Ann. Inst. Statist. Math.
Vol. 51, No. 2, 231–251 (1999)

ON THE ESTIMATION OF JUMP POINTS IN SMOOTH CURVES

IRENE GIJBELS[1]*, PETER HALL[1,2] AND ALOÏS KNEIP[1]*

[1] *Institut de Statistique, Université Catholique de Louvain, 20 Voie du Roman Pays, B-1348 Louvain-la-Neuve, Belgium*
[2] *Centre for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia*

(Received May 15, 1997; revised November 5, 1997)

Abstract. Two-step methods are suggested for obtaining optimal performance in the problem of estimating jump points in smooth curves. The first step is based on a kernel-type diagnostic, and the second on local least-squares. In the case of a sample of size n the exact convergence rate is n^{-1}, rather than $n^{-1+\delta}$ (for some $\delta > 0$) in the context of recent one-step methods based purely on kernels, or $n^{-1}(\log n)^{1+\delta}$ for recent techniques based on wavelets. Relatively mild assumptions are required of the error distribution. Under more stringent conditions the kernel-based step in our algorithm may be used by itself to produce an estimator with exact convergence rate $n^{-1}(\log n)^{1/2}$. Our techniques also enjoy good numerical performance, even in complex settings, and so offer a viable practical alternative to existing techniques, as well as providing theoretical optimality.

Key words and phrases: Bandwidth, curve estimation, change point, diagnostic, discontinuity, kernel, least squares, nonparametric regression.

1. Introduction

Jump points in otherwise-continuous treatment effects arise when the treatment changes suddenly, often without warning or planning. Examples occur in the context of medical monitoring, and are typically addressed using methods of signal processing. One novel, contemporary approach of this type, based on wavelets, has been suggested recently by Wang (1995) (see also Raimondo (1996)), who has also provided an excellent review of literature on jump-point estimation. His method is tailored to the case of a continuously observed signal, with Gaussian white noise, although it has a variant in the setting of a signal sampled at equally spaced time points, with Normally distributed noise. There, Wang's approach produces an

* Gijbels and Kneip were supported by "Projet d'Actions de Recherche Concertées", No. 93/98–164.

estimator which converges at rate $n^{-1}(\log n)^{1+\delta}$, where n denotes the number of sampled points within a given interval, and δ is a positive number.

Results such as these raise the question of just how accurately the position of a jump point may be estimated, and of whether procedures that achieve optimal rates are practicable. In the present paper we show that even in a nonparametric setting, a convergence rate of n^{-1} is available under mild conditions on the error distribution (only a little more severe than existence of finite variance); that this rate is optimal in a minimax sense; and that procedures which achieve the rate are definitely practicable.

It is possible to refine Wang's estimator and obtain a convergence rate of $n^{-1}(\log n)^{1/2}$, at least in the context of Normal errors. This may be done by regarding the spatial indices of empirical wavelet coefficients as defined in the continuum, not just at points that are integer multiples of integer powers of $\frac{1}{2}$; and closely monitoring the way in which the coefficients behave as functions of this continuous variable. The nature of wavelets introduces practical difficulties, however. Wavelets tend to have either rough, fractal-like graphs, or smooth graphs with many turning points. This makes it difficult to identify the spatial location where the appropriate maximum absolute value of an empirical wavelet coefficient occurs, and so to achieve the rate $n^{-1}(\log n)^{1/2}$ in practice.

An alternative technique, appropriate for error distributions other than the Normal, is to employ Wang's wavelet-based estimator as a diagnostic for obtaining a rough idea of the region where a jump point lies, and then use other tools to identify the jump point more accurately. That is the approach suggested in the present paper. For the reasons noted in the previous paragraph we employ kernel rather than wavelet methods, however, to obtain our first crude estimator. We show that the kernel diagnostic leads to a convergence rate of $n^{-1}(\log n)^{1/2}$, provided the error distribution has finite moment generating function; and that if the diagnostic is combined with a local least-squares fitting step in order to refine its performance, then it enjoys the convergence rate n^{-1}, and under substantially weaker conditions. We show that in the least-squares part of the algorithm it is adequate to fit curves that are locally constant, although more sophisticated techniques could be employed. The resulting estimator enjoys good numerical performance.

Our method has obvious generalizations to estimating points of discontinuity in densities, where it produces identical convergence rates. In both settings, once the locations of jump discontinuities have been estimated, left- and right-hand limits at those points may be estimated using kernel methods based on right- and left-handed kernels, respectively. It is straightforward to generalize our techniques to estimators of jump points in the k-th derivative of a regression mean, for general $k \geq 0$. Diagnostics based on kernel estimators of the $(k+1)$-st derivative of the regression mean are employed to obtain a rough idea of the positions of jumps, and are then refined by fitting k-th order polynomials in small neighbourhoods on either side of the jump points. The convergence rate is $n^{-1/(2k+1)}$. For the sake of brevity, and since practical motivation for estimating jumps in the zero-th derivative is considerably stronger than that for derivatives of higher order, we do not provide details here.

Loader (1996) has recently suggested a one-step jump-point estimator that attains the $O_p(n^{-1})$ rate in the case of Gaussian errors and regularly-spaced design. His approach is founded on ideas from likelihood-based inference in a parametric model—hence the assumption of Gaussian errors. Müller and Song (1997) have also proposed a jump-point estimator which attains the $O_p(n^{-1})$ rate for regularly-spaced design. They assume particularly mild conditions on the error distribution, however, and like us, employ a two-stage procedure. Their first stage is based on a "generic" pilot estimator, which they show may be taken to be of the kernel type. Their second stage refines the estimate in the first stage by comparing heights of function estimators on either side of a candidate value for the jump point.

Wang (1995) offers such a good literature survey that we note here only those items that involve methods which are relatively near to ones in the present paper, as well as some not mentioned by Wang. Korostelev (1987) considered estimation of jump points in a Gaussian white noise model, and introduced in that setup a two-step procedure which has some similarities with our approach. The important contribution in our approach, which is proposed in a general setting, is the least squares step which allows for generalizations to fitting *locally* other parametric models (for example polynomials of a certain degree). The work of Korostelev and Tsybakov (1993) on optimal methods for image analysis is in concept close to ours, although this is arguable. They provide a detailed discussion of likelihood-based methods for jump-point detection in the context of data that are generated either in the continuum or discretely. In their account of one-dimensional problems, Korostelev and Tsybakov (1993) are interested predominantly in optimality in a parametric sense, and their approach to obtaining optimal performance is unavailable in the nonparametric setting that we address. Nevertheless, the convergence rates that they derive under parametric assumptions are identical to ours under nonparametric ones. Korostelev and Tsybakov ((1993), p. 45) do suggest that some of their ideas may have nonparametric counterparts, but the assumptions that they have in mind, such as their condition (1.50), would be quite restrictive in a statistical as distinct from image-analytic setting.

Recent work of a more statistical nature includes that of Müller (1992) and Eubank and Speckman (1993), who employ kernel methods. Theoretical analysis not unlike the one given in the present paper shows that their techniques, in the forms that they discussed, achieve convergence rates of only $n^{-1+\delta}$ for some $\delta > 0$. A conference proceedings edited by Carlstein *et al.* (1994) addresses a wide variety of estimation problems in the context of change and jump points. It includes an article by Eubank and Speckman (1994), extending ideas of Eubank and Speckman (1993) to the estimation of jumps in derivatives. Techniques based on kernel estimators are also used by Hall and Titterington (1992) and Wu and Chu (1993). The paper of McDonald and Owen (1986), treating general but particularly sophisticated curve estimation methods for functions with jumps, employs (like ours) local least-squares methods in parts of its estimation algorithm.

Section 2 will introduce our techniques and discuss their properties, including theoretical and numerical performance. Details of technical arguments will be given in Section 3.

2. Methodology

2.1 *Introduction and summary*

Assume that data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are generated by the model

$$Y_i = g(X_i) + \epsilon_i, \quad 1 \leq i \leq n, \tag{2.1}$$

where the function g is smooth except for a known finite number of jump discontinuities, and the errors ϵ_i are independent and identically distributed. In our theoretical analysis and simulation study, the design points X_i will be taken to be either regularly spaced or ordered values of independent and identically distributed random variables. In each case, the design is assumed to be on the interval $\mathcal{I} = [0, 1]$.

Subsection 2.2 suggests and discusses our kernel-based jump-point diagnostic, and Subsection 2.3 proposes methods that combine it with least-squares fitting. Theoretical properties of our methods are outlined in Subsection 2.4, and numerical performance is summarized in Subsection 2.5.

