ELSEVIER

Statistics & Probability Letters 43 (1999) 309–316

STATISTICS & PROBABILITY LETTERS

Rates of convergence for the pre-asymptotic substitution bandwidth selector

Jianqing Fan[a,1], Li-Shan Huang[b,*]

[a]Department of Statistics, Chinese University of Hong Kong, Shatin, Hong Kong
[b]Department of Statistics, Florida State University, Tallahassee, FL 32306-3033, USA

Received March 1997

Abstract

An effective bandwidth selection method for local linear regression is proposed in Fan and Gijbels [1995, J. Roy. Statist. Soc. Ser. B, 57, 371–394]. The method is based on the idea of the pre-asymptotic substitution and has been tested extensively. This paper investigates the rate of convergence of this method. In particular, we show that the relative rate of convergence is of order $n^{-2/7}$ if the locally cubic fitting is used in the pilot stage, and the rate of convergence is $n^{-2/5}$ when the local polynomial of degree 5 is used in the pilot fitting. The study also reveals a marked difference between the bandwidth selection for nonparametric regression and that for density estimation: The plug-in approach for the latter case can admit the root-n rate of convergence while for the former case the best rate is of order $n^{-2/5}$.

Keywords: Bandwidth selection; Convergence rate; Local polynomial regression; Kernel density estimation

1. Introduction

Local polynomial regression is a curve estimation method by fitting locally weighted polynomial regression. Recent work on local polynomial regression includes Fan (1993), Hastie and Loader (1993), Ruppert and Wand (1994), and the monographs by Wand and Jones (1995), Simonoff (1996), Fan and Gijbels (1996) and the references therein. One critical step for using local regression methods is the choice of the smoothing parameter h, or "bandwidth", which controls the degree of smoothing. Different values of the bandwidth result in different estimated curves. One possible method for selecting an appropriate bandwidth is to plot out several estimated curves with different bandwidths and choose one estimate subjectively. This subjective method has two serious drawbacks: It is hard to process the data beyond the domain of our visualization and the method

* Corresponding author.

[1] On leave from the University of North Carolina at Chapel Hill and supported by the NSF grant DMS9504414 and the NSA grant 96-1-0015.

PII: S0167-7152(98)00271-5

cannot be automated. Thus, it is desirable to provide a data-driven bandwidth that suits practical needs and has a good convergence rate.

The problem of bandwidth selection has stimulated much research especially in kernel density estimation. The approaches include cross-validation and "plug-in" methods; see the review paper by Jones et al. (1996). In particular, Hall et al. (1991) propose a $n^{-1/2}$-consistent bandwidth selector for kernel density estimator where n is the sample size. Based on a similar idea, Chiu (1991a,b) constructs $n^{-1/2}$-rate bandwidth selectors for the kernel-type estimators of density and regression functions. Ruppert et al. (1995) give a bandwidth selection method for local linear regression based on the "plug-in" idea. Hart and Yi (1996) focus on a non-random design model and study a cross-validation type bandwidth selector with rate $n^{-3/10}$, while Schucany (1995) study local bandwidth selection for Priestly-Chao kernel estimators (Priestly and Chao, 1972) with possible extension to local linear regression. Opsomer (1995) proposes a "plug-in" bandwidth selection method for bivariate additive models with local linear fitting. Fan and Gijbels (1995) study the bandwidth selection problem for local polynomial fitting and propose an idea of pre-asymptotic assessment of the bias and variance expressions. This approach saves the effort of estimating unknown terms related to the design density in the asymptotic expansion of the optimal bandwidth. While this idea has been empirically demonstrated to be powerful, no formal theory has yet been established.

