Statistica Sinica 3(1993), 557-576

NONPARAMETRIC FUNCTION ESTIMATION AND BANDWIDTH SELECTION FOR DISCONTINUOUS REGRESSION FUNCTIONS

J. S. Wu and C. K. Chu

Tamkang University and National Tsing Hua University

Abstract: For nonparametric regression, where the regression function has discontinuity points, the kernel regression estimator and cross-validation are known to be affected by discontinuity. This effect is precisely quantified through the mean average square error (MASE) for the kernel regression estimator and a limiting distribution for the cross-validated bandwidth. An approach is proposed to adjust for the effect of discontinuity on kernel regression estimation and bandwidth selection. The resulting kernel regression estimator and cross-validation are further analyzed by the MASE and a limiting distribution, respectively. Simulation studies show that the asymptotic results are applicable to reasonable sample sizes.

Key words and phrases: Nonparametric regression, kernel regression estimator, cross-validation, discontinuous regression function, mean average square error, asymptotic normality.

1. Introduction

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. Kernel regression estimators are local weighted averages of the response variables. The weights assigned to the observations are calculated from a given function called the kernel function. The width of the neighborhood in which averaging is performed is called the bandwidth or smoothing parameter. The magnitude of bandwidth controls the smoothness of the resulting estimate of the regression function. Choosing a suitable value of bandwidth is the essence of the smoothing problem.

Currently, the results on kernel regression estimation and bandwidth selection given in the literature are usually derived under the assumption that the regression function has two continuous derivatives. In this case, for asymptotic properties of kernel regression estimators, see, for example, the monographs by Eubank (1988), Müller (1988), and Härdle (1990, 1991). Also, for bandwidth selection, cross-validation introduced by Clark (1975) is an attractive method. It

takes the minimizer of the cross-validation score as an estimate of the optimal bandwidth, the minimizer of the mean average square error (MASE) of the kernel regression estimator. For asymptotic properties of the cross-validated bandwidth and asymptotic equivalence of some popular data-driven bandwidth selectors to cross-validation, see, for example, Rice (1984) and Härdle, Hall, and Marron (1988). For other bandwidth selectors, see also Marron (1988), a survey paper, and references given therein.

In practice, however, the regression function may have discontinuity points. For example, consider the cases of studying the impact of advertising, the effect of medicine, and the influence of sudden changes in governmental policies and international relationships. These actions may cause effect instantly. But the times at which these actions cause effect are not known. See Shiau (1985) and McDonald and Owen (1986) for many interesting examples where regression functions are not continuous. See also Yin (1988) and Wu and Chu (1991a) for estimating locations of discontinuity points of the regression function. For a detailed discussion of the effect of discontinuity on the kernel density estimator and cross-validation in the related field of kernel density estimation, see, for example, Van Eeden (1985), Cline and Hart (1989), and Van Es (1990).

For the case that the regression function has discontinuity points, the MASE for the kernel regression estimator and a central limit theorem (CLT) for the cross-validated bandwidth are given in Section 3. This MASE quantifies the effect of discontinuity on the kernel regression estimator by showing the minimum order of magnitude of the MASE. Also, this CLT quantifies the effect of discontinuity on cross-validation by giving the rate of convergence of the cross-validated bandwidth. The minimum MASE of the kernel regression estimator and the rate of convergence of the cross-validated bandwidth are of larger and of smaller orders than those given in Härdle, Hall, and Marron (1988) for the case that the regression function has two continuous derivatives, respectively. These results are the same as those given in Van Eeden (1985), Cline and Hart (1989), and Van Es (1990) for estimating a discontinous density function.

To adjust for the effect of discontinuity on the kernel regression estimator and cross-validation, an immediate remedy is to estimate the regression function and to construct the cross-validation score on subintervals separated by estimates of locations of discontinuity points. For this, if the number of discontinuity points is known, using either of the methods in Yin (1988) and Wu and Chu (1991a), then locations of discontinuity points can be estimated accurately, in the sense of the rate of strong consistency. Based on these estimates of locations of discontinuity points, the resulting kernel regression estimator and cross-validation are further analyzed by the MASE and a CLT in Section 3. The results obtained are the same as those given in Härdle, Hall, and Marron (1988) for the case that the

regression function has two continuous derivatives.

The organization of this paper is as follows. Section 2 describes the regression settings, including precise formulation of the proposed approach for estimating the regression function and choosing the value of bandwidth when the regression function has discontinuity points. Section 3 gives the theoretical results of this paper. Section 4 contains simulation results which give additional insight into what the theoretical results mean. For applications of the proposed approach, we suggest taking the bandwidth needed by either of the methods in Yin (1988) and Wu and Chu (1991a) for estimating locations of discontinuity points as the cross-validated bandwidth. Simulation studies show that the performance of the resulting kernel regression estimator is good, in the sense of the sample MASE. Finally, sketches of the proofs are given in Section 5.

2. Regression Settings

In this paper, the equally spaced fixed design nonparametric regression model is considered. The regression model is given by

$$Y_{n,i} = m(x_{n,i}) + \epsilon_{n,i}, \tag{2.1}$$

for $i = 0, 1, \ldots, n$. Here m is an unknown discontinuous regression function defined on the interval $[0, 1]$ (without loss of generality), $x_{n,i}$ are equally spaced fixed design points, i.e. $x_{n,i} = i/n$, $\epsilon_{n,i}$ are unobservable regression errors, and $Y_{n,i}$ are noisy observations of the regression function m at the design points $x_{n,i}$. In the following, for simplicity of notation, let Y_i, x_i, and ϵ_i denote $Y_{n,i}$, $x_{n,i}$, and $\epsilon_{n,i}$, respectively.

The discontinuous regression function m in (2.1) is defined by

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

where r is a continuous function defined on the interval $[0, 1]$ and ψ is a step function defined by $\psi(x) = \sum_{k=1}^{q} d_k I_{[t_k,1]}(x)$ for $x \in [0, 1]$. Here q is a positive integer representing the number of discontinuity points, t_k are locations of discontinuity points, and d_k are nonzero real numbers representing jump sizes of m at t_k. For simplicity of notation, let $t_0 = 0$, $t_{q+1} = 1$, $|d_1| > |d_2| > \cdots > |d_q|$, and the distance between any two of these t_j, for $j = 0, 1, \ldots, q + 1$, be greater than δ, where δ is an arbitrarily small positive constant.

To estimate $m(x)$, the kernel regression estimator introduced by Nadaraya (1964) and Watson (1964) is considered. To deal with boundary effects on the kernel regression estimator, the method of projected data in Wu and Chu (1991b) is applied. For a detailed discussion of boundary effects, see, for example, Section

4.3 in Müller (1988) and Section 4.4 in Härdle (1990). The method of projected data and the Nadaraya-Watson estimator are introduced in the following.

We now introduce the method of projected data in Wu and Chu (1991b). Given the kernel functions $L(x) = (-6 - 12x) \cdot I_{[-1,0]}(x)$ and $R(x) = -1 \cdot L(-x)$, the bandwidth g, and the observations Y_i at $x_i \in [\kappa, \tau]$, a subinterval of $[0, 1]$, the projected data Y_i^P at $x_i = i/n \in [2\kappa - \tau, 2\tau - \kappa]$ are defined by

$$Y_i^P = \begin{cases} Y_{2A-i} + 2\hat{m}_{gL}(\kappa)(x_i - \kappa) & \text{for } i = 2A - B, 2A - B + 1, \ldots, A - 1 \\ Y_i & \text{for } i = A, A + 1, \ldots, B \\ Y_{2B-i} + 2\hat{m}_{gR}(\tau)(x_i - \tau) & \text{for } i = B + 1, B + 2, \ldots, 2B - A. \end{cases} \tag{2.3}$$

Here A and B denote the minimum and the maximum subindices i for $x_i = i/n \in [\kappa, \tau]$, respectively. Note that $\hat{m}_{gL}(\kappa)$ and $\hat{m}_{gR}(\tau)$ are defined by

$$\hat{m}_{gL}(\kappa) = n^{-1} \Sigma_{i:x_i \in [\kappa,\tau]} L_g(\kappa - x_i) Y_i,$$
$$\hat{m}_{gR}(\tau) = n^{-1} \Sigma_{i:x_i \in [\kappa,\tau]} R_g(\tau - x_i) Y_i.$$

Here and throughout this paper, the upper index P stands for projection and the notation $f_g(\cdot)$ denotes $g^{-1} f(\cdot/g)$ for any given function f and bandwidth g. For formulation of Y_i^P and the choice of L and R, see (2.3) and Remark 3.2 in Wu and Chu (1991b), respectively.

