The Annals of Statistics
1993, Vol. 21, No. 3, 1545–1566

KERNEL-TYPE ESTIMATORS OF JUMP POINTS AND VALUES OF A REGRESSION FUNCTION[1]

BY J. S. WU AND C. K. CHU

Tamkang University and Tsing Hua University

In the fixed-design nonparametric regression model, kernel-type estimators of the locations of jump points and the corresponding sizes of jump values of the regression function are proposed. These kernel-type estimators are analyzed with almost sure results and limiting distributions. Using the limiting distributions, we are able to test the number of jump points and give asymptotic confidence intervals for the sizes of jump values of the regression function. Simulation studies demonstrate that the asymptotic results hold for reasonable sample sizes.

1. Introduction. In applications of regression methods, we are often interested in the locations of jump points and the corresponding sizes of jump values of the regression function. For example, when studying the impact of advertising, the time at which this action takes effect could effectively be modeled by the location of a jump point and the magnitude of the effect of this action is measured by the size of the jump. If we ignore the existence of the jump point, then we may make a serious error in drawing inferences about the process under study. For this, see Figure 1 where it is difficult to distinguish visually from the simulated data (solid squares) alone that the underlying regression function (dotted curves) has a jump point at $x = \frac{1}{2}$. Note that, in the neighborhood of this jump point, the difference between the regression function and the moving weighted average of the data (dashed curve) is large. Here the weights assigned to the observations are proportional to the heights of the closely spaced dotted curve at the bottom.

In practice, a suitable parametric method may not be available to estimate the locations of jump points and the corresponding sizes of jump values of the regression function. Whenever there is no appropriate parametric method available, we may start from nonparametric regression. Nonparametric regression is a smoothing method for recovering the regression function and its characteristics from noisy data. The simplest and most widely used regression smoothers are based on kernel methods. For the asymptotic properties of kernel estimators, see the monographs by Eubank (1988), Müller (1988) and Härdle (1990, 1991) for the case that the regression function has no jump point. For the effects of jump points on the asymptotic behaviors of kernel

Received July 1990; revised September 1992.
[1]This research is part of the Ph.D. dissertation of the first author, under the supervision of the second at the Tsing Hua University. The research for the second author was supported by the National Science Council under contract NSC 80-0208-M007-32.
AMS 1991 *subject classifications*. Primary 62G05; secondary 62G20.
Key words and phrases. Kernel estimator, nonparametric regression, jump point, size of jump value, strong consistency, asymptotic normality.

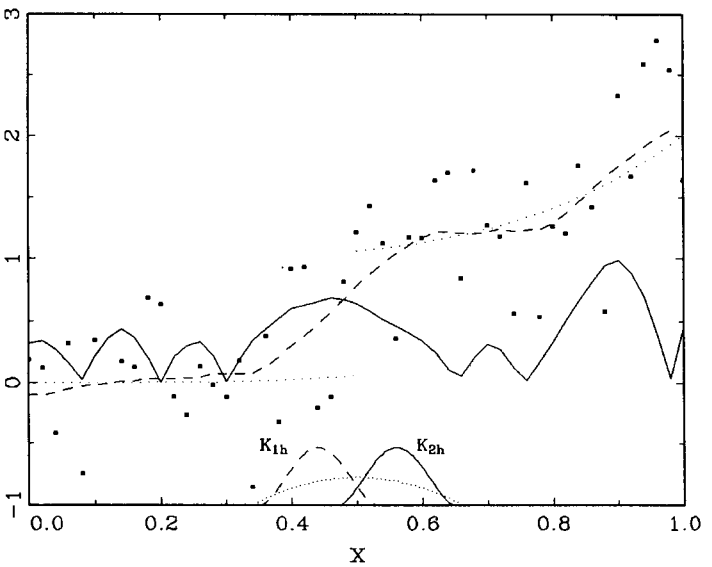

FIG. 1. *Plot of the discontinuous regression function (dotted curves), 50 independent observations (solid squares), $|J(x)|$ (solid curve) and the moving weighted average of the observations (dashed curve), where the weights assigned to the observations are proportional to the heights of the closely spaced dotted curve at the bottom.*

estimators and those on an optimally chosen bandwidth, see, for example, Wu and Chu (1993). For these, see also van Eeden (1985), van Es (1992) and Cline and Hart (1991) in the related field of kernel density estimation.

For the case that the number of jump points is known, the locations of jump points and the corresponding sizes of jump values of the regression function and its derivatives can be estimated by the smoothing algorithms proposed by Shiau (1985) and Speckman (1988) along with obtaining a regression function estimate. McDonald and Owen (1986), Chiu (1987) and Hall and Titterington (1992) introduce smoothing algorithms that can produce discontinuous output. Based on local averages of the response variables, Yin (1988) gives strongly consistent estimators for the number of jump points, the locations of jump points and the corresponding sizes of jump values of the regression function. Also, Müller (1992) gives weakly consistent estimators for the locations of jump points and the corresponding sizes of jump values and the rates of global L^p convergence of kernel estimators adjusted to estimates of the locations of jump points. In the field of edge detection in image analysis, Lee (1990) proposes smoothing algorithms to estimate the locations of jump points and the corresponding sizes of jump values of the regression function and its derivatives. See also the references given therein for more applications of jump detection.

In this paper we will use kernel estimators to construct estimators of the locations of jump points and the corresponding sizes of jump values of the

regression function. The motivation and the precise formulation of the proposed kernel-type estimators are described in Section 2. The almost sure behaviors and limiting distributions of these kernel-type estimators are given in Section 3. Using the limiting distributions, we are able to test the number of jump points and give asymptotic confidence intervals for the sizes of jump values of the regression function. The choices of kernel functions and values of bandwidths for constructing these kernel-type estimators are discussed. Section 4 contains simulation studies which give additional insight into what the theoretical results mean. Finally, sketches of the proofs are given in Section 5.

2. Regression settings and kernel-type estimators. In this paper the equally spaced fixed-design nonparametric regression model is considered. The regression model is given by

$$Y_i = m(x_i) + \varepsilon_i, \tag{2.1}$$

for $i = 1, 2, \ldots, n$. Here m is the unknown regression function defined on the interval $[0, 1]$ (without loss of generality), x_i are equally spaced fixed-design points, that is, $x_i = i/n$, ε_i are independent and identically distributed (iid) regression errors with mean 0 and variance σ^2, $0 < \sigma^2 < \infty$, and Y_i are noisy observations of m at x_i.

The regression function m in (2.1) is defined by

$$m(x) = r(x) + \psi(x). \tag{2.2}$$

Here r is a continuous function defined on the interval $[0, 1]$ and ψ is a step function defined by $\psi(x) = \sum_{j=1}^{p} d_j I_{[t_j, 1]}(x)$ for $x \in [0, 1]$. Note that p is a nonnegative integer representing the number of jump points of m, t_j are locations of jump points and d_j are nonzero real numbers representing the sizes of jump values of m at t_j. If p is 0, then m is a continuous function. For simplicity of presentation, let $d_{p+1} = 0$, $|d_j| > |d_{j+1}|$ and $t_j \in [\delta, 1 - \delta]$ for $j = 1, 2, \ldots, p$, and let the distance between any two of these t_j be greater than δ. Here δ is an arbitrarily small positive constant. The purpose of this paper is to use the observations Y_i to discover the value of p, the locations of jump points t_j and the sizes of jump values d_j of the regression function m in (2.2).

To construct the kernel-type estimators $\hat{t}_j$ and $\hat{d}_j$ of t_j and d_j, respectively, for $j = 1, 2, \ldots, \rho$, we consider the kernel estimator proposed by Gasser and Müller (1979). Here ρ is a given positive integer since the value of p is unknown. Given the kernel function K as a probability density function and the bandwidth h, then the Gasser–Müller estimator is defined by

$$\hat{m}(x) = \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K_h(x - z)\, dz, \tag{2.3}$$

for $x \in (0, 1)$, where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $s_0 = 0$, $s_i = (x_i + x_{i+1})/2$ for $i = 1, 2, \ldots, n - 1$ and $s_n = 1$.

The rest of this section is devoted to deriving $\hat{t}_j$ and $\hat{d}_j$ for $j = 1, 2, \ldots, \rho$. The ideas behind $\hat{t}_j$ and $\hat{d}_j$ will be shown in Figures 1 through 3 which

represent simulated regression settings defined in Section 4. In these figures the simulated data were generated by (2.1) with $n = 50$ and $\sigma^2 = (\frac{1}{2})^2$. To discover t_j, we consider the magnitude of the difference between two kernel estimators. Let

$$J(x) = \hat{m}_1(x) - \hat{m}_2(x), \tag{2.4}$$

for $x \in (0, 1)$, where $\hat{m}_1(x)$ and $\hat{m}_2(x)$ are Gasser–Müller estimators with different kernel functions K_1 and K_2, respectively, and the same bandwidth h. In the following, the value of $|J(x)|$ will be analyzed in each of the two cases that $m(x)$ has no jump point and multiple jump points.

