Statistica Sinica **9**(1999), 775-794

AUTOMATIC BANDWIDTH CHOICE IN A SEMIPARAMETRIC REGRESSION MODEL

Sheng-Yan Hong

Northwestern University

Abstract: Speckman (1988) proposed a kernel smoothing method to estimate the parametric component β in the semiparametric regression model $y = x^\tau\beta + g(t) + e$, and showed that this kernel smoothing estimator is $\sqrt{n}$-consistent for a certain deterministic bandwidth choice. However, the important issue of automatic bandwidth choice in this semiparametric setting has not been examined. This paper studies the asymptotic behavior of the bandwidth choice based on a general bandwidth selector which covers such well known data-driven methods as GCV and CV. This automatic bandwidth choice is proved to be asymptotically optimal and its asymptotic normality is established. The resulting data-driven kernel smoothing estimator of β is then showed to be still $\sqrt{n}$-consistent. A simulation study is performed to compare small sample behaviors of various commonly used bandwidth selectors in this semiparametric setting, and a real data example is given.

Key words and phrases: Asymptotic normality, automatic bandwidth choice, data-driven estimator, kernel smoothing, semiparametric regression model, $\sqrt{n}$-consistency.

1. Introduction

Consider the following semiparametric regression model

$$y_i = x_i^\tau\beta + g(t_i) + e_i, \qquad 1 \le i \le n, \tag{1.1}$$

where $x_i = (x_{i1}, \ldots, x_{ip})^\tau$ and $t_i \in [0,1]$ are covariates, $\beta = (\beta_1, \ldots, \beta_p)^\tau$ is a p-vector of unknown parameters, g is an unknown smooth function, and $\{e_i\}$ are i.i.d. errors with mean 0 and variance $\sigma^2 > 0$. This model, also called the partial linear model, was proposed in Wahba (1984) and Engle, Granger, Rice and Weiss (1986) and has received considerable attention in the last decade.

Primary concern is to estimate the parameter of interest β with usual parametric convergence rate $n^{-1/2}$. The first approach is the partial spline smoothing proposed in Engle, Granger, Rice and Weiss (1986) and Wahba (1984). However, this method suffers the problem of undersmoothing (Rice (1986)), that is, the partial spline smoothing estimate of β cannot attain the $n^{-1/2}$ convergence rate unless the nonparametric component g is undersmoothed. This problem has been overcome by the kernel smoothing method proposed by Speckman (1988). The kernel smoothing estimator of β is of the following form

$$\hat{\beta}_h = (\tilde{X}^\tau \tilde{X})^{-1} \tilde{X}^\tau \tilde{Y}, \tag{1.2}$$

where ($I = I_n$ being $n \times n$ identity matrix)

$$X = (x_1, \ldots, x_n)^\tau, \quad \tilde{X} = (\tilde{x}_1, \ldots, \tilde{x}_n)^\tau = (I - W(h))X,$$
$$Y = (y_1, \ldots, y_n)^\tau, \quad \tilde{Y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^\tau = (I - W(h))Y,$$

with

$$W(h) = (K_{nh}(t_i, t_j)),$$

where K_{nh} is associated with a kernel function and the bandwidth $h = h_n > 0$.

Speckman (1988) showed that the asymptotic normality of $\hat{\beta}_h$ (which yields $\sqrt{n}$-consistency) and the optimal nonparametric convergence rate of $\hat{g}_h$ can be simultaneously achieved for a certain nonrandom bandwidth choice. See Hong and Cheng (1992, 1994) for other asymptotic properties of $\hat{\beta}_h$. From a practical point of view, however, we are more concerned with asymptotic properties when the bandwidth h is chosen by some data-driven methods, such as the generalized cross-validation (GCV) proposed by Craven and Wahba (1979). Although this issue has been extensively studied in the context of nonparametric regression, much less has been done in the present semiparametric regression setting. To my knowledge, the only relevant references are Speckman (1988) and Chen and Shiau (1994). Speckman (1988) gave a weak GCV theorem for the kernel smoothing method as in Craven and Wahba (1979). Chen and Shiau (1994) obtained $\sqrt{n}$ consistency for the estimator of β based on a two-stage partial spline smoothing with the smoothing parameter chosen by GCV or Mallows' C_L criterion (Mallows (1973)). The method considered in Chen and Shiau (1994) depends strongly on the existence of a common orthonormal basis for the spline smoothing matrix and is not applicable to the kernel smoothing setting.

In this paper, we study two basic questions. Are commonly used bandwidth selection methods such as GCV and (delete-one) CV applicable here? Is Speckman's estimator $\hat{\beta}_h$ still $\sqrt{n}$-consistent when the bandwidth h is chosen by one of these selectors? As in nonparametric setting, when we look at the first question, we are paticularly interested in the so-called asymptotic optimality (see Section 2 for the definition) and convergence rates of the data-driven bandwidth choice. These are investigated in Section 2 where a general bandwidth selector is introduced. A simulation study is presented which compares the small sample behaviors of several bandwidth selectors, including GCV and CV. In Section 3, the second question is answered by establishing asymptotic normality, and an application to a real data set is given. Section 4 contains some technical lemmas used in the proofs of our main results.

2. Automatic Bandwidth Choice

In this section, a general bandwidth selector is defined and asymptotic properties of its minimizer are studied.

2.1. A general bandwidth selector

Let $Y_1, \ldots, Y_n$ be independent observations with unknown means $\mu_1, \ldots, \mu_n$ and common variance σ^2. Write $Y = (Y_1, \ldots, Y_n)^\tau$ and $\mu = (\mu_1, \ldots, \mu_n)^\tau$. Suppose that to estimate μ, a class of linear estimators $\hat{\mu}(h) = S(h)Y$, indexed by $h \in \Lambda$, is proposed. Here $S(h)$ is an $n \times n$ 'hat' matrix. Our objective is to select from Λ the optimal h_{ASE} which minimizes the *Average square error* (*ASE*)

$$L_n(h) = n^{-1}\|\mu - \hat{\mu}(h)\|^2.$$

However, since h_{ASE} cannot be computed without knowing μ, the $L_n(h)$ must be estimated from the data and then minimized with respect to h in Λ to obtain an estimator of h_{ASE}. Many such data-based criteria have the form

$$G(h) = \Xi(h)n^{-1}\|(I - S(h))Y\|^2, \tag{2.1}$$

where $\Xi(h)$ is a correction factor which may be random or nonrandom. Usually $\Xi(h)$ depends on h through the *trace* of $S(h)$. Examples include

(a) GCV: $\Xi_{GCV}(h) = (1 - n^{-1}trS(h))^{-2}$;
(b) AIC (Akaike (1974)): $\Xi_{AIC}(h) = \exp\{2n^{-1}trS(h)\}$;
(c) FPE (Akaike (1970)): $\Xi_{FPE}(h) = (1 + n^{-1}trS(h))/(1 - n^{-1}trS(h))$;
(d) S (Shibata (1981)): $\Xi_S(h) = 1 + 2n^{-1}trS(h)$;
(e) T (Rice (1984)): $\Xi_T(h) = (1 - 2n^{-1}trS(h))^{-1}$.