The aim of this paper is to investigate the rate of convergence for the pre-asymptotic substitution method in local linear regression. We first expand in Theorem 1 the asymptotic bias and variance into the second order. This expansion is a necessary technical device for establishing rate of convergence of the implicitly defined pre-asymptotic substitution bandwidth selector, and enables us to understand why the best relative rate of convergence for the plug-in bandwidth selector is at most of order $n^{-2/5}$. This marks a salient contrast with the bandwidth selection problem in the density estimation setting, where it is shown that the root-n bandwidth selector can be constructed via a plug-in method (Hall et al., 1991). Our study shows that the pre-asymptotic substitution method has convergence rate $n^{-2/7}$ if the local polynomial of degree 3 is used in the pilot fitting and has rate at least $n^{-2/5}$ if the local polynomial of degree 5 is used in the pilot fitting. One technical challenge here is that the bandwidth selector is defined implicitly.

The paper is organized as follows. Section 2 sets up some necessary notation on the local polynomial fitting. The asymptotic expansion of the optimal bandwidth for local linear regression is given in Section 3. In Section 4 we study the rate of convergence for the pre-asymptotic substitution method. The technical proofs are collected in Section 5.

2. Local polynomial fitting

Let (X_i, Y_i), $i = 1, \dots, n$, denote independent data generated from a random design model:

$$Y = m(X) + \sigma(X)\varepsilon, \qquad E(\varepsilon) = 0,\ \mathrm{var}(\varepsilon) = 1, \tag{2.1}$$

where $m(\cdot)$ is an unknown regression function, ε is an error variable, and X is independent of ε. The location-scale model (2.1) is postulated for the ease of comprehension. It is not critical to our development. Let $K(\cdot)$ be a kernel function, usually a symmetric probability density function, and h be the bandwidth. Then, a local polynomial fit of order p at a point x is to find the solution of $\beta(x) = (\beta_0(x), \dots, \beta_p(x))^{\mathrm{T}}$ to the following weighted least-squares problem:

$$\min_{\beta} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j \right)^2 K\left(\frac{X_i - x}{h} \right). \tag{2.2}$$

The dependence of β on x is suppressed for simplicity. It is easy to see that (2.2) can be written as

$$\min_{\beta} (\boldsymbol{Y} - \boldsymbol{X}\beta)^{\mathrm{T}} W (\boldsymbol{Y} - \boldsymbol{X}\beta),$$

where

$$\boldsymbol{X} = \begin{pmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^p \end{pmatrix}, \quad W = \mathrm{diag}\left(K\left(\frac{X_i - x}{h}\right)\right)_{n\times n} \tag{2.3}$$

and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. Hence the solution $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^{\mathrm{T}}$ to Eq. (2.2) can be expressed as

$$\hat{\beta} = (\boldsymbol{X}^{\mathrm{T}} W \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathrm{T}} W \boldsymbol{Y}. \tag{2.4}$$

Note that $\hat{\beta}_v$ estimates the parameter $\beta_v(x) = m^{(v)}(x)/v!$. When $p = 1$, the case of local linear regression, the solution $\hat{\beta}_0(x)$, denoted by $\hat{m}_l(x)$, is the local linear regression estimator for $m(x)$, and can be expressed explicitly as

$$\hat{m}_l(x) = \frac{\sum_{i=1}^n \{S_{n,2} - (X_i - x)S_{n,1}\} K((X_i - x)/h) Y_i}{\sum_{i=1}^n \{S_{n,2} - (X_i - x)S_{n,1}\} K((X_i - x)/h)},$$

where $S_{n,j} = \sum_{i=1}^n (X_i - x)^j K((X_i - x)/h)$, $j = 0, 1, \ldots$. Again, the dependence of $S_{n,j}$ on x is suppressed.

3. Asymptotic expansion

Consider fitting local linear regression on an interval $[a, b]$. A criterion that measures the discrepancy between the estimated curve $\hat{m}_l(x)$ and the true regression function is the conditional weighted mean integrated square error (MISE) given by

$$M(h) = E\left\{ \int (\hat{m}_l(x) - m(x))^2 w(x)\,\mathrm{d}x | X_1, \ldots, X_n \right\}, \tag{3.1}$$

where $w(\cdot)$ is a nonnegative weight function with support $[a, b]$. Define the optimal bandwidth h_{OPT} to be the minimizer of the $M(h)$. The first-order expansion of h_{OPT} is well known; see for example Fan and Gijbels (1995), and Ruppert et al. (1995) with $w(x) = f(x)$, where $f(\cdot)$ is the marginal density of the independent variable X. The goal of this section is to derive the second order expansion, which will serve as a technical device for studying the pre-asymptotic bandwidth selector defined implicitly.