If $m(x)$ has no jump point, under the usual regularity conditions, then $\hat{m}_1(x)$ and $\hat{m}_2(x)$ are uniformly strongly consistent estimators of $m(x)$ for $x \in [\delta, 1 - \delta]$. For this, see, for example, Theorem 3 of Cheng and Lin (1981). In this case, the magnitude of $|J(x)|$ is of small order for $x \in [\delta, 1 - \delta]$. Figure 2 shows simulation results for a continuous regression function $m(x)$ (dotted curve), the average of $|J(x)|$ (solid curve) over 1000 data sets and vertically rescaled $K_{1h}(x - \frac{1}{2})$ and $K_{2h}(x - \frac{1}{2})$ (dashed curve and solid curve at the bottom, respectively). Note that the small and approximately constant magnitude of $|J(x)|$ in Figure 2 indicates the continuity of the underlying regression function.

On the other hand, if $m(x)$ has jump points, then, by (2.1) and (2.2), $J(x)$ consists of two components. One is the difference between the two kernel estimators applied to the sum of the continuous function $r(x)$ and the regression errors ε_i. The other is the same difference for the step function $\psi(x)$. As above, the magnitude of the former component is of small order. We now consider the magnitude of the latter component. Since $\psi(x)$ is a step

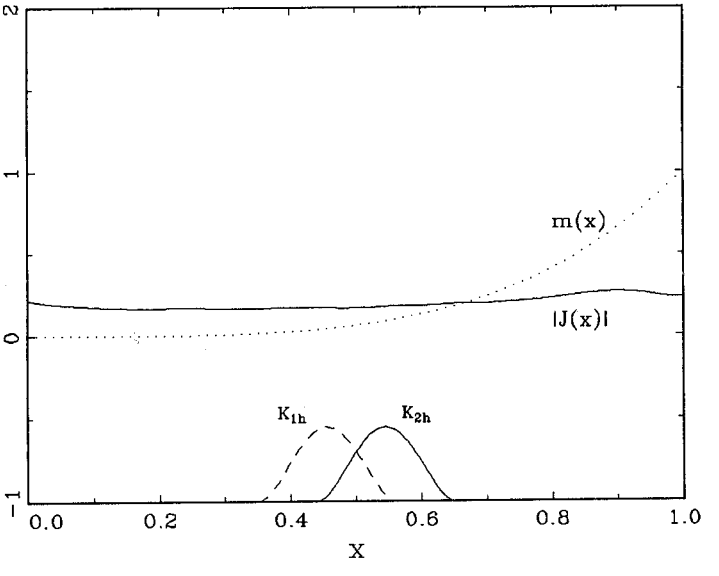

FIG. 2. *The magnitude of* $|J(x)|$ *in the case of a continuous regression function* $m(x)$.

function, to show locations of t_j by the magnitude of the latter component, we add the following conditions. Let K_1 and K_2 be compactly supported in the interval $[-1, 1]$ and satisfy $K_1(x) = K_2(-x)$ for all x and $\int_0^1 K_1 \neq \int_0^1 K_2$. For these K_1 and K_2, through a straightforward calculation, the magnitude of the latter component is symmetric about t_j and convex downward on some neighborhood of t_j for each $j = 1, 2, \ldots, p$. Also, outside of the union of these neighborhoods of t_j, the magnitude is of small order. Note that this neighborhood of t_j is the union of the intervals where $K_{1h}(x - t_j)$ and $K_{2h}(x - t_j)$ are supported for each $j = 1, 2, \ldots, p$. Based on these characteristics of the magnitudes of the two components, the local maximizers of $|J(x)|$ are good estimators of jump points, in some sense. Since K_1 and K_2 are supported in $[-1, 1]$, the widths of the above neighborhoods of t_j are less than or equal to $2h$. Combining this result with the fact that $|d_j| > |d_{j+1}|$ for $j = 1, 2, \ldots, p$, we propose to take $\hat{t}_j$ as maximizers of $|J(x)|$ over the sets A_j, where

$$A_j = [\delta, 1 - \delta] - \bigcup_{k=1}^{j-1} [\hat{t}_k - 2h, \hat{t}_k + 2h],$$

for $j = 1, 2, \ldots, \rho$.

Figure 1 shows $|J(x)|$ (solid curve) derived from the simulated data (solid squares) and vertically scaled $K_{1h}(x - t_1)$ and $K_{2h}(x - t_2)$ (dashed curve and solid curve at the bottom, respectively). Note that the maximizer of $|J(x)|$ over the interior $[0.2, 0.8]$ of the interval $[0, 1]$ shows the location of t_1 accurately. Figure 3 also shows simulation results for a discontinuous regression function $m(x)$ (dotted curves) with $t_1 = \frac{1}{2}$ and $d_1 = 1$, averages of $|J(x)|$ (solid curve),

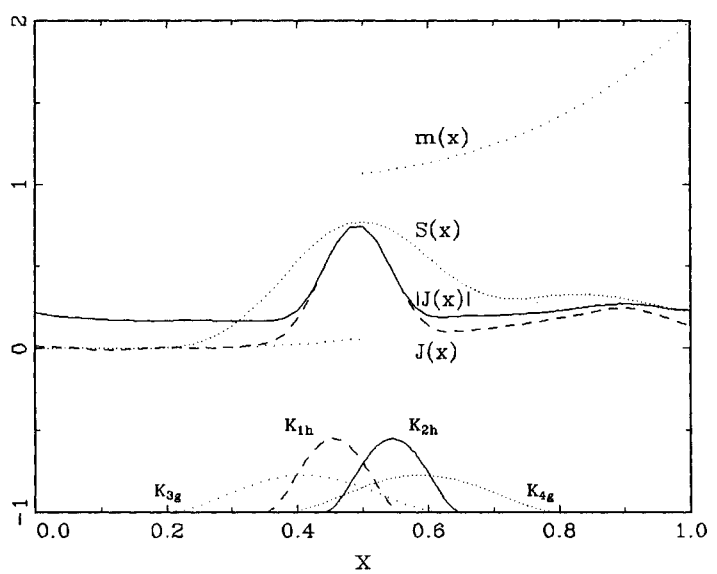

FIG. 3. *The magnitudes of* $|J(x)|$, $J(x)$ *and* $S(x)$ *in the case of a discontinuous regression function* $m(x)$.

$J(x)$ (dashed curve) and $S(x)$ (closely spaced dotted curve), over 1000 data sets and vertically rescaled $K_{1h}(x - t_1)$, $K_{2h}(x - t_1)$, $K_{3g}(x - t_1)$ and $K_{4g}(x - t_1)$ (dashed curve, solid curve, dotted curve and closely spaced dotted curve at the bottom, respectively). Here $S(x)$, K_{3g} and K_{4g} will be defined later. In Figure 3, $|J(x)|$ is approximately symmetric about t_1 and convex downward on the neighborhood of t_1. Also, outside of this neighborhood of t_1, $|J(x)|$ is of small and approximately constant magnitude. This neighborhood of t_1 is the union of the intervals where $K_{1h}(x - t_1)$ and $K_{2h}(x - t_1)$ are supported. The maximizer of $|J(x)|$ shows the location of t_1 accurately.

To estimate d_j, based on the above t_j, an immediate idea is to take the rescaled $J(\hat{t}_j)$ as $\hat{d}_j$ for each $j = 1, 2, \ldots, \rho$. Here the scale factor for $J(\hat{t}_j)$ is the ratio of d_j to $J(t_j)$ for each $j = 1, 2, \ldots, \rho$. This idea is indicated in Figure 3. Figure 3 shows that $J(t_1) \neq d_1$. By this, $J(\hat{t}_1)$, an estimator of $J(t_1)$, should be rescaled by the ratio of d_1 to $J(t_1)$ to properly estimate d_1. However, in this paper, we do not adopt this idea. To estimate d_j, we propose to take the rescaled $S(\hat{t}_j)$ as $\hat{d}_j$ for each $j = 1, 2, \ldots, \rho$, where

$$S(x) = \hat{m}_3(x) - \hat{m}_4(x), \tag{2.5}$$

for $x \in (0, 1)$, and where $\hat{m}_3(x)$ and $\hat{m}_4(x)$ are Gasser–Müller estimators with kernel functions K_3 and K_4, respectively, and the same bandwidth g. Here K_3 and K_4 satisfy the above conditions given on K_1 and K_2, respectively, and the value of g is of larger order than that of h. In the following, we shall use the above example to illustrate the effects of magnitudes of h and g on these two estimators of d_j.