Bandwidth selection based on these data-driven methods has been examined by many researchers in the context of nonparametric regression. See Härdle and Marron (1985), Härdle, Hall and Marron (1988) and references therein. Härdle, Hall and Marron (1988) observed that, in the kernel nonparametric regression, each of the above factors is of the form $1 + 2K(0)(nh)^{-1} + O((nh)^{-2})$.

In the present semiparametric setting, by taking expectation conditionally on $\{x_i, t_i\}$, it is easy to see that the hat matrix is of the form

$$S(h) = W(h) + P_{\tilde{X}}(I - W(h)), \tag{2.2}$$

where $P_{\tilde{X}} = \tilde{X}(\tilde{X}^\tau\tilde{X})^{-1}\tilde{X}^\tau$ is a projection matrix.

Since the basic requirement on the bandwidth h in a large sample study is that $h \to 0$ and $nh \to \infty$, it is reasonable to choose the index set $\Lambda_n = [(n\delta_n)^{-1},\ \delta_n]$, where $\delta_n \to 0$ can be arbitrarily slow.

It turns out that $n^{-1}trS(h)$ can be approximated by (see Lemma 4.7)

$$n^{-1}trS(h) = K(0)/(nh) + p/n + O_p(n^{-1/2}r(h))$$

uniformly over $h \in \Lambda_n$, where $r(h)$ is defined in (2.8). So it is easy to see that each of the above factors has Taylor expansion

$$\Xi(h) = 1 + 2K(0)/(nh) + 2p/n + O_p(r^{3/2}(h)).$$

Therefore we consider the general criterion (2.1) with $\Xi(h)$ being of the form

$$\Xi(h) = 1 + 2K(0)/(nh) + a_n + O(r^{3/2}(h)) \tag{2.3}$$

uniformly over $h \in \Lambda_n$, where $a_n = O(n^{-1})$ is independent of h. (The term $O(\cdot)$ in (2.3) is replaced by $O_p(\cdot)$ if $\Xi(h)$ is random). Note that this general bandwidth selector essentially includes the (delete-one) cross-validation proposed by Clark (1975),

$$CV(h) = n^{-1}\sum_{i=1}^{n}(y_i - x_i^\tau \hat{\beta}_h^{(i)} - \hat{g}_{2h}^{(i)}(t_i))^2,$$

where $\hat{\beta}_h^{(i)}$ and $\hat{g}_{2h}^{(i)}(t)$ are "leave one out" versions of $\hat{\beta}_h$ and $\hat{g}_{2h}(t)$ respectively. In fact, one can show that

$$\frac{CV(h)}{n^{-1}\|(I - S(h))Y\|^2} = 1 + \frac{2K(0)}{nh} + O_p(r^{3/2}(h))$$

uniformly over $h \in \Lambda_n$.

Let $\hat{h}_G$ and h_{ASE} be the minimizers of $G(h)$ and $L_n(h)$ in Λ_n, respectively. The data-driven bandwidth $\hat{h}_G$ is called *asymptotically optimal* if

$$L_n(\hat{h}_G)/L_n(h_{ASE}) \xrightarrow{P} 1. \tag{2.4}$$

An alternative to the performance criterion $L_n(h)$ is the *conditional mean average square error* (*CMASE*)

$$R_n(h) = E(L_n(h)|x,t) = n^{-1}\|(I - S(h))g\|^2 + n^{-1}\sigma^2 tr(S^\tau(h)S(h)),$$

the expectation being taken conditionally on $\{x_i, t_i\}$. Let h_{CMASE} be the minimizer of $R_n(h)$ in Λ_n. Note that the usual *mean average square error* (*MASE*), which is the mean of $R_n(h)$, has no explicit expression in this semiparametric setting. Hence we consider $R_n(h)$ here.

2.2. Asymptotic properties of $\hat{h}_G$

For simplicity, we assume throughout this paper that $t_i = i/n$, $i = 1, \ldots, n$. For a symmetric kernel function $K(\cdot)$, the weight K_{nh} is taken to be

$$K_{nh}(t, t') = \frac{1}{nh}K\Big(\frac{t - t'}{h}\Big), \tag{2.5}$$

as proposed by Priestley and Chao (1972). Suppose, as is common in this setting, that $\{x_i\}$ and $\{t_i\}$ are related via the regression model

$$x_{ij} = g_j(t_i) + \eta_{ij}, \quad 1 \le i \le n, \ \ 1 \le j \le p, \tag{2.6}$$

where the g_j's are smooth functions and $\{\eta_i\} = \{(\eta_{i1}, \ldots, \eta_{ip})^\tau\}$ are i.i.d. error vectors with zero mean and positive definite covariance matrix Σ. It is assumed that $\{\eta_i\}$ and $\{e_i\}$ are independent. We need the following conditions.

(C1) The kernel function K is symmetric with compact support and, for some integer $k \ge 2$,

$$\int K(t)t^r dt = \begin{cases} 1, & \text{if } r = 0, \\ 0, & \text{if } 1 \le r < k, \\ a_k \ne 0, & \text{if } r = k. \end{cases}$$

(C2) $g(t)$ and $g_j(t), 1 \le j \le p$, are k times continuously differentiable.

(C3) $E(e_1^4) < \infty$ and $E\|\eta_1\|^4 < \infty$.

Let

$$r_n(h) = n^{-1} g^\tau (I - W(h))^2 g + n^{-1}\sigma^2 tr(W^2(h)), \tag{2.7}$$

$$r(h) = d_k h^{2k} + b\sigma^2/(nh), \tag{2.8}$$

$$c_{1k} = \Big(b\sigma^2/(2kd_k)\Big)^{1/(2k+1)}, \tag{2.9}$$

where

$$b = \int K^2(t)dt, \qquad d_k = \Big(\frac{a_k}{k!}\Big)^2 \int (g^{(k)}(t))^2 dt.$$

Note that the unique minimizer of $r(h)$ over $h > 0$ is $h_0^* = c_{1k} n^{-1/(2k+1)}$. Also it is well known that

$$\sup_{h \in \Lambda_n} |r_n(h)/r(h) - 1| = o(1). \tag{2.10}$$

Thus, letting h_0 be the minimizer of $r_n(h)$ over Λ_n, we have $h_0/h_0^* \longrightarrow 1$ and

$$n^{(2k-2)/(2k+1)} r_n''(h_0) \longrightarrow (2k+1)b\sigma^2/c_{1k}^3. \tag{2.11}$$

Theorem 2.1. *Suppose that conditions* (C1)-(C3) *hold and that the kernel K is k times continuously differentiable. Then $\hat{h}_G$ is asymptotically optimal with respect to $L_n(h)$ and $R_n(h)$, respectively. Also we have $\hat{h}/h_0^* \xrightarrow{\text{P}} 1$ for $\hat{h} = h_{ASE}$, h_{CMASE} and $\hat{h}_G$.*

The next theorem deals with the convergence rates of $\hat{h}_G$ and $L_n(\hat{h}_G)$, which shows that the relative convergence rate of $\hat{h}_G$ to h_{ASE} is slower than (half, actually) that of $L_n(\hat{h}_G)$ to $L_n(h_{ASE})$. Let

$$\sigma_4^2 = 8\sigma^4 c_{1k}^{-3} \int \left(K(u) + uK'(u)\right)^2 du + 4c_{1k}^{2k-2} d_k \sigma^2. \tag{2.12}$$

Theorem 2.2. *Suppose that conditions* (C1)-(C3) *hold and that the kernel* K *is* $(k+2)$ *times continuously differentiable. Then we have*

$$n^{1/(2(2k+1))}\left(\hat{h}_G/h_{ASE}-1\right)\xrightarrow{D}N(0,\ \sigma_1^2), \tag{2.13}$$

$$n^{1/(2k+1)}\left(L_n(\hat{h}_G)/L_n(h_{ASE})-1\right)\xrightarrow{D}k\sigma_1^2\chi_1^2, \tag{2.14}$$

where $\sigma_1^2=c_{1k}^4\sigma_4^2/((2k+1)b\sigma^2)^2$.