To estimate $m(x)$ on the subinterval $[\kappa, \tau]$ of $[0, 1]$, the Nadaraya-Watson estimator $\hat{m}(x)$ is as follows. Given the projected data Y_i^P in (2.3), the kernel function K which is chosen to be a probability density function, and the bandwidth $h = g/1.572$, define $\hat{m}(x)$ for $x \in [\kappa, \tau]$ by

$$\hat{m}(x) = \frac{n^{-1} \Sigma_{i:x_i \in [2\kappa-\tau, 2\tau-\kappa]} K_h(x - x_i) Y_i^P}{n^{-1} \Sigma_{i:x_i \in [2\kappa-\tau, 2\tau-\kappa]} K_h(x - x_i)}, \tag{2.4}$$

(if the denominator of $\hat{m}(x)$ is zero, take $\hat{m}(x) = 0$). If $m(x)$ has two continuous derivatives on $[\kappa, \tau]$, then the magnitude of bias of $\hat{m}(x)$, for $x \in [\kappa, \tau]$, is of order h^2. For this and the factor 1.572, see Section 3 and Remark 3.2 in Wu and Chu (1991b), respectively. For another formulation of the projected data Y_i^P, see Hall and Wehrly (1991).

The performance of the regression function estimator $\hat{m}(x)$ in (2.4) is measured by the MASE defined by

$$d_M(h) = E\left[n^{-1} \sum_{j=0}^{n} \left(\hat{m}(x_j) - m(x_j) \right)^2 \right]. \tag{2.5}$$

The optimal bandwidth h_M for constructing $\hat{m}(x)$ is taken as the minimizer of $d_M(h)$.

In practice, however, the value of h_M is not available because the quantity depends on the unknown $m(x)$. Since the value of h_M can not calculated, cross-validation is designed to choose the bandwidth by minimizing the cross-validation score $\mathrm{CV}(h)$ defined by

$$\mathrm{CV}(h) = n^{-1} \sum_{j=0}^{n} \left(\hat{m}_j(x_j) - Y_j \right)^2. \tag{2.6}$$

Here $\hat{m}_j(x_j)$ is the "leave-1-out" version of $\hat{m}(x_j)$, i.e. the observation (x_j, Y_j) is left out in constructing $\hat{m}(x_j)$, for each $j = 0, 1, \ldots, n$. More specifically, $\hat{m}_j(x_j)$ is defined by

$$\hat{m}_j(x_j) = \frac{n^{-1} \Sigma_{i:x_i \in [-1,2], i \neq j} K_h(x_j - x_i) Y_i^P}{n^{-1} \Sigma_{i:x_i \in [-1,2], i \neq j} K_h(x_j - x_i)},$$

(if the denominator of $\hat{m}_j(x_j)$ is zero, take $\hat{m}_j(x_j) = 0$), where Y_i^P are the projected data derived from Y_i at $x_i \in [0, 1]$. Let $\hat{h}_{\mathrm{CV}}$ denote the minimizer of $\mathrm{CV}(h)$. Based on (2.1), the asymptotic values of $d_M(h)$ and h_M and the asymptotic behavior of $\hat{h}_{\mathrm{CV}}$ will be studied in Section 3.

Note that the above $\hat{m}(x)$, $d_M(h)$, h_M, $\mathrm{CV}(h)$, and $\hat{h}_{\mathrm{CV}}$ are defined for the case that $m(x)$ in (2.2) is not assumed to be discontinuous. When $m(x)$ in (2.2) is assumed to have unknown discontinuity points, the proposed approach for estimating the regression function and choosing the value of bandwidth is given in the following. If the value of q is known, then t_k can be estimated by either of the methods in Yin (1988) and Wu and Chu (1991a). In this paper, we shall use the kernel type estimators $\hat{t}_k$ in Wu and Chu (1991a) to estimate t_k. Let $\hat{t}_0 = 0$ and $\hat{t}_{q+1} = 1$. These kernel type estimators and their asymptotic behavior will be given at the end of this section.

Based on the estimates $\hat{t}_k$ of t_k, (2.3), and (2.4), the proposed approach is to estimate $m(x)$ independently on each subinterval $[\hat{t}_{k-1}, \hat{t}_k]$ for $k = 1, 2, \ldots, q+1$. Let $\hat{m}^P(x)$ denote the result obtained, i.e. for each $k = 1, 2, \ldots, q+1$, and $x \in [\hat{t}_{k-1}, \hat{t}_k]$,

$$\hat{m}^P(x) = \frac{n^{-1} \Sigma_i^k K_h(x - x_i) Y_i^{Pk}}{n^{-1} \Sigma_i^k K_h(x - x_i)}, \tag{2.7}$$

(if the denominator of $\hat{m}^P(x)$ is zero, take $\hat{m}^P(x) = 0$). Here Y_i^{Pk} are the projected data derived from Y_i at $x_i \in [\hat{t}_{k-1}, \hat{t}_k]$ and $\Sigma_i^k = \sum_{k=1}^{q+1} \Sigma_{i:x_i \in [2\hat{t}_{k-1} - \hat{t}_k, 2\hat{t}_k - \hat{t}_{k-1}]}$. The MASE of $\hat{m}^P(x)$ is given by

$$d_M^P(h) = E\left[n^{-1} \sum_{j=0}^{n} \left(\hat{m}^P(x_j) - m(x_j) \right)^2 \right]. \tag{2.8}$$

Let h_M^P denote the minimizer of $d_M^P(h)$.

On the other hand, based on (2.3) and (2.4), the proposed approach is to choose the bandwidth by minimizing $\mathrm{CV}^P(h)$ defined by

$$\mathrm{CV}^P(h) = n^{-1} \sum_{k=1}^{q+1} \Sigma_{j:x_j \in [\hat{t}_{k-1}, \hat{t}_k]} \left(\hat{m}_j^P(x_j) - Y_j \right)^2. \tag{2.9}$$

Here, for each $x_j \in [\hat{t}_{k-1}, \hat{t}_k]$, $\hat{m}_j^P(x_j)$ is the "leave-1-out" version of $\hat{m}^P(x_j)$, i.e.

$$\hat{m}_j^P(x_j) = \frac{n^{-1} \Sigma_{i,i \neq j}^k K_h(x_j - x_i) Y_i^{Pk}}{n^{-1} \Sigma_{i,i \neq j}^k K_h(x_j - x_i)},$$

(if the denominator of $\hat{m}_j^P(x_j)$ is zero, take $\hat{m}_j^P(x_j) = 0$). Let $\hat{h}_{\mathrm{CV}}^P$ denote the minimizer of $\mathrm{CV}^P(h)$. The asymptotic values of $d_M^P(h)$ and h_M^P and the asymptotic behavior of $\hat{h}_{\mathrm{CV}}^P$ will be studied in Section 3.

