Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: https://www.tandfonline.com/loi/gnst20

A simple root n bandwidth selector for nonparametric regression

Siegfried Heiler & Yuanhua Feng

To cite this article: Siegfried Heiler & Yuanhua Feng (1998) A simple root n bandwidth selector for nonparametric regression, Journal of Nonparametric Statistics, 9:1, 1-21, DOI: 10.1080/10485259808832733

To link to this article: https://doi.org/10.1080/10485259808832733

Nonparametric Statistics, Vol. 9, pp. 1–21
Reprints available directly from the publisher
Photocopying permitted by license only

A SIMPLE ROOT n BANDWIDTH SELECTOR FOR NONPARAMETRIC REGRESSION

SIEGFRIED HEILER and YUANHUA FENG

Department of Economics and Statistics, University of Konstanz, Universitaetsstrasse 10, D-78434 Konstanz, Germany

(Received 28 February 1997; In final form 30 May 1997)

The purpose of this paper is to investigate data-driven bandwidth selection for nonparametric regression based on a double-smoothing procedure. It will be shown that the best convergence rate can be achieved by kernel regression with non-negative kernels in both pilot smoothing and as well as in main smoothing. The asymptotic results are given for a naive kernel estimator with an equally spaced design, but they can also be used for other kernel estimators or for locally weighted regression. Three variates of data-driven bandwidth selectors for local linear regression are proposed. One of them, $\hat{h}_{DSI}$, is root n consistent. The performance of these bandwidth selectors is studied through simulation. They are also compared with the bandwidths selected by the R criterion of Rice and the true ASE optimal bandwidth (h_{ASE}). In spite of satisfactory performances of all bandwidth selectors, the root n one turns out to be the best in theory as well as in practice.

Keywords: Bandwidth choice; double-smoothing; plug-in; local linear regression

INTRODUCTION AND MOTIVATION

The development of computer facilities and the fact that in practice, parametric regression is often not suitable for adequately fitting curves to many data sets, has led to rapid developments in the field of nonparametric regression. Recent developments in this field may be found in the monographs of Müller [24], Härdle [10], Wand and Jones [32], and Fan and Gijbels [6]. Effective use of these methods requires the choice of the bandwidths or smoothing parameters. This is one of the most important aspects of nonparametric regression. In this paper

we focus on the selection of a global bandwidth for univariate fixed design nonparametric regression.

In the related field of nonparametric density estimation, there has been major progress made in recent years in data-driven bandwidth selection (Jones, Marron and Sheather [17, 18]; Cao, Cuevas and González-Manteiga [2]). Jones, Marron and Sheather [17] grouped the existing methods into *first generation* and *second generation* ones. For information about first generation methods, refer to the survey by Marron [20]. The second generation methods, including various new plug-in methods (e.g., Park and Marron [26]; Sheather and Jones [31]), smoothed cross-validation (Hall, Marron and Park [14]), smoothed bootstrap (Marron [22] and Cao [1]) and some root n convergent methods (Jones, Marron and Park [16]; Marron [21]), are far superior to the better known first generation methods.

Most first generation methods in the context of nonparametric regression can be found in Rice [27] and Härdle, Hall and Marron [11]. Developement of second generation methods in this field is still at the initial stages. See Gasser, Kneip and Köhler [7], Chiu [3], Härdle, Hall and Marron [12] and Ruppert, Sheather and Wand [28] for some proposed second generation bandwidth selectors. The proposal in Härdle, Hall and Marron [12] is a *Double-Smoothing* (DS) procedure. The DS bandwidth selectors often have a high convergence rate. Under certain conditions the bandwidth selectors of Härdle, Hall and Marron [12] are root n consistent. Further, this proposal does not directly depend on asymptotic considerations and hence can be used for bandwidth selection of a general linear smoother, e.g., locally weighted regression, without difficulty. This method has already been successfully applied to bandwidth selection of time series decomposition with locally weighted regression (Heiler and Feng [15]). But there is a hurdle to the actual use of this methodology, i.e., one has to choose a pilot bandwidth. The goal of this paper is to improve DS and to give a data-driven selection procedure of the pilot bandwidth.