We require the following technical assumptions.

(A1) The regression function $m(\cdot)$ admits a continuous fourth derivative on $[a, b]$.

(A2) The design density $f(\cdot)$ has a second continuous derivative and is bounded away from 0 on $[a, b]$.

(A3) The bandwidth h lies in an interval $(\alpha_1 n^{-t_1}, \alpha_2 n^{-t_2})$ such that $(\gamma_1 n^{-1/5}, \gamma_2 n^{-1/5}) \subset (\alpha_1 n^{-t_1}, \alpha_2 n^{-t_2})$, for some positive constants α_1, α_2, γ_1, γ_2, t_1 and t_2.

(A4) The kernel function $K(\cdot)$ is a bounded symmetric probability density defined on a compact interval. Further, $K(\cdot)$ has a bounded second derivative almost everywhere.

(A5) The variance function $\sigma^2(\cdot)$ has a bounded second derivative on $[a, b]$.

Theorem 1. *Under Conditions* (A1)–(A5), *the asymptotic bias and variance of* $\hat{m}_l(x)$, $x \in [a, b]$, *are*

$$\begin{aligned} E(\hat{m}_l(x)|X_1, \ldots, X_n) - m(x) = {} & h^2 \mu_2 \beta_2(x) + h^4\{(\mu_2^2 - \mu_4)\beta_2(x)(f'^2(x)/f^2(x) \\ & - f''(x)/(2f(x))) + \beta_4(x)\mu_4\} + \mathrm{o}(h^4) + \mathrm{O}_{\mathrm{P}}(n^{-1/2}h^{3/2}) \end{aligned} \tag{3.2}$$

and

$$\begin{aligned}\operatorname{var}(\hat{m}_l(x)|X_1,\dots,X_n) = {} & n^{-1}h^{-1}\sigma^2(x)v_0/f(x) + n^{-1}h\{(2v_0\mu_2+v_2)\sigma^2(x)f'^2(x)/f^3(x)\\ & -v_0\mu_2\sigma^2(x)f''(x)/f^2(x) + v_2g''(x)/(2f^2(x))\\ & -2v_2f'(x)g'(x)/f^3(x)\} + \mathrm{o}(n^{-1}h) + \mathrm{O_P}(n^{-3/2}h^{-3/2}),\end{aligned} \tag{3.3}$$

where $g(x) = f(x)\sigma^2(x)$, $\mu_j = \int u^jK(u)\,\mathrm{d}u$, *and* $v_j = \int u^jK^2(u)\,\mathrm{d}u$, $j = 0, 1, \dots$

The proof of this theorem is given in Section 5. Expressions (3.2) and (3.3) hold uniformly for h in the interval specified by Condition (A3) if the $\mathrm{O_P}$ terms are replaced by $\mathrm{O_P}(n^{-1/2}h^{3/2}\log n)$ and $\mathrm{O_P}(n^{-3/2}h^{-3/2}\log n)$ respectively, since $S_{n,j}$ converges uniformly in h with an inflation factor of $\log n$ in the $\mathrm{O_p}$-term (see (5.2)).