In Figure 3, $S(x)$ was derived in the case of $g = 2h$, $K_3 = K_1$ and $K_4 = K_2$. Note that the maximum value of $S(x)$ occurred at t_1, and is approximately equal to that of $J(x)$, which also occurred at t_1. Hence, the scale factors for $S(\hat{t}_1)$ and $J(\hat{t}_1)$ determined by the ratios of d_1 to $S(t_1)$ and $J(t_1)$, respectively, are approximately equal in magnitude. Note also that, for each x in the neighborhood of t_1, the magnitude of the difference between $S(x)$ and $S(t_1)$ is smaller than that between $J(x)$ and $J(t_1)$. This neighborhood of t_1 is the union of the intervals where $K_{3g}(x - t_1)$ and $K_{4g}(x - t_1)$ are supported. By these two results, to estimate d_1, if $\hat{t}_1$ falls into this neighborhood of t_1, then the rescaled $S(\hat{t}_1)$ is of better performance than the rescaled $J(\hat{t}_1)$, in the sense that the former is closer to d_1 than the latter.

Finally, the asymptotic behaviors of the proposed kernel-type estimators $\hat{t}_j$ and $\hat{d}_j$ of t_j and d_j, respectively, will be studied in Section 3.

3. Results. In this section we will study the asymptotic behaviors of $\hat{t}_j$ and $\hat{d}_j$ for $j = 1, 2, \ldots, \rho$. For these, using the regression model (2.1), $|J(x)|$ in (2.4) and $S(x)$ in (2.5), we impose the following assumptions:

(A.1) The function $r(x)$ in (2.2) is Lipschitz continuous.

(A.2) The regression errors ε_i are iid random variables with mean 0, with variance σ^2, $0 < \sigma^2 < \infty$, and with the lth absolute moment finite for some $l > 2$.

(A.3) The kernel function K_2 supported on the interval $[-\lambda, 1]$, $\lambda \in [0, 1]$, is a probability density function with $\int_{-\lambda}^{1} zK_2 \neq 0$. The first derivative $K_2^{(1)}$ of K_2 is square integrable and Lipschitz continuous. Also, $K_2^{(1)}(0) \neq 0$. Here and throughout this paper, the notation $f^{(j)}$ denotes the jth derivative of the function f. The kernel function K_1 is defined by $K_1(z) = K_2(-z)$ for all z. The kernel functions K_1 and K_2 satisfy $\int_0^1 K_2 - \int_0^\lambda K_1 \neq 0$. Furthermore, there is a constant $\eta > 0$ such that $|\int_0^{C_n}(K_1 - K_2)| > \eta \cdot c_n^2$ for any sequence c_n of positive real numbers converging to 0 as $n \to \infty$.

(A.4) The kernel function K_4 supported on the interval $[-\omega, 1]$, $\omega \in [0, 1]$, is a square-integrable and Lipschitz-continuous probability density function with $\int_{-\omega}^{1} zK_4 \neq 0$. The kernel function K_3 is defined by $K_3(z) = K_4(-z)$ for all z. The kernel functions K_3 and K_4 satisfy $\int_0^1 K_4 - \int_0^\omega K_3 \neq 0$.

(A.5) The total number of observations in this regression setting is n with $n \to \infty$. The bandwidths $h = h_n$ and $g = g_n$ satisfy $h \to 0$ with $nh \to \infty$, and $g \to 0$ with $ng \to \infty$ as $n \to \infty$.

We first give the formulation of $\hat{d}_j$. Based on the above assumptions and the regression model (2.1), through a straightforward calculation,

$$E[S(x)] = \tilde{S}(x) + \sum_{j=1}^{p} d_j S_j(x), \tag{3.1}$$

for $x \in (0, 1)$, where

$$\begin{aligned}
\tilde{S}(x) &= \sum_{i=1}^{n} r(x_i) \int_{s_{i-1}}^{s_i} \left(K_{3g}(x - z) - K_{4g}(x - z)\right) dz \\
&= O(g), \\
S_j(x) &= \sum_{i=1}^{n} I_{[t_j, 1]}(x_i) \int_{s_{i-1}}^{s_i} \left(K_{3g}(x - z) - K_{4g}(x - z)\right) dz \\
&= \int_{(t_j - x)/g}^{1} (K_4 - K_3) + O\left((ng)^{-1}\right).
\end{aligned}$$

Based on (3.1),

$$E\left[S(t_j)\right] = d_j \cdot \left[\int_0^1 K_4 - \int_0^\omega K_3\right] + O\left(g + (ng)^{-1}\right),$$

for $j = 1, 2, \ldots, p$. According to this result, d_j, the rescaled $S(\hat{t}_j)$, are defined by

$$\hat{d}_j = c_\omega \cdot S\left(\hat{t}_j\right),$$

for $j = 1, 2, \ldots, \rho$. Here the scale factor c_ω is defined by

$$c_\omega = \left[\int_0^1 K_4 - \int_0^\omega K_3\right]^{-1}.$$

Let γ and θ be positive constants, where $\theta \in (0, \frac{1}{2})$, β_n be a sequence of positive real numbers diverging to ∞ as $n \to \infty$, $\hat{d}^*$ denote the supremum of $|S(x)|$ over the interval $[a, b]$, where $0 < a < b < 1$, and $\Lambda_j = ((n/h)^{1/2}(\hat{t}_j - t_j), (ng)^{1/2}(\hat{d}_j - d_j))^T$ for $j = 1, 2, \ldots, \rho$. Here the notation T stands for the transpose of a matrix. Theorems 1 and 2 give the asymptotic behaviors of $\hat{t}_j$ and $\hat{d}_j$ in the cases of $p \geq \rho \geq 1$ and $\rho > p \geq 0$, respectively. Theorem 3 gives the limiting distribution of $\hat{d}^*$ for $p = 0$. The proofs of these theorems are given in Section 5. In these theorems the conditions given on the values of h, g, γ and β_n include:

$$(n^{(l-1)/l}h^{1+2\theta})^{-1} = O(1), \tag{B.1}$$

$$(nh^{1+4\theta})^{-1} = O(1), \tag{B.2}$$

$$n^{\gamma}h^{1+\theta}(g\beta_n)^{-1} = o(1), \tag{B.3}$$

$$n^{\gamma}g(\beta_n \log n)^{-1} = o(1), \tag{B.4}$$

$$n^{2\gamma-1}(g\beta_n)^{-1} = O(1), \tag{B.5}$$

$$n^{-2+\gamma+(1/l)}h^{-2}(\beta_n \log n)^{-1} = o(1), \tag{B.6}$$

$$n^{-1+\gamma+(1/l)}h^{-1}\beta_n^{-1/2} \leq 2K_1^M, \tag{B.7}$$

for n sufficiently large, where K_1^M denotes the maximum value of K_1,

$$n^{(1/2)+(1/(l-1))}h^{(1/2)+(l+1)\theta}(\log n)^{1+l} = o(1), \tag{B.8}$$

$$nh^3 = o(1), \tag{B.9}$$

$$n^{(1/2)+(1/(l-1))}h^{(l+1)(\theta+1)}g^{-l-(1/2)}(\log n)^{1+l} = o(1), \tag{B.10}$$

$$n^{1/2}h^{1+\theta}g^{-1/2}\log n = o(1), \tag{B.11}$$

$$ng^3 = o(1), \tag{B.12}$$

$$hg^{-1} = o(1). \tag{B.13}$$

THEOREM 1. *In the case of $p \geq \rho \geq 1$, under the above assumptions, if* (B.1) *and* (B.2) *hold, then*

$$P\left(|\hat{t}_j - t_j| > h^{(1+\theta)}\log n \ i.o.\right) = 0, \tag{3.2}$$

for $j = 1, 2, \ldots, \rho$. If (B.1) *through* (B.7) *hold, then*

$$n^{\gamma}(\beta_n \log n)^{-1}|\hat{d}_j - d_j| \to 0 \quad a.s., \tag{3.3}$$

for $j = 1, 2, \ldots, \rho$. If (B.1), (B.2) *and* (B.8) *through* (B.13) *hold, then*

$$\Lambda_j \Rightarrow N\left((0,0)^T, \sigma^2\begin{pmatrix} d_j^{-2}U & 0 \\ 0 & V \end{pmatrix}\right), \tag{3.4}$$

for $j = 1, 2, \ldots, \rho$, *and these* Λ_j *are asymptotically independent, where*

$$U = \left[\int \left(K_1^{(1)} - K_2^{(1)}\right)^2\right]\left[2K_2^{(1)}(0)\right]^{-2},$$

$$V = c_\omega^2 \int (K_3 - K_4)^2.$$

THEOREM 2. *In the case of* $\rho > p \geq 0$, *under the above assumptions, if* (B.1) *through* (B.7) *hold, then*

$$n^\gamma (\beta_n \log n)^{-1} \hat{d}_j \to 0 \quad a.s., \tag{3.5}$$

for $j = p + 1, p + 2, \ldots, \rho$.