In section 2 we extend the results of Härdle, Hall and Marron [12] following the ideas in the paper of Jones, Marron and Park [16], we present some special cases which provide a class of fast bandwidth selectors. It is shown that the best convergence rate $n^{-1/2}$ can be achieved by kernel regression with nonnegative kernels in both pilot smoothing and in main smoothing. Hence we call such a bandwidth

selector a simple root n one, which involves the use of high order kernels only when one selects an unknown constant in the pilot bandwidth. In section 3 the data-driven procedure for selecting the pilot bandwidth is described. As a by-product of this root n procedure we obtain a direct plug-in bandwidth selector, which is similar to the proposal of Ruppert, Sheather and Wand [28]. Section 4 gives the simulation results of the performances of the proposed bandwidth selectors for local linear regression. Some concluding remarks are made in section 5.

THE DS PROCEDURE AND ITS EXTENSION

The idea of double-smoothing goes back at least to Müller [23]. DS bandwidth selectors were studed by Härdle, Hall and Marron [12] and as a result of their research they arrived at some important asymptotic properties. In their proposal [12] a constant pilot bandwidth g is used. Jones, Marron and Park [16] proposed the use of a pilot bandwidth of the form $g = Cn^{\nu}h^{\delta}$ is the smoothed cross-validation procedure discussed by Hall, Marron and Park [14], where C, ν and δ are constants, which influence the performance of the bandwidth selector and must be chosen beforehand. The authors also allowed for a so-called nonstochastic term in the estimator of the mean integrated squared error to improve the performance of the bandwidth selector. We transfer these ideas into DS in order to obtain a class of fast bandwidth selectors for nonparametric regression.

We consider in this paper a nonparametric model with *equally spaced design*

$$Y_i = m(x_i) + \varepsilon_i, \quad 1 \le i \le n, \tag{1}$$

where $x_i = (i-0.5)/n$ and the errors are i.i.d. random variables with $E(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = \sigma^2$. Our goal is to estimate the mean function $m(\cdot)$ from these n observations. The DS procedure can be used for a general linear smoother, but for simplicity we use the so-called *naive* kernel estimator

$$\hat{m}_h(x) = (nh)^{-1} \sum_{j=1}^{n} Y_j K[(x - x_j)/h],$$

where K is a kernel of order r (i.e., $\int K(u)du=1$, $\int u^p K(u)du=0$ for $0<p<r$ and $\neq 0$ for $p=r$) and h is the bandwidth. Observing that for equally spaced designs the well known Nadaraya-Watson estimator (Nadaraya [25] and Watson [33]), the Gasser-Müller estimator (Gasser and Müller [8]) and local polynomial fitting (Cleveland [4]) give exactly the same asymptotic performance, theorem 1 also holds for these estimators.

The Mean Averaged Squared Error (MASE) is considered as a distance between $\hat{m}(x)$ and $m(x)$,

$$M = M(h) = n^{-1} \sum_i^* E[\hat{m}(x_i) - m(x_i)]^2,$$

where $\sum_i^*$ denotes summing over indices i such that $c < x_i < d$, where $0<c<d<1$. c and d are used in order to remove boundary effects (see [12]). In the simulation study in section 4, $c=0.1$ and $d=0.9$ are used. h_0, the minimizer of M, is taken as the optimal bandwidth. It is well known that h_0 is asymptotically approximated by the asymptotically optimal bandwidth $h_{\mathrm{AM}} = c_0 n^{-1/(2r+1)}$, where here and in the sequel,

$$c_0 = \left(\frac{(r!)^2}{2r} \cdot \frac{(d-c)\sigma^2 \int K^2(u)du}{\int_c^d (m^{(r)}(x))^2 dx (\int u^r K(u)du)^2} \right)^{1/(2r+1)} \tag{1}$$

In the sequel we describe the DS procedure (see also [12]). For DS we need a main smoothing

$$\hat{m}(x) = \hat{m}_h(x) = (nh)^{-1} \sum_{j=1}^n Y_j K[(x - x_j)/h] = \sum_{j=1}^n w_{jh} Y_j,$$

with kernel K and bandwidth h, and a pilot smoothing

$$\hat{m}_g(x) = (ng)^{-1} \sum_{j=1}^n Y_j L(x - x_j)/g] = \sum_{j=1}^n w_{jg} Y_j,$$

with kernel L and bandwidth g, which are allowed to be different from K and h. We assume that the kernels are of orders r and s, respectively, and define

$$\kappa_r (-1)^r (r!)^{-1} \int u^r K(u)du$$

and

$$\lambda_s(-1)^s(s!)^{-1}\int u^s L(u)du.$$

It is well known that the MASE splits up into a variance part and a bias part. The variance part of $M(h)$ is given by

$$V = V(h) = n^{-1}\sum_i^* \operatorname{var}[\hat{m}(x_i)] = n^{-1}\sigma^2\sum_i^*\sum_{j=1}^n w_{jh}(x_i)^2.$$