We now close this section by introducing the kernel type estimators $\hat{t}_k$ in Wu and Chu (1991a) and their asymptotic behavior, for the case that the value of q is known. To estimate t_k, given the kernel functions $K_1(x) = (0.4857 - 3.8560x + 2.8262x^2 - 19.1631x^3 + 11.9952x^4) \cdot I_{[-1,0.2012]}(x)$ and $K_2(x) = K_1(-x)$ and the bandwidth ω, define $J(x)$ for $x \in [0,1]$ by

$$J(x) = \hat{m}_{\omega K_1}(x) - \hat{m}_{\omega K_2}(x), \tag{2.10}$$

where

$$\hat{m}_{\omega F}(x) = \frac{n^{-1} \Sigma_{i:x_i \in [-1,2]} F_\omega(x - x_i) Y_i^P}{n^{-1} \Sigma_{i:x_i \in [-1,2]} F_\omega(x - x_i)},$$

and where F denotes K_1 and K_2, in each case. For the motivation of $J(x)$ and the choice of K_1 and K_2, see (2.4) and Remark 2 in Wu and Chu (1991a), respectively. In this case, $\hat{t}_k$ are taken as maximizers of $|J(x)|$ over the sets $I_k = [0,1] - \bigcup_{j=1}^{k-1} \left[\hat{t}_j - \frac{\delta}{3}, \hat{t}_j + \frac{\delta}{3} \right]$, for $k = 1, 2, \ldots, q$. For the asymptotic behavior of $\hat{t}_k$, based on Theorem 1 in Wu and Chu (1991a), $\hat{t}_k$ are asymptotically independent and

$$|\hat{t}_k - t_k| < \omega^{1+\theta} \cdot (\log \theta) \qquad \text{almost surely,} \tag{2.11}$$

$$(n/\omega)^{1/2} \left(\hat{t}_k - t_k \right) \Rightarrow N(0, V_k), \tag{2.12}$$

as $n \to \infty$, for $k = 1, 2, \ldots, q$, where $\theta \in (0, \frac{1}{2})$ is a given constant and

$$V_k = \sigma^2 \frac{\int \left(K_1^{(1)}(x) - K_2^{(1)}(x) \right)^2 dx}{\left[2 d_k K_2^{(1)}(0) \right]^2},$$

and where $K_1^{(1)}$ and $K_2^{(1)}$ denote the first derivatives of K_1 and K_2, respectively.

3. Results

In this section, we study the asymptotic values of $d_M(h)$, h_M, $d_M^P(h)$, and h_M^P and the asymptotic behavior of $\hat{h}_{\mathrm{CV}}$ and $\hat{h}_{\mathrm{CV}}^P$. For these, in addition to the assumptions given in Section 2, we impose the following assumptions:

(A.1) The function $r(x)$ in (2.2) has a uniformly continuous and square integrable second derivative $r^{(2)}(x)$ on the interval (0,1).

(A.2) The kernel function K is a square integrable probability density function with support contained in the interval $[-1,1]$. Also K is symmetric about zero and the second derivative of K is Lipschitz continuous.

(A.3) The regression errors ϵ_i are independent and identically distributed random variables with mean 0, variance σ^2, and all other moments finite.

(A.4) The total number of observations in this regression setting is n, with $n \to \infty$.

(A.5) The minimizers $\hat{h}_{\mathrm{CV}}$ and $\hat{h}_{\mathrm{CV}}^P$ of $\mathrm{CV}(h)$ and $\mathrm{CV}^P(h)$, respectively, are searched in the interval $H_n = [\alpha n^{-1+\rho}, \beta n^{-\rho}]$, for $n = 1, 2, \ldots$. Here the positive constants α and ρ are arbitrarily small and β large. The bandwidth ω in (2.10) is also selected in H_n and satisfies the conditions $n^{-1}\omega^{-(1+\theta)} = o(1)$ and $\omega^{1+\theta} = o(h^2)$, where $\theta \in (0, \frac{1}{2})$ is a given constant.

Under the above assumptions, it is shown in Section 5 that $d_M(h)$ can be asymptotically expressed as

$$d_M(h) = \alpha_1 \cdot n^{-1}h^{-1} + \beta_\psi \cdot h + o(n^{-1}h^{-1} + h), \tag{3.1}$$

where

$$\alpha_1 = \sigma^2 \int K(x)^2 dx,$$

$$\beta_\psi = \left[\sum_{k=1}^{q} d_k^2\right] \int C_K(x)^2 dx.$$

Here, the function $C_f(x)$ is defined by $\int_{-a}^{x} f(u)du$ for $-a \le x \le 0$ and $(-1)\int_{x}^{a} f(u)\,du$ for $0 < x \le a$, when the function f with support on the interval $[-a, a]$, $0 < a \le 1$, is given. For the components of MASE, $\alpha_1 \cdot n^{-1}h^{-1}$ and $\beta_\psi \cdot h$ represent the variance and the squared bias, respectively. A consequence of (3.1) is that the minimum order of magnitude of $d_M(h)$ is $n^{-1/2}$ which is arrived at when the value of h is of order $n^{-1/2}$. However, it is larger than the minimum order of magnitude of the MASE, $n^{-4/5}$, given in Härdle, Hall, and Marron (1988) for the case that the regression function has two continuous derivatives.

The following (3.2) and Theorem 1 show the effect of discontinuity of the regression function on the asymptotic behavior of $\hat{h}_{\mathrm{CV}}$. The proof Theorem 1 is given in Section 5. By these results, the magnitude of $\hat{h}_{\mathrm{CV}}$ is of order $n^{-1/2}$ and the rate of convergence of $\hat{h}_{\mathrm{CV}}$ to h_M is of order $n^{-3/4}$. These orders for the cross-validated bandwidth are smaller than those, $n^{-1/5}$ and $n^{-3/10}$, given in Härdle, Hall, and Marron (1988) for the case that the regression function has two continuous derivatives, respectively.

According to (3.1), by a straightforward calculation, the optimal bandwidth h_M can be asymptotically expressed as

$$h_M = \ell_\psi \cdot n^{-1/2} \cdot (1 + o(1)), \tag{3.2}$$

where

$$\ell_\psi = [\alpha_1/\beta_\psi]^{1/2} = \left[\sigma^2 \int K(x)^2 dx \Big(\sum_{k=1}^{q} d_k^2 \Big)^{-1} \Big(\int C_K(x)^2 dx \Big)^{-1}\right]^{1/2}.$$

Theorem 1. *Under the above assumptions, if* $\alpha_1 > 0$, *then*

$$\hat{h}_{\mathrm{CV}}/h_M \;\to\; 1 \quad \textit{almost surely}, \tag{3.3}$$

$$n^{1/4}\big(\hat{h}_{\mathrm{CV}}/h_M - 1\big) \;\Rightarrow\; N(0, \mathrm{VAR}_\psi), \tag{3.4}$$

$$n^{1/2}\big(d_M(\hat{h}_{\mathrm{CV}})/d_M(h_M) - 1\big) \;\Rightarrow\; \frac{1}{2} \cdot \mathrm{VAR}_\psi \cdot \chi_1^2, \tag{3.5}$$

as $n \longrightarrow \infty$, *where*

$$\mathrm{VAR}_\psi = \Big[\sigma^2 \Big/ \sum_{k=1}^{q} d_k^2\Big]^{1/2} \cdot \psi_K,$$

and where

$$\begin{aligned}\psi_K \;=\; & \Big[\int C_K(x)^2 dx \int K(x)^2 dx\Big]^{-3/2} \cdot \Big[2 \int C_K(x)^2 dx \int (K * (K - G)(x) \\ & -(K - G)(x))^2 dx + \int (C_{G-2K} * K(x) + C_K * G(x) \\ & +C_{G-K}(x))^2 dx \int K(x)^2 dx\Big].\end{aligned}$$

Here and throughout this paper, $G(x) = (-x)K^{(1)}(x)$, and $*$ means convolution.

We now study the performance of the proposed approach to kernel regression estimation and bandwidth selection. For this, in addition to the above assumptions, add

(A.6) The value of q in (2.2) is known.

Under the above assumptions, it is shown in Section 5 that $d_M^P(h)$ can be asymptotically expressed as

$$d_M^P(h) = \alpha_1 \cdot n^{-1}h^{-1} + \beta_r \cdot h^4 + o(n^{-1}h^{-1} + h^4), \tag{3.6}$$

where the coefficient α_1 has been given in (3.1) and

$$\beta_r = (1/4)\Big(\int x^2 K(x)dx\Big)^2 \int r^{(2)}(x)^2 dx.$$

This asymptotic value of $d_M^P(h)$ is the same as that of the MASE given in Härdle, Hall, and Marron (1988) for the case that the regression function has two continuous derivatives.