THEOREM 3. *Based on the above assumptions, if* $l > 3$, $m(x)$ *is Lipschitz continuous on the above interval* $[a, b]$, *and* $g = n^{-\kappa}$, $\kappa \in (\frac{1}{3}, 1 - 2/l)$, *then*

$$P\left(\hat{d}^* < a_n + b_n x\right) \to \exp(-2\exp(-x)), \tag{3.6}$$

where

$$a_n = \left[(ng)^{-1}\sigma^2 c_\omega^2 \int (K_3 - K_4)^2\right]^{1/2}\left[g_{ab} + g_{ab}^{-1}\left(\log\left(3(4\pi)^{-1/2}\right)\right)\right],$$

$$b_n = \left[(ng)^{-1}\sigma^2 c_\omega^2 \int (K_3 - K_4)^2\right]^{1/2} g_{ab}^{-1},$$

and where

$$g_{ab} = \left[2\log((b - a)/g)\right]^{1/2}.$$

We now close this section with some remarks.

REMARK 1 (Strong convergence rates of the proposed kernel-type estimators of the sizes of jump values). Based on (B.4) and (B.5), the maximum value of γ in (3.3) and (3.5) is $\frac{1}{3}$ which is arrived at when the value of g is of order $n^{-1/3}$ in each case. Combining this result with the conditions that $\theta \in (0, \frac{1}{5})$, $l \geq (1 + 4\theta)/(2\theta)$ and the value of h is of order $n^{-\zeta}$, where $(\frac{2}{3})$ $(1 + 2\theta)(1 + \theta)^{-1} \leq \zeta \leq (1 + 2\theta)(1 + 4\theta)^{-1}$, then the rate of strong consistency of $\hat{d}_j$ in (3.3) and (3.5) is $n^{-1/3}(\beta_n \log n)$ in each case. This rate of strong consistency is the same as that of uniformly strong consistency of kernel estimators of a Lipschitz-continuous regression function as given in Theorem 3 of Cheng and Lin (1981).

REMARK 2 (Choices of kernel functions and values of bandwidths for constructing the proposed kernel-type estimators of the locations of jump points and the corresponding sizes of jump values). For the case of $p \geq \rho \geq 1$, to estimate t_j, a possible approach for the practical choice of the value of h and the kernel functions K_1 and K_2 is based on an analogue of the mean square error. Using (5.7) and the above assumptions, through a straightfor-

ward calculation, the asymptotic mean square error (AMSE) of $\hat{t}_j$ can be expressed as

$$\text{AMSE}\left[\hat{t}_j\right] = O(h^4) + n^{-1}hd_j^{-2}\sigma^2 U, \tag{3.7}$$

for $j = 1, 2, \ldots, \rho$. For the components of AMSE, the first and the second terms on the right-hand side of (3.7) represent the asymptotic bias-square and the asymptotic variance, respectively. By (3.7), the optimal values h_j^* of h for the estimation of t_j are of the minimum order of the values of h satisfying (B.1), (B.2), (B.8) and (B.9). However, this minimum order is not available since it depends on the unknown value of l. Therefore, the choice of these optimal values h_j^* needs further study. In practice, if a too small value of h is inserted into $|J(x)|$, then t_j might pick locations of outliers (observations with large regression errors) as estimates of t_j. This result is caused by the fact that if the averages in $|J(x)|$ are made with too few observations, then the effects of outliers appear on the magnitude of $|J(x)|$. By (B.9), the value of h is of smaller order than $n^{-1/3}$. Combining this result with (3.7), then the magnitude of the second term on the right-hand side of (3.7) is of larger order than that of the first term. According to this, if there are kernel functions K_1^* and K_2^* which minimize the value of U over K_1 and K_2 in (A.3), then K_1^* and K_2^* are the optimal kernel functions of K_1 and K_2, respectively, in the sense of AMSE. Unfortunately, by Jensen's inequality, the minimum value of U over K_1 and K_2 in (A.3) is not attainable. Hence, we suggest choosing K_1 and K_2 by minimizing the value of U over the class of fourth-degree polynomials satisfying the conditions given in (A.3). Through a straightforward calculation, the minimum value of U over this class can be arrived at by choosing K_1 and K_2 as

$$K_2(x) = (0.4857 + 3.8560x + 2.8262x^2 - 19.1631x^3 + 11.9952x^4) \times I_{[-0.2012,\,1]}(x),$$

and $K_1(x) = K_2(-x)$ for all x. To estimate d_j, apply the same idea to g, K_3 and K_4. Using (5.8) and the above assumptions, through a straightforward calculation, the AMSE of $\hat{d}_j$ can be expressed as

$$\text{AMSE}\left[\hat{d}_j\right] = c_\omega^2 \tilde{S}(t_j)^2 + n^{-1}g^{-1}c_\omega^2\sigma^2\int(K_3 - K_4)^2, \tag{3.8}$$

for $j = 1, 2, \ldots, \rho$. Theoretically, if there exist a value g_j^* and the kernel functions K_{3j}^* and K_{4j}^* which minimize (3.8) over g in (B.10) and (B.11) and K_3 and K_4 in (A.4), then g_j^* is the optimal value of g, and K_{3j}^* and K_{4j}^* are the optimal kernel functions of K_3 and K_4, respectively, in the sense of AMSE. By (3.8) and the order of magnitude of $\tilde{S}(t_j)$ in (3.1), the values of g_j^* are of the same order $n^{-1/3}$ for $j = 1, 2, \ldots, \rho$. Note that g_j^*, K_{3j}^* and K_{4j}^* depend on the unknown factors l, t_j, r and σ^2. In practice, we may plug estimates of these unknown factors into (3.8) to get estimates of g_j^*, K_{3j}^* and K_{4j}^*. However, the performance of these estimates of g_j^*, K_{3j}^* and K_{4j}^* needs further study.

REMARK 3 (Minimum asymptotic variances of the proposed kernel-type estimators of the sizes of jump values). Using the Jensen inequality and the Cauchy–Schwarz inequality, through a straightforward calculation, the value of V in (3.4) has a lower bound, $V \geq 2$, over K_3 and K_4 in (A.4). This lower bound on the value of V can be arrived at by choosing $K_3(z) = I_{[-1,0]}(z)$ and $K_4(z) = I_{[0,1]}(z)$. Note that these rectangular kernels K_3 and K_4 exhibit jump points at endpoints of their support. In general, kernel functions with jump points will lead to bad finite sample behaviors of kernel estimators. For this, see, for example, subsection 5.3 of Müller (1988) and subsection 2.1 of Härdle (1991).

REMARK 4 (The hypothesis test of the number of jump points). Using (3.6), we can test whether the regression function $m(x)$ has jump points on the interval $[\delta, 1-\delta]$. Replacing a, b and σ^2 in (3.6) with δ, $1-\delta$ and a consistent estimator $\hat{\sigma}^2$, $\hat{\sigma}^2 = \sigma^2 + o_p((ng)^{-1/2})$, respectively, and applying Slutsky's theorem, through a straightforward calculation, (3.6) becomes

$$P\left(\hat{d}^* < \hat{a}_n + \hat{b}_n x\right) \to \exp(-2\exp(-x)). \tag{3.9}$$

Here $\hat{a}_n$ and $\hat{b}_n$ are coefficients a_n and b_n with parameters a, b and σ replaced by δ, $1-\delta$ and $\hat{\sigma}$, respectively, in each case. From (3.9), the test of the null hypothesis H_0: $p = 0$ against the alternative hypothesis H_1: $p > 0$ is available. For $l \geq 4$, according to the facts that Y_i are independent, $m(x)$ has the finite number of jump points and $r(x)$ is Lipschitz continuous, then $\hat{\sigma}^2$ can be constructed as a trimmed mean, that is,

$$\hat{\sigma}^2 = (2(n-1-2q))^{-1} \sum_{i=2+q}^{n-q} \xi_i = \sigma^2 + O_p(n^{-1/2}), \tag{3.10}$$

where the last part follows through a straightforward calculation. Here ξ_i, for $i = 2, 3, \ldots, n$, denote the rearranged $(Y_i - Y_{i-1})^2$, and ξ_i are in ascending order. The number q is determined by the experimenter. In this case, the large jumps caused by the noise and the regression function can be left out in constructing $\hat{\sigma}^2$. For the untrimmed version of $\hat{\sigma}^2$, see, for example, Rice (1984). Based on (3.2) and (3.6), we can test the number of jump points of $m(x)$ on the interval $[\delta, 1-\delta]$ by the following approach. Let $\hat{d}_j^*$ denote the supremum of $|S(x)|$ over the sets A_j, where

$$A_j = [\delta, 1-\delta] - \bigcup_{k=1}^{j-1} \left[\hat{t}_k - 2h, \hat{t}_k + 2h\right] = \bigcup_{k=1}^{j} A_{j,k},$$

and where $A_{j,k}$ are disjoint subintervals of A_j for $k = 1, 2, \ldots, j$ and $j = 1, 2, \ldots$. If $j = p$, given $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_p$, through a straightforward calculation, then

$$P\left(\hat{d}_{p+1}^* < x\right) \approx \prod_{k=1}^{p+1} \exp\left(-2\exp\left(-(x - a_{n,p,k})/b_{n,p,k}\right)\right). \tag{3.11}$$

Here $a_{n,p,k}$ and $b_{n,p,k}$ are coefficients of a_n and b_n in (3.6) with a and b, the

endpoints of the interval $[a, b]$, replaced by the endpoints of the subinterval $A_{p+1,k}$, respectively, for each $k = 1, 2, \ldots, p+1$. The notation $L_n \approx R_n$ denotes $L_n/R_n \to 1$ as $n \to \infty$. Replacing σ^2 in $a_{n,p,k}$ and $b_{n,p,k}$ with $\hat{\sigma}^2$ in (3.10), in each case, then the test of the null hypothesis H_0: $p = j$ against the alternative hypothesis H_1: $p > j$ can be performed.