Similarly, the following analog of Theorem 2.2 holds.

Theorem 2.3. *Under the conditions of Theorem* 2.2 *we have*

$$n^{-1/(2(2k+1))}\left(\hat{h}_G/h_{CMASE}-1\right)\xrightarrow{D}N(0,\ \sigma_2^2),$$

$$n^{-1/(2k+1)}\left(R_n(\hat{h}_G)/R_n(h_{CMASE})-1\right)\xrightarrow{D}k\sigma_2^2\chi_1^2.$$

Here $\sigma_2^2=c_{1k}^4\sigma_5^2/((2k+1)b\sigma^2)^2$ *with*

$$\sigma_5^2=\frac{8\sigma^2}{c_{1k}^3}\int(K-L_1-K*K+K*L_1)^2,$$

$L_1(u)=-uK'(u)$, *and* $*$ *denoting convolution.*

Remark 2.1. The covariates t need not be equally spaced. They can be generated by some density. Also, the covariates t can be multivariate (q-dimensional, say). In this case, the weight K_{nh} of (2.5) is replaced by its multivariate version

$$K_{nh}(t,\,t')=\frac{1}{nh^q}\prod_{j=1}^{q}K\Big(\frac{t_j-t_j'}{h}\Big),$$

where t_j and t_j' are the jth component of the vectors t and t', respectively. Theorems 2.1-2.3 still hold with appropriate changes in constants and rates of convergence.

Remark 2.2. The Priesley-Chao weight (2.5) can be replaced by other weights, such as Nadaraya-Waston kernel weights.

Our results show that the bandwidth choice based on $G(h)$ has the same asymptotic performances as in the nonparametric model

$$y_i=g(t_i)+e_i,\qquad 1\le i\le n, \tag{2.15}$$

obtained from (1.1) with $\beta=0$. See Härdle and Marron (1985) and Härdle, Hall and Marron (1988). This is not surprising because the estimator $\hat{\beta}_h$ is

$\sqrt{n}$-consistent, hence the bandwidth choice here is essentially a nonparametric problem. In view of this, it should be mentioned that recent developments of bandwidth selection methodology in nonparametric settings provide several more efficient alternatives to the method used in this paper (see Härdle, Hall and Marron (1992), Jones, Marron and Sheather (1996a,b) and references therein). For example, note the plug-in method (Gasser, Kneip and Köhler (1991) and Hall, Sheather, Jones and Marron (1991)) or, even better, the solve-the-equation method (Sheather and Jones (1991)). Furthermore, the traditional kernel regression and the bandwidth methodology developed for this situation have been extended to local polynomial regression (see Fan and Gijbels (1995), Ruppert, Sheather and Wand (1995) and references therein). It is expected that the merits of these methods continue to the present semiparametric setting, though details might be more complicated.

2.3. Proofs of theorems

The proofs make use of a series of lemmas in the next section. To simplify notation, we use $o_p^*(\cdot)$ ($O_p^*(\cdot)$) to indicate "$o_p(\cdot)$ ($O_p(\cdot)$) holds uniformly over $h \in \Lambda_n$" and write W and S for $W(h)$ and $S(h)$, respectively.

Proof of Theorem 2.1. By (2.2) and Lemmas 4.5 and 4.3 with $\nu = k$ and $\alpha = \alpha_1 = k/(4k+1)$, one can easily get

$$n^{-1}\|(I-S)g\|^2 = n^{-1}g^\tau(I-W)^2 g + O_p^*(r^2(h)),$$

$$n^{-1}g^\tau(I-S)^\tau Se = o_p^*(r(h)h^{(1-\varepsilon)/2}),$$

$$n^{-1}e^\tau S^\tau Se = \sigma^2 n^{-1}tr(W^2) + \xi_n + o_p^*(r(h)h^\alpha),$$

where

$$\xi_n = n^{-2}e^\tau\eta\Sigma^{-1}\eta^\tau e = O_p(n^{-1}) = o_p^*(r(h)h^\alpha).$$

Hence by (2.10),

$$\begin{aligned} L_n(h) &= n^{-1}\|(I-S)g\|^2 + n^{-1}e^\tau S^\tau Se - 2n^{-1}g^\tau(I-S)^\tau Se \\ &= r_n(h) + o_p^*(r(h)h^\alpha) = r(h) + o_p^*(r(h)). \end{aligned} \tag{2.16}$$

Moreover, by Lemmas 4.3 and 4.5,

$$\begin{aligned} n^{-1}\|(I-S)Y\|^2 &= L_n(h) + n^{-1}e^\tau e + 2n^{-1}g^\tau(I-S)e - 2n^{-1}e^\tau Se \\ &= L_n(h) + n^{-1}e^\tau e - 2K(0)\sigma^2/(nh) + o_p^*(r(h)h^\alpha). \end{aligned} \tag{2.17}$$

Thus it is easily seen that

$$\begin{aligned} G(h) &= \Big(L_n(h) + n^{-1}e^\tau e - 2K(0)\sigma^2/(nh) + o_p^*(r(h)h^\alpha)\Big) \\ &\quad \times \Big(1 + 2K(0)/(nh) + a_n + O(r^{3/2}(h))\Big) \\ &= r(h) + n^{-1}e^\tau e + o_p^*(r(h)). \end{aligned} \tag{2.18}$$

Theorem 2.1 follows from (2.16) and (2.18).