2.2 *Diagnostic*

We suggest using as a diagnostic a function that is proportional to an estimator of the first derivative of either the regression mean or its product with the design density. Arguably the simplest quantity of this type is

$$D(x, h) = D_1(x, h) = (nh^2)^{-1} \sum_{i=1}^{n} K'\{(x - X_i)/h\} Y_i, \tag{2.2}$$

where K is a kernel function and h a bandwidth. It may be interpreted as the first derivative of the numerator of a Nadaraya-Watson kernel estimator. Alternatively, one could differentiate the entire estimator, obtaining

$$D(x, h) = D_2(x, h) = \frac{\partial}{\partial x} \left(\frac{\sum_{i=1}^{n} K\{(x - X_i)/h\} Y_i}{\sum_{i=1}^{n} K\{(x - X_i)/h\}} \right). \tag{2.3}$$

The diagnostic D_1 is recommended for equally-spaced design, whereas D_2 is preferable for stochastic design, although in first-order asymptotic terms both perform equally well. In the present context we would argue that it is not a good idea to use a diagnostic based on local linear smoothing, since that method is relatively sensitive to the problem of sparse design (see Seifert and Gasser (1996)).

We propose plotting $|D(x, h)|$ as a function of x, for a range of values of h, and identifying jumps as points x in the vicinity of which $|D(x, h)|$ is consistently large for a range of values of h. Properties of D are easily translated into explicit estimators of the positions of jumps, as follows. Here and in some of our later discussion we shall suppose for the sake of simplicity that there is a single discontinuity of g, occurring at an unknown point $x_0 \in (0, 1)$. The case of two or more jumps will be discussed in Subsection 2.5. Let $(-v, v)$ denote the support of K.

Given a sequence $\{\eta_n\}$ which decreases more slowly than $n^{-1}\log n$, let $\tilde{x}(h)$ denote the point which produces a global maximum of $|D(\cdot, h)|$ in an interval slightly smaller than $\mathcal{I}$ (to avoid edge effects). Put

$$\tilde{x}_-(h) = \sup_{h_1 \in [h, \eta_n]} \{\tilde{x}(h_1) - 2vh_1\}, \qquad \tilde{x}_+(h) = \inf_{h_1 \in [h, \eta_n]} \{\tilde{x}(h_1) + 2vh_1\},$$

for $h \le \eta_n$. Let $\tilde{h}$ denote the infimum of values h such that $\tilde{x}_-(h) \le \tilde{x}_+(h)$, and define $\hat{x}_0 = \tilde{x}(\tilde{h})$.

2.3 *Enhancement of the diagnostic*

The diagnostic may be enhanced in several ways. One is by iteration, as follows. First use the diagnostic to identify an interval of width $2v\tilde{h} = O_p(n^{-1}\log n)$ within which x_0 lies with high probability. This interval contains $n' = O_p(\log n)$ data values. Rescale it to an interval of length 1, and apply the diagnostic again, this time using n' instead of n. The resulting estimator of x_0 will have convergence rate $n'^{-1}(\log n')^{1/2}$ on the new scale, and $n^{-1}(\log n) \times n'^{-1}(\log n')^{1/2} = O\{n^{-1}(\log_2 n)^{1/2}\}$ on the original scale, where the subscript 2 indicates the second recursion of the logarithm function. A third iteration will produce an estimator with convergence rate $O_p\{n^{-1}(\log_3 n)^{1/2}\}$, and so on.

The envelope of these rates, $O_p(n^{-1})$, may be achieved by using the diagnostic to identify a relatively wide interval in which the discontinuity lies with high probability, and then applying least-squares to estimate the position of the jump within the interval, as follows. Let $h = cn^{-\alpha}$ denote a bandwidth, with $c > 0$ and $0 < \alpha < 1$. Let $\tilde{x}_0^*$ denote the point at which $|D(\cdot, h)|$ is maximized, in an interval a little shorter than $\mathcal{I}$ (to avoid edge effects). Assuming as before that the kernel K is supported on $(-v, v)$, pretend that g is a step function on $(\tilde{x}_0^* - 2vh, \tilde{x}_0^* + 2vh)$, and use least squares to estimate it. That is, we estimate the jump discontinuity to occur between design points X_{i_0} and X_{i_0+1}, where i_0 is chosen to minimize the sum of squares

$$\sum_{i_1 \le i \le i_0} \left\{ Y_i - (i_0 - i_1 + 1)^{-1} \sum_{i_1 \le j \le i_0} Y_j \right\}^2 + \sum_{i_0+1 \le i \le i_2} \left\{ Y_i - (i_2 - i_0)^{-1} \sum_{i_0+1 \le j \le i_2} Y_j \right\}^2, \tag{2.4}$$

and $\{i_1, i_1 + 1, \ldots, i_2\}$ is the set of integers i such that $X_i \in (\tilde{x}_0^* - 2vh, \tilde{x}_0^* + 2vh)$.

2.4 *Theoretical results*

To simplify matters we analyze the case of a single jump discontinuity, at $x_0 \in (0, 1)$, since that of several peaks differs only in notational complexity. Assume K has support $(-v, v)$, and put $\mathcal{I}_h = [vh, 1 - vh]$. To be completely explicit in the definition of $\hat{x}_0$, let the global maximum of $|D(\cdot, h)|$, which defines $\tilde{x}(h)$, be taken over $x \in \mathcal{I}_h$. We begin by developing theory describing the performance of this

estimator, and, more generally, the fluctuations of the curves $D(\cdot, h)$ for a range of sizes of bandwidth.

Next we introduce regularity conditions. Since there is a single peak at x_0 then we assume that the regression mean in the model at (2.1) may be expressed as $g(x) = g_1(x) + g_2(x)I(x > x_0)$, where g_1 and g_2 are smooth functions. More specifically, we make the following assumptions: g_1 and g_2 have bounded derivatives on $\mathcal{I}$, and $g_2(x_0) \neq 0$; K is compactly supported and has two Hölder-continuous derivatives on $\mathcal{I}$, $K(0) > K(x) \geq 0$ for all $x \neq 0$, $K'(0) = 0$, and $K''(0) < 0$; the errors ϵ_i are independent and identically distributed with zero mean and a distribution that has finite moment generating function in some neighbourhood of the origin; and the design points X_i are either regularly spaced, in which case they are taken equal to $(i + c)/n$ for a constant $c \in [-1, 0]$, or they are ordered (in increasing size) values of independent and identically distributed random variables whose distribution is supported on $\mathcal{I}$ and has a density that has a bounded derivative and is bounded away from zero on $\mathcal{I}$. In the latter case the X_i's are assumed independent of the errors. Collectively we call these conditions (C_1).

To obtain a lower bound to the rate of convergence we ask that the tails of the error distribution not be too short. Specifically, we assume that there exist constants $A_1, A_2 > 0$ such that $P(|\epsilon_i| > u) \geq \exp(-A_1 u^2)$ for all $u > A_2$. This condition is of course satisfied if the errors are Gaussian. We call it (C_2).

Now we describe limit theory for D and $\hat{x}_0$. Let $B, B_1, B_2 > 0$, and as before, let $\{\eta_n\}$ denote a sequence of constants decreasing to zero more slowly than $n^{-1} \log n$. Define $p_n(B)$ to be the probability that for each $h \in [Bn^{-1} \log n, \eta_n]$, the maximum of $|D(x, h)|$ on $\mathcal{I}_h$ occurs at a point x satisfying $|x - x_0| \leq vh$; let $q_n(B_1, B_2)$ be the probability that the maximum of $|D(x, h)|$ over all $(x, h) \in \mathcal{I}_h \times [B_1 n^{-1} \log n, \eta_n]$ occurs when $|x - x_0| \leq B_2 n^{-1} (\log n)^{1/2}$; let $r_n(B_1, B_2)$ equal the probability that the maximum of $|D(x, h)|$ over all $(x, h) \in \mathcal{I}_h \times [B_1 n^{-1} \log n, \eta_n]$ occurs when $|x - x_0| > B_2 n^{-1} (\log n)^{1/2}$; and let $s_n(B)$ equal the probability that $|\hat{x}_0 - x_0| \leq Bn^{-1} (\log n)^{1/2}$.