REMARK 5 (Bandwidth selection for the hypothesis test of the number of jump points). In practice, using (3.9) and (3.11) to test the null hypothesis H_0: $p = j$ against the alternative hypothesis H_1: $p > j$ for some $j \geq 0$, the choice of the value of g and the kernel functions K_3 and K_4 needs further study. By (5.10), if a too small value of g is used in $|S(x)|$, then the effects of outliers appear in the magnitude of $|S(x)|$. In this case, a large jump caused by the noise might be taken as the value of the test statistic $\hat{d}^*_{j+1}$. On the other hand, if a too large value of g is used in $|S(x)|$, then the value of $\hat{d}^*_{j+1}$ is affected severely by the quantity $\tilde{S}(x)$ in (3.1).

REMARK 6 (Asymptotic confidence intervals for the sizes of jump values). For $p \geq \rho \geq 1$ and $l \geq 4$, using the asymptotic normalities of $\hat{d}_j$ in (3.4), replacing σ^2 with $\hat{\sigma}^2$ in (3.10) and applying Slutsky's theorem, through a straightforward calculation, then

$$P\left((ng)^{1/2}(\hat{\sigma}^2 V)^{-1/2}\left(\hat{d}_j - d_j\right) < x\right) \to \Phi(x), \tag{3.12}$$

for $j = 1, 2, \ldots, \rho$. Here Φ is the distribution function of a standard normal random variable. Based on (3.12), asymptotic confidence intervals of d_j are available for $j = 1, 2, \ldots, \rho$.

REMARK 7 (Applications of the proposed kernel-type estimators). The results of this paper can be applied directly to the heteroscedastic regression model. For example, we might be interested to check whether the variance function Γ of the independent regression errors ε_i, that is, $\Gamma(x_i) = \text{var}[\varepsilon_i]$ for all i, has jump points. In this case, the locations of jump points and the corresponding sizes of jump values of Γ can be estimated by the following approach. Set $\hat{\Gamma}_i = (\frac{1}{2})(Y_{2i} - Y_{2i-1})^2$ for $i = 1, 2, \ldots, [n/2]$. Applying $|J(x)|$ in (2.4) and $S(x)$ in (2.5) to $\hat{\Gamma}_i$, results similar to Theorems 1, 2 and 3 follow, through a straightforward calculation. Note that this method will be inefficient since not all information available on the differences is utilized. In the case where the variance function Γ is continuous, see, for example, Carroll (1982) and Müller and Stadtmüller (1987) for the estimation of Γ.

REMARK 8 (Construction of the proposed kernel-type estimators in the unequally spaced fixed-design case). The proposed kernel-type estimators can be constructed in the case of unequally spaced fixed design. Suppose that the unequally spaced fixed-design points x_i satisfy the conditions $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq 1$ and $x_i = i/n + O(n^{-1})$ uniformly. An extreme case of this design is that some observations are allowed to be made at the same design

point. Given this design, if some design points are closer to their neighbors than others are, then the asymptotic variance of the Gasser–Müller estimator $\hat{m}(x)$ in (2.3) might be increased. For this, see, for example, Chu and Marron (1991). To deal with this drawback to $\hat{m}(x)$, we propose a new version of the Gasser–Müller estimator $\hat{m}^*(x)$ defined by

$$\hat{m}^*(x) = \sum_{i=1}^{n} Y_i \int_{(i-1)/n}^{i/n} K_h(x - z)\, dx.$$

Under the above assumptions, through a straightforward calculation, $\hat{m}^*(x)$ in this unequally spaced fixed-design case has the same AMSE as $\hat{m}(x)$ in the equally spaced fixed-design case. By this, replacing $\hat{m}(x)$ by $\hat{m}^*(x)$ in Section 2, the motivation for constructing $\hat{t}_j$ and $\hat{d}_j$ is still available in this unequally spaced fixed-design case. Finally, the performance of the resulting estimators of t_j and d_j needs further study.

4. Simulations. To investigate the practical implications of the asymptotic results of $\hat{t}_j$ and $\hat{d}_j$ presented in Section 3, an empirical study was carried out. The simulated regression settings and those given in Figures 1 through 3 are introduced in the following. The sample size was $n = 50$. The regression model (2.1) and the Gasser–Müller estimator (2.3) were considered. The continuous function $r(x)$ and the step function $\psi(x)$ in (2.2) were $r(x) = x^4$ and $\psi(x) = I_{[1/2,1]}(x)$ for $x \in [0, 1]$. Two regression functions $m(x)$ were considered. One $m(x)$ had no jump point, that is, $m(x) = r(x)$. The other $m(x)$ had one jump point, that is, $m(x) = r(x) + \psi(x)$, where the location of the jump point was $t_1 = \frac{1}{2}$ and the corresponding size of the jump value was $d_1 = 1$. The regression errors ε_i were pseudo independent normal random variables $N(0, \sigma^2)$, where $\sigma^2 = (\frac{1}{2})^2$. Based on (2.1), 1000 independent sets of the observations Y_i were generated for each $m(x)$. For the latter $m(x)$, given this large value of $\sigma^2 = (\frac{1}{2})^2$, the location of the jump point $t_1 = \frac{1}{2}$ was not always distinguishable visually from the data alone. For this, see, for example, Figure 1. The kernel functions K_1 and K_2 and K_3 and K_4 were those given in Remarks 2 and 3, respectively. Finally, the values of δ in (2.2) and q in (3.10) were $\delta = \frac{1}{5}$ and $q = 2$. The choice of the values of δ and q was made arbitrarily.

For each data set, to test whether the underlying regression function has jump points, 21 equally spaced values of g on the interval $[0.02, 0.48]$ were chosen. For each data set and each value of g, the values of $|S(x)|$ were calculated on the grid $u_i = i/(3n)$ for $i = 0, 1, \ldots, 3n$. The maximum value $\hat{d}^*$ of $|S(x)|$ on the interval $[\delta, 1 - \delta]$ was calculated. After evaluation on the grid, a one-step interpolation improvement was done, with the result taken as $\hat{d}^*$. When $\hat{d}^*$ was obtained, the values of $\hat{a}_n$ and $\hat{b}_n$ in (3.9) and $\hat{\sigma}^2$ in (3.10) were calculated. Based on (3.9) and given $\alpha = 0.05$, the test of the null hypothesis H_0: $p = 0$ against the alternative hypothesis H_1: $p > 0$ was performed. For the two cases of $m(x)$, Figure 4 shows the number of times N out of the 1000 data sets that the null hypothesis H_0: $p = 0$ was rejected for each given value of g.

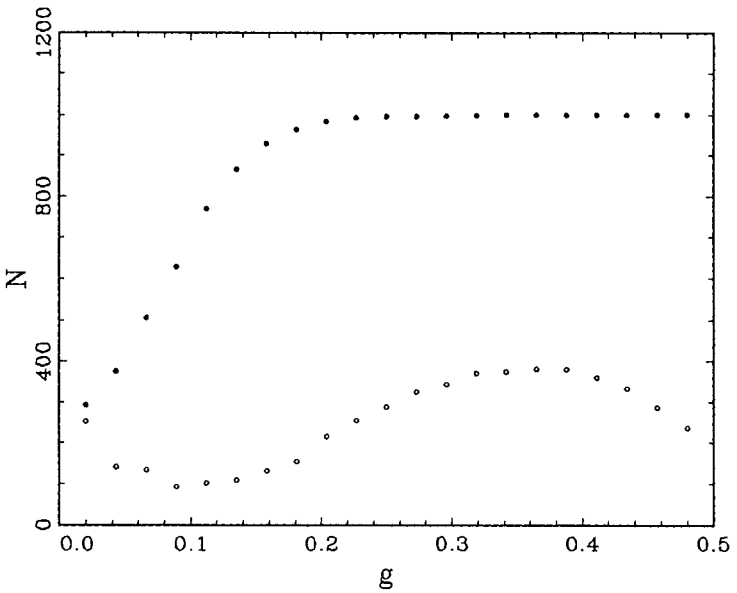

FIG. 4. *The number of times N out of the* 1000 *data sets that the null hypothesis* H_0: $p = 0$ *was rejected when the value of g was inserted into* $|S(x)|$ *in the cases of the continuous regression function (circles) and the discontinuous regression function (solid circles).*

In Figure 4, for the case that $m(x)$ had no jump point, the values of N (circles) showed that the empirical Type I error was at least 10%, although $\alpha = 0.05$. This poor performance of the hypothesis test of $p = 0$ might be caused by the small sample size $n = 50$ and the slow convergence of $\hat{d}^*$ toward the double exponential as given in (3.6). On the other hand, for the case that $m(x)$ had one jump point, the performance of the hypothesis test of $p = 0$ was poor as the value of g was small, since the corresponding value of N (solid circle) was small. As the value of g increased, the value of N increased and the performance of the hypothesis test was improved. In this example of the discontinuous regression function, based on the increasing values of N, there was no empirically best value of g for the test of $p = 0$.