Proof of Theorem 2.2. The proof is similar in spirit to that of Härdle, Hall and Marron (1988). Denote for any small $\varepsilon > 0$

$$\Lambda_\varepsilon = \{h : |h/h_0^* - 1| \le \varepsilon\},$$
$$l_n(h) = n^{-1}\|g - W(g+e)\|^2,$$
$$D(h) = l_n(h) - El_n(h) = l_n(h) - r_n(h),$$
$$\delta(h) = 2n^{-1} < g - W(g+e), e > +2K(0)e^\tau e/(n^2 h).$$

By Lemma 4.5 we see, similarly to (2.16), that

$$L_n(h) = l_n(h) + \xi_n + O_p^*(r^2(h)) + o_p^*(n^{-1/2}r(h)).$$

Since $P\{h_{ASE} \in \Lambda_\varepsilon\} \to 1$ by Theorem 2.1, differentiating $L_n(h)$ for $h \in \Lambda_\varepsilon$ gives

$$\begin{aligned} 0 = L_n'(h_{ASE}) &= l_n'(h_{ASE}) + o_p\Big(n^{-(\frac{1}{2}+\frac{2k-1}{2k+1})}\Big) \\ &= r_n''(\triangle)(h_{ASE} - h_0) + D'(h_{ASE}) + o_p(n^{-\frac{6k-1}{2(2k+1)}}), \quad (2.19) \end{aligned}$$

where $\triangle$ is between h_{ASE} and h_0 (recall that h_0 is the minimizer of $r_n(h)$ defined in (2.7)). On the other hand,

$$\begin{aligned} D(h) &= (n^{-1}e^\tau W^2 e - \sigma^2 n^{-1} tr(W^2)) - 2n^{-1}g^\tau(I-W)We \\ &= D_1(h) - D_2(h). \qquad (2.20) \end{aligned}$$

Let

$$L_1(u) = -uK'(u), \qquad L_2(u) = -uL_1'(u),$$
$$L_{nl}(t,t') = \frac{1}{nh}L_l\Big(\frac{t-t'}{h}\Big), \qquad W_l = (L_{nl}(t_i,t_j)), \quad l = 1,2.$$

Then we have

$$\begin{aligned} D_1'(h) &= -2h^{-1}(n^{-1}e^\tau((W-W_1)W)e - \sigma^2 n^{-1} tr((W-W_1)W)) \\ &= -2h^{-1}D_{11}(h), \\ D_1''(h) &= 2h^{-2}D_{11}(h) + 2h^{-2}\Big\{n^{-1}e^\tau[(W-2W_1+W_2)W + (W-W_1)^2]e \\ &\qquad -\sigma^2 n^{-1} tr[(W-2W_1+W_2)W + (W-W_1)^2]\Big\} \\ &= 2h^{-2}D_{11}(h) + 2h^{-2}D_{12}(h). \end{aligned}$$

Note that both $L_1(u)$ and $L_2(u)$ still satisfy condition (C1). So, applying Lemma 4.1(ii) to $D_{11}(h)$ with $\nu = k+1$ and to $D_{12}(h)$ with $\nu = k$, respectively,

$$hD_1'(h) = o_p^*(r(h)h^{(k+3)/(4k+5)}), \qquad h^2 D_1''(h) = o_p^*(r(h)h^{k/(4k+1)}).$$

Similarly,

$$hD_2'(h) = o_p^*\left(r(h)h^{(1-\varepsilon)/2}\right), \quad h^2 D_2''(h) = o_p^*\left(r(h)h^{(1-\varepsilon)/2}\right).$$

With these facts, (2.11), (2.19) and (2.20) imply that

$$h_{ASE} - h_0 = o_p\left(n^{-\frac{1}{2k+1}(2-\frac{3k+2}{4k+5})}\right),$$

$$\begin{aligned} D'(h_{ASE}) &= D'(h_0) + D''(\triangle)(h_{ASE} - h_0) \\ &= D'(h_0) + o_p\left(n^{-\frac{2k-\rho}{2k+1}}\right), \end{aligned}$$

where $\rho = \frac{3k+2}{4k+5} - \frac{k}{4k+1} < \frac{1}{2}$. Consequently,

$$c_{1k}^{-3}(2k+1)b\sigma^2 n^{-\frac{2k-2}{2k+1}}(h_{ASE} - h_0) + D'(h_0) = o_p\left(n^{-\frac{2k-\rho}{2k+1}}\right). \tag{2.21}$$

Now, the arguments leading to (2.19) and (2.21) can be easily modified to prove that for some $\triangle$ between $\hat{h}_G$ and h_0,

$$0 = G'(\hat{h}_G) = r_n''(\triangle)(\hat{h}_G - h_0) + D'(\hat{h}_G) + \delta'(\hat{h}_G) + O_p^*(n^{-\frac{3k-1}{2k+1}}),$$

$$c_{1k}^{-3}(2k+1)b\sigma^2 n^{-\frac{2k-2}{2k+1}}(\hat{h}_G - h_0) + D'(h_0) + \delta'(h_0) = o_p\left(n^{-\frac{2k-\rho}{2k+1}}\right).$$

The rest of proof is similar to Theorem 1 of Härdle, Hall and Marron (1988).

2.4. Simulation study

Here is a simulation study comparing the small sample behavior of the six bandwidth selectors introduced in Section 2.1. The simulation data are generated according to the following model:

$$y_i = x_i + m_2(t_i) + e_i, \qquad 1 \le i \le n, \tag{2.22}$$

where $x_i = m_1(t_i) + \eta_i$ and t_i's are equispaced on [.1, .9], with $e_i \sim N(0, .25)$ and $\eta_i \sim N(0, .01)$. The two regression functions are

$$m_1(x) = x^3(1-x)^3 \quad \text{and} \quad m_2(x) = x/(x^2+1). \tag{2.23}$$

Note that the true parameter is $\beta = 1$. The kernel K is taken to be the one used in Härdle, Hall and Marron (1988)

$$K(x) = \frac{15}{8}(1-4x^2)^2 I_{(|x|\le .5)}.$$

Table 1 compares the ratios of error criteria for these bandwidth selectors. Its entries show the number of times out of 100 that either $L_n(\hat{h}_G)/L_n(h_{ASE})-1$, or $R_n(\hat{h}_G)/R_n(h_{CMASE})-1$, exceeded the value of the column heading. Table 2 compares the ratios of bandwidths selected by these bandwidth selectors to the "optimal" bandwidth based on either the ASE or $CMASE$ criterion. Its entries show the number of times out of 100 that either $|\hat{h}_G/h_{ASE}-1|$, or $|\hat{h}_G/h_{CMASE}-1|$, exceeded the value of the column heading. The sample size is $n = 50$.

Table 1. Number of exceedances of the column headings (by ratios of error criteria) for various bandwidth selectors: 100 data sets of size 50 from the model (2.22) along with (2.23).

		0.05	0.1	0.15	0.2	0.25	0.3	0.5	0.7	0.9	1.1
GCV	ASE	37	20	15	12	10	9	3	1	1	0
	CMASE	25	10	3	2	2	0	0	0	0	0
AIC	ASE	37	21	14	13	10	10	3	1	1	0
	CMASE	25	12	4	2	2	0	0	0	0	0
FPE	ASE	37	21	14	13	10	10	3	1	1	0
	CMASE	25	12	4	2	2	0	0	0	0	0
S	ASE	37	23	16	13	10	10	4	1	1	0
	CMASE	28	15	7	4	3	2	0	0	0	0
T	ASE	35	22	15	11	10	9	3	1	1	0
	CMASE	24	9	3	2	0	0	0	0	0	0
CV	ASE	92	89	86	84	80	78	64	43	27	18
	CMASE	96	93	92	92	92	90	0	0	0	0

The tables reveal that GCV, AIC, FPE, Shibata's S and Rice's T perform nearly the same. As we know in the nonparametric regression, the CV method is subject to a great deal of sample variability, in the sense that for different data sets from the same distributions, it may give much different results (Marron (1989)). This drawback is also present in the simulation: the numbers of exceedances in both ASE and $CMASE$ rows for the CV method are dramatically larger than those for all other methods, indicating much greater variations of the bandwidths and the values of ASE and $CMASE$ among these 100 data sets. On the other hand, it seems that the $CMASE$ criterion is better estimated by other bandwidth selectors than CV. This is understandable. Since $CMASE$ is the average of ASE over all possible y values generated by given $\{x_i, t_i\}$ based on (1.1), there is extra variability in the ASE criterion due to the randomness of y. That $CMASE$ is worse than ASE for the CV method is probably due to the large sample variability of CV.