THEOREM 2.1. *Assume* (C_1), *and let* D *be defined by either* (2.2) *or* (2.3). *Then*

$$\text{(a)} \lim_{B \to \infty} \liminf_{n \to \infty} p_n(B) = 1, \qquad \text{(b)} \lim_{B_2 \to \infty} \liminf_{B_1 \to \infty} \liminf_{n \to \infty} q_n(B_1, B_2) = 1.$$

If in addition (C_2) *holds then*

$$\text{(c)} \lim_{B_2 \to 0} \limsup_{B_1 \to 0} \limsup_{n \to \infty} r_n(B_1, B_2) = 0, \qquad \text{(d)} \lim_{B \to \infty} \liminf_{n \to \infty} s_n(B) = 1.$$

The theorem provides theoretical support for the properties noted in our simulation study in Subsection 2.5; see for example Fig. 1. In particular, parts (a) and (c) of the theorem show that the function $|D(\cdot, h)|$ has its peak close to x_0, except in the case of bandwidths h of size $Cn^{-1} \log n$ for a small constant C, where erratic fluctuations of stochastic error cause peaks to arise some distance from x_0. Part (b) shows that for large C, the pair (x, h) that maximizes $D(\cdot, \cdot)$

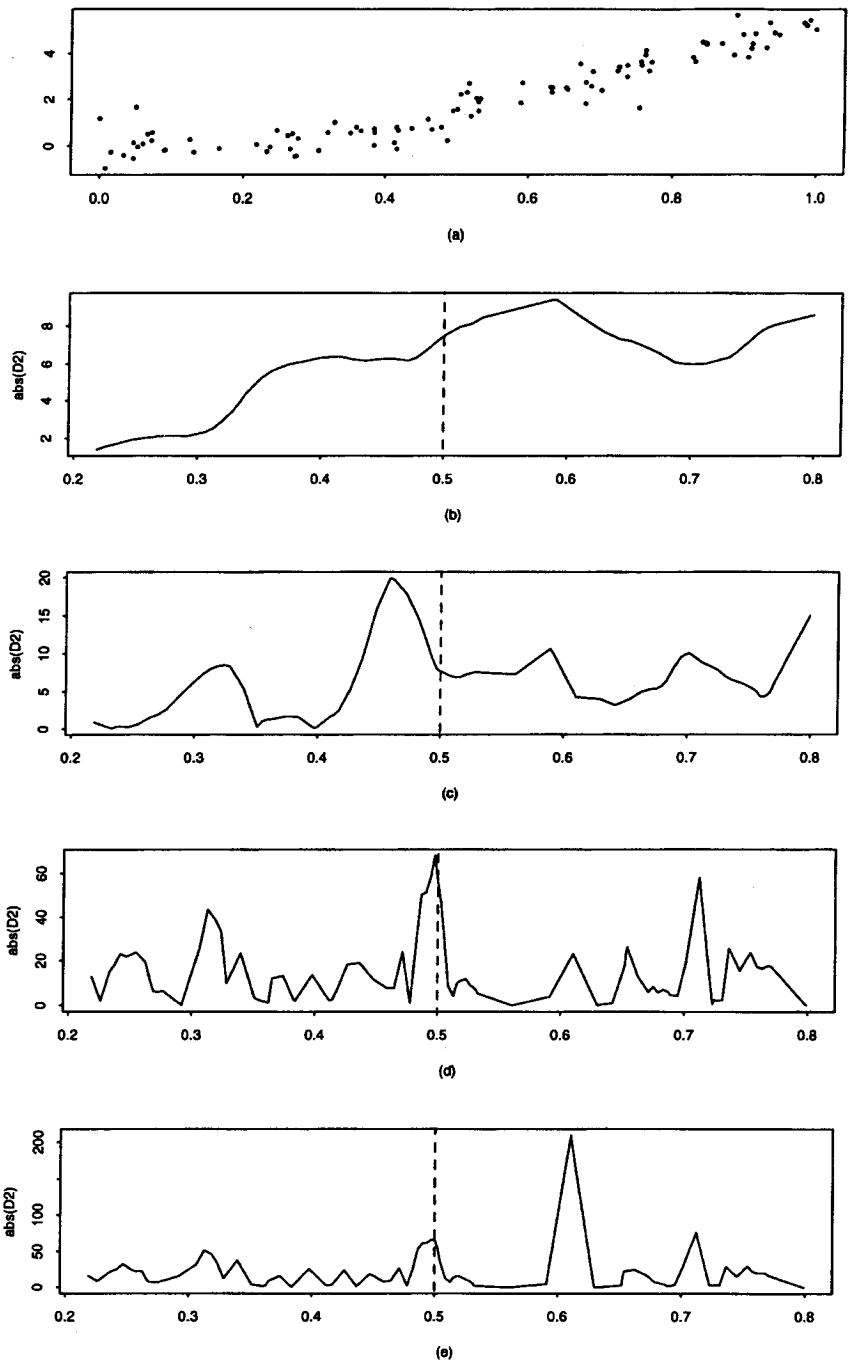

Fig. 1. Performance of the diagnostic in the case $g = g_1$, for $n = 100$ and $\sigma = 0.5$. Panel (a) is a scatterplot of a typical data set, and panels (b)–(e) depict plots of $|D_2(\cdot, h)|$ for $h = h_0 = 0.2$, $h = 0.08$, $h = \tilde{h} = 0.02188$ and $h = 0.02$, respectively.

over $\{(x,h) : x \in \mathcal{I}_h \text{ and } Cn^{-1}\log n \le h \le \eta_n\}$ is such that the x-component is within order $n^{-1}(\log n)^{1/2}$ of x_0 with high probability. Part (d) refines this into an explicit rate-of-convergence result for the estimator $\hat{x}_0$, under the additional condition (C_2).

In the case of Gaussian errors the $O_p(n^{-1})$ rate of convergence evidenced by (d) may be shown to be best possible for this particular estimator, in the sense that

$$\lim_{B_1 \to 0} \limsup_{n \to \infty} s_n(B_1) = 0.$$

For heavy-tailed error distributions, such as Student's t with d degrees of freedom, the rate may be shown to be no better than $O_p(n^{-1+\delta})$, where $\delta > 0$ is a decreasing function of d.

Next we describe performance of the second estimator suggested in Subection 2.3, derived using least squares. For explicitness, use some rule (which may be arbitrarily chosen) for breaking any ties that might arise in defining i_0 by minimizing the series at (2.4), and put $\hat{x}_0^* = \frac{1}{2}(X_{i_0} + X_{i_0+1})$. By way of regularity conditions, assume as in Subsection 2.3 that $h = cn^{-\alpha}$, where $c > 0$ and $0 < \alpha < 1$. (Recall that the first step in constructing $\hat{x}_0^*$ is to find the local maximum, $\tilde{x}_0^*$, of the diagnostic $|D(\cdot, h)|$ on $\mathcal{I}_h$, with $h = cn^{-\alpha}$. If the error distribution has finite moment generating function then we may take $h = c(n)n^{-1}\log n$ for any sequence $c(n) \to \infty$ such that $c(n)/n = O(n^{-\delta})$ for some $\delta > 0$.) Assume the same conditions on g and the design set $\{X_1, \ldots, X_n\}$ as in (C_1); let K have two bounded derivatives and be supported on $(-v, v)$, and satisfy $K(0) > K(x) \ge 0$ for all $x \ne 0$; and suppose the errors ϵ_i are independent and identically distributed with zero mean and $E|\epsilon_i|^\beta < \infty$, for some $\beta > \max\{2, 1/(1-\alpha)\}$. Collectively we call these conditions (C_3).

Given $B > 0$, define $t_n(B)$ to equal the probability that $|\hat{x}_0^* - x_0| \le Bn^{-1}$. Our next result shows that $\hat{x}_0^*$ converges to x_0 at rate n^{-1}.

THEOREM 2.2. *Assume* (C_3), *and let* D *be defined by either* (2.2) *or* (2.3). *Then*

$$\lim_{B \to \infty} \liminf_{n \to \infty} t_n(B) = 1.$$

An identical conclusion may be reached if r-th degree polynomials are fitted locally on either side of the jump point, provided that g has at least $r+1$ derivatives on either side. There are many other modifications of the algorithm which enjoy the convergence rate n^{-1}. One of these will be explored in Subsection 2.5.

It is easy to see that the rate n^{-1} is minimax optimal, for either regularly spaced or stochastic design. The "average" spacings between design points are of order n^{-1}, and so jump discontinuities cannot be determined with greater accuracy than this.

As in the case of the estimator $\hat{x}_0$, the limiting distribution of the rescaled estimation error $n(\hat{x}_0^* - x_0)$ is particularly complex and depends intimately on the distribution of ϵ_i.

2.5 *Numerical performance*

We conducted extensive simulations in a variety of settings, leading to the following conclusions.

(a) In cases where design is regularly spaced, the diagnostics D_1 and D_2 perform virtually identically; but if design is stochastic then D_2 is preferable, since it is better calibrated to counteract fluctuations in the density of design points.

(b) Performance of both $\hat{x}_0$ and $\hat{x}_0^*$ deteriorates as the tail weight of the error distribution increases. The deterioration is less marked in the case of $\hat{x}_0^*$, however.

(c) If there is a jump discontinuity at x_0, if g approaches and departs from the jump at not too steep an angle, and if g is not particularly steep in other places in $\mathcal{I}$, then the method based on fitting step functions performs well, and the estimator $\hat{x}_0^*$ is a little better than $\hat{x}_0$, even in small samples. If $g(x)$ approaches or departs steeply, however, then fitting straight line segments (rather than simply horizontal lines) on either side of the jump can improve performance significantly in small to moderate samples. We expect, but have not attempted to verify, that performance in general settings could be further enhanced by adaptive choice of the degrees of polynomials fitted on either side of the jump.

