)}{2Ck\log m}$$

then the probability of rejecting the null hypothesis when $g = g_n$ equals

$$1 - \left(1 - \pi_1^{r_0}\right)^m = 1 - \exp\left\{(\tfrac{1}{2})^{\eta_m} c \log(1-\alpha)\right\} + o(1) \le 1 - (1-\alpha)^c + o(1)\,.$$

Therefore, the probability of rejecting H_0' can only be made arbitrarily close to 1 by choosing c sufficiently large, or equivalently, by choosing δ_n greater than a sufficiently large constant multiple of $(n/k)\,(\log n)^{-1}$. The converse result, that by taking c to be sufficiently large we force the probability of correctly rejecting H_0' to be sufficiently close to 1, follows in like manner. Both results are true uniformly in n.

Acknowledgement

This research was carried out while the first, third and fourth author were visiting the Centre for Mathematics and its Applications at the Australian National University. The hospitality and the financial support of the Centre are gratefully acknowledged. The research of Gijbels was partially supported by 'Projet d'Actions de Recherche Concertées', No. 98/03–217. The authors thank Professor Adrian Bowman for making the programming code on the 'critical bandwidth' testing approach available to them.

REFERENCES

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

CHAUDHURI, P., AND MARRON, J.S. (1997). SiZer for exploration of structures in curves. Manuscript.

CHENG, M.-Y., AND HALL, P. (1997). Calibrating the excess mass and dip tests of modality. *J. Royal Statist. Society, Series B*, **60**, 579–589.

FISHER, N.I., MAMMEN, E., AND MARRON, J.S. (1994). Testing for multimodality. *Comp. Stat. Dat. Anal.* **18**, 499–512.

FRIEDMAN, J.H., AND TIBSHIRANI, R.J. (1984). The monotone smoothing of scatterplots. *Technometrics*, **26**, 243–250.

HALL, P., AND HECKMAN, N. (1998). Testing for monotonicity of a regression mean without selecting a bandwidth. Manuscript.

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

HEINRICH, L. (1985). Non-uniform estimates, moderate and large deviations in the central limit theorem for m-dependent random variables. *Math. Nachr.*, **121**, 107–121.

MAMMEN, E. (1991). Estimating a smooth 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 estimates. *Probab. Theory Rel. Fields*, **91**, 115–132.

MINNOTTE, M.C. (1997). Nonparametric testing of the existence of modes. *Ann.*

Statist., **25**, 1646–1660.

MINNOTTE, M.C., AND SCOTT, D.W. (1993). The mode tree: a tool for visualization of nonparametric density estimates. *J. Comput. Graphical Statist.*, **2**, 51–68.

MUKERJEE, H. (1988). Monotone nonparametric regression. *Ann. Statist.*, **16**, 741–750.

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

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

RAMSAY, J.O. (1998). Estimating smooth monotone functions. *J. Royal Statist. Society, Series B*, **60**, 365–375.

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), Amsterdam: North-Holland, 823–836.

SILVERMAN, B.W. (1981). Using kernel density estimates to investigate multimodality. *J. Royal Statist. Society, Series 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), Cambridge: Cambridge University Press, 248–259.

TURLACH, B.A. (1997). Constrained smoothing splines revisited. *Technical Report* SSR97-008, Australian National University, Centre for Mathematics and its Applications.

Table 1: *Power Values for T_{run} test under Models I and II.*

Model	I & II; $a = 0$	I; $a = 0.1$	I; $a = 0.2$	II; $a = 0.075$	II; $a = 0.15$
Power	0.048	0.774	1.000	0.985	1.000

Table 2: *Power Values for T_{run} and Bowman et al. test under Model III.*

	$a = 0$			$a = 0.25$			$a = 0.45$		
σ	0.025	0.05	0.1	0.025	0.05	0.1	0.025	0.05	0.1
T_{run}	0.000	0.000	0.000	0.106	0.037	0.014	1.000	1.000	0.805
Bowman	0.002	0.008	0.004	0.994	0.748	0.174	1.000	1.000	0.874