It should be mentioned that the CV method is not always worse. For example, we reran the simulation above but instead of using the functions at (2.23), we used the following:

$$m_1(x) = (x+2)^2 \quad \text{and} \quad m_2(x) = x(1-x)^4. \tag{2.24}$$

Table 2. Number of exceedances of the column headings (by the distance between the ratios of the data-driven bandwidths to the optimal bandwidths and 1) for various bandwidth selectors: 100 data sets of size 50 from the model (2.22) along with (2.23).

		0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1
GCV	ASE	31	22	17	12	10	10	9	6	5	5
	CMASE	34	15	5	2	1	0	0	0	0	0
AIC	ASE	33	22	17	12	10	10	9	6	5	5
	CMASE	35	15	6	2	1	0	0	0	0	0
FPE	ASE	33	22	17	12	10	10	9	6	5	5
	CMASE	35	15	6	2	1	0	0	0	0	0
S	ASE	34	22	16	12	9	9	8	5	5	5
	CMASE	39	19	8	4	3	1	0	0	0	0
T	ASE	31	22	17	11	10	10	9	6	5	5
	CMASE	33	14	5	2	0	0	0	0	0	0
CV	ASE	92	91	90	87	81	67	13	0	0	0
	CMASE	96	95	92	92	92	89	69	0	0	0

Table 3. Number of exceedances of the column headings (by ratios of error criteria) for various bandwidth selectors: 100 data sets of size 50 from the model (2.22) along with (2.24).

		0.05	0.1	0.15	0.2	0.25	0.3	0.5	0.7	0.9	1.1
GCV	ASE	68	61	57	51	43	37	26	14	11	11
	CMASE	59	47	21	0	0	0	0	0	0	0
AIC	ASE	68	61	56	52	43	38	26	14	10	10
	CMASE	61	49	23	0	0	0	0	0	0	0
FPE	ASE	68	62	57	53	44	38	26	14	11	11
	CMASE	60	48	21	0	0	0	0	0	0	0
S	ASE	71	63	56	52	47	42	25	14	12	11
	CMASE	66	52	27	0	0	0	0	0	0	0
T	ASE	65	59	54	50	40	36	25	14	11	11
	CMASE	55	42	18	0	0	0	0	0	0	0
CV	ASE	70	59	53	51	44	41	27	15	13	13
	CMASE	58	50	25	0	0	0	0	0	0	0

The results are shown in Tables 3.4 and 3.5, set up the same as Tables 3.2 and 3.3, respectively. We can see that the performance of the CV method is now nearly the same as that of all the others. These simulations indicate that the behavior of the CV method is likely to be more sensitive to the model specification (the forms of the regression functions $m_1(x)$ and $m_2(x)$, in particular), and hence less stable than the other methods. Therefore CV should be used with caution. In light of this, we recommand using the GCV method to choose the bandwidth in this semiparametric setting.

Table 4. Number of exceedances of the column headings (by the distance between the ratios of the data-driven bandwidths to the optimal bandwidths and 1) for various bandwidth selectors: 100 data sets of size 50 from the model (2.22) along with (2.24).

		0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1
GCV	ASE	62	50	37	30	25	22	22	20	15	13
	CMASE	54	50	33	21	0	0	0	0	0	0
AIC	ASE	63	49	38	27	23	21	21	18	14	11
	CMASE	60	53	37	26	0	0	0	0	0	0
FPE	ASE	64	50	38	28	24	22	22	19	15	12
	CMASE	59	52	36	25	0	0	0	0	0	0
S	ASE	66	49	37	29	22	21	20	17	13	11
	CMASE	63	55	39	29	0	0	0	0	0	0
T	ASE	61	49	38	32	27	24	22	20	15	13
	CMASE	52	42	29	16	0	0	0	0	0	0
CV	ASE	64	50	39	28	23	22	21	19	16	14
	CMASE	55	51	38	28	0	0	0	0	0	0

3. Data-Driven Estimator $\hat\beta_{\hat h_G}$

3.1. $\sqrt{n}$-Consistency

For a certain deterministic bandwidth choice h, Speckman (1988) showed that the kernel smoothing estimator $\hat\beta_h$ can attain the usual parametric rate $O(n^{-1/2})$. In fact, he proved that

$$\sqrt{n}(\hat\beta_h - \beta) \xrightarrow{\mathrm{D}} N(0,\ \sigma^2\Sigma^{-1}),$$

where Σ is the covariance matrix of $\eta_1 = (\eta_{11}, \ldots, \eta_{1p})^\tau$ in (2.6). Then a natural question arises: is the $O(n^{-1/2})$ rate still attainable for the data-driven bandwidth choice $\hat h_G$, i.e., is $\hat\beta_{\hat h_G}$ $\sqrt{n}$-consistent? The following theorem shows that the same asymptotic normality holds for $\hat\beta_{\hat h_G}$.

Theorem 3.1. *Under the conditions of Theorem* 2.1 *we have*

$$\sqrt{n}(\hat\beta_{\hat h_G} - \beta) \xrightarrow{\mathrm{D}} N(0,\ \sigma^2\Sigma^{-1}). \tag{3.1}$$

Proof. We have the decomposition

$$\sqrt{n}(\hat\beta_{\hat h_G} - \beta) = \frac{1}{\sqrt{n}}\Sigma_{n2}^{-1}(\hat h_G)\eta^\tau(I - W(\hat h_G))^2 g$$

$$
\begin{aligned}
&+\frac{1}{\sqrt{n}}\Sigma_{n2}^{-1}(\hat{h}_G)G^\tau(I-W(\hat{h}_G))^2 g \\
&+\frac{1}{\sqrt{n}}(\Sigma_{n2}^{-1}(\hat{h}_G)-\Sigma^{-1})\eta^\tau(I-W(\hat{h}_G))^2 e \\
&+\frac{1}{\sqrt{n}}(\Sigma_{n2}^{-1}(\hat{h}_G)-\Sigma^{-1})G^\tau(I-W(\hat{h}_G))^2 e \\
&-\frac{1}{\sqrt{n}}\eta^\tau(2W(\hat{h}_G)-W^2(\hat{h}_G))e+\frac{1}{\sqrt{n}}\Sigma^{-1}\eta^\tau e \\
=&\sum_{j=1}^{5} B_j(\hat{h}_G)+\frac{1}{\sqrt{n}}\Sigma^{-1}\eta^\tau e.
\end{aligned}
$$

By Lemmas 4.2-4.4 we have, uniformly over $h \in \Lambda_\varepsilon$,

$$B_j(h) = o_p(n^{1/2}r(h)) = o_p(1), \quad j = 1, 3, 4, 5,$$

$$B_2(h) = O_p(n^{1/2}h^{2k}) = o_p(1).$$

Thus, since $P\{\hat{h}_G \in \Lambda_\varepsilon\} \longrightarrow 1$ by Theorem 2.1,

$$B_j(\hat{h}_G) = o_p(1), \quad 1 \le j \le 5.$$

The convergence in (3.1) then follows from the classical CLT.