Writing $\theta_2 = \int \beta_2^2(x)w(x)\,\mathrm{d}x$, we show in Theorem 2 that the optimal bandwidth

$$h_{\mathrm{OPT}} = a_1n^{-1/5} + \mathrm{O_P}(n^{-3/5}) \tag{3.4}$$

with

$$a_1 = \left(v_0\int\sigma^2(x)f^{-1}(x)w(x)\,\mathrm{d}x/4\mu_2^2\theta_2\right)^{1/5}.$$

We have attempted to expand h_{OPT} further into $h_{\mathrm{OPT}} = a_1n^{-1/5} + a_2n^{-3/5} + \mathrm{O_P}(n^{-4/5})$ so that the relative rate between h_{OPT} and its first two leading terms in the asymptotic expansion is of order $\mathrm{o}(n^{-1/2})$. However, the results are contrary to our expectation. As shown in Huang (1995), unlike the coefficient a_1 in the leading term, the coefficient a_2 involves already stochastic components and hence (3.4) is the best expansion we can have. The main reason for this is due to the intrinsic difficulty of approximating elements in the design matrix $(\boldsymbol{X}^{\mathrm{T}}W\boldsymbol{X})$. A typical component $S_{n,j}$ in the design matrix admits approximation (5.2), whose stochastic error is of size $(nh)^{-1/2}$. These stochastic error terms enter into the coefficient a_2.

The above remarks reveal that in the nonparametric regression setting, it is not possible to construct a root-n consistent bandwidth selector using the conventional plug-in idea as in Hall et al. (1991). This marks a major difference between the density estimation and the nonparametric regression.

Theorem 2. *Under Conditions* (A1)–(A5), *the optimal bandwidth admits the following expansion*:

$$\frac{(h_{\mathrm{OPT}} - h_{\mathrm{O}})}{h_{\mathrm{OPT}}} = \mathrm{O_P}(n^{-2/5}),$$

where $h_{\mathrm{O}} = a_1n^{-1/5}$.

4. Pre-asymptotic substitution bandwidth selector

In this section, we study the asymptotic performance of the pre-asymptotic substitution bandwidth selector. The homoscedastic model ((2.1) with $\sigma^2(x) = \sigma$) is adopted here for simplicity.

The basic idea stems from Fan and Gijbels (1995). Instead of using the asymptotic bias and variance, their procedure involves assessing the bias and variance via

$$\operatorname{bias}(\hat{m}_l(x)|X_1,\dots,X_n) = e_0^{\mathrm{T}}(X_l^{\mathrm{T}}WX_l)^{-1}X_l^{\mathrm{T}}W(\boldsymbol{m} - X_l(\beta_0,\beta_1)^{\mathrm{T}}),$$

$$\operatorname{var}(\hat{m}_l(x)|X_1,\dots,X_n) = \sigma^2e_0^{\mathrm{T}}(X_l^{\mathrm{T}}WX_l)^{-1}(X_l^{\mathrm{T}}W^2X_l)(X_l^{\mathrm{T}}WX_l)^{-1}e_0, \tag{4.1}$$

where X_l is the design matrix for fitting a local linear regression at point x (i.e. a specific case of (2.3) with $p=1$), $\boldsymbol{m} = (m(X_1),\dots,m(X_n))^{\mathrm{T}}$ and $e_0 = (1,0)^{\mathrm{T}}$. Here, as before, $\beta_j = m^{(j)}(x)/j!$. Note that the unknown terms in (4.1) are $(\boldsymbol{m} - X_l(\beta_0, \beta_1)^{\mathrm{T}})$ and the variance parameter σ^2. Estimated MSE for $\hat{m}_l(x)$ can then be obtained by assessing only $(\boldsymbol{m} - X_l(\beta_0, \beta_1)^{\mathrm{T}})$ and σ^2. Estimation of the conditional variance σ^2 is studied by Rice (1984), Hall et al. (1990), and Ruppert et al. (1995), among others, where one can easily construct estimators of the parametric rate. Hence, the estimation of variance is not an objective of this study. To approximate $(\boldsymbol{m} - X_l(\beta_0, \beta_1)^{\mathrm{T}})$, a Taylor expansion around the point x is used to obtain

$$\boldsymbol{m} - X_l(\beta_0, \beta_1)^{\mathrm{T}} \approx \begin{pmatrix} \beta_2(X_1 - x)^2 + \beta_3(X_1 - x)^3 \\ \vdots \\ \beta_2(X_n - x)^2 + \beta_3(X_n - x)^3 \end{pmatrix}. \tag{4.2}$$

From (4.1) and (4.2), we need only estimate σ^2, $\beta_2(x)$, and $\beta_3(x)$. Following the development of the local polynomial fitting, we use a local polynomial approximation of degree p with a pilot bandwidth g to estimate σ^2, $\beta_2(x)$, and $\beta_3(x)$. Note that the difference of this approach from the "plug-in" method is that we substitute estimates into pre-asymptotic expressions (4.1) and (4.2) instead of their asymptotic counterparts. As a result, we do not need to estimate the unknown terms involving the design density although it is present in the asymptotic expansion.