We now describe the calculation of $\hat{t}_1$ and $\hat{d}_1$ for the case that $m(x)$ had one jump point. For each data set, to estimate t_1, the above 21 equally spaced values of g chosen for calculating $\hat{d}^*$ were taken as the values of h. For each data set and each value of h, the values of $|J(x)|$ were calculated on the above grid u_i. The maximizer $\hat{t}_1$ of $|J(x)|$ over the interval $[\delta, 1 - \delta]$ was calculated. After evaluation on the grid, a one-step interpolation improvement was done, with the result taken as $\hat{t}_1$. When $\hat{t}_1$ was obtained, the 21 values of g were inserted into $c_\omega \cdot S(\hat{t}_1)$ separately to derive 21 values of $\hat{d}_1$.

The absolute values of the sample biases, the sample standard deviations and the sample mean square errors (MSEs) of the ratio $\hat{t}_1/t_1$ and the sample

TABLE 1

The equally spaced values of h and g (first column), the absolute values of the sample biases (second column), the sample standard deviations (third column), the sample MSEs of $\hat{t}_1/t_1$ (fourth column), where $\hat{t}_1$ was derived by $|J(x)|$ with the value of h in the row, the sample MSEs of $\hat{d}_1/d_1$ with $g = h$ (fifth column), the MSEs of $\hat{d}_1/d_1$ with $g = g^*$ (sixth column) and the minimizers g^* (seventh column) of the 21 values of the sample MSEs of $\hat{d}_1/d_1$, where $\hat{d}_1$ were derived by $c_\omega \cdot S(\hat{t}_1)$ with the value of h in the row and the 21 values of g in the first column separately

Values of h and g	Absolute value of the bias of $\hat{t}_1/t_1$	Variance of $\hat{t}_1/t_1$	ME of $\hat{t}_1/t_1$	MSE of $\hat{d}_1/d_1$ with $g = h$	MSE of $\hat{d}_1/d_1$ with $g = g^*$	Value of g^*
0.020	0.0039	0.3114	0.0970	3.0200	0.1019	0.4645
0.043	0.0082	0.2724	0.0743	1.1440	0.1195	0.4041
0.066	0.0035	0.2292	0.0526	0.5061	0.1103	0.3707
0.089	0.0065	0.1844	0.0341	0.1932	0.0851	0.3052
0.112	0.0053	0.1578	0.0249	0.1153	0.0723	0.2856
0.135	0.0020	0.1344	0.0181	0.0777	0.0653	0.2640
0.158	0.0017	0.1215	0.0148	0.0653	0.0579	0.2458
0.181	0.0049	0.1166	0.0136	0.0602	0.0558	0.2428
0.204	0.0038	0.1085	0.0118	0.0562	0.0548	0.2471
0.227	0.0051	0.1110	0.0123	0.0552	0.0542	0.2507
0.250	0.0081	0.1115	0.0125	0.0535	0.0525	0.2599
0.273	0.0126	0.1147	0.0133	0.0537	0.0532	0.2681
0.296	0.0183	0.1175	0.0141	0.0563	0.0554	0.2668
0.319	0.0218	0.1152	0.0137	0.0599	0.0534	0.2141
0.342	0.0269	0.1144	0.0138	0.0647	0.0541	0.2378
0.365	0.0323	0.1134	0.0139	0.0699	0.0518	0.2370
0.388	0.0377	0.1093	0.0134	0.0797	0.0484	0.2381
0.411	0.0454	0.1068	0.0135	0.0889	0.0493	0.2452
0.434	0.0548	0.1057	0.0142	0.1018	0.0474	0.2615
0.457	0.0651	0.1025	0.0147	0.1130	0.0460	0.2629
0.480	0.0743	0.0975	0.0150	0.1230	0.0450	0.2676

MSEs of the ratio $\hat{d}_1/d_1$ were summarized. The sample bias of $\hat{\tau}$ was taken as the average of the 1000 values of $\hat{\tau} - 1$. Here $\hat{\tau}$ denotes the ratios $\hat{t}_1/t_1$ and $\hat{d}_1/d_1$, in each respective case. The sum of the sample variance and the sample bias-square was taken as the sample MSE. For each given value of h, there were 21 values of the sample MSEs of $\hat{d}_1/d_1$. Among these 21 values of the sample MSEs of $\hat{d}_1/d_1$, the minimizer g^* was calculated. For each h, when g^* was obtained, the MSE of $\hat{d}_1/d_1$ with $g = g^*$ was calculated. The numeric results are given in Table 1.

We now analyze the performance of $\hat{t}_1$. As the value of h (first column), $h \in [0.02, 0.19]$, increased, the performance of $\hat{t}_1$ was improved, since the magnitude of the bias (second column) and the standard deviation (third column) of $\hat{t}_1/t_1$ essentially decreased. This result was caused by the fact that the magnitudes of the effects of outliers on the value of $|J(x)|$ decreased as the

value of h increased. As the value of h increased further, both the magnitude of the bias and the standard deviation of $\hat{t}_1/t_1$ increased. However, the latter decreased as the value of h, $h \in [0.3, 0.48]$, increased. This result was caused by the fact that the large magnitude of bias had the effect of shifting $\hat{t}_1$ to one side of the selection interval $[\delta, 1 - \delta]$. In this example, based on the MSEs of $\hat{t}_1/t_1$ (fourth column), the empirically best value h_1^* of h for the estimation of t_1 was $h_1^* = 0.2119$ derived by a one-step interpolation improvement.

We now show the performance of $\hat{d}_1$, $c_\omega \cdot S(\hat{t}_1)$. For each value of h given in the first column, the MSE of $\hat{d}_1/d_1$ with $g = h$, the MSE of $\hat{d}_1/d_1$ with $g = g^*$ and g^* were given in the fifth, sixth and seventh columns, respectively. These values decreased as the value of h, $h \in [0.02, 0.19]$, increased. As the value of h, $h \in [0.19, 0.3]$, increased, $\hat{d}_1$ derived by using $g = h$ or $g = g^*$ gave nearly the same performance, since the corresponding MSEs of $\hat{d}_1/d_1$ were approximately equal in magnitude. As the value of h, $h \in [0.3, 0.48]$, increased further, the MSE of $\hat{d}_1/d_1$ with $g = h$ increased, but that with $g = g^*$ decreased. Hence, the performance of $\hat{d}_1$ with $g = h$ deteriorated. This drawback to $\hat{d}_1$ with $g = h$ could be improved by using the bandwidth g^* in $\hat{d}_1$. For this, see the corresponding MSEs of $\hat{d}_1/d_1$ with $g = g^*$ which were smaller than those with $g = h$.

For each positive integer j, to test the null hypothesis H_0: $p = j$ against the alternative hypothesis H_1: $p > j$, by (3.11), we should cut out the regions $[\hat{t}_k - 2h, \hat{t}_k + 2h]$ for $k = 1, 2, \ldots, j - 1$. In this case, if the value of h is chosen as $h_1^* = 0.2119$, the empirically best value of h for the estimation of t_1, then the test statistic $\hat{d}_j^*$ given in Remark 4 cannot be calculated since the set A_j is empty. To make the set A_j of enough length such that $\hat{d}_j^*$ can be calculated, if a small value of h is given, then the resulting $\hat{d}_j^*$ will suffer from the effects of outliers. For this, see Remark 5. Therefore, the hypothesis test of $p = j$ was not performed in this simulation study.