The following (3.7) and Theorem 2 give the asymptotic behavior of $\hat{h}_{CV}^P$. The proof of Theorem 2 is given in Section 5. Based on (3.6), by a straightforward calculation, the value of h_M^P can be asymptotically expressed as

$$h_M^P = \ell_r \cdot n^{-1/5} \cdot (1 + o(1)), \tag{3.7}$$

where

$$\ell_r = [\alpha_1/(4\beta_r)]^{1/5} = \left[\sigma^2 \int K(x)^2 dx \Big(\int x^2 K(x)dx\Big)^{-2} \Big(\int r^{(2)}(x)^2 dx\Big)^{-1}\right]^{1/5}$$

Theorem 2. *Under the above assumptions, if* $\alpha_1 > 0$ *and* $\beta_r > 0$, *then*

$$\hat{h}_{\mathrm{CV}}^P / h_M^P \;\to\; 1 \quad \textit{almost surely}, \tag{3.8}$$

$$n^{1/10}\big(\hat{h}_{\mathrm{CV}}^P / h_M^P - 1\big) \;\Rightarrow\; N(0, \mathrm{VAR}_r), \tag{3.9}$$

$$n^{1/5}\big(d_M^P(\hat{h}_{\mathrm{CV}}^P)/d_M^P(h_M^P) - 1\big) \;\Rightarrow\; 2 \cdot \mathrm{VAR}_r \cdot \chi_1^2, \tag{3.10}$$

as $n \longrightarrow \infty$, *where*

$$\mathrm{VAR}_r = \left[\sigma^2 / \int r^{(2)}(x)^2 dx\right]^{1/5} \cdot r_K,$$

and where

$$r_K = (8/25)\frac{\int (K * (K - G)(x) - (K - G)(x))^2 dx}{\left[\Big(\int K(x)^2 dx\Big)^9 \Big(\int x^2 K(x)dx\Big)^2\right]^{1/5}}.$$

By these results, the asymptotic behavior of $\hat{h}_{\mathrm{CV}}^P$ is the same as that of the cross-validated bandwidth given in Härdle, Hall, and Marron (1988) for the case that the regression function has two continuous derivatives.

We now close this section by the following remarks.

Remark 3.1. If $\alpha_1 = 0$, then h_M, $\hat{h}_{\mathrm{CV}}$, h_M^P, and $\hat{h}_{\mathrm{CV}}^P$, are equal to the left-end of H_n for all large n. On the other hand, if $\beta_r = 0$, then h_M^P and $\hat{h}_{\mathrm{CV}}^P$ are equal to the right-end of H_n for all large n.

Remark 3.2. To apply the proposed approach, the number q of discontinuity points of $m(x)$ can be obtained from prior knowledge about the process under study. Also, it can be obtained from testing hypotheses about the value of q by Theorem 3 and Remark 4 in Wu and Chu (1991a), and estimated by the method in Yin (1988). Let ℓ be the estimate of q produced by either of these two approaches. Based on these values of q and ℓ, the performance of the proposed approach to kernel regression estimation and bandwidth selection is given in the following. If $\ell \geq q \geq 0$, then (3.6) through (3.10) still hold. On the other hand, if $\ell < q$, then (3.6) through (3.10) become (3.1) through (3.5) with the coefficient $\sum_{k=1}^{q} d_k^2$ replaced by $\sum_{k=\ell+1}^{q} d_k^2$.

Remark 3.3. To estimate t_k, the choice of the value of ω in (2.10) in practice is now discussed. In the case of $q \geq 1$, we suggest taking ω as $\hat{h}_{\mathrm{CV}}$. In this case, the value of $\hat{h}_{\mathrm{CV}}$ is of order $n^{-1/2}$. Also, by (3.6), the minimum order of magnitude of the MASE of $\hat{m}^P(x)$ is arrived at when the value of h is of order $n^{-1/5}$. Given the values of h and ω of the orders $n^{-1/5}$ and $n^{-1/2}$, respectively, the conditions on ω given in (A.5) are satisfied for any value of $\theta \in (0, 1/2)$. Simulation studies in Section 4 show that the performance of $\hat{t}_k$, the estimate of t_k, derived by choosing ω as $\hat{h}_{\mathrm{CV}}$ is good, in the sense of the sample MASE of $\hat{m}^P(x)$.

4. Simulations

To investigate the practical implications of the asymptotic results presented in Section 3, an empirical study was carried out. We first introduce the simulated regression settings. The sample size was $n = 100$. The regression model (2.1) and the kernel regression estimator (2.4) were considered. The continuous function r and the step function ψ in (2.2) were $r(x) = x^4$ and $\psi(x) = I_{[1/2,1]}(x)$, for $x \in [0, 1]$. In this simulation study, we took the value $q = 1$. It was assumed to be known in advance. The location of the discontinuity point was $t_1 = 1/2$ and the corresponding jump size was $d_1 = 1$. The kernel function K was $K(x) = (3/4)(1 - x^2)$, for $x \in [-1, 1]$. See Section 5.4 in Müller (1988) for properties of this kernel function K. Given this kernel function K, we have $\int C_K(x)^2 dx = 0.118$. The regression errors ϵ_i were pseudo independent normal random variables $N(0, \sigma^2)$, where $\sigma = 1/2$. Based on (2.1), 100 independent sets of the observations Y_i were generated. Given this large value of $\sigma = 1/2$, the location of the discontinuity point $t_1 = 1/2$ of the regression function was not always distinguishable visually,

from the data alone. For this, see Figure 1 where stars denote one set of simulated data. To produce the projected data Y_i^P, we took the value of bandwidth $g = 1.572 \cdot h$ where the bandwidth h was applied to $\hat{m}(x)$ and $\hat{m}^P(x)$, in each case. If $g < 0.05$, then it was taken as 0.05. On the other hand, if $g > 0.2$, then it was taken as 0.2. This restriction on the value of g was based on the fact that the magnitudes of biases of $\hat{m}_{gL}(\kappa)$ and $\hat{m}_{gR}(\tau)$ are proportional to the value of g^2. Hence, to produce the projected data Y_i^P properly, the value of g should not be large. However, if the value of g is too small, then there is few observations which can be used by $\hat{m}_{gL}(\kappa)$ and $\hat{m}_{gR}(\tau)$.

The calculation of $d_M(h)$, h_M, $\mathrm{CV}(h)$, and $\hat{h}_{\mathrm{CV}}$ is now introduced. For each data set, the values of $d_A(h)$ and $\mathrm{CV}(h)$ were calculated on an equally spaced grid of 16 values in the interval $[0.03, 0.48]$. Here $d_A(h)$ was defined by

$$d_A(h) = n^{-1} \sum_{j=0}^{n} \Big(\hat{m}(x_j) - m(x_j)\Big)^2. \tag{4.1}$$

The value of $d_M(h)$ was empirically approximated by averaging $d_A(h)$ over the 100 simulated data sets, for each given value of h. The minimizers h_M and $\hat{h}_{\mathrm{CV}}$ of $d_M(h)$ and $\mathrm{CV}(h)$, respectively, were calculated. After evaluation on the grid, a one step interpolation improvement was performed, with the results taken as the selected bandwidths. If these functions had multiple minimizers on the grid, the algorithm chose the smallest one, respectively (the choice could be made arbitrarily).

We now introduce the calculation of $\hat{t}_1$, the estimate of t_1. For each data set, to estimate t_1, the value of ω in (2.10) was chosen as $\omega = \hat{h}_{\mathrm{CV}}$. By this, the values of $|J(x)|$ were calculated at the design points x_i. Given the value $\delta = 0.05$ in (2.2), the maximizer $\hat{t}_1$ of $|J(x)|$ over the set $[\delta, 1-\delta]$ was calculated. After evaluation at the design points x_i, a one step interpolation improvement was performed, with the result taken as the estimate of t_1.

Based on the above estimate of t_1, following the same algorithm for calculating $d_A(h)$, $d_M(h)$, h_M, $\mathrm{CV}(h)$, and $\hat{h}_{\mathrm{CV}}$, the values of $d_A^P(h)$, $d_M^P(h)$, h_M^P, $\mathrm{CV}^P(h)$, $\hat{h}_{\mathrm{CV}}^P$, $d_M^T(h)$, and h_M^T were calculated on the above equally spaced grid of h. Here $d_A^P(h)$ was defined by

$$d_A^P(h) = n^{-1} \sum_{j=0}^{n} \Big(\hat{m}^P(x_j) - m(x_j)\Big)^2.$$

Also $d_M^T(h)$ and h_M^T were defined the same as $d_M^P(h)$ and h_M^P with $\hat{t}_k$ replaced by t_k, in each case. Note that $d_M^T(h)$ denotes the MASE function of $\hat{m}(x)$ in the case that locations of discontinuity points in (2.2) are known. In the following, the results of the simulation study are presented.