With the above pre-asymptotic substitution, we obtain an estimate of $\widehat{\mathrm{MSE}}(x;h)$ from (4.1) and (4.2). Consider the estimated mean integrated squared error (MISE)

$$\widehat{\mathrm{MISE}}(h) \equiv \hat{M}(h) = \int \widehat{\mathrm{MSE}}(x;h) w(x)\,\mathrm{d}x, \tag{4.3}$$

where w is a non-negative weight function on $[a,b]$ with three bounded derivatives. A data-driven bandwidth selector $\hat{h}_{\mathrm{OPT}}$ is the one that minimizes $\hat{M}(h)$.

For the simplicity of implementation, Fan and Gijbels (1995) consider a pilot fitting of the local cubic polynomial with a pilot bandwidth g selected by the residual squares criterion. For theoretical consideration, we take a non data-driven optimal pilot bandwidth g, which is of size $n^{-1/7}$. The convergence rate of this pre-asymptotic substitution method is depicted in part (1) of Theorem 3. Clearly, the pilot degree $p=3$ does not explore fully the smoothness condition of the regression function $m(\cdot)$. The rate of convergence can be improved upon if one uses the pilot fitting of order $p=5$. The result is summarized in part (2) of Theorem 3.

Theorem 3. *Suppose $\hat{\sigma}^2$ is a root-n consistent estimator of σ^2. Assume that $h=h(n) \to 0$, $nh + \log(h) \to \infty$, as $n \to \infty$, and the weight function satisfies $w^{(i)}(a) = w^{(i)}(b) = 0$, for $i = 0,1,2,3$. Under conditions* (A1) – (A4),

1. *if $p=3$ and $g = d_1 n^{-1/7}$ in the pilot estimation for a positive constant d_1, then*

$$\frac{(\hat{h}_{\mathrm{OPT}} - h_{\mathrm{OPT}})}{h_{\mathrm{OPT}}} = O_P(n^{-2/7}); \tag{4.4}$$

2. *if $p=5$ and $g \in (d_2 n^{-3/25}, d_3 n^{-1/10})$ in the pilot estimation for some positive constants d_2 and d_3, then*

$$\frac{(\hat{h}_{\mathrm{OPT}} - h_{\mathrm{OPT}})}{h_{\mathrm{OPT}}} = O_P(n^{-2/5}). \tag{4.5}$$

The proof is given in Section 5.

5. Proofs

This section outlines the key idea of the proof. Details can be found in Huang (1995).

Proof of Theorem 1. We begin by estimating the conditional bias. The expectation of the local linear regression estimator is

$$E\{\hat{m}_l(x)|X_1,\dots,X_n\} = e_0^{\mathrm{T}}(X_l^{\mathrm{T}}WX_l)^{-1}X_l^{\mathrm{T}}W\boldsymbol{m}.$$

A Taylor's expansion of $\boldsymbol{m}(\cdot)$ gives

$$\boldsymbol{m} = \begin{pmatrix} m(x)+m'(x)(X_1-x)+\cdots+m^{(4)}(x)(X_1-x)^4/4!+\mathrm{o}\{(X_1-x)^4\} \\ \vdots \\ m(x)+m'(x)(X_n-x)+\cdots+m^{(4)}(x)(X_n-x)^4/4!+\mathrm{o}\{(X_n-x)^4\} \end{pmatrix},$$