Following the idea of DS the bias at each point x_i is estimated by

$$\begin{aligned}\hat{b}(x_i) &= \sum_{k=1}^n w_{kh}(x_i)\hat{m}_g(x_k) - \hat{m}_g(x_i)\\ &= \sum_{k=1}^n a_k\hat{m}_g(x_k),\end{aligned} \tag{2}$$

where

$$a_k = \begin{cases} w_{kh}, & k \neq i,\\ w_{kh} - 1, & k = i.\end{cases}$$

$\hat{b}(x_i)$ can be written as a linear combination of the observations too. With the notation

$$A_j(x) = n\left[\sum_{k=1}^n w_{kh}(x)w_{jg}(x_k) - w_{jg}(x)\right],$$

one obtains

$$\hat{b}(x_i) = n^{-1}\sum_{j=1}^n A_j(x_i)Y_j.$$

The systematic bias part of $M(h)$, $B = B(h) = \sum_i^* b(x_i)^2$, is estimated by

$$\hat{B} = \hat{B}(h) = n^{-1}\sum_i^* \hat{b}(x_i)^2.$$

There is a variance term, $n^{-3}\sigma^2 \sum_i^* \sum_j A_j(x_i)^2$, in this estimate, which was subtracted in [12]. This terms plays the same role as the nonstochastic term in Jones and Sheather [19]. These authors showed that the nonstochastic term should be taken into account in order to improve the performance of the plug-in bandwidth selector (Sheather and Jones [31]). The same idea is used by Jones, Marron and Park [16] in order to improve the performance of the smoothed cross-validation bandwidth selector proposed by Hall, Marron and Park [14]. It will be shown later that the performance of the DS bandwidth selector can also often be improved if this term is included. To handle this variance term we introduce an indicator variable Δ which takes the value 0 when this term is subtracted, as in [12], and 1 when it is included. Hence the final estimate of $M(h)$ is

$$\hat{M}(h) = \hat{V} + \hat{B} - (1-\Delta)n^{-3}\hat{\sigma}^2 \sum_i^* \sum_{j=1}^n A_j(x_i)^2, \quad \Delta = 0, 1, \qquad (3)$$

where $\hat{\sigma}^2$ is an estimator of σ^2 and $\hat{V} = n^{-1}\hat{\sigma}^2 \sum_i^* \sum_j w_{jh}(x_i)^2$. The DS estimator of h_0 is $\hat{h}$, the minimizer of (3). Note that the estimate of the variance part does not involve pilot smoothing. With respect to this, the DS and the smoothed bootstrap differ.

The asymptotic properties of $\hat{h}$ are described by theorem 1 under the following assumptions:

- *A1* K and L are compactly supported kernels of orders r and s respectively, K' and $L^{(r+1)}$ are bounded.
- *A2* Let $r' = \max(r, s)$. Assume that $m^{(r+r')}$ is continuous on (0, 1).
- *A3* $\hat{\sigma}^2$ is root n consistent σ^2, i.e., $E\{(\hat{\sigma}^2 - \sigma^2)^2\} = \alpha n^{-1} + o(n^{-1})$, where α is a positive constant which depends on the selected estimator.
- *A4* The pilot bandwidth is of the form $g = Cn_\nu\, h^\delta$, and $g/h \to \infty$.

A1–*A3* are the same as in [12]. *A4* is an additional assumption on the form of the pilot bandwidth (see [12], pp. 232–233, and [14], pp. 5–9, for the reason of the assumption $g/h \to \infty$). Common conditions, under which the kernel estimators are consistent, are also assumed.