3.2. An application to diabetes data

The data come from a study (Sockett, Daneman, Clarson and Ehrich (1987)) of factors affecting patterns of insulin-dependent diabetes mellitus in children. The objective was to investigate the dependence of the level of serum C-peptide on various other factors in order to understand the patterns of residual insulin secretion. The response variable is C-peptide concentration (pmol/ml) at diagnosis, and the predictors are age and base deficit, a measure of acidity. These two predictors are a subset of those used in the original study. The data scatterplot is shown in Figure 1. Two observations in the original data set are excluded as outliers because of their unusually large absolute values of *base deficit*. Note that while the plot of the response variable, *C-peptide*, versus one predictor, *base deficit*, shows a roughly linear relationship between them, we see a nonlinear pattern in the plot of *C-peptide* versus *age*. Thus a semiparametric regression model with *base deficit* as its linear component and *age* as its nonparametric component is fitted using the kernel smoothing method with

$$K_{nh}(t, t_i) = K(\frac{t-t_i}{h}) / \sum_{j=1}^{n} K(\frac{t-t_j}{h}),$$

where the kernel $K(t) = 15(1-t^2)I_{(|x|\le 1)}/16$. The GCV-selected bandwidth is 4.9033 and the estimated linear coefficient of *base deficit* is 0.0549. Figure 2 shows the fitted regression surface. We can see that the general trend in each variable revealed in the scatterplot Figure 1 is well represented.

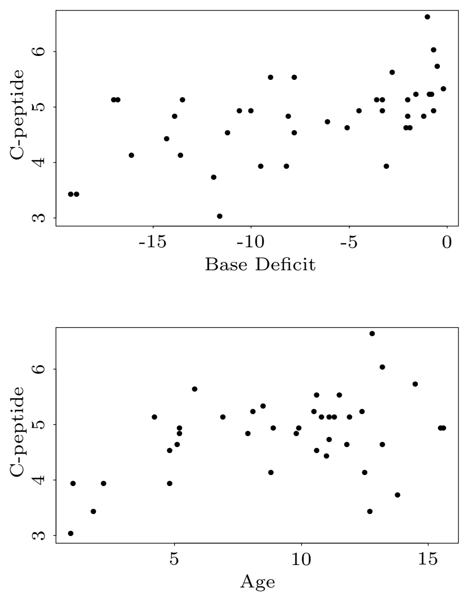

Figure 1. Scatterplot of diabetes data.

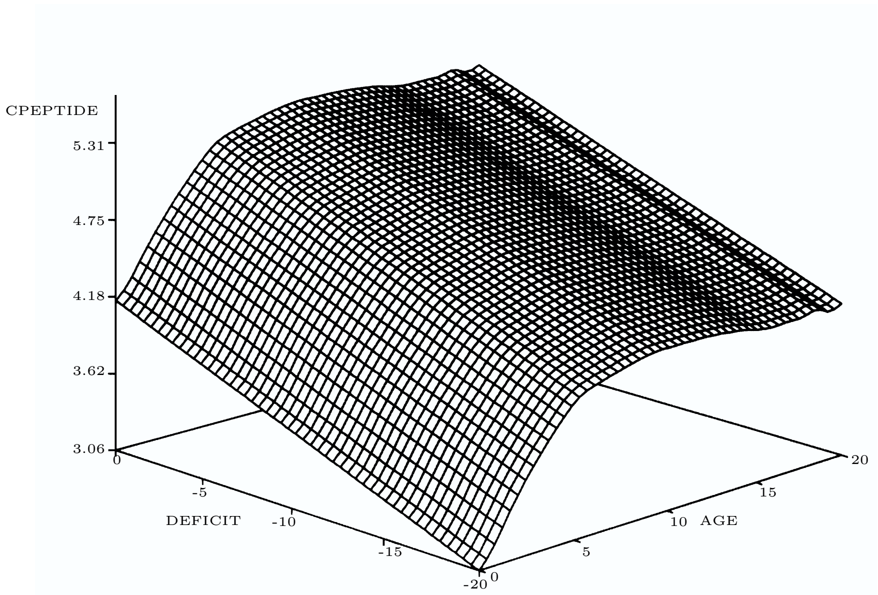

Figure 2. The Fitted Regression Surface for the diabetes data.

4. Technical Lemmas

In this section we give some lemmas used in the proofs of Theorems 2.1 and 2.2. In the sequel C denotes a generic constant which may differ at each appearance.

Lemma 4.1. *Suppose that $A(h) = (A_{ij}(h))$ is an $n \times n$ matrix, satisfying*
(a) $|A_{ij}(h)| \le C(nh)^{-1}$ *for each* $1 \le i, j \le n$,
(b) $A_{ij}(h) = 0$ *if* $|i - j| \ge Cnh$,
(c) *For some* $\nu \ge 1$, *each* $A_{ij}(h)$ *is* ν *times continuously differentiable, and* $B_{ij}^{(l)}(h) \triangleq h^l A_{ij}^{(l)}(h)$, *all satisfy (a) and (b),* $1 \le l \le \nu$.
Suppose $\{u_i\}$ *and* $\{v_i\}$ *are two independent sequences of i.i.d. variables with mean zero and finite 4th moments, with* $u = (u_1, \ldots, u_n)^\tau$ *and* $v = (v_1, \ldots, v_n)^\tau$. *Then the following results hold uniformly over* $h \in \Lambda_n$.

(i) *If* $f(t)$ *is a bounded function on* $[0,\ 1]$ *satisfying*

$$|f^\tau(I - A(h))| \le Ch^k, \qquad \text{for any } h \in \Lambda_n, \tag{4.1}$$

then for any small $\varepsilon > 0$

$$n^{-1} f^\tau (I - A(h))u = o_p(r(h)h^{(1-\varepsilon)/2}).$$

(ii) *If* $\nu \ge 2k/3$, *then for* $\alpha_1 = \frac{3\nu - 2k}{4\nu + 1}$, $\alpha_2 = \frac{\nu}{2\nu + 1}$ *and* $\alpha = min(\alpha_1, \alpha_2)$,

$$n^{-1} \sum_{1 \le j \ne s \le n} A_{js}(h) u_j u_s = o_p(r(h)h^{\alpha_1}), \tag{4.2}$$

$$n^{-1} \sum_{j=1}^{n} A_{jj}(h) \left(u_j^2 - E(u_1^2)\right) = o_p(r(h)h^{\alpha_2}),$$

$$n^{-1} u^\tau A(h) u - n^{-1} E(u_1^2) tr(A(h)) = o_p(r(h)h^{\alpha}),$$

$$n^{-1} u^\tau A(h) v = o_p(r(h)h^{\alpha}).$$

Proof. We only prove (4.2) here. The proofs of others are similar in spirit. Let

$$h_i = (1 + b_n)^i (n\delta_n)^{-1}, \qquad 0 \le i \le i_n = \log(n\delta_n^2)/\log(1 + b_n),$$

where $b_n = \varepsilon^2 n^{-\frac{2k + \alpha_1}{\nu(2k+1)}}$ for an arbitrarily small $\varepsilon > 0$. We have

$$\sup_{h \in \Lambda_n} \left| (nr(h)h^{\alpha_1})^{-1} \sum_{1 \le j \ne s \le n} A_{js}(h) u_j u_s \right|$$

$$\le 2 \max_{0 \le i \le i_n} \left| (nr(h_i)h_i^{\alpha_1})^{-1} \sum_{1 \le j \ne s \le n} A_{js}(h_i) u_j u_s \right|$$

$$
\begin{aligned}
&+2\max_i \sup_{h_i\le h\le h_{i+1}} \left|(nr(h_i)h_i^{\alpha_1})^{-1}\sum_{1\le j\ne s\le n}(A_{js}(h)-A_{js}(h_i))u_ju_s\right| \\
&= T_1+T_2. \qquad (4.3)
\end{aligned}
$$