Figure 1 shows one set of simulated observations (stars), the regression function (dashed curves), and the regression function estimates derived by $\hat{m}(x)$ with $h = \hat{h}_{\mathrm{CV}}$ (solid curve) and $\hat{m}^P(x)$ with $\omega = \hat{h}_{\mathrm{CV}}$ and $h = \hat{h}^P_{\mathrm{CV}}$ (dotted curves). For this data set, $\hat{h}_{\mathrm{CV}} = 0.0748$. Given the value of ω for estimating t_1 as this $\hat{h}_{\mathrm{CV}}$, we had $\hat{t}_1 = 0.4876$ and $\hat{h}^P_{\mathrm{CV}} = 0.1711$. Note that the regression function estimate derived by $\hat{m}(x)$ with $h = \hat{h}_{\mathrm{CV}}$ is very rough and does not show the discontinuity of the regression function. On the other hand, the regression function estimate derived by $\hat{m}^P(x)$ with $\omega = \hat{h}_{\mathrm{CV}}$ and $h = \hat{h}^P_{\mathrm{CV}}$ is smooth on the subintervals $[0, \hat{t}_1]$ and $[\hat{t}_1, 1]$ and shows the discontinuity of the regression function clearly. Note also that the magnitude of bias of $\hat{m}^P(x)$ in the neighborhood of $x = t_1$ is larger than those at $x = 0$ and $x = 1$. This magnitude of bias of $\hat{m}^P(x)$ in the neighborhood of $x = t_1$ was caused by the gap between t_1 and $\hat{t}_1$. The gap between $\hat{t}_1$ and t_1 caused one of the two estimates $\hat{m}_{gL}(\hat{t}_1)$ and $\hat{m}_{gR}(\hat{t}_1)$ to have larger magnitude of bias than the other. Hence the performance of the projected data at $x = \hat{t}_1$ was inferior to that at $x = 0$ and $x = 1$, in the sense of the magnitude of bias of $\hat{m}^P(x)$.

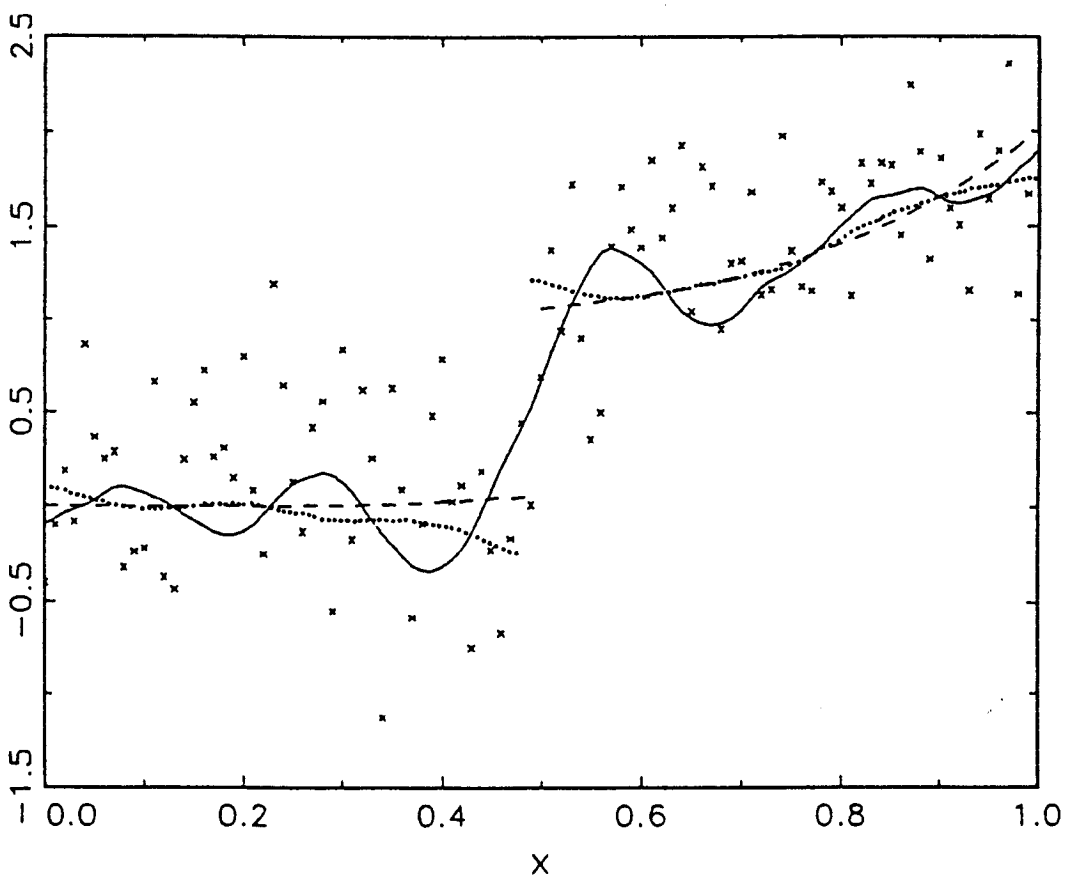

Figure 1. Plot of one simulated data set (stars), the regression function $m(x)$ (dashed curves), and the regression function estimates derived by $\hat{m}(x)$ with $h = \hat{h}_{\mathrm{CV}}$ (solid curve) and $\hat{m}^P(x)$ with $\omega = \hat{h}_{\mathrm{CV}}$ and $h = \hat{h}^P_{\mathrm{CV}}$ (dotted curves).

Figure 2 shows $d_M(h)$ (solid curve), $d_M^P(h)$ derived by using $\omega = \hat{h}_{CV}$ (dotted curve), and $d_M^T(h)$ (dashed curve). The location of the star on each curve denotes that of the minimizer of the corresponding curve. In view of locations of these stars, the value of h_M is significantly smaller than those of h_M^P and h_M^T. Also the value of $d_M(h_M)$ is larger than those of $d_M^P(h_M^P)$ and $d_M^T(h_M^T)$. The difference between the values of $d_M^P(h_M^P)$ and $d_M^T(h_M^T)$ is caused by the gaps between $\hat{t}_1$ and t_1, over the 100 simulated data sets. These gaps between $\hat{t}_1$ and t_1 explain the magnitudes of variance and bias of $\hat{m}^P(x)$ in the neighborhood of t_1.