It can be shown that there exist positive constants c_1 and c_2 such that

$$M''(h_0) \simeq c_1(nh_0^3)^{-1} \simeq c_2 h_0^{2r-2}.$$

THEOREM 1 *Under A1–A4,*

$$(\hat{h}-h_0)/h_0 = \gamma_1(\hat{\sigma}^2-\sigma^2)+(\gamma_2 n^{-2}g_0^{-(4r+1)}+\gamma_3 n^{-1})^{1/2}Z_n+\gamma_4 g_0^s + \gamma_5 g_0^{2s}+o(g_0^{2s})+\Delta\gamma_6(n^{-1}g_0^{1(2r+1)})(1+o(1)), \tag{4}$$

where Z_n *is asymptotically normal* $N(0, 1)$, $g_0 = Cn^{\nu}h_0^{\delta}$ *and* γ_i, $i=1,\ldots,6$, *are constants, which are given by*

$$\gamma_1 = c_1^{-1}(d-c)\int K^2,$$
$$\gamma_2 = c_2^{-2}(d-c)[2r-(2r+1)\delta]^2\kappa_r^4\sigma^4\int[\int L^{(r)}(y)L^{(r)}(y+z)dy]^2dz,$$
$$\gamma_3 = 16c_2^{-2}r^2\kappa_r^4\sigma^2\int_c^d(m^{(2r)})^2,$$
$$\gamma_4 = -2c_2^{-1}(2r+s\delta)\kappa_r^2\lambda_s\int_c^d m^{(r)}m^{(r+s)},$$
$$\gamma_5 = -2c_2^{-1}(r+s\delta)\kappa_r^2\lambda_s^2\int_c^d(m^{(r+s)})^2 \quad \textit{and}$$
$$\gamma_6 = c_2^{-1}(d-c)[2r-(2r+1)\delta]\sigma^2k_r^2\int(L^{(r)})^2.$$

The constants γ_1 and γ_3 are not affected by the selection of g and are the same as the ones in [12]. If $\delta=0$ and $\Delta=0$, theorem 1 is the same as theorem 1 in [12], where $(\gamma_5 g_0^{2s}) = o(\gamma_4 g_0^s)$ holds because $g_0\to 0$ as $n\to\infty$. This relationship is true as long as $\delta\neq -2r/s$. When $\delta=-2r/s$ we obtain $\gamma_4=0$ and the fourth term in theorem 1 vanishes. The proof of theorem 1 above is similar to the proof of theorem 1 in [12] and is hence omitted. We only explain the differences between them (for the meanings of the symbols see [12]). On the one hand the proof is simplified because of the use of a naive kernel estimator with an equally spaced design. On the other hand it is complicated at some points: firstly, the case $\Delta=1$ adds an extra tern T_5, say, to D_2 in [12]; secondly, the term of order $(h^{2r}g^{2s})$ in the approximation of T_1 is not negligible when $\delta=-2r/s$; and thirdly, the pilot bandwidth g depends on h unless $\delta=0$.

Theorem 1 can be used to derive good choices of C, ν and δ in $g_0 = Cn^{\nu}h_0^{\delta}$. The optimal choices of ν and δ induce a linear constraint between them. Hence if one of them is given, the other can easily be obtained. In the following we discuss some special cases. Because $\delta=-2r/s$ is a critical value we consider the cases of $\delta=-r/s$ and $\delta\neq -2r/s$. Note that the first three terms in theorem 1 give the asymptotic variance while the other terms give the asymptotic bias of

$(\hat{h}-h_0)/h_0$ which can be combined to give an asymptotic mean squared error of $(\hat{h}-h_0)/h_0$,

$$\text{AMSE} = \gamma_1^2 \alpha n^{-1} + \gamma_2 n^{-2} g_0^{-(4r+1)} + \gamma_3 n^{-1} + [\gamma_4 g_0^s + \gamma_5 g_0^{2s} + \Delta \gamma_6 n^{-1} g_0^{-(2r+1)}]^2.$$

For given δ, C has to be chosen so as to minimize AMSE. When $\Delta = 1$, then the dominant term of AMSE is reduced to the squared bias part, and in this case C is chosen to minimize this part.

Case 1 $\delta \neq -2r/s$, $\Delta = 0$. Now the best choices are obtained by balancing the second term of the variance part and the first term of the bias part of AMSE:

$$\nu = \frac{\delta}{2r+1} - \frac{2}{4r+2s+1}, \quad C = c_0^{-\delta} \left(\frac{(4r+1)\gamma_2}{2s\gamma_4^2} \right)^{1/(4r+2s+1)}$$

The resulting rate of convergence is

$$(\hat{h} - h_0)/h_0 \sim \begin{cases} n^{-(4s+1)/(4r+4s+2)}, & \text{if } s < 2r+1, \\ n^{-1/2}, & \text{if } s \geq 2r+1. \end{cases}$$

If $\delta = 0$, this gives the same results as in Remarks 2 and 3 in [12]. In particular, the rate of convergence is $n^{-4/13}$ when $r = s = 2$.