5. Sketches of the proofs. The following results and notation will be used in this section. By the regression model (2.1) and the Gasser–Müller estimator (2.3), $J(x)$ and $S(x)$ can be decomposed into

$$J(x) = J_V(x) + \tilde{J}(x) + \sum_{j=1}^{p} d_j J_j(x),$$

$$S(x) = S_V(x) + \tilde{S}(x) + \sum_{j=1}^{p} d_j S_j(x),$$

where $\tilde{S}(x)$ and $S_j(x)$ have been given in (3.1) and $S_V(x)$ is defined as $\tilde{S}(x)$ with $r(x_i)$ replaced by ε_i. Here $J_V(x)$, $\tilde{J}(x)$ and $J_j(x)$ are defined as $S_V(x)$, $\tilde{S}(x)$ and $S_j(x)$ with K_3, K_4 and g replaced by K_1, K_2 and h, respectively, in each case.

Let z_i, $i \in Z$, denote partition points of $[0, 1]$ satisfying $z_i - z_{i-1} = n^{-2}$, Ψ the interval $[\delta, 1 - \delta]$,

$$\Psi^* = \{i: z_i \in \Psi\}, \Psi_j = \left\{x: x \in \Psi \quad \textit{and} \quad \prod_{k=1}^{j} I_{[|x-t_k| > h^{1+\theta}(\log n)]} = 1\right\},$$

$$\Psi_j^* = \{i: z_i \in \Psi_j\}$$

and z_j^* the partition point satisfying $|z_j^* - t_j| = \min\{|z_i - t_j|: z_i \in [\delta, 1 - \delta]\}$ for $j = 1, 2, \ldots, p$.

PROOF OF THEOREM 1.

We first give the proof of (3.2). The proof for $\hat{t}_1$ is complete by showing

$$P\left(\sup_{x \in \Psi_1} |J(x)| \geq |J(z_1^*)| \text{ i.o.}\right) = 0. \tag{5.1}$$

To check (5.1), by (A.1), (A.3), (B.2) and $|d_j| > |d_{j+1}|$ and $J_j(x) = 0$, $x \notin [t_j - h, t_j + h]$ for $j = 1, 2, \ldots, p$, through a straightforward calculation,

$$\sup_{x \in \Psi_1} |\tilde{J}(x)| = O(h),$$

$$|d_1 J_1(z_1^*)| - \sup_{x \in \Psi_1} \left|\sum_{j=1}^{p} d_j J_j(x)\right|$$

$$= \left|d_1 \int_0^{h^{\theta}(\log n)} (K_1 - K_2)\right| + O(n^{-1}h^{-1}) \geq 4C + O(n^{-1}h^{-1}),$$

where

$$C = \tfrac{1}{4}|d_1|\eta h^{2\theta}(\log n)^2.$$

Combining these two results with the decomposition of $J(x)$, through a straightforward calculation,

$$\sup_{x \in \Psi_1} |J(x)| - |J(z_1^*)|$$

$$\leq 2 \sup_{i \in \Psi^*} |J_V(z_i)| + \sup_{i \in \Psi_1^*} \sup_{|x - z_i| \leq n^{-2}} |J_V(x) - J_V(z_i)|$$

$$- 4C + O(h + n^{-1}h^{-1}).$$

By this inequality, the proof of (5.1) is complete by showing that

$$P\left(\sup_{i \in \Psi^*} |J_V(z_i)| \geq C \text{ i.o.}\right) = 0,$$

$$P\left(\sup_{i \in \Psi_1^*} \sup_{|x - z_i| \leq n^{-2}} |J_V(x) - J_V(z_i)| + O(h + n^{-1}h^{-1}) \geq 2C \text{ i.o.}\right) = 0.$$

These proofs are essentially the same as (2.1) of Cheng and Lin (1981). Hence, the proof for $\hat{t}_1$ is complete.

We now give the proof for $\hat{t}_2$. The proofs for the rest of $\hat{t}_j$ follow similarly. Since the distance between any two of t_j, $j = 1, 2, \ldots, p$, is greater than δ and $h = o(1)$, then, for sufficiently large n, we have $|z_2^* - t_1| > 3h$. Using this result and the property of $\hat{t}_1$ in (3.2),

$$P\left(z_2^* \in \left[\hat{t}_1 - 2h, \hat{t}_1 + 2h\right] \text{ i.o.}\right) = 0.$$

Following essentially the same proof of (5.1), through a straightforward calculation,

$$P\left(\sup_{x \in \Psi_2} |J(x)| \geq |J(z_2^*)| \text{ i.o.}\right) = 0.$$

According to the property of $\hat{t}_1$ in (3.2) and the fact that

$$|\hat{t}_2 - t_1| \geq |\hat{t}_2 - \hat{t}_1| - |\hat{t}_1 - t_1| \geq 2h - |\hat{t}_1 - t_1|,$$

then

$$P\left(|\hat{t}_2 - t_1| < h \text{ i.o.}\right) = 0.$$

Combining these results with the definition of Ψ_2,

$$P\left(|\hat{t}_2 - t_2| > h^{1+\theta}(\log n) \text{ i.o.}\right) = 0.$$

Hence, the proof of (3.2) is complete.

We now give the proof of (3.3). Here we only give the proof for $\hat{d}_1 - d_1$. The proofs for the rest of $\hat{d}_j - d_j$ follow similarly. Using the decomposition of $S(x)$, (A.1) and (A.4) and subtracting and adding the term $\sum_{j=1}^{p} d_j S_j(t_1)$, $\hat{d}_1 - d_1$ can be asymptotically expressed as

$$\begin{aligned} \hat{d}_1 - d_1 = {} & c_\omega S_V(\hat{t}_1) + c_\omega \sum_{j=1}^{p} d_j \left(S_j(\hat{t}_1) - S_j(t_1)\right) \\ & + \left[c_\omega \sum_{j=1}^{p} d_j S_j(t_1) - d_1\right] + O(g). \end{aligned} \tag{5.2}$$

Multiplying the second term on the right-hand side of (5.2) by $I_{[|\hat{t}_1 - t_1| \geq h^{1+\theta}(\log n)]} + I_{[|\hat{t}_1 - t_1| < h^{1+\theta}(\log n)]}$ and combining the result with (A.4), through a straightforward calculation, it becomes

$$\begin{aligned} & \left| c_\omega \sum_{j=1}^{p} d_j \left(S_j(\hat{t}_1) - S_j(t_1)\right) \right| \\ & \quad = \left| c_\omega \sum_{j=1}^{p} d_j \left(S_j(\hat{t}_1) - S_j(t_1)\right) \right| \cdot I_{[|\hat{t}_1 - t_1| \geq h^{1+\theta}(\log n)]} + O\left(h^{1+\theta} g^{-1}(\log n)\right). \end{aligned}$$

Combining this result with (5.2), (3.1), (B.2) through (B.4), the property of $\hat{t}_1$ in (3.2) and (2.1) of Cheng and Lin (1981), through a straightforward calculation, the proof of (3.3) is complete.

We now give the proof of (3.4). Here we only give the proof of the asymptotic normality for Λ_1. The proofs for the rest of Λ_j follow similarly. By the

decomposition of $J(x)$, the proof of the asymptotic normality for $\hat{t}_1$ is based on

$$0 = J^{(1)}(\hat{t}_1) = J_V^{(1)}(\hat{t}_1) + \tilde{J}^{(1)}(\hat{t}_1) + \sum_{j=1}^{p} d_j J_j^{(1)}(\hat{t}_1). \tag{5.3}$$

By Riemann summation, (A.1), and (A.3), through a straightforward calculation,

$$\sup_{x \in \Psi} \left| \tilde{J}^{(1)}(x) \right| = O(n^{-1}h^{-2} + 1), \tag{5.4}$$

$$J_j^{(1)}(x) = h^{-1}(K_1 - K_2)\big((x - t_j)/h\big) + O(n^{-1}h^{-2}), \tag{5.5}$$

for $x \in \Psi$ and $j = 1, 2, \ldots, p$. Multiplying (5.5) by $I_{[|\hat{t}_1 - t_1| \geq h^{1+\theta}(\log n)]} + I_{[|\hat{t}_1 - t_1| < h^{1+\theta}(\log n)]}$ and combining the result with $J_j(x) = 0$, $x \notin [t_j - h, t_j + h]$ for $j = 1, 2, \ldots, p$, $\sum_{j=1}^{p} d_j J_j^{(1)}(\hat{t}_1)$ in (5.3) can be asymptotically expressed as

$$\begin{aligned} \sum_{j=1}^{p} d_j J_j^{(1)}(\hat{t}_1) = {} & \sum_{j=1}^{p} d_j J_j^{(1)}(\hat{t}_1) \cdot I_{[|\hat{t}_1 - t_1| \geq h^{1+\theta}(\log n)]} \\ & + \left[d_1 h^{-1}(K_1 - K_2)\big((\hat{t}_1 - t_1)/h\big) + O(n^{-1}h^{-2}) \right] \\ & \times I_{[|\hat{t}_1 - t_1| < h^{1+\theta}(\log n)]}. \end{aligned}$$