By a Taylor expansion,

$$
\begin{aligned}
T_2 &\le C\max_i \sum_{l=1}^{\nu-1} b_n^l \left|(nr(h_i)h_i^{\alpha_1})^{-1}\sum_{1\le j\ne s\le n} B_{js}^{(l)}(h_i)u_ju_s\right| \\
&\quad + Cb_n^{\nu}\max_i \sup_{h_i\le h\le h_{i+1}} \left|(nr(h_i)h_i^{\alpha_1})^{-1}\sum_{1\le j\ne s\le n} B_{js}^{(\nu)}(\triangle_{js})u_ju_s\right| \\
&= T_{21}+T_{22}, \qquad (4.4)
\end{aligned}
$$

where $\triangle_{js}$ is between h and h_i. From conditions (a)-(c),

$$
\begin{aligned}
T_{22} &\le Cb_n^{\nu}\max_i\left((n^2r(h_i)h_i^{1+\alpha_1})^{-1}\sum_{0<|j-s|<Cnh_i}(|u_ju_s|-E|u_ju_s|)\right) \\
&\quad + Cb_n^{\nu}\left(\inf_{h>0}(r(h)h^{\alpha_1})\right)^{-1}.
\end{aligned}
$$

Obviously the second term of the right hand side above tends to zero as $n\to\infty$. The first term is $o_p(1)$ by the Cauchy inequality. To handle T_1, let

$$
\Lambda_m = \begin{cases} [mCnh_i+1,\ (m+1)Cnh_i], & \text{if } 0\le m\le m_i=(Ch_i)^{-1}-1, \\ \emptyset, & \text{if } m<0 \text{ or } m>m_i, \end{cases}
$$

and $\Lambda_{mj}=\Lambda_m-\{j\}$. Then we have

$$
\begin{aligned}
T_1 &= 2\max_i\left|(nr(h_i)h_i^{\alpha_1})^{-1}\sum_{m=0}^{m_i}\sum_{j\in\Lambda_m}\sum_{s\in\Lambda_{m-1}\cup\Lambda_{mj}\cup\Lambda_{m+1}} A_{js}(h_i)u_ju_s\right| \\
&\le 2\max_i\left|(nr(h_i)h_i^{\alpha_1})^{-1}\sum_{m=0}^{m_i}\sum_{j\in\Lambda_m}\sum_{s\in\Lambda_{mj}} A_{js}(h_i)u_ju_s\right| \\
&\quad +2\max_i\left|(nr(h_i)h_i^{\alpha_1})^{-1}\sum_{m=1}^{m_i}\sum_{j\in\Lambda_m}\sum_{s\in\Lambda_{m-1}} A_{js}(h_i)u_ju_s\right| \\
&\quad +2\max_i\left|(nr(h_i)h_i^{\alpha_1})^{-1}\sum_{m=0}^{m_i-1}\sum_{j\in\Lambda_m}\sum_{s\in\Lambda_{m+1}} A_{js}(h_i)u_ju_s\right| \\
&= T_{11}+T_{12}+T_{13}. \qquad (4.5)
\end{aligned}
$$

Write

$$z_{mi} = \sum_{j\in\Lambda_m}\sum_{s\in\Lambda_{mj}} A_{js}(h_i)u_j u_s,$$

$$z'_{mi} = z_{mi}I(|z_{mi}| \le h_i^{-\gamma}) - E\left(z_{mi}I(|z_{mi}| \le h_i^{-\gamma})\right),$$

and $z''_{mi} = z_{mi} - z'_{mi}$, where $\gamma \in (1/2,\ 1-\alpha_1)$. It follows from condition (a) and $|\Lambda_m| \le Cnh_i$ that for each i and m, $E(z_{mi})^4 \le C$. Hence it is easy to see that

$$\begin{aligned} E(\sum_m z''_{mi})^4 &\le C\sum_m E(z''_{mi})^4 + C(\sum_m E(z''_{mi})^2)^2 \\ &\le C\gamma_n h_i^{-1} + Ch_i^{4\gamma-2}, \end{aligned}$$

where $\gamma_n = o(1)$. Consequently

$$\begin{aligned} &P\left\{\max_i \left|(nr(h_i)h_i^{\alpha_1})^{-1}\sum_m z''_{mi}\right| \ge \varepsilon\right\} \\ &\le C\sum_i (nr(h_i)h_i^{\alpha_1})^{-4}E(\sum_m z''_{mi})^4 \longrightarrow 0. \end{aligned} \tag{4.6}$$

On the other hand, since by condition (b)

$$\sum_m E(z'_{mi})^2 \le \sum_m \sum_{j\in\Lambda_m}\sum_{s\in\Lambda_{mj}} A^2_{js}(h_i)E(u_j)^2E(u_s)^2 \le Ch_i^{-1},$$

Bernstein's Inequality gives

$$\begin{aligned} &P\left\{\max_i \left|(nr(h_i)h_i^{\alpha_1})^{-1}\sum_m z'_{mi}\right| \ge \varepsilon\right\} \\ &\le 2\sum_i \exp\left\{-C\,(nr(h_i)h_i^{\alpha_1})^2 \Big/ \left(\sum_m E(z'_{mi})^2 + C(nr(h_i)h_i^{\alpha_1})h_i^{-\gamma}\right)\right\} \\ &\le 2i_n\left(\exp\{-Cn^{-(1-2\alpha_1)/(2k+1)}\} + \exp\{-Cn^{-(1-\gamma-\alpha_1)/(2k+1)}\}\right) \le Cn^{-2}, \end{aligned}$$

Putting this together with (4.6) gives $T_{11} = o_p(1)$. As for T_{12}, we have

$$\begin{aligned} T_{12} \le 2\max_i &\left|(nr(h_i)h_i^{\alpha_1})^{-1}\sum_{m:\,even} u_{mi}\right| \\ &+2\max_i \left|(nr(h_i)h_i^{\alpha_1})^{-1}\sum_{m:\,odd} u_{mi}\right| = T'_{12} + T''_{12}, \end{aligned}$$

where

$$u_{mi} = \sum_{j\in\Lambda_m}\sum_{s\in\Lambda_{m-1}} A_{js}(h_i)u_j u_s.$$

Now, similar to T_{11}, both T'_{12} and T''_{12} are $o_p(1)$. Hence $T_{12} = o_p(1)$. Similarly $T_{13} = o_p(1)$. Thus by (4.5) $T_1 = o_p(1)$. The same argument leads to $T_{21} = o_p(1)$, thus (4.2) follows.