Case 2 $\delta \neq -2r/s$, $\Delta = 1$. We assume that $\delta \neq 2r/(2r/(2r+1)$, i.e., $\gamma_6 \neq 0$. If $\delta = 2r/(2r+1)$, then it is the same as in case 1. In this case the asymptotically best choices come from trading off or balancing the first and the third terms in the bias part of ASME,

$$\nu = \frac{\delta}{2r+2} - \frac{1}{2r+s+1}$$

and

$$C = \begin{cases} c_0^{-\delta} (-\gamma_6/\gamma_4)^{1/(2r+s+1)}, & \text{when } \gamma_4\gamma_6 < 0, \\ c_0^{-\delta} \left(\frac{2r+1}{2} \frac{\gamma_6}{\gamma_4} \right)^{1/(2r+s+1)}, & \text{when } \gamma_4\gamma_6 > 0. \end{cases}$$

The resulting rate of convergence for the first subcase is $n^{-(2s+1)/(4r+2s+2)}$ if $s < 2r$ or $n^{-1/2}$ if $s \geq 2r$. And the resulting rate of

convergence for the second subcase is $n^{-s/(2r+s+1)}$ if $s < 2r+1$ or $n^{-1/2}$ if $s \geq 2r+1$ or $n^{-1/2}$ if $s \geq 2r+1$. We see that with $\Delta = 1$, we obtain a slightly higher rate of convergence when $\gamma_4\gamma_6 > 0$.

Case 3 $\delta \neq -2r/s$, $\Delta = 0$. The asymptotic best choices are

$$\nu = \frac{8r^2 + 6rs + 2r + 2s}{s(2r+1)(4s+4r+1)}, \quad C = c_0^{2r/2}\left(\frac{(4r+1)\gamma_2}{4s\gamma_5^2}\right)^{1/(4r+4s+1)}$$

The resulting rate of convergence is

$$(\hat{h} - h_0)/h_0 \sim \begin{cases} n^{-4s/(4r+4s+2)}, & \text{if } s < r+1, \\ n^{-1/2}, & \text{if } s \geq r+1. \end{cases}$$

For the special case $r = s = 2$ the rate of convergence is $n^{-8/17}$.

Case 4 $\delta = -2r/s$, $\Delta = 1$. Observing that now $\gamma_5 > 0$, $\gamma_6 < 0$, the asymptotically best choices come from trading off the second and the third terms in the bias part of ASEM,

$$\nu = \frac{4r^2 + 6rs + 2r + s}{s(2r+1)(2s+2s+1)}, \quad C = c_0^{2r/s}(-\gamma_6/\gamma_5)^{1/(2r+2s+1)}.$$

The resulting rate of convergence is

$$(\hat{h} - h_0)/h_0 \sim \begin{cases} n^{-(4s+1)/(4r+4s+2)}, & \text{if } s < r, \\ n^{-1/2}, & \text{if } s \geq r. \end{cases}$$

In this case the best rate of convergence can be achieved with $r = s = 2$. (i.e., with symmetric positive kernels in both pilot smoothing and main smoothing). This provides a simple root n bandwidth selector for non-parametric regression. Since in the DS procedure s should be equal to or larger than r, the bandwidth selector in case 4 should always be root n consistent.

The choice of C in case 1, the second subcase of case 2 and case 3 does not affect the rate of convergence. But if C is not correctly selected in the first subcase of case 2 the resulting rate of convergence will be reduced to that for the second subcase of case 2. In case 4 when $C \neq c_0^{2r/s}(-\gamma_6/\gamma_5)^{1/(2r+2s+1)}$ the rate of convergence is reduced to $n^{-2s/(2r+2s+1)}$ if $s \leq r$ or $n^{-1/2}$ if $s > r$. Now the rate of convergence for

$\Delta = 1$ is a little slower than that for $\Delta = 0$. However if $\delta = -2r/s$ and $s > r$, then the DS bandwidth selector is always a root n estimator and the rate of convergence in this case does not depend on C and Δ. We prefer to use $\Delta = 1$ because of its greater computational simplicity. If $\delta \neq -2r/s$, the rate of convergence does not depend on δ. Hence the choice of $\delta = 0$, as in [12], is also favorable, because in this case the pilot bandwidth is a constant. For fixed r, a larger s leads to a higher rate of convergence. The choice of C, which is more difficult, will be discussed in the next section.