Combining this result with (5.3), (5.4) and (A.3) and applying Taylor's theorem to $K_1 - K_2$, through a straightforward calculation, (5.3) becomes

$$\begin{aligned} 0 = {} & J_V^{(1)}(\hat{t}_1) + \left[\sum_{j=1}^{p} d_j J_j^{(1)}(\hat{t}_1) \right] \cdot I_{[|\hat{t}_1 - t_1| \geq h^{1+\theta}(\log n)]} \\ & + 2d_1 h^{-2}\left[K_2^{(1)}(0) + O(h^{\theta}(\log n)) \right](\hat{t}_1 - t_1) \\ & \times I_{[|\hat{t}_1 - t_1| < h^{1+\theta}(\log n)]} + O(n^{-1}h^{-2} + 1). \end{aligned} \tag{5.6}$$

Giving partition points on $[0, 1]$ such that the distance between every two consecutive partition points is $2n^{-(1/2)-(1/(l-1))}h^{1/2}$ and using Theorem 2 of Whittle (1960) and (B.8), through a straightforward calculation,

$$\sup_{|x - t_1| \leq h^{1+\theta}(\log n)} \left| J_V^{(1)}(x) - J_V^{(1)}(t_1) \right| = o_p(n^{-1/2}h^{-3/2}).$$

Combining this result with (5.6) and the property of $\hat{t}_1$ in (3.2), (5.6) becomes

$$\hat{t}_1 - t_1 = \frac{J_V^{(1)}(t_1) + o_p(n^{-1/2}h^{-3/2}) + O(n^{-1}h^{-2} + 1)}{2d_1 h^{-2} K_2^{(1)}(0) + O(h^{\theta - 2}(\log n))}. \tag{5.7}$$

By the Lindeberg–Feller theorem, through a straightforward calculation,

$$n^{1/2}h^{3/2}J_V^{(1)}(t_1) \Rightarrow N\left(0, \sigma^2 \int \left(K_1^{(1)} - K_2^{(1)}\right)^2\right).$$

Combining this result with (5.7), (B.9) and (A.5), the proof of the asymptotic normality for $\hat{t}_1$ is complete.

The proof of the asymptotic normality for $\hat{d}_1$ is now given. Using the decomposition of $S(x)$ and following essentially the same proof of (5.7), through a straightforward calculation,

$$\hat{d}_1 - d_1 = c_\omega S_V(t_1) + c_\omega \tilde{S}(t_1) + o_p(n^{-1/2}g^{-1/2}) + O(h^{1+\theta}g^{-1}(\log n)). \tag{5.8}$$

By the Lindeberg–Feller theorem, through a straightforward calculation,

$$(ng)^{1/2}S_V(t_1) \Rightarrow N\left(0, \sigma^2 \int (K_3 - K_4)^2\right).$$

Combining this result with (5.8), (3.1), (B.10) through (B.12) and (A.5), the proof of the asymptotic normality for $\hat{d}_1$ is complete. By (5.7), (5.8) and the Cramér–Wold device, through a straightforward calculation, the asymptotic normality of Λ_1 follows.

We now give the proof of the asymptotic independence between Λ_j, $j = 1, 2, \ldots, \rho$. Following essentially the same proofs of (5.7) and (5.8), through a straightforward calculation, we have

$$\hat{t}_j - t_j = \frac{J_V^{(1)}(t_j) + o_p(n^{-1/2}h^{-3/2}) + O(n^{-1}h^{-2} + 1)}{2d_j h^{-2}K_2^{(1)}(0) + O(h^{\theta-2}(\log n))},$$

$$\hat{d}_j - d_j = c_\omega S_V(t_j) + o_p(n^{-1/2}g^{-1/2}) + O(g + h^{1+\theta}g^{-1}(\log n)).$$

Based on these two results and (B.13), the limiting distribution of Λ_j depends on the observations Y_i whose design points $x_i \in [t_j - g, t_j + g]$ for $j = 1, 2, \ldots, \rho$. Since the distance between any two of t_j, for $j = 1, 2, \ldots, p$, is assumed to be greater than δ, then, for sufficiently large n, the intervals $[t_j - g, t_j + g]$ for $j = 1, 2, \ldots, \rho$ are disjoint. This result implies Λ_j are asymptotically independent. Hence, the proof of (3.4) is complete, that is, the proof of Theorem 1 is complete. □

PROOF OF THEOREM 2. Based on the decomposition of $S(x)$, (A.1) and (A.4),

$$S(\hat{t}_j) = S_V(\hat{t}_j) + \sum_{k=1}^{p} d_k S_k(\hat{t}_j) + O(g), \tag{5.9}$$

for $j = p + 1, p + 2, \ldots, \rho$. By (3.2),

$$P\left(\hat{t}_j \in \bigcup_{k=1}^{p} [t_k - h, t_k + h] \text{ i.o.}\right) = 0,$$

for $j = p + 1, p + 2, \ldots, \rho$. Combining this result with the fact that $S_k(x) = 0$, $x \notin [t_k - h, t_k + h]$, for $k = 1, 2, \ldots, p$,

$$P\left(\sum_{k=1}^{p} d_k S_k(\hat{t}_j) \neq 0 \text{ i.o.}\right) = 0,$$

for $j = p + 1, p + 2, \ldots, \rho$. Combining this result with (5.9), (B.4) and (2.1) of

Cheng and Lin (1981), through a straightforward calculation, the proof of Theorem 2 is complete. □

PROOF OF THEOREM 3. Let $\hat{d}^{**}$ denote the supremum of $|S_V(x)|$ for $x \in [a, b]$. By the Lipschitz continuity of m, K_3 and K_4, $\sup_{x \in \Psi} |\tilde{S}(x)| = O(g)$. Combining this result with the decomposition of $S(x)$,

$$\hat{d}^{*} = \hat{d}^{**} + O(g). \tag{5.10}$$

Following essentially the same proof of Theorem 7 of Stadtmüller (1986), through a straightforward calculation,

$$P\left(\hat{d}^{**} < a_n + b_n x\right) \to \exp(-2\exp(-x)).$$

By (B.12) and the orders of the magnitudes of a_n and b_n, $g = o(a_n)$ and $g = o(b_n)$. Combining this result with (5.10), $\hat{d}^{*}$ and $\hat{d}^{**}$ have the same limiting distribution. Hence, the proof of Theorem 3 is complete. □

Acknowledgments. We gratefully thank the two referees, an Associate Editor and the Editor for their many valuable comments which substantially improved the presentation.

REFERENCES

CARROLL, R. J. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10** 1224–1233.

CHENG, K. F. and LIN, P. E. (1981). Nonparametric estimation of a regression function. *Z. Wahrsch. Verw. Gebiete* **57** 223–233.

CHIU, S. T. (1987). A feature-preserving smoother which preserves abrupt changes in mean. Unpublished manuscript.

CHU, C. K. and MARRON, J. S. (1991). Choosing a kernel regression estimator. *Statist. Sci.* **6** 404–436.

CLINE, D. B. H. and HART, J. D. (1991). Kernel estimation of densities with discontinuities or discontinuous derivatives. *Statistics* **22** 69–84.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

GASSER, T. and MÜLLER, H. G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 23–68. Springer, New York.

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

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press.

HÄRDLE, W. (1991). *Smoothing Techniques: With Implementation in S*. Springer, Berlin.

LEE, D. (1990). Coping with discontinuities in computer vision: their detection, classification, and measurement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12** 321–344.

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

MÜLLER, H. G. (1988). *Nonparametric Analysis of Longitudinal Data. Lecture Notes in Statist.* **46**. Springer, Berlin.

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

MÜLLER, H. G. and STADTMÜLLER, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15** 610–625.

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

SHIAU, J. H. (1985). Smoothing spline estimation of functions with discontinuities. Ph.D. dissertation, Dept. Statistics, Univ. Wisconsin, Madison.

SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50** 413–436.

STADTMÜLLER, U. (1986). Asymptotic properties of nonparametric curve estimates. *Period. Math. Hungar.* **17** 83–108.

VAN EEDEN, C. (1985). Mean integrated squared error of kernel estimators when the density and its derivatives are not necessarily continuous. *Ann. Inst. Statist. Math.* **37** 461–472.

VAN ES, B. (1991). Asymptotics for least squares cross-validation bandwidths in non-smooth cases. Unpublished manuscript.

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

WU, J. S. and CHU, C. K. (1993). Nonparametric function estimation and bandwidth selection for discontinuous regression functions. *Statist. Sinica*. To appear.

YIN, Y. Q. (1988). Detection of the number, locations and magnitudes of jumps. *Comm. Statist. Stochastic Models* **4** 445–455.

DEPARTMENT OF MATHEMATICS
TAMKING UNIVERSITY
TAIPEI, TAIWAN
REPUBLIC OF CHINA

INSTITUTE OF STATISTICS
NATIONAL TSING HUA UNIVERSITY
HSINCHU, TAIWAN 30043
REPUBLIC OF CHINA