and hence

$$\mathrm{bias}(\hat{m}_l(x)|X_1,\dots,X_n) = e_0^{\mathrm{T}}\begin{pmatrix} S_{n,0} & S_{n,1} \\ S_{n,1} & S_{n,2}\end{pmatrix}^{-1}\begin{pmatrix} \beta_2S_{n,2}+\beta_3S_{n,3}+\beta_4S_{n,4} \\ \beta_2S_{n,3}+\beta_3S_{n,4}+\beta_4S_{n,5}\end{pmatrix} + r(x,X_1,\dots,X_n), \tag{5.1}$$

where $r(\cdot)$ denotes the remainder terms. Using a similar argument as in the proof of Theorem 1 of Fan et al. (1996), one can show that

$$S_{n,j} = nh^{j+1}(s_j^* + \mathrm{O_P}(a_n)), \tag{5.2}$$

where $s_j^* = f(x)\mu_j + hf'(x)\mu_{j+1} + h^2f''(x)\mu_{j+2}/2 + \mathrm{o}(h^2)$ and $a_n = (nh)^{-1/2}$. It follows from (5.1) and (5.2) that

$$\mathrm{bias}(\hat{m}_l(x)|X_1,\dots,X_n) = e_0^{\mathrm{T}}\left(\begin{pmatrix} s_0^* & s_1^* \\ s_1^* & s_2^*\end{pmatrix} + \mathrm{O_P}(a_n)\right)^{-1} h^2\begin{pmatrix} \beta_2s_2^*+h\beta_3s_3^*+h^2\beta_4s_4^*+\mathrm{O_P}(a_n) \\ \beta_2s_3^*+h\beta_3s_4^*+h^2\beta_5s_5^*+\mathrm{O_P}(a_n)\end{pmatrix}.$$

It is easy to see that

$$\begin{pmatrix} s_0^* & s_1^* \\ s_1^* & s_2^*\end{pmatrix} = f(x)\begin{pmatrix} 1 & 0 \\ 0 & \mu_2\end{pmatrix} + hf'(x)\begin{pmatrix} 0 & \mu_2 \\ \mu_2 & 0\end{pmatrix} + h^2f''(x)\begin{pmatrix} \mu_2/2 & 0 \\ 0 & \mu_4/2\end{pmatrix} + \mathrm{o}(h^2).$$

Using the fact that for square matrices A, B, and C with A being invertible,

$$\{A+hB+h^2C+\mathrm{o}(h^2)\}^{-1} = A^{-1} - hA^{-1}BA^{-1} - h^2A^{-1}CA^{-1} + h^2A^{-1}BA^{-1}BA^{-1} + \mathrm{o}(h^2),$$

the bias expression in Theorem 1 is obtained with some matrix algebra.

For the variance term of the locally linear fit, we have

$$\mathrm{var}\{\hat{m}_l(x)|X_1,\dots,X_n\} = e_0^{\mathrm{T}}(X_l^{\mathrm{T}}WX_l)^{-1}(X_l^{\mathrm{T}}W\Sigma(x)WX_l)(X_l^{\mathrm{T}}WX_l)^{-1}e_0,$$

where $\Sigma(x) = \mathrm{diag}\{\sigma^2(X_1),\dots,\sigma^2(X_n)\}$. A typical element of the matrix $(X_l^{\mathrm{T}}W\Sigma(x)WX_l)$ is of form

$$R_{n,j} = \sum_{i=1}^{n}(X_i-x)^j\sigma^2(X_i)K^2\left(\frac{X_i-x}{h}\right).$$

By the approximation

$$R_{n,j} = nh^{j+1}(r_j^* + \mathrm{O_P}(a_n)),$$

where $r_j^* = g(x)v_j + hg'(x)v_{j+1} + h^2g''(x)v_{j+2}/2 + \mathrm{o}(h^2)$, it follows that

$$\mathrm{var}\{\hat{m}_l(x)|X_1,\dots,X_n\}$$
$$= \frac{1}{nh}e_0^{\mathrm{T}}\left(\begin{pmatrix} s_0^* & s_1^* \\ s_1^* & s_2^*\end{pmatrix} + \mathrm{O_P}(a_n)\right)^{-1}\left(\begin{pmatrix} r_0^* & r_1^* \\ r_1^* & r_2^*\end{pmatrix} + \mathrm{O_P}(a_n)\right)\left(\begin{pmatrix} s_0^* & s_1^* \\ s_1^* & s_2^*\end{pmatrix} + \mathrm{O_P}(a_n)\right)^{-1} e_0.$$