Remark Theorem 1 can also be extended to the Gasser-Müller estimator in the case of a regular fixed design. But this will not be discussed here.

THE PROPOSED DATA-DRIVEN DS PROCEDURE

The DS procedure discussed above is a data-driven procedure only if one has a data-driven selector $\hat{g}$ of the pilot bandwidth g, which leads to another bandwidth selection problem. This is a hurdle to the actual use of DS and was an open question in [12]. In this section we discuss the data-driven selection of the pilot bandwidth g for local polynomial fitting. We particularly have in mind the local linear estimator. Now the kernel functions are the so-called equivalent kernels as defined in Ruppert and Wand [29]. If local linear fitting is used in both pilot smoothing and main smoothing we have $s = r = 2$. We consider two special cases: (1) the bandwidth selector, $\hat{h}_{\mathrm{DS0}}$, in case 2 of section 2 with $\delta = 0$ and (2) the bandwidth selector, $\hat{h}_{\mathrm{DS1}}$, in case 4 of section 2 with $\delta = 2r/s = -2$. In both cases $\Delta = 1$. $\hat{h}_{\mathrm{DS1}}$ is root n consistent. The rate of convergence of $\hat{h}_{\mathrm{DS0}}$ is at least $n^{-2/7}$. The pilot bandwidth for $\hat{h}_{\mathrm{DS0}}$ does not depend on h, therefore the procedure for $\hat{h}_{\mathrm{DS0}}$ is faster than that for $\hat{h}_{\mathrm{DS1}}$.

The choice of pilot bandwidths for $\hat{h}_{\mathrm{DS0}}$ and $\hat{h}_{\mathrm{DS1}}$ is equivalent to the choice of the constants γ_4, γ_5, γ_6 and c_0. The only unknown term in γ_6 is the variance σ^2. The unknown terms in γ_4, γ_5 and c_0 are σ^2, θ_{22}, θ_{24} and θ_{44}, where θ_{lk} is the integral

$$\theta_{kl} = \int_c^d m^{(k)}(x) m^{(l)}(x) dt, \quad k, l \geq 0.$$

To estimate σ^2, a second-order difference-based estimator of the variance proposed by Gasser, Sroka and Jennen-Steinmetz [9] is used. This estimator $\hat{\sigma}^2$, is in accordance with $A3$ of theorem 1, because it is already a root n consistent estimator of σ^2. A first-order difference-based estimator of σ^2 can be found in Rice [27]. Hall and Maron [13] proposed a variance estimator in nonparametric regression based on the mean square of a sequence of residuals. Their idea was extended to nonparametric variance estimators based on local least squares by Ruppert, Sheather and Wand [28] (see also Fan and Gijbels [5]). The implementation of these variance estimators requires selection of another bandwidth (see Hall and Marron [13], and Ruppert, Sheather and Wand [28]). That is the reason why we do not use this method in the simulation.

The estimation of θ_{lk} is studied by Ruppert, Sheather and Wand [28] for $k+l$ even in the context of locally weighted regression. These authors suggested that one can estimate θ_{kl} by local polynomial estimation of derivatives with another bandwidth, say α_{kl}. If we use local polynomials of order 5, assuming that m has 8 continuous derivatives and that $n\alpha_{kl}^{k+l+1} \to \infty$ as $n \to \infty$, according to (3.1) and (3.2) in Ruppert, Sheather and Wand [28], the so-called MSE-optimal bandwiths for estimating θ_{22}, θ_{24} and θ_{44} are:

$$\alpha_{22} \simeq C_{22}(K)\left[\frac{\sigma^2(d-c)}{|\theta_{26}|n}\right]^{1/9}, \tag{5}$$

where

$$C_{22}(K) = \begin{cases} C_{22}^{I}(K), & \theta_{26} < 0, \\ C_{22}^{II}(K), & \theta_{26} > 0, \end{cases}$$