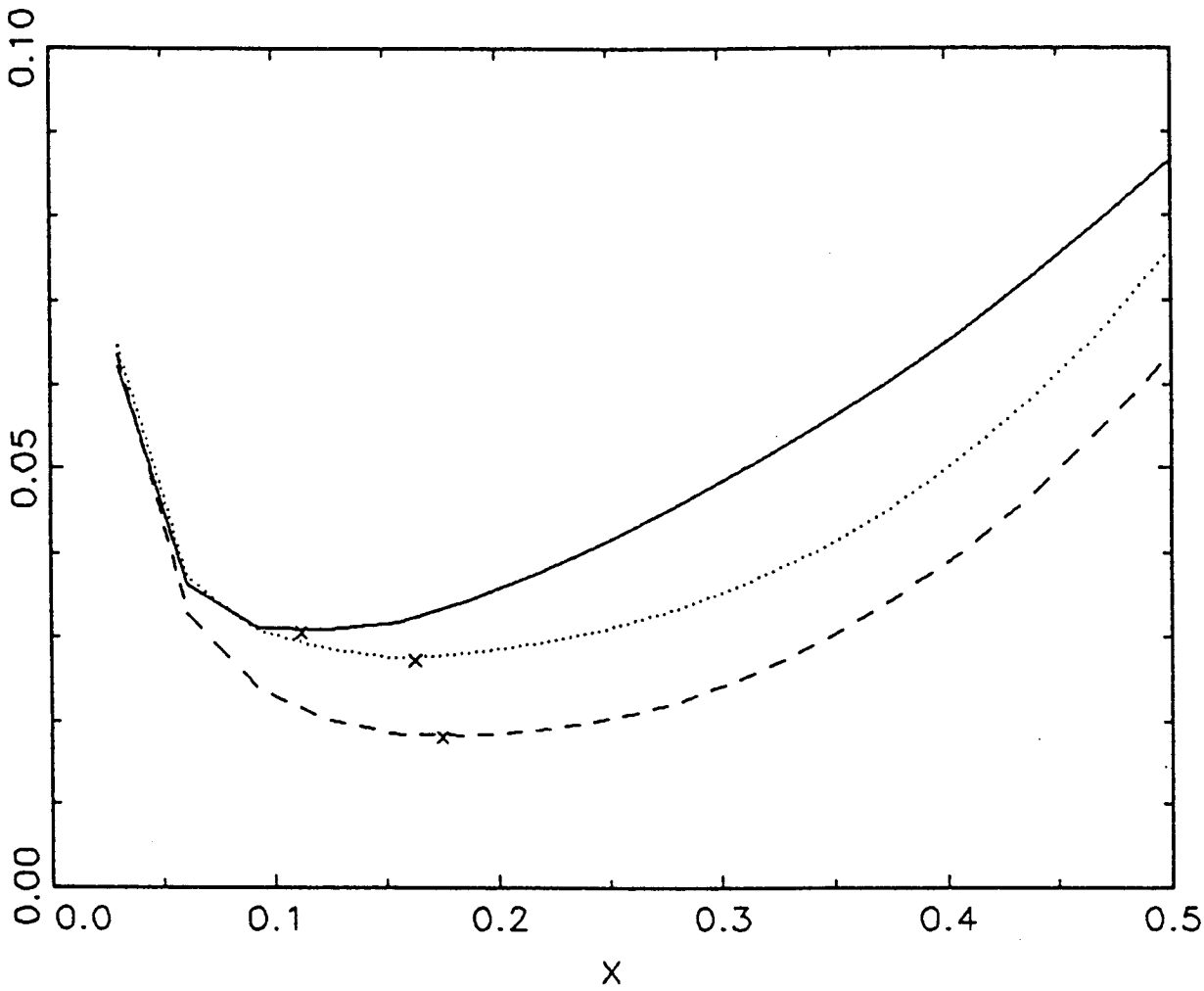

Figure 2. The MASE functions $d_M(h)$ (solid curve), $d_M^P(h)$ with $\omega = \hat{h}_{CV}$ (dotted curve), and $d_M^T(h)$ (dashed curve). Here the location of the star on each curve denotes that of the minimizer of the curve.

Figure 3 shows the regression function (dashed curves) and the average of regression function estimates derived by $\hat{m}(x)$ with $h = \hat{h}_{CV}$ (solid curve) over the 100 simulated data sets. Here bold dotted curves denote 1*STD bands around the average. This STD was taken as the sample standard deviation over the 100 regression function estimates. Note that, for $x \in [0.42, 0.58]$, the 1*STD bands do not contain the regression function $m(x)$, and the magnitude of bias of $\hat{m}(x)$ is large. Hence, the continuity of the regression function should be checked before the regression function is recovered by kernel regression estimators. For this, see Theorem 3 and Remark 4 in Wu and Chu (1991a).

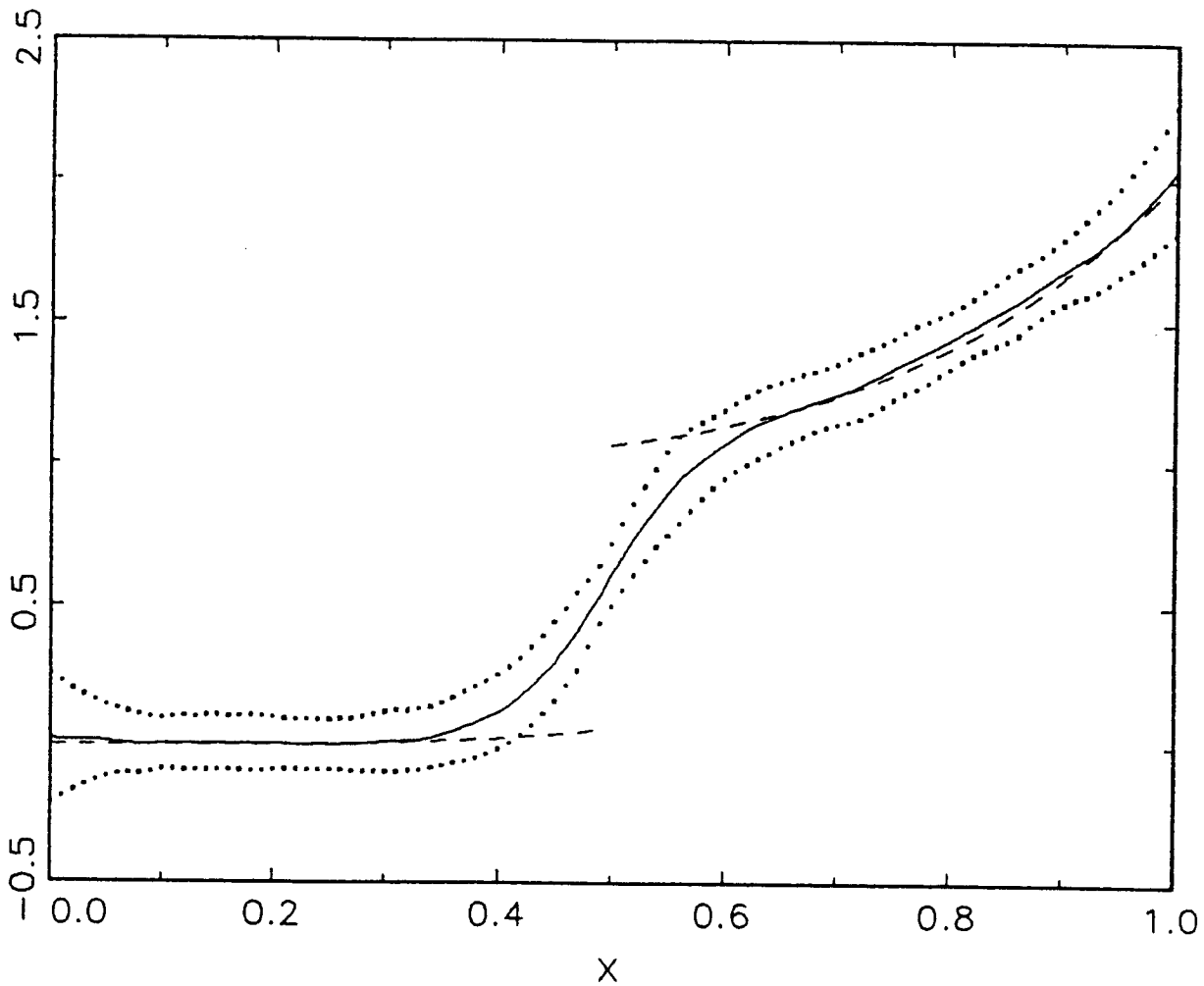

Figure 3. The regression function $m(x)$ (dashed curves), the average of the regression function estimates derived by $\hat{m}(x)$ with $h = \hat{h}_{CV}$ (solid curve), and 1*STD bands (bold dotted curves).

Finally, Figure 4 shows the regression function (dashed curves) and the average of regression function estimates derived by $\hat{m}^P(x)$ with $\omega = \hat{h}_{\mathrm{CV}}$ and $h = \hat{h}^P_{\mathrm{CV}}$ (dotted curve), over the 100 simulated data sets. Bold dotted curves denote 1*STD bands around the average. Note that these 1*STD bands contain the regression function $m(x)$ for $x \in [0, 1]$. Also, this average of regression function estimates shows the discontinuity of the regression function at $x = 1/2$ clearly. However, in the neighborhood of $x = t_1$, the magnitudes of STD and bias of $\hat{m}^P(x)$ are larger than those for x outside this neighborhood. These large magnitudes of STD and bias of $\hat{m}^P(x)$ were caused by the gaps between $\hat{t}_1$ and t_1 over the 100 simulated data sets.