(d) If g has several different steep sections, or if it possesses more than one jump, then difficulty can be experienced keeping track of different local maxima of $|D|$ as h decreases. A simple algorithm, introduced below, overcomes this problem.

(e) If there are $k \geq 1$ jumps and k is known then their positions may be identified, and then enhanced using local least-squares, as in the case $k = 1$. If k is not known, however, then preliminary identification requires selection of a threshold for the diagnostic, and for that we do not yet have a general recommendation.

We do no more than summarize our numerical work here. In particular, we report only results in the case of uniform stochastic design, using the diagnostic D_2 with biweight kernel $K(x) = (1 - x^2)^2$ for $|x| \leq 1$; employing least-squares fitting of horizontal lines (the method addressed in Theorem 2.2); using $n = 100$, 250 or 500 and $\sigma = 0.1$ or 0.5 (with σ^2 the variance of the errors); and considering three functions, $g_1(x) = 4x^2 + I(x > \frac{1}{2})$, $g_2(x) = \cos\{8\pi(0.5 - x)\} - 2\cos\{8\pi(0.5 - x)\}I(x > 0.5)$ and

$$g_3(x) = \begin{cases} \exp\{-2(x - 0.35)\} - 1 & \text{if } x \in [0, 0.35) \\ \exp\{-2(x - 0.35)\} & \text{if } x \in [0.35, 0.65) \\ \exp\{2(x - 0.65)\} + \exp(-0.6) - 2 & \text{if } x \in [0.65, 1]. \end{cases}$$

Both g_1 and g_2 have a single jump discontinuity of size 1 at $x_0 = \frac{1}{2}$, whereas g_3 has two jump discontinuities, both of size 1 and occurring at $x_0 = 0.35$ and 0.65. All bias and standard deviation approximations in the tables were computed by averaging over the results of 200 simulations.

When calculating $\tilde{h}$, as a prelude to deriving $\hat{x}_0$, we suggest searching over a discrete set of bandwidths $h_i = h_0\rho^i$ for $i \geq 0$, where $h_0 > 0$ and $0 < \rho < 1$. We took $\rho = 0.9$, and $h_0 = 0.1$ or 0.2, throughout. (More generally, one could take h_0 to be an empirical bandwidth selector for smooth parts of the curve.) When $g = g_1$ and the error distribution was Normal $N(0, 0.1)$, choosing $h_0 = 0.2$ and

Table 1. Performance of $\hat{x}_0$ and $\hat{x}_0^*$ when $g = g_1$, for Normally distributed errors.

σ	n	$\hat{x}_0$			$\hat{x}_0^*$		
		bias × 100	s.d. × 100	% in [0.4, 0.6]	bias × 100	s.d. × 100	% in [0.4, 0.6]
0.1	100	0.2	4.0	95	0.1	1.8	100
	250	−0.1	1.1	100	0.0	2.1	99
	500	0.0	0.5	100	0.0	0.1	100
0.5	100	1.6	9.3	74	0.8	5.9	90
	250	1.0	4.6	94	0.3	1.6	100
	500	0.0	1.5	100	0.1	0.6	100

$\rho = 0.9$ gave average values (and standard deviations) of $\tilde{h}$, for sample sizes 100, 250 and 500, equal to 0.024 (0.007), 0.012 (0.003) and 0.007 (0.002), respectively. To compute $\hat{x}_0^*$ we took $h = 2\tilde{h}$ in the algorithm for finding $\tilde{x}_0^*$, the factor 2 arising from the fact that $\tilde{h}$ is already at the threshold, in a sense, and we wished to be above that value.

Table 1 lists biases, standard deviations, and the percentage of estimates in the interval $[0.4, 0.6]$, for the estimators $\hat{x}_0$ and $\hat{x}_0^*$, in the case of Normally distributed errors and for $g = g_1$. The gradient of this function is only 4 at the jump, and never exceeds 8 on $\mathcal{I} = [0, 1]$. Therefore it is to be expected that both estimators will perform well. In almost all cases $\hat{x}_0^*$, is a little better than $\hat{x}_0$. Results are similar for Student's t errors, except that error rates for both methods increase somewhat, and the superiority of $\hat{x}_0^*$ over $\hat{x}_0$ is more evident. In the case of Student's t with 3 or 6 degrees of freedom, and with $n = 100$, 250 or 500, the increase in root mean squared error for either $\hat{x}_0$ or $\hat{x}_0^*$ is usually less than three fold when $\sigma = 0.1$, but can be six fold (or occasionally more) when $\sigma = 0.5$.

For the parameter settings used to produce Table 1, Fig. 1 depicts a typical scatterplot of the data, and graphs of the function $|D_2(\cdot, h)|$ for several values of h. Observe that when h is either too small (meaning that asymptotically, $h = Cn^{-1} \log n$ for small C) or too large (asymptotically, $h = C$), the plot of $|D_2(\cdot, h)|$ shows large peaks well away from x_0.

The function g_2 presents special difficulties, since as well as having a jump, the absolute value of its gradient is as large as $8\pi \approx 25$ in eight distinct places. As a result, $|D_2(\cdot, h_i)|$ exhibits many local maxima, and (depending on choice of bandwidth) the largest of these may represent a steep gradient rather than the jump. We suggest below a four-step algorithm for overcoming this difficulty. It monitors all the local maxima of $|D_2(\cdot, h_i)|$ in $(vh_0, 1 - vh_0)$, until one or more clearly provide a dominant improvement over the case $i = 0$; and it selects the locations of those as preliminary estimates of the sites of jumps. See Fig. 2. The algorithm uses the definition of h_i given two paragraphs above.

Step 1. *Initialization.* Let M denote the number of local maxima of $|D_2(\cdot, h_0)|$ on $(vh_0, 1 - vh_0)$, and let $\{\xi_{0j}, 1 \le j \le M\}$ denote the points at which

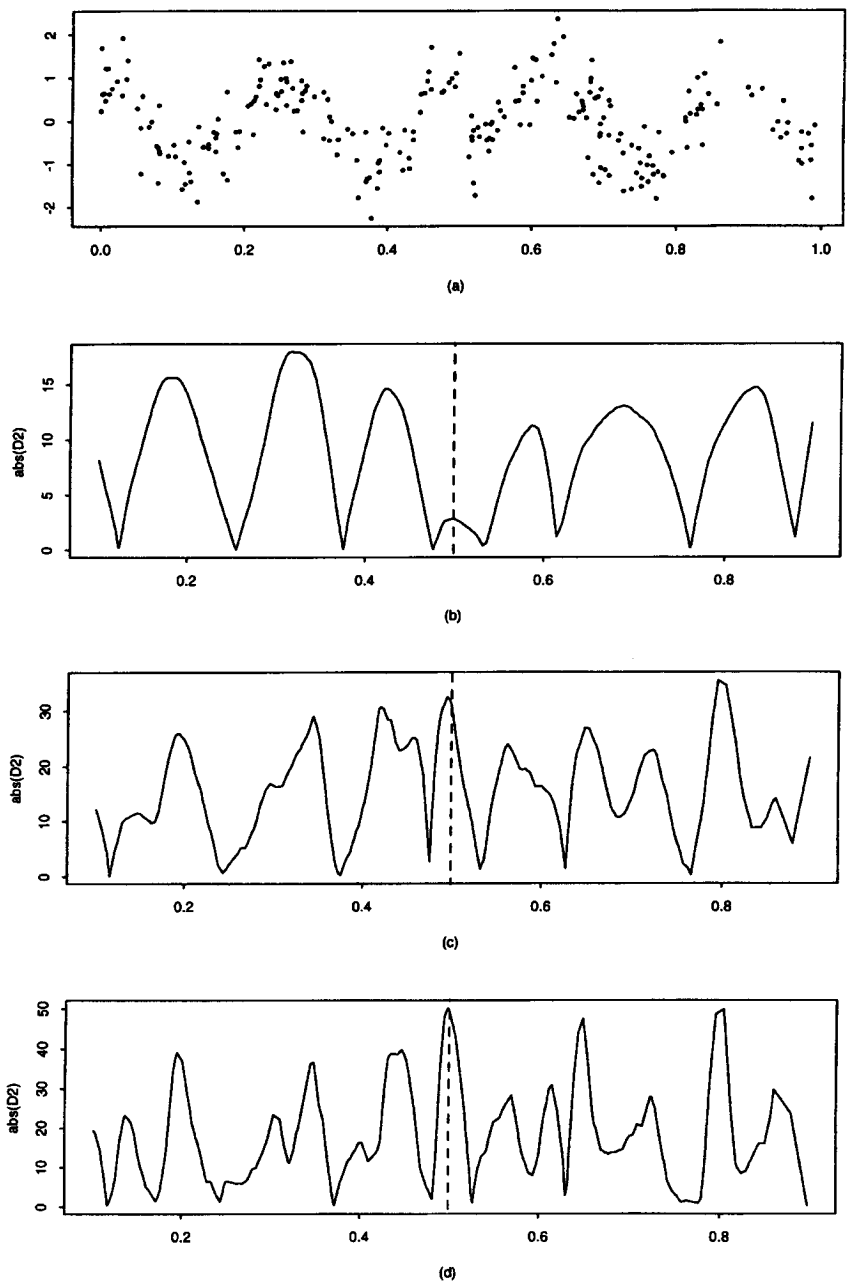

Fig. 2. Performance of the diagnostic in the case $g = g_2$, for $n = 250$ and $\sigma = 0.5$, and using the four-step algorithm. Panel (a) is again a scatterplot of a typical data set, while panels (b)–(d) illustrate plots of $|D_2(\cdot, h)|$ for $h = h_0 = 0.1$, $h = 0.05$ and $h = h_{\hat{i}} = 0.03487$, respectively.