Let

$$g = (g(t_1), \ldots, g(t_n))^\tau, \quad e = (e_1, \ldots, e_n)^\tau, \quad \eta = (\eta_1, \ldots, \eta_n)^\tau,$$

$$G_i = (g_i(t_1), \ldots, g_i(t_n))^\tau, \quad G = (G_1, \ldots, G_p).$$

Lemma 4.2. *Suppose that condition $(C1)$ holds and that K is ν times continuously differentiable. Then for each $s \geq 1$, $W^s(h)$ satisfies conditions $(a) - (c)$ of Lemma* 4.1. *Also, the condition* (4.1) *in Lemma* 4.1*(i) holds for $f = g$ or G and $A(h) = W^s(h)$.*

Proof. Let $w_{ij,s}(h)$ denote the $(i,j)th$ element of $W^s(h)$. Obviously $w_{ij,1}(h)$ satisfies (a)-(c). Since for $s \geq 2$

$$w_{ij,s}(h) = \sum_{l=1}^{n} w_{il,1}(h) w_{lj,s-1}(h),$$

the first result follows from induction. The second is a standard result in nonparametric kernel regression.

The following three lemmas follow from Lemmas 4.1 and 4.2.

Lemma 4.3. *Suppose that conditions* (C1)-(C3) *hold and that K is ν times continuously differentiable. Then we have, uniformly over $h \in \Lambda_n$, that for each $s \geq 1$,*

(i) *For any small $\varepsilon > 0$, $f = g$ or G and $u = e$ or η,*

$$n^{-1} f^\tau (I - W(h))^s u = o_p(r(h) h^{(1-\varepsilon)/2}).$$

(ii) *If $\nu \geq 2k/3$, then*

$$n^{-1} u^\tau W^s(h) u - n^{-1} \mathrm{Var}\,(u_1) tr(W^s(h)) = o_p(r(h) h^\alpha),$$

$$n^{-1} \eta^\tau W^s(h) e = o_p(r(h) h^\alpha),$$

where $u = e$ or η and $\alpha = \min(\alpha_1,\ \alpha_2)$ is as in Lemma 4.1*(ii).*

For $s \geq 1$, let

$$\Sigma_{ns}(h) = n^{-1} X^\tau (I - W(h))^s X.$$

Lemma 4.4. *Under the assumptions of Theorem* 2.1 *we have, uniformly over $h \in \Lambda_n$, that for each $s \geq 2$,*

$$\Sigma_{ns}(h) - \Sigma = O_p(n^{-1/2} + r(h)), \tag{4.7}$$

$$\Sigma_{ns}^{-1}(h) - \Sigma^{-1} = O_p(n^{-1/2} + r(h)). \tag{4.8}$$

Lemma 4.5. *Under the assumptions of Theorem* 2.1 *we have, uniformly over* $h \in \Lambda_n$*, that for* $s \geq 0$*,*

$$\begin{aligned} n^{-1}tr(S(h)) &= K(0)/(nh) + p/n + o_p(n^{-1/2}r(h)), \\ n^{-1}tr(S^\tau(h)S(h)) &= n^{-1}tr(W^2(h)) + p/n + o_p(n^{-1/2}r(h)) \\ &= (nh)^{-1}\int K^2(t)dt + p/n + o_p(r(h)). \end{aligned}$$

$$n^{-1}g^\tau(I - W(h))P_{\tilde{X}}(I - W(h))g = O_p(r^2(h)),$$

$$n^{-1}g^\tau(I - W(h))P_{\tilde{X}}(I - W(h))^s e = o_p(r^2(h)) + O_p(n^{-1/2}r(h)),$$

$$n^{-1}e^\tau(I - W(h))P_{\tilde{X}}(I - W(h))^s e = n^{-2}e^\tau\eta\Sigma^{-1}\eta^\tau e + o_p(r^2(h) + n^{-1/2}r(h)).$$

Acknowledgement

I am very grateful to the Associate Editor and the referees' for the helpful comments and suggestions that have dramatically improved the paper. This paper is based on a part of my Ph.D. dissertation. I would like to thank my advisor, Professor Thomas A. Severini, for his guidance and advice. Thanks are also due to Professor Bruce D. Spencer for his encouragement and support. The simulation study was conducted on the computers supported by NSF under the grant DMS-9505799.

References

Akaike, H. (1970). Statistical predictor information. *Ann. Inst. Statist. Math.* **22**, 203-217.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Auto. Control* **19**, 716-723.

Chen, H. and Shiau, J. H. (1994). Data-driven efficient estimator for a partially linear model. *Ann. Statist.* **22**, 211-237.

Clark, R. M. (1975). A calibration curve for radio cabon dates. *Antiguity* **49**, 251-266.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377-403.

Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81**, 310-320.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *JRSS Ser. B* **57**, 371-394.

Gasser, T., Kneip, A. and Köhler, W. (1991). A fast and flexible method for automatic smoothing. *JASA* **86**, 643-652.

Hall, P., Sheather. S. J., Jones, M. C. and Marron, S. J. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78**, 263-269.

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

Härdle, W., Hall, P. and Marron, J. S. (1992). Regression smoothing parameters that are not far from their optimum. *JASA* **87**, 227-233.

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

Hong, S. Y. and Cheng, P. (1992). Berry-Esseen rate for the estimator of parametric components in a partial linear model. *Technical Report, Institute of System Sciences, Academia Sinica.*

Hong, S. Y. and Cheng, P. (1994). Convergence rates of estimators of parameters in a semi-parametric regression model. *Chinese J. Appl. Prob. Statist.* **10**, 62-71.

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996a). A brief survey of bandwidth selection for density estimation. *JASA* **91**, 401-407.

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996b). Progress in data-based bandwidth selection for kernel density selection. *Computational Statistics* **11**, 337-381.

Marron, J. S. (1989). Automatic smoothing parameter selection: a survey. *Empirical Econ.* **13**, 187-208.

Priestley, M. B. and Chao, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34**, 385-392.

Rice, J. (1986). Convergence rates for partially spline models. *Statist. Probab. Lett.* **4**, 203-208.

Ruppert, D., Sheather, S. J. and Wand, M. P (1995). An effective bandwidth selector for local least squares regression. *JASA* **90**, 1257-1270.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *JRSS Ser. B* **53**, 683-690.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

Sockett, E. B., Daneman, D., Clarson, C. and Ehrich, R. M. (1987). Factors affecting and patterns of residual insulin secretion during the first year of type I (insulin dependent) diabetes mellitus in children. *Diabet.* **30**, 453-459.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50**, 413-456.

Wahba, G. (1984). Cross validated spline methods for the estimation of multivariate functions from data on functionals. In *Statistics: An Appraisal, Proc. 50th Anniversary Conf. Iowa State Statistical Laboratory* (Edited by H. A. David and H. T. David), 205-235. Iowa State University Press, Ames.

Department of Statistics, Northwestern University, Evanston, IL 60208.

E-mail: syhong@nwu.edu

(Received May 1997; accepted July 1998)