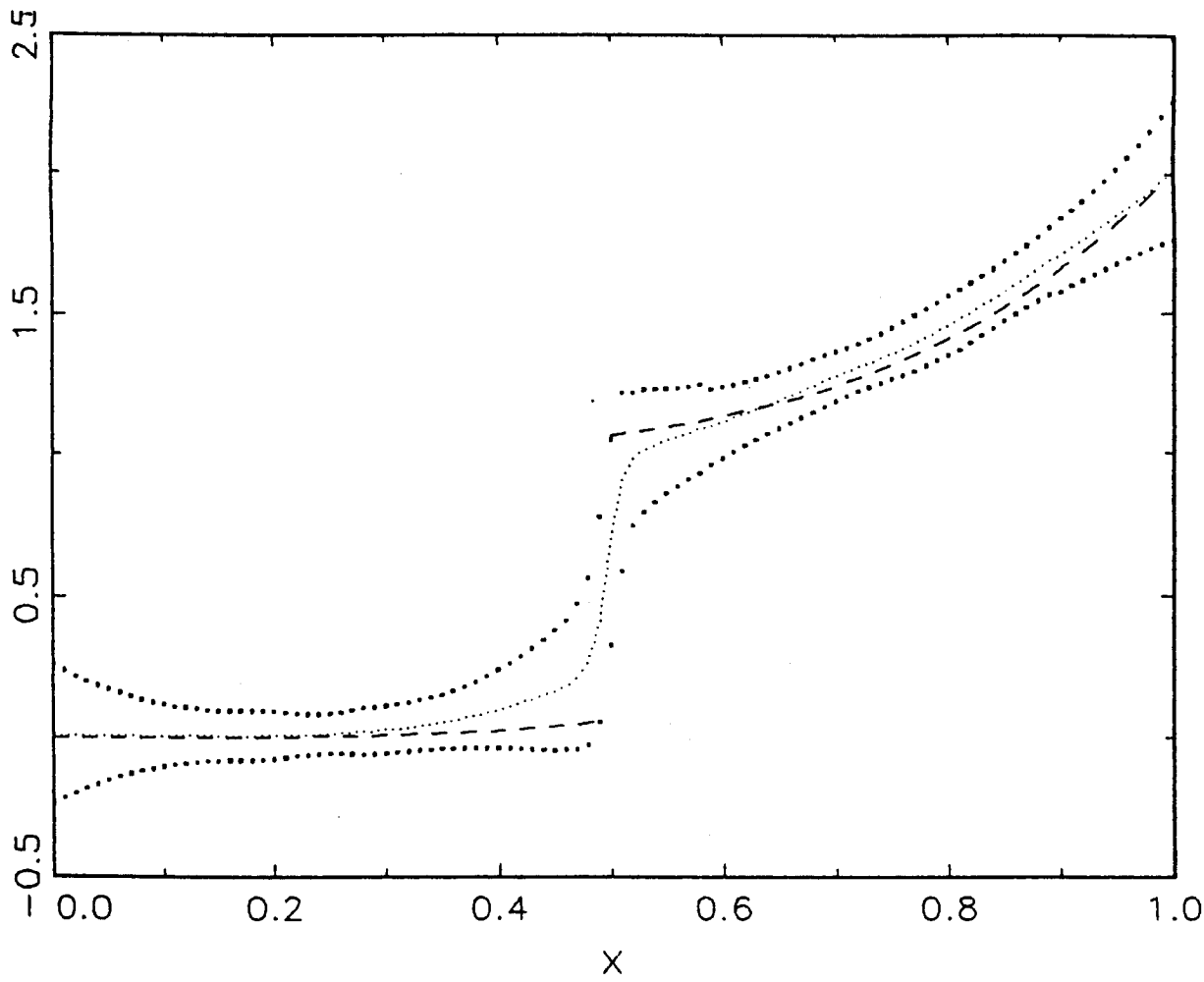

Figure 4. The regression function $m(x)$ (dashed curves), the average of the regression function estimates derived by $\hat{m}^P(x)$ with $\omega = \hat{h}_{\mathrm{CV}}$ and $h = \hat{h}^P_{\mathrm{CV}}$ (dotted curves), and 1*STD bands (bold dotted curves).

5. Sketches of the Proofs

The following notation and results will be used in this section. Let $X_n = o_u(a_n)$ denote that, as $n \to \infty$, $|X_n/a_n| \to 0$ almost surely, and uniformly on H_n if a_n involves $h \in H_n$. For each $x_j \in [h, 1-h]$, under the above assumptions, using Riemann summation and Theorem 2 in Whittle (1960), and by a straightforward calculation, we have the following asymptotic results:

$$n^{-1}\sum_{i=0}^{n} K_h(x_j - x_i) = 1 + O(n^{-1}h^{-1}),$$

$$n^{-2}\sum_{i=0}^{n} K_h(x_j - x_i)^2 = n^{-1}h^{-1}\int K(x)^2 dx + O(n^{-2}h^{-2}),$$

$$\begin{aligned} b_{rj} &= \Big[n^{-1}\sum_{i=0}^{n} K_h(x_j - x_i)(r(x_i) - r(x_j))\Big] \Big/ \Big[n^{-1}\sum_{i=0}^{n} K_h(x_j - x_i)\Big] \\ &= \frac{1}{2}h^2 r^{(2)}(x_j)\int x^2 K(x)dx + o(h^2) + O(n^{-1}), \end{aligned}$$

$$\begin{aligned} b_{mj} &= \Big[n^{-1}\sum_{i=0}^{n} K_h(x_j - x_i)(m(x_i) - m(x_j))\Big] \Big/ \Big[n^{-1}\sum_{i=0}^{n} K_h(x_j - x_i)\Big] \\ &= \begin{cases} b_{rj} & \text{if } K_h(x_j - t_k) = 0 \text{ for all } k \\ b_{rj} + d_k C_K((x_j - t_k)/h) + O(n^{-1}h^{-1}) & \text{if } K_h(x_j - t_k) \neq 0 \text{ for some } k, \end{cases} \end{aligned}$$

$$E\Big[\Big(\sum_{j=0}^{n}\epsilon_j\Big)^{2k}\Big] = O(n^k), \quad \text{for } k = 1, 2, \ldots,$$

$$v_j = \frac{n^{-1}\sum_{i=0}^{n} K_h(x_j - x_i)\epsilon_i}{n^{-1}\sum_{i=0}^{n} K_h(x_j - x_i)} = o_u\Big((nh)^{-2/5}\Big) \quad \text{or} \quad O_p\Big((nh)^{-1/2}\Big).$$

Proof of (3.1). Based on the performance of the projected data Y_i^P, the magnitudes of variance and bias of $\hat{m}(x)$ for each $x \in [0, h]$ and $[1-h, 1]$ are of the orders $n^{-1}h^{-1}$ and h^2, respectively. Using these results, $d_M(h)$ can be asymptotically expressed as

$$d_M(h) = n^{-1}\sum_{j:x_j \in [h,1-h]} E[v_j^2] + n^{-1}\sum_{j:x_j \in [h,1-h]} b_{mj}^2 + O(n^{-1} + h^5).$$

Combining this result with the above asymptotic results and the continuity of C_K^2, by a straightforward calculation, the proof of (3.1) is complete.

Proof of Theorem 1. We first give the proof of (3.3). It is based on the expansion of CV(h). Through adding and subtracting the terms $\hat{m}(x_j)$ and $m(x_j)$, CV(h) can be expressed as

$$\mathrm{CV}(h) = n^{-1}\sum_{j=0}^{n}\epsilon_j^2 + d_M(h) + D(h) + \mathrm{Cross}(h) + \mathrm{Remainder}(h), \tag{5.1}$$

where

$$\begin{aligned} D(h) &= d_A(h) - d_M(h), \\ \mathrm{Cross}(h) &= (-2)n^{-1}\sum_{j=0}^{n}\Big(\hat{m}_j(x_j) - m(x_j)\Big)\epsilon_j, \\ \mathrm{Remainder}(h) &= n^{-1}\sum_{j=0}^{n}\Big(\hat{m}_j(x_j) - \hat{m}(x_j)\Big)\Big(\hat{m}_j(x_j) + \hat{m}(x_j) - 2m(x_j)\Big), \end{aligned}$$

and where $d_A(h)$ has been given in (4.1).