Table 2. Performance of $\check{x}_0$ and $\check{x}_0^*$ when $g = g_2$, for Normally distributed errors.

σ	n	$\check{x}_0$			$\check{x}_0^*$		
		bias x 100	s.d. x 100	% in [0.4, 0.6]	bias x 100	s.d. x 100	% in [0.4, 0.6]
0.1	100	−1.0	14.4	70	−1.4	16.1	60
	250	−0.1	3.8	98	−0.1	4.0	98
	500	0.0	0.5	100	0.0	0.1	100
0.5	100	−1.5	15.8	64	−2.0	17.4	52
	250	−0.3	7.4	92	−0.4	7.5	92
	500	0.0	0.7	100	0.0	0.5	100

they are achieved.

Step 2. *Iteration.* Given a set $\{\xi_{ij}, 1 \le j \le M\}$ of local maxima of $|D_2(\cdot, h_i)|$ in $(vh_0, 1-vh_0)$, let $\xi_{i+1,j}$ denote the local maximum of $|D_2(\cdot, h_{i+1})|$ that is nearest to ξ_{ij}.

Step 3. *Termination.* Stop the algorithm at iteration $i = \check{\imath}$, say, where the number of data values in some interval $(x - h_i, x + h_i) \subseteq (vh_0, 1 - vh_0)$ first falls below a predetermined value ν. If we seek the locations of k jump points $x_1, \ldots, x_k$, we take the preliminary estimates $\check{x}_1, \ldots, \check{x}_k$ of these to be those values of $\xi_{\check{\imath}j}$ for which $|D_2(\xi_{\check{\imath}j}, h_{\check{\imath}}) - D_2(\xi_{0j}, h_0)|$ is one of the k largest.

Step 4. *Least squares.* Refine these preliminary estimates to $\check{x}_1^*, \ldots, \check{x}_k^*$, by least-squares fitting of an l-th degree polynomial within $(\check{x}_j - vh_{\check{\imath}}, \check{x}_j + vh_{\check{\imath}})$ for $1 \le j \le k$.

It is straightforward to modify the proof of Theorem 2.2 to show that if $\nu = cn^{1-\alpha}$ then, under identical conditions except for the assumption of $k \ge 1$ jumps, $\check{x}_j^* - x_j = O_p(n^{-1})$ for $1 \le j \le n$. (If the error distribution has finite moment generating function then $\nu = c(n) \log n$, for $c(n) \to \infty$ such that $c(n) = O(n^\delta)$, is permissible.) In the case of the function g_2 and Normal errors, Table 2 describes the result of applying this method with $h_i = 0.1\ 0.9^i$, $\nu = \frac{1}{2}(\log n)^2$, $k = 1$ and $l = 0$. The estimators $(\check{x}_0, \check{x}_0^*)$ do well at selecting the correct local maximum, and significantly better than the estimators $(\hat{x}_0, \hat{x}_0^*)$ (for which results are not tabulated here). Since the first derivative of g_2 is so great on either side of the jump then the least-squares fitting step makes hardly any difference to the result, although for larger n it gives noticeable improvement.

The case of g_3 is similar, as Table 3 illustrates. There, all parameter settings are as for Table 2, except that now $h_0 = 0.2$ and of course $g = g_3$ and $k = 2$. Since g_3 changes relatively slowly near the jumps then the least-squares fitting step does improve accuracy.

Table 3. Performance of $\check{x}_0$ and $\check{x}_0^*$ when $g = g_3$, for Normally distributed errors.

σ	n	$\check{x}_0$				$\check{x}_0^*$			
		0.35		0.65		0.35		0.65	
		bias × 100	s.d. × 100	bias × 100	s.d. × 100	bias × 100	s.d. × 100	bias × 100	s.d. × 100
0.1	100	0.3	1.8	−0.8	2.9	0.1	0.7	−0.2	2.1
	250	0.0	0.9	0.0	0.9	0.0	0.2	0.0	0.3
	500	0.0	0.6	0.0	0.5	0.0	0.1	0.0	0.1
0.5	100	0.8	4.2	−1.7	4.3	0.2	3.9	−1.3	4.0
	250	0.1	2.7	−0.9	3.8	0.2	2.7	−0.9	3.6
	500	0.1	1.1	−0.2	1.6	0.0	0.8	−0.2	1.6

3. Technical arguments

The symbols $C, C_1, C_2, \ldots$ will denote positive constants.

3.1 *Proof of Theorem* 2.1

For the sake of brevity we give the proof only in the case $D = D_1$ (see (2.2)). The case of D_2 is similar. Indeed, note that if we define

$$A_1(x,h) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\} Y_i \quad \text{and}$$

$$A_2(x,h) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\},$$

then $D_2(x,h) = A_3(x,h) - A_4(x,h)$, where

$$A_3(x,h) = D_1(x,h)/A_2(x,h) \quad \text{and}$$
$$A_4(x,h) = \{\partial A_2(x,h)/\partial x\} A_1(x,h)/A_2(x,h).$$

It may be shown that if h is of size equal to a large constant multiple of $n^{-1} \log n$, or larger, then with probability tending to 1 as $n \to \infty$, the supremum of $|A_3(\cdot, h)|$ dominates that of $|A_4(\cdot, h)|$. Therefore, the technical argument for deriving properties of D_2 is in effect that for deriving properties of D_1.

Step 1. *Bias contributions.* Let f denote the design density in the case of stochastic design, and put $f \equiv 1$ for regularly spaced design. It may be proved by Taylor expansion and integral approximations to series that for either regularly spaced or stochastic design,

$$\beta(x) = E\{D(x,h)\} = h^{-1} f(x_0) g_2(x_0) K\{(x - x_0)/h\} + O\{(nh^2)^{-1} + (n^2h^3)^{-1} + 1\}, \tag{3.1}$$

uniformly in $x \in \mathcal{I}_h$, $0 < h \leq 1$ and $n \geq 1$.

Step 2. *Stochastic contributions*. In the case of regularly spaced design there is a single stochastic contribution,

$$\xi_1(x) = (nh^2)^{-1} \sum_{i=1}^{n} K'\{(x - X_i)/h\}\epsilon_i.$$

Then, $D = \beta + \xi_1$. With stochastic design, however, there is also a contribution from the design, through the term

$$\xi_2(x) = (nh^2)^{-1} \sum_{i=1}^{n} [K'\{(x - X_i)/h\}g(X_i) - h^2\beta(x)].$$

In this setting, $D = \beta + \xi_1 + \xi_2$.