and

$$C_{22}^{I}(K) = \left[\frac{450R(K_{2,5})}{|\mu_6(K_{2,5})|}\right]^{1/9}, \quad C_{22}^{II} = \left[\frac{360R(K_{2,5})}{|\mu_6(K_{2,5})|}\right]^{1/9};$$

$$\alpha_{24} \simeq C_{24}(K)\left[\frac{\sigma^2(d-c)}{|\theta_{26}|n}\right]^{1/9}, \tag{6}$$

where

$$C_{24}(K) = \begin{cases} C_{24}^{I}(K), & \theta_{26} < 0, \\ C_{24}^{II}(K), & \theta_{26} > 0, \end{cases}$$

and

$$C_{24}^{I}(K) = \left[\frac{2520|\int K_{2,5}K_{4,5}|}{|\mu_6(K_{4,5})|}\right]^{1/9}, \quad C_{24}^{II} = \left[\frac{720|\int K_{2,5}K_{4,5}|}{|\mu_6(K_{4,5})|}\right]^{1/9};$$

and

$$\alpha_{44} \simeq C_{44}(K)\left[\frac{\sigma^2(d-c)}{|\theta_{46}|n}\right]^{1/11}, \tag{7}$$

where

$$C_{44}(K) = \begin{cases} C_{44}^{I}(K), & \theta_{46} < 0, \\ C_{44}^{II}(K), & \theta_{46} > 0, \end{cases}$$

and

$$C_{44}^{I}(K) = \left[\frac{360R(K_{4,5})}{|\mu_6(K_{4,5})|}\right]^{1/11}, \quad C_{44}^{II} = \left[\frac{1620(K_{4,5}K_{4,5})}{|\mu_6(K_{4,5})|}\right]^{1/11},$$

where $R(K_{\nu,5}) = \int K_{\nu,5}^2$, $\mu_6(K_{\nu,5}) = \int u^6 K_{\nu,5}(u)$, and where $K_{\nu,5}$, $\nu = 2$ or 4, is the equivalent kernel for estimating the ν-th derivative of $m(x)$ with a local polynomial of order 5 (Ruppert and Wand [29]). The values of C_{22}, C_{24} and C_{44} for some common kernels are given in Table I.

We see that in order to estimate α_{22}, α_{24} and α_{44} we have to estimate θ_{26} and θ_{46}. Though this again is a bandwidth selection problem. The dependence of $\hat{\alpha}_{22}$, $\hat{\alpha}_{24}$ and $\hat{\alpha}_{44}$ on $\hat{\theta}_{26}$ or $\hat{\theta}_{46}$ is less important. Now we can use a kernel estimator with bandwith selected by a simple pilot method to estimate θ_{26} and θ_{46}. Hence the following data-driven procedure is proposed: 1. Estimate σ^2 by using a second-order difference-based variance estimator; 2. Estimate θ_{26} and θ_{46} by using a

TABLE I Kernel Dependent Constants for Some Common Kernels

kernel	*Uniform*	*Epanechnikov*	*Quartic*	*Triweight*	*Gaussian*
C_{22}^{I}	3.7200	4.0179	4.3535	4.6751	1.2207
C_{22}^{II}	3.6289	3.9195	4.2469	4.5606	1.1908
C_{24}^{I}	4.0179	4.2938	4.6391	4.9750	1.2941
C_{24}^{II}	3.4958	3.7359	4.0363	4.3285	1.1260
C_{44}^{I}	3.3231	3.5392	3.8167	4.0884	1.0576
C_{44}^{II}	3.8100	4.0578	4.3760	4.6874	1.2125

local polynomial of order 7 with the bandwidth selected by the R criterion (Rice [27]); 3. Plug $\hat{\sigma}^2$, $\hat{\theta}_{26}$ and $\hat{\theta}_{46}$ into (5)–(7) to obtain $\hat{\alpha}_{22}$, $\hat{\alpha}_{24}$ and $\hat{\alpha}_{44}$; 4. Estimate θ_{22}, θ_{24} and θ_{44} by using a local polynomial of order 5 with the bandwidths $\hat{\alpha}_{22}$, $\hat{\alpha}_{24}$ and $\hat{\alpha}_{44}$, respectively; 5. Plug $\hat{\sigma}^2$, $\hat{\theta}_{22}$, $\hat{\theta}_{24}$ and $\hat{\theta}_{44}$ into γ_4, γ_5, γ_6 and c_0 to obtain the estimates of them and to obtain the estimates