Under the above assumptions and asymptotic results, using the fact that there are no boundary effects on $\hat{m}(x)$ for $x \in [0, h]$ and $[1-h, 1]$ and following essentially the same proofs of (5.3) and Lemmas 3 and 4 in Härdle and Marron (1985) in the random design setting, we have

$$D(h) = o_u(d_M(h)), \tag{5.2}$$

$$\mathrm{Cross}(h) = o_u(d_M(h)), \tag{5.3}$$

$$\mathrm{Remainder}(h) = o_u(d_M(h)). \tag{5.4}$$

Combining these results with (5.1) yields

$$\mathrm{CV}(h) = n^{-1}\sum_{j=0}^{n}\epsilon_j^2 + d_M(h) + o_u(d_M(h)). \tag{5.5}$$

Applying Taylor's theorem to (5.5) and combining the result with the consequence of (5.5) that the values of $\hat{h}_{\mathrm{CV}}$ and h_M are of the order $n^{-1/2}$ yields

$$\hat{h}_{\mathrm{CV}}/h_M = 1 + o_u(1).$$

Hence the proof of (3.3) is complete.

Based on (5.1) and the fact that there are no boundary effects on $\hat{m}(x)$ for $x \in [0, h]$ and $[1-h, 1]$, the proofs of (3.4) and (3.5) are essentially the same as those

of Theorems 1 and 2 in Härdle, Hall, and Marron (1988). The only difference between these proofs concerns on the variance of $D^{(1)}(h) + \text{Cross}^1(h)$. In our case, under the above assumptions and asymptotic results, by a straightforward calculation, this variance can be asymptotically expressed as

$$\begin{aligned}
&\text{Var}\Big(D^{(1)}(h) + \text{Cross}^{(1)}(h)\Big) \\
&= 8n^{-2}h^{-3}\sigma^4 \int \Big(K * (K-G)(x)dx - (K-G)(x)\Big)^2 dx + 4n^{-1}h^{-1}\sigma^2\Big[\sum_{k=1}^{q} d_k^2\Big] \\
&\quad \int \Big(C_{G-2K} * K(x) + C_K * G(x) + C_{G-K}(x)\Big)^2 dx + o(n^{-2}h^{-3} + n^{-1}h^{-1}).
\end{aligned}$$

Under the above assumptions, using this asymptotic variance and following the proofs of Theorems 1 and 2 in Härdle, Hall, and Marron (1988), the proofs of (3.4) and (3.5) are complete. Thus the proof of Theorem 1 is complete.

Proof of (3.6). By virtue of (2.12) and (A.5), we have

$$|\hat{t}_k - t_k| < \omega^{1+\theta} = o(h^2) \quad \text{almost surely.}$$

Based on this result, the distance between $\hat{t}_k$ and t_k is much smaller than the values of bandwidths h and g. Hence, the effect of d_k on the projected data Y_i^{Pk} derived from Y_i at $x_i \in [\hat{t}_{k-1}, \hat{t}_k]$ is negligible, in the sense of the bias of $\hat{m}^P(\hat{t}_{k-1})$ and $\hat{m}^P(\hat{t}_k)$. Combining this result with the above asymptotic results, the performance of the projected data on dealing with boundary effects, and the boundedness of $m(x)$ for $x \in [0,1]$, the proof of (3.6) is complete.

Proof of Theorem 2. We first give the proof of (3.8) which is based on the expansion of $\text{CV}^P(h)$. Through adding and subtracting the terms $\hat{m}^P(x_j)$ and $m(x_j)$ and using the results given in the proof of (3.6), $\text{CV}^P(h)$ can be expressed as

$$\begin{aligned}
\text{CV}^P(h) &= n^{-1}\sum_{j=0}^{n} \epsilon_j^2 + d_M^P(h) + D^P(h) + \text{Cross}^P(h) \\
&\quad + \text{Remainder}^P(h) + o_u(n^{-1}h^{-1} + h^4), \qquad (5.6)
\end{aligned}$$

where

$$\begin{aligned}
D^P(h) &= d^P(h) - d_M^P(h), \\
\text{Cross}^P(h) &= (-2)n^{-1}\Sigma_{j:x_j \in B_J}\Big(\hat{m}_j^P(x_j) - m(x_j)\Big)\epsilon_j, \\
\text{Remainder}^P(h) &= n^{-1}\Sigma_{j:x_j \in B_J}\Big(\hat{m}_j^P(x_j) - \hat{m}^P(x_j)\Big)\Big(\hat{m}_j^P(x_j) + \hat{m}^P(x_j) - 2m(x_j)\Big),
\end{aligned}$$

and where

$$B_J = [0,1] - \bigcup_{k=1}^{q} [\hat{t}_k - \omega, \hat{t}_k + \omega],$$

$$d^P(h) = n^{-1}\Sigma_{j:x_j \in B_J}\left(\hat{m}^P(x_j) - m(x_j)\right)^2.$$

Under the above assumptions and asymptotic results, following the same proofs of (5.3) and Lemmas 3 and 4 in Härdle and Marron (1985) in the random design setting, by a straightforward calculation, $D^P(h)$, $\text{Cross}^P(h)$, and $\text{Remainder}^P(h)$ are of the order $o_u\left(d_M^P(h)\right)$. Combining these results with (5.6) yields

$$\text{CV}^P(h) = n^{-1}\sum_{j=0}^{n} \epsilon_j^2 + d_M^P(h) + o_u\left(d_M^P(h)\right). \tag{5.7}$$

Applying Taylor's theorem to (5.7) and combining the result with the consequence of (5.7) that the values of $\hat{h}_{\text{CV}}^P$ and h_M^P are of the order $n^{-1/5}$ yields

$$\hat{h}_{\text{CV}}^P / h_M^P = 1 + o_u(1).$$

Hence the proof of (3.8) is complete.

Given (5.6), the proofs of (3.9) and (3.10) are essentially the same as those of Theorems 1 and 2 in Härdle, Hall, and Marron (1988). Hence the proof of Theorem 2 is complete.

Acknowledgements

The research of the second author was supported by the National Science Council under the contract NSC80-0208-M007-32. We gratefully thank the referees, the associate editor, and the editor for their many valuable comments which substantially improved the presentation.

References

Clark, R. M. (1975). A calibration curve for radiocarbon data. *Antiquity* **49**, 251–266.

Cline, D. B. H. and Hart, J. D. (1989). Kernel estimation of densities with discontinuities or discontinuous derivatives. *Statistics* **22**, 69–84.

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

Hall, P. and Wehrly, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *J. Amer. Statist. Assoc.* **86**, 665–672.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press.

Härdle, W. (1991). *Smoothing Techniques: With Implementation in S.* Springer Series in Statistics, Springer-Verlag, New York.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83**, 86–101.

Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13**, 1465–1481.

Marron, J. S. (1988). Automatic smoothing parameter selection: A survey. *Empirical Econom.* **13**, 187–208.

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 Statistics* **46**, Springer-Verlag, Berlin.

Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9**, 141–142.

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, Department of Statistics, University of Wisconsin, Madison, Wisconsin.

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**, Part A, 461–472.

Van Es, B. (1990). Asymptotics for least squares cross-validation bandwidths in non-smooth cases. To appear in *Ann. Statist.*

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser.A* **26**, 359–372.

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. (1991a). Kernel type estimators of jump points and values of a regression function. To appear in *Ann. Statist.*

Wu, J. S. and Chu, C. K. (1991b). Modification for boundary effects and jump points in nonparametric regression. To appear in *J. Nonparamet. Statist.*

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

Department of Mathematics, Tamkang University, Taipei 25103, Taiwan.
Institute of Statistics, National Tsing Hua University, Hsinchu 30043, Taiwan.

(Received July 1991; accepted January 1993).