Step 2(i). *Approximation to the process* ξ_1. If the errors are not Normally distributed, we approximate to ξ_1 using partial sums constructed in the Hungarian fashion. Define $S_i = \sum_{j \le i} \epsilon_j$, and let σ^2 denote the variance of ϵ_i. There exists a standard Brownian motion W, and positive constants c_1, c_2 and c_3, depending only the error distribution, such that, with

$$\Delta_1 = \max_{1 \le i \le n} |S_i - \sigma W(i)|,$$

we have for $j = 1$,

$$P(\Delta_j > c_1 \log n + x) \le c_2 \exp(-c_3 x) \tag{3.2}$$

for all $x > 0$. See for example Shorack and Wellner ((1986), p. 66 ff.). Putting $w_i = W(i) - W(i-1)$,

$$\begin{aligned} \delta_{i1}(x) &= |K'\{(x - X_i)/h\} - K'\{(x - X_{i-1})/h\}| \\ &\le \delta_{i2}(x) = C_1 h^{-1}(X_i - X_{i-1}) \\ &\quad \cdot \{I(|x - X_{i-1}| \le C_2 h) + I(|x - X_i| \le C_2 h)\} \end{aligned} \tag{3.3}$$

(where C_1 and C_2 are constants depending only on K, and $X_0 = 0$), and

$$\xi_3(x) = (nh^2)^{-1} \sum_{i=1}^{n} K'\{(x - X_i)/h\}\sigma w_i,$$

we have

$$nh^2|\xi_1(x) - \xi_3(x)| - K(x/h)\Delta_1 \le 2\Delta_1 \sum_{i=1}^{n} \delta_{i1}(x) \le 2\Delta_1 \sum_{i=1}^{n} \delta_{i2}(x). \tag{3.4}$$

We bound the right-hand side using somewhat different arguments in the cases of regularly spaced and stochastic design, respectively, and treat only the latter here. There, first define $\mathcal{E}_1(C)$ to be the event

$$\left\{ \sum_{i=1}^{n} I(|x - X_i| \le C_2 h) > Cnh \text{ for some } x \in \mathcal{I} \text{ or } h \in [n^{-1} \log n, 1] \right\}.$$

Use Bernstein's inequality, and an approximation to $\mathcal{E}_1(C)$ on a lattice of pairs (x, h), of fineness $O(n^{-k})$ for fixed but arbitrarily large $k > 0$, to show that if $C_3 > 0$ is given then $C_4 > 0$ may be chosen so large that $P\{\mathcal{E}_1(C_4)\} \le n^{-C_3}$ for all sufficiently large n. Let F denote the distribution function corresponding to density f. If the complement of $\mathcal{E}_1(C_4)$ holds then

$$h\sum_{i=1}^{n}\delta_{i2}(x) \le 2C_1\max{}^*\sum_{i=i_1}^{i_2}(X_i - X_{i-1}) \le 2C_1(\sup f^{-1})\max{}^*(U_{i_2} - U_{i_1-1}),$$

where $\max^*$ denotes the maximum over integer pairs (i_1, i_2) such that $1 \le i_1 \le i_2 \le n$ and $i_2 - i_1 + 1 \le C_4nh$, and where $U_i = F(X_i)$ are the order statistics of a Uniform random sample on the interval $\mathcal{I}$. Let $\mathcal{E}_2(C)$ denote the event that $U_{i_2} - U_{i_1-1} > Ch$ for some pair (i_1, i_2) such that $1 \le i_1 \le i_2 \le n$ and $i_2 - i_1 + 1 \le C_4nh$, and some $h \in [n^{-1}\log n, 1]$. It may be proved from the "other Hungarian construction" (see Theorem 3, p. 495 of Shorack and Wellner (1986)) that $C_5 > 0$ may be chosen so large that $P\{\mathcal{E}_2(C_5)\} \le n^{-C_3}$ for all sufficiently large n. Combining the results in this paragraph we deduce that if $C_3 > 0$ is given then C_6 may be chosen so large that

$$P\left\{\sum_{i=1}^{n}\delta_{i2}(x) \le C_6 \text{ for all } x \in \mathcal{I} \text{ and all } h \in [n^{-1}\log n, 1]\right\} > 1 - n^{-C_3} \tag{3.5}$$

for all sufficiently large n.

From (3.2)–(3.5) we deduce that in the cases of regularly spaced and stochastic design, given $C_3 > 0$ there exists $C_7 > 0$ such that for $j = 1$,

$$P\{|\xi_j(x) - \xi_{j+2}(x)| \le C_7(nh^2)^{-1}\log n \text{ for all } x \in \mathcal{I} \text{ and all } h \in [n^{-1}\log n, 1]\} \ge 1 - n^{-C_3} \tag{3.6}$$

for all sufficiently large n.

Step 2(ii). *Approximation to the process* ξ_2. Here the design is assumed to be stochastic. The "other Hungarian construction" (Shorack and Wellner (1986), p. 495) may be employed to show that there exists a standard Brownian Bridge B such that, defining

$$\xi_4(x) = n^{-1/2}h^{-2}\int B\{F(x - hz)\}d\{K'(z)g(x - hz)\},$$

(3.6) holds for $j = 2$ and some $C_7 = C_7(C_3) > 0$.

Step 2(iii). *Upper bounds for the processes* ξ_1 *and* ξ_2. Fernique's lemma (Marcus (1970)) may be employed to prove that for each given $C_3 > 0$ there exists $C_8 > 0$ so large that for $j = 3$ and 4,

$$P[|\xi_j(x)| \le C_8\{(nh^3)^{-1}\log n\}^{1/2} \text{ for all } x \in \mathcal{I} \text{ and all } h \in [n^{-1}\log n, 1]] \ge 1 - n^{-C_3}. \tag{3.7}$$

In view of (3.6), which holds for $j = 1$ and 2, (3.7) is also valid for $j = 1$ and 2.

Step 3. *Upper bound to convergence rate.* We may write $D = \beta + \xi_5$, where $\xi_5 = \xi_1$ if design is regularly spaced, and $\xi_5 = \xi_1 + \xi_2$ in the case of stochastic design. Furthermore, from (3.1) we have $\beta = \beta_1 + \beta_2$, where $\beta_1(x) = h^{-1} f(x_0) g_2(x_0) K\{(x - x_0)/h\}$ and

$$\sup_{x \in \mathcal{I}_h} |\beta_2(x)| \le C_9(\delta)\{(nh^2)^{-1} + 1\} \tag{3.8}$$

for all $h \in [n^{-1}, 1]$. In view of (3.7) and (3.8), for each $C_3, \eta > 0$ and $\delta \in (0, \min(x_0, 1 - x_0))$ there exists $C_{10} > 0$ such that for all sufficiently large n,

$$P\{|\xi_5(x)| \le \eta h^{-1} \text{ for all } x \in \mathcal{I} \text{ and all } h \in [C_{10} n^{-1} \log n, 1]\} \ge 1 - n^{-C_3},$$
$$\sup_{x \in \mathcal{I}_h} |\beta_2(x)| \le \eta h^{-1} \text{ for all } h \in [n^{-1} \log n, 1].$$

Together these results imply that if η_n denotes any sequence of positive numbers decreasing to zero more slowly than $n^{-1} \log n$, if $\eta > 0$, and if the support of K is contained within $(-v, v)$, then

$$\lim_{C \to \infty} \liminf_{n \to \infty} P\Bigg\{ \sup_{|x - x_0| \le vh} |D(x,h)| > \sup_{x \in \mathcal{I}_h, |x - x_0| > vh} |D(x,h)| \text{ for all } h \in [Cn^{-1} \log n, \eta_n]; \text{ and the maximum of } |D(x,h)| \text{ over all } x \in \mathcal{I}_h \text{ and all } h \in [Cn^{-1} \log n, \eta_n] \text{ occurs at a pair } (x,h) \text{ satisfying } |x - x_0| \le vh \text{ and } h \in [Cn^{-1} \log n, C(1+\eta) n^{-1} \log n] \Bigg\} = 1. \tag{3.9}$$

Part (a) of the theorem follows. Result (3.9) also implies that if $\tilde{h}$ is as suggested in the definition of $\hat{x}_0$ then

$$\liminf_{C \to \infty} \liminf_{n \to \infty} P(\tilde{h} \le Cn^{-1} \log n) = 1. \tag{3.10}$$

In view of (3.9), if $h_1 = Cn^{-1} \log n$ and C is large then with high probability the value of x that maximizes $|D(x,h)|$ over $x \in \mathcal{I}_h$ and $h \in [Cn^{-1} \log n, \eta_n]$ is within $vh_1 = O(n^{-1} \log n)$ of x_0. In the next step we show that it actually occurs within $O\{n^{-1} (\log n)^{1/2}\}$ of x_0.

Step 4. *Refined upper bound to convergence rate.* Our argument is by Taylor expansion. Let $\alpha = (x - x_0)/h$, where we shall assume that $|\alpha| < v$ and $h \in \mathcal{H} = [C_{11} n^{-1} \log n, C_{12} n^{-1} \log n]$ for positive constants $C_{11} < C_{12}$. Observe that $\beta_1(x) = h^{-1} f(x_0) g_2(x_0) \{K(0) - \frac{1}{2}\alpha^2 K''(0) + \alpha^2 \rho(\alpha)\}$, where the function ρ is bounded in $|\alpha| \le v$, and satisfies $\rho(\alpha) \to 0$ as $\alpha \to 0$. Also, $\xi_5(x) = \xi_5(x_0) + \alpha R(x)$, where $R(x) = \xi_5'(x_0 + \alpha\theta)$ for some $0 \le \theta \le 1$, and $R(x) = O_p\{(nh^3)^{-1/2}\}$

uniformly in $|x - x_0| \le vh$ and $h \in \mathcal{H}$. It follows that, uniformly in $|\alpha| \le v$ and $h \in \mathcal{H}$,

$$D(x,h) = h^{-1} f(x_0) g_2(x_0) \left\{ K(0) - \frac{1}{2}\alpha^2 K''(0) + \alpha^2 \rho(\alpha) \right\} + O_p\{|\alpha|(nh^3)^{-1/2}\}.$$

Hence, the pair $(\bar{x}, \bar{h})$ that maximizes $|D(x,h)|$ over $(-v,v) \times \mathcal{H}$ has the property that $\alpha = \alpha(\bar{x}) = O_p\{(\log n)^{-1/2}\}$. Hence, $\bar{x} - x_0 = \bar{h}\alpha = O_p\{n^{-1}(\log n)^{1/2}\}$. Equivalently,

$$\liminf_{C_{13}\to\infty} \liminf_{n\to\infty} P\{|\bar{x} - x_0| \le C_{13} n^{-1} (\log n)^{1/2}\} = 1. \tag{3.11}$$

Results (3.9) and (3.11) together imply part (b) of the theorem.