The asymptotic variance follows by some matrix computations. □

Proof of Theorem 2. For simplicity, denote the asymptotic bias and variance in Theorem 1 as

$$\mathrm{bias}(\hat{m}_l(x)|X_1,\ldots,X_n) = h^2 b_0(x) + \mathrm{O_P}(h^4 + n^{-1/2}h^{3/2}),$$

$$\mathrm{var}(\hat{m}_l(x)|X_1,\ldots,X_n) = n^{-1}h^{-1}v_0(x) + \mathrm{O_P}(n^{-1}h + n^{-3/2}h^{-3/2});$$

hence the conditional mean integrated square error is given by

$$M(h) = h^4\int b_0^2(x)w(x)\,\mathrm{d}x + n^{-1}h^{-1}\int v_0(x)w(x)\,\mathrm{d}x + \mathrm{O_P}(b_n),$$

where $b_n = h^6 + n^{-1/2}h^{7/2} + n^{-1}h + n^{-3/2}h^{-3/2}$. The above expression holds uniformly (except inflating a $\log n$ factor as remarked after Theorem 1) in $h \in (\alpha_1 n^{-t_1}, \alpha_2 n^{-t_2})$, the range specified in Condition (A3). The last statement implies that h_{OPT} is of order $n^{-1/5}$. Using the same arguments as in the case of obtaining the expression for $M(h)$, we have

$$M'(h) = 4h^3\int b_0^2(x)w(x)\,\mathrm{d}x - n^{-1}h^{-2}\int v_0(x)w(x)\,\mathrm{d}x + \mathrm{O_P}(b_n/h)$$

and

$$M''(h) = 12h^2\int b_0^2(x)w(x)\,\mathrm{d}x + 2n^{-1}h^{-3}\int v_0(x)w(x)\,\mathrm{d}x + \mathrm{O_P}(h^4 + n^{-1/2}h^{3/2} + n^{-3/2}h^{-7/2}).$$

Since h_{OPT} minimizes $M(h)$, $M'(h_{\mathrm{OPT}}) = 0$, and by a Taylor's expansion,

$$M'(h_{\mathrm{OPT}}) = M'(h_{\mathrm{O}}) + (h_{\mathrm{O}} - h_{\mathrm{OPT}})M''(\tilde{h}) = 0,$$

where $\tilde{h}$ lies between h_{O} and h_{OPT} and hence $\tilde{h} = \mathrm{O}(n^{-1/5})$. It is easy to check that $M'(h_{\mathrm{O}}) = \mathrm{O_P}(n^{-1})$ and $M''(\tilde{h}) = cn^{-2/5}(1 + \mathrm{o}(1))$ for some constant $c > 0$. Thus

$$(h_{\mathrm{OPT}} - h_{\mathrm{O}}) = M'(h_{\mathrm{O}})/M''(\tilde{h}) = \mathrm{O_P}(n^{-3/5}).$$

The theorem follows. □

Proof of Theorem 3. Let $\hat{h}_{\mathrm{OPT}}$ denote the optimal bandwidth that minimizes $\hat{M}(h)$ and $\hat{h}_{\mathrm{O}} = \hat{h}_{\mathrm{O}}(\hat{\beta}_2, \hat{\sigma})$ be defined similarly to h_{O} with the terms $\beta_2(x)$ and σ^2 substituted by their corresponding estimates. We first show that $\hat{h}_{\mathrm{OPT}}$ can be approximated well by $\hat{h}_{\mathrm{O}}$. For the given pilot bandwidth g, $\hat{\beta}_2(x)$ and $\hat{\beta}_3(x)$ are uniformly (in $x \in [a,b]$) consistent estimators of their counterparts and hence are stochastically bounded. Using the arguments that lead to Theorem 2, it can be verified that

$$\frac{(\hat{h}_{\mathrm{OPT}} - \hat{h}_{\mathrm{O}})}{\hat{h}_{\mathrm{OPT}}} = \mathrm{O_p}(n^{-2/5}).$$

Combining this with Theorem 2, we have

$$\frac{(\hat{h}_{\mathrm{OPT}} - h_{\mathrm{OPT}})}{h_{\mathrm{OPT}}} = \frac{(\hat{h}_{\mathrm{O}} - h_{\mathrm{O}})}{h_{\mathrm{OPT}}} + \mathrm{o_p}(n^{-2/5}). \tag{5.3}$$

By inspecting the differences between $\hat{h}_{\mathrm{O}} - h_{\mathrm{O}}$, the convergence rate is dictated by that of $\hat{\theta}_2 - \theta_2$. It follows from Theorem 4.1 of Huang and Fan (1995) that when g is of order $n^{-1/7}$ and $p = 3$,

$$\hat{\theta}_2 - \theta_2 = \mathrm{O_P}(n^{-2/7}),$$

and for $p = 5$ and $g \in (d_2 n^{-3/25}, d_3 n^{-1/10})$,

$$\hat{\theta}_2 - \theta_2 = \mathrm{O_P}(n^{-2/5}).$$

The conclusion follows from the above two observations and (5.3). □

References

Chiu, S.T., 1991a. Some stabilized bandwidth selectors for nonparametric regression. Ann. Statist. 19, 1528–1546.

Chiu, S.T., 1991b. Bandwidth selection for kernel density estimation. Ann. Statist. 19, 1883–1905.

Fan, J., 1993. Local linear regression smoothers and their minimax efficiency. Ann. Statist. 21, 196–226.

Fan, J., Gijbels, I., 1995. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. J. Roy. Statist. Soc. Ser. B 57, 371–394.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall, London.

Fan, J., Gijbels, I., Hu, T.-C., Huang, L.-S., 1996. An asymptotic study of variable bandwidth selection for local polynomial regression. Statist. Sinica 6, 113–127.

Hall, P., Sheather, S.J., Jones, M.C., Marron, J.S., 1991. On optimal data-based bandwidth selection in kernel density estimation. Biometrika 78, 263–271.

Hart, J.D., Yi, S., 1996. One-sided cross-validation. Technical Report, Department of Statistics, Texas A & M University.

Hastie, T.J., Loader, C., 1993. Local regression: automatic kernel carpentry (with discussion). Statist. Sci. 8, 120–143.

Huang, L.-S., 1995. On nonparametric estimation and goodness-of-fit. Dissertation, Mimeo Series #2333, Department of Statistics, University of North Carolina at Chapel Hill.

Huang, L.-S., Fan, J., 1995. Nonparametric estimation of quadratic functionals. Technical Report #905, Department of Statistics, Florida State University.

Jones, M.C., Marron, J.S., Sheather, S.J., 1996. A brief survey of bandwidth selection for density estimation. J. Amer. Statist. Assoc. 91, 401–407.

Opsomer, J.D., 1995. Optimal bandwidth selection for fitting an additive model by local polynomial regression. Dissertation, Cornell University.

Priestly, M.B., Chao, M.T., 1972. Nonparametric function fitting. J. Roy. Statist. Soc. Ser. B 34, 385–392.

Ruppert, D., Sheather, S.J., Wand, M.P., 1995. An effective bandwidth selector for local least squares regression. J. Amer. Statist. Assoc. 90, 1257–1270.

Ruppert, D., Wand, M.P., 1994. Multivariate locally weighted least squares regression. Ann. Statist. 22, 1346–1370.

Schucany, W.R., 1995. Adaptive bandwidth choice for kernel regression. J. Amer. Statist. Assoc. 90, 535–540.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York.

Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman & Hall, London.