Step 5. *Lower bound.* Let $\check{x}(h)$ denote the value of x that maximizes $|D(\cdot,h)|$ on $\mathcal{I}_h$. We shall prove that under condition (C_2),

$$\lim_{C_{14}\to 0} \liminf_{C\to 0} \liminf_{n\to\infty} P\{|\check{x}_0(Cn^{-1}\log n) - x_0| > C_{14}\} = 1. \tag{3.12}$$

This implies that

$$\lim_{C\to 0} \liminf_{n\to\infty} P(\tilde{h} > Cn^{-1}\log n) = 1. \tag{3.13}$$

Parts (c) and (d) of the theorem follow from (3.9) and (3.11). In the case of part (d) we need to show that with high probability, $\tilde{h}$ in the definition of $\hat{x}_0$ lies in the interval $\mathcal{H}$ defined in Step 4, for small C_{11} and large C_{12}. That this is indeed true follows from (3.10) and (3.13).

For the sake of simplicity we shall treat only the case of regularly spaced design. If $Y_1, \ldots, Y_m$ are independent random variables and $P(Y_i > y) \ge \exp(-\zeta_1 y^2)$ for all i and all $y > \zeta_2 > 0$, then if $y > m^{1/2}\zeta_2$,

$$P\left(m^{-1/2} \sum_{i=1}^{m} Y_i > y \right) \ge \prod_{i=1}^{m} P(Y_i > m^{-1/2} y) \ge \exp(-\zeta_1 y^2). \tag{3.14}$$

For each $x \in \mathcal{I}_h$, the quantity $nh^2\xi_5(x)$ equals a weighted sum of approximately $2vnh$ independent values of ϵ_i, the weights being of course derived from K. From this result and (3.14) (implied in the present context by (C_2)) it follows that if $h = Cn^{-1}\log n$ then for each $\zeta_3 > 0$ and some $\zeta_4 > 0$ not depending on ζ_3,

$$\inf_{x\in\mathcal{I}_h} P\{|nh^2\xi_5(x)| > \zeta_3 (nh\log n)^{1/2}\} \ge \exp(-\zeta_3^2 \zeta_4 \log n)$$

for all sufficiently large n. Hence, since the variables $\xi_5(2jvh)$, $j \ge 1$, are independent, then, writing j_1 and j_2 for the integer parts of $\{\delta/(2vh)\}+1$ and $(1-\delta)/(2vh)$

respectively,

$$P\left[\sup_{x\in\mathcal{I}_h}|\xi_5(x)| \le \zeta_3\{(nh^3)^{-1}\log n\}^{1/2}\right]$$
$$\le \prod_{j=j_1}^{j_2} P[|\xi_5(2jvh)| \le \zeta_3\{(nh^3)^{-1}\log n\}^{1/2}]$$
$$\le \prod_{j=j_1}^{j_2}\{1-\exp(-\zeta_3^2\zeta_4\log n)\} \le (1-n^{-\zeta_3^2\zeta_4})^{\{(1-2\delta)/(2vh)\}-2}.$$

For large n the right-hand side is dominated by $(1-n^{-\zeta_3^2\zeta_4})^{C_{12}n/\log n}$, where C_{12} is any constant strictly less than $(1-2\delta)/(2vC)$. Therefore, if ζ_3 is chosen so small that $\zeta_3^2\zeta_4 < 1$, and $h = Cn^{-1}\log n$, then for all $C > 0$,

$$P\left[\sup_{x\in\mathcal{I}_h}|\xi_5(x)| > \zeta_3\{(nh^3)^{-1}\log n\}^{1/2}\right] \to 1 \tag{3.15}$$

as $n \to \infty$.

A minor modification of the argument leading to (3.15) shows that for any sufficiently small $\zeta_3 > 0$, and for $h = Cn^{-1}\log n$ and all $C > 0$,

$$\lim_{\zeta_5\to 0}\liminf_{n\to\infty} P\Big[\sup_{x\in\mathcal{I}_h}|\xi_5(x)| > \zeta_3\{(nh^3)^{-1}\log n\}^{1/2}, \text{ and the supremum occurs at a point } x \text{ such that } |x-x_0| > \zeta_5\Big] = 1 \tag{3.16}$$

for each $C > 0$. The desired result (3.12) follows from (3.1) and (3.16).

3.2 *Proof of Theorem* 2.2

Again, we confine attention to the case where the diagnostic D is the function D_1, defined at (2.2).

Step (i). *Performance of diagnostic.* We outline a proof of the following analogue of (3.9) for the single bandwidth $h = cn^{-\alpha}$ used here:

$$P\left\{\sup_{|x-x_0|\le vh}|D(x,h)| > \sup_{x\in\mathcal{I}_h,|x-x_0|>vh}|D(x,h)|\right\} \to 1. \tag{3.17}$$

This result implies that

$$P(|\tilde{x}_0^* - x_0| \le vh) \to 1. \tag{3.18}$$

Result (3.1) continues to hold under the present weaker conditions on K, since it does not require a second derivative. Under the condition $E|\epsilon_i|^\beta < \infty$, rather

than existence of a finite moment generating function, the embedding result (3.2) in the case $j = 1$ should be weakened to

$$\max_{1 \le i \le n} |S_i - \sigma W(i)| = O_p(n^{(1/\beta)+\eta})$$

for each $\eta > 0$. See for example Shorack and Wellner ((1986), pp. 60–61). This leads to the following weaker version of (3.6) (for $j = 1$):

$$\sup_{x \in \mathcal{I}} |\xi_1(x) - \xi_3(x)| = O_p(n^{(1/\beta)+\eta-1}h^{-2}). \tag{3.19}$$

In the case $j = 2$, (3.6) holds without change; and (3.7) is valid for $j = 1$ and 2, without alteration. (We define $\xi_1, \ldots, \xi_5$ as in the proof of Theorem 2.1.) When $j = 2$ we obtain from (3.6) and (3.7) that

$$\sup_{x \in \mathcal{I}} |\xi_2(x)| = O_p[\{(nh^3)^{-1} \log n\}^{1/2} + (nh^2)^{-1} \log n];$$

and by (3.7) in the case $j = 1$, and (3.19), we have

$$\sup_{x \in \mathcal{I}} |\xi_1(x)| = O_p\{n^{(1/\beta)+\eta-1}h^{-2} + (nh^2)^{-1} \log n\}.$$

Combining the last two displayed formulae, and assuming that $\beta > 1/(1 - \alpha)$, we see that for $j = 1$ and 2, and hence for $j = 5$, $\eta > 0$ may be chosen so small that the result

$$\sup_{x \in \mathcal{I}} |\xi_j(x)| = o_p(h^{-1})$$

obtains. The claimed formula (3.17) follows from this and (3.1).

Step (ii). *Performance of least squares.* Let $\mathcal{U}$ denote the set of all pairs (u_1, u_2) such that $u_1 \in [x_0 - 3vh, x_0 - vh]$ and $u_2 \in [x_0 + vh, x_0 + 3vh]$, and let $\hat{x}_0^*(u_1, u_2)$ denote the version of $\hat{x}_0^*$ defined by minimizing the sum of squares at (2.4) when $\{i_1, i_1 + 1, \ldots, i_2\}$ is redefined to equal the set of integers i such that $X_i \in [u_1, u_2]$. In view of (3.18), the proof of the theorem will be complete if we show that

$$\lim_{B \to \infty} \liminf_{n \to \infty} P\left\{ \sup_{(u_1,u_2) \in \mathcal{U}} |\hat{x}_0^*(u_1, u_2) - x_0| \le Bn^{-1} \right\} = 1. \tag{3.20}$$

The proof will be given only in outline.

It is straightforward to show that for each $\eta > 0$,

$$P\left\{ \sup_{(u_1,u_2) \in \mathcal{U}} |\hat{x}_0^*(u_1, u_2) - x_0| \le \eta \right\} \to 1.$$

Therefore, in establishing (3.20) it suffices to confine attention to the case where the series at (2.4) is minimized over integers i_0 that satisfy $|F(i_0/n) - x_0| \le \eta_n$

for a sequence η_n converging to zero arbitrarily slowly (certainly more slowly than h), where F is the distribution function of the design density. (We take $F(u) \equiv u$ if the design is regularly spaced.)

Given $(u_1, u_2) \in \mathcal{U}$, let $i_1, i_1 + 1, \ldots, i_2$ be the set of indices i such that $X_i \in [u_1, u_2]$, let $\mathcal{J}_0 = \{i_0 : |F(i_0/n) - x_0| \le \eta_n\}$, and let $\mathcal{J} = \mathcal{J}(u_1, u_2)$ denote the set of triples (i_0, i_1, i_2) arising in this way, with $i_1 \le i_0 < i_2$. We estimate the change point in the data set $(X_{i_1}, Y_{i_1}), \ldots, (X_{i_2}, Y_{i_2})$ by minimizing over $i_0 \in \mathcal{J}_0$ the sum of squares,

$$S(i_0) = \sum\nolimits_1 (Y_j - \bar{Y}_1)^2 + \sum\nolimits_2 (Y_j - \bar{Y}_2)^2,$$

where $\sum_1$ and $\sum_2$ denote summation over indices j with $i_1 \le j \le i_0$ and $i_0 + 1 \le j \le i_2$, respectively, and where $\bar{Y}_1 = (i_0 - i_1 + 1)^{-1} \sum_1 Y_j$ and $\bar{Y}_2 = (i_2 - i_0)^{-1} \sum_2 Y_j$, with $i_1 \le i_0 < i_2$. We estimate the changepoint to occur at $\frac{1}{2}(X_{\hat{i}_0} + X_{\hat{i}_0+1})$ if $S(\hat{i}_0) \le S(i_0)$ for all $i_1 \le i_0 < i_2$. (Ties for the definition of $\hat{i}_0$ may be broken in an arbitrary manner.)

Suppose the changepoint x_0 really lies between $X_{i_{00}}$ and $X_{i_{00}+1}$. With probability tending to 1, $i_1 \le i_{00} < i_2$ uniformly in all triples $(i_0, i_1, i_2) \in \mathcal{J}(u_1, u_2)$ and all $(u_1, u_2) \in \mathcal{U}$, and so we may suppose without loss of generality that this inequality holds. Let $\sum_{01}$ and $\sum_{02}$ denote summation over integers j satisfying $i_1 \le j \le i_{00}$ and $i_{00} + 1 \le j \le i_2$, respectively, and put $m = i_0 - i_{00}$, $\Delta = \sum_1 Y_j - \sum_{01} Y_j$, $\delta_1 = m/(i_{00} - i_1 + 1)$, $\delta_2 = m/(i_2 - i_{00})$, $A_1 = (i_{00} - i_1 + 1)^{-1} \sum_{01} Y_j$ and $A_1 = (i_2 - i_{00})^{-1} \sum_{01} Y_j$. We may prove that in this notation, $S(i_{00}) - S(i_0) = T + U$, where $T = (A_1 - A_2)\{2\Delta - m(A_1 + A_2)\}$ and

$$\begin{aligned} U = & -2\Delta\{A_1(\delta_1 - \delta_1^2 + \delta_1^3 - \cdots) + A_2(\delta_2 + \delta_2^2 + \delta_2^3 + \cdots)\} \\ & + \Delta^2\{(i_{00} - i_1 + 1)^{-1}(1 - \delta_1 + \delta_1^2 - \cdots) \\ & + (i_2 - i_{00})^{-1}(1 + \delta_2 + \delta_2^2 + \cdots)\} \\ & + (i_{00} - i_1 + 1)(\delta_1^2 - \delta_1^3 + \cdots)A_1^2 + (i_2 - i_{00})(\delta_2^2 + \delta_2^3 + \cdots)A_2^2. \end{aligned}$$

Given any large constant $C_{15} > 1$, and a statistic W such as $|U/m|$, let $\sup^* W$ denote the supremum of W over all triples $(i_0, i_1, i_2) \in \mathcal{J}(u_1, u_2)$, all $(u_1, u_2) \in \mathcal{U}$ and all integers m with $1 \le |m| \le C_{15} n \eta_n$. Let $\sup_C^*$ denote the same supremum when the range of m's is restricted to $C \le |m| \le C_{15} n \eta_n$, for some $C \ge 1$. Put $a_1 = g(x_0-)$ and $a_2 = g(x_0+)$, and let $\mathcal{X} = \{X_1, \ldots, X_n\}$ denote the set of design points. It may be proved that for each $\zeta > 0$, the conditional probabilities $P(\sup^* |U/m| > \zeta \mid \mathcal{X})$, $P(|A_1 - a_1| > \zeta \mid \mathcal{X})$ and $P(|A_2 - a_2| > \zeta \mid \mathcal{X})$ all converge to zero in probability. Hence, the unconditional probabilities converge to zero as well. Therefore, defining $V = (a_1 - a_2)\{2\Delta - m(a_1 + a_2)\}$, we have

$$P(\sup^* |[\{S(i_{00}) - S(i_0)\} - V]/m| > \zeta \mid \mathcal{X}) \to 1$$

in probability. Furthermore, for each $\zeta > 0$,

$$\lim_{C \to \infty} \liminf_{n \to \infty} P\{\sup\nolimits_C^* m^{-1} V \le -(a_1 - a_2)^2 + \zeta\} = 1.$$

Combining the last two displayed formulae we see that for all $\zeta > 0$,

$$\lim_{C\to\infty} \liminf_{n\to\infty} P[\sup\nolimits_C^* m^{-1}\{S(i_{00}) - S(i_0)\} \le -(a_1 - a_2)^2 + \zeta] = 1.$$

Hence, the probability $\pi_n(C)$ that the minimum of $S(i_0)$ over $(i_0, i_1, i_2) \in \mathcal{J}(u_1, u_2)$, $(u_1, u_2) \in \mathcal{U}$ and integers m with $|m| \le C_{15} n\eta_n$, occurs with $|m| = |i_0 - i_{00}| \le C$, satisfies

$$\lim_{C\to\infty} \liminf_{n\to\infty} \pi_n(C) = 1. \tag{3.21}$$

Formula (3.20), in the modified form suggested in the paragraph below that result, follows from (3.21).

Acknowledgements

The helpful comments of two reviewers are gratefully acknowledged.

REFERENCES

Carlstein, E., Müller, H.-G. and Siegmund, D. (1994). *Change-Point Problems*, Institute of Mathematical Statistics Lecture Note Series Vol. 23, Hayward, California.

Eubank, R. L. and Speckman, P. L. (1993). Confidence bands in nonparametric regression, *J. Amer. Statist. Assoc.*, **88**, 1287–1301.

Eubank, R. L. and Speckman, P. L. (1994). Nonparametric estimation of functions with jump discontinuities, *Change-Point Problems* (eds. E. Carlstein, H.-G. Müller and D. Siegmund), Institute of Mathematical Statistics Lecture Note Series Vol. 23, 130–144, Hayward, California.

Hall, P. and Titterington, D. M. (1992). Edge-preserving and peak-preserving smoothing, *Technometrics*, **34**, 429–440.

Korostelev, A. P. (1987). On minimax estimation of a discontinuous signal, *Theory Probab. Appl.*, **32**, 727–730.

Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image Reconstruction*, Lecture Notes in Statistics Vol. 82, Springer, Berlin.

Loader, C. L. (1997). Change-point estimation using nonparametric regression, *Ann. Statist.*, **24**, 1667–1678.

Mcdonald, J. A. and Owen, A. B. (1986). Smoothing with split linear fits, *Technometrics*, **28**, 195–208.

Marcus, M. B. (1970). A bound for the distribution of the maximum of continuous Gaussian processes, *Ann. Math. Statist.*, **41**, 305–309.

Müller, H.-G. (1992). Change-points in nonparametric regression analysis, *Ann. Statist.*, **20**, 737–761.

Müller, H.-G. and Song, K.-S. (1997). Two-stage change-point estimators in smooth regression models, *Statist. Probab. Lett.*, **34**, 323–335.

Raimondo, M. (1996). Modèles en rupture, situations non ergodique et utilisation de méthode d'Ondelette, Ph.D. Thesis, University of Paris VII.

Seifert, B. and Gasser, T. (1996). Finite-sample analysis of local polynomials: analysis and solutions, *J. Amer. Statist. Assoc.*, **91**, 267–275.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*, Wiley, New York.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets, *Biometrika*, **82**, 385–397.

Wu, J. S. and Chu, C. K. (1993). Kernel type estimators of jump points and values of a regression function, *Ann. Statist.*, **21**, 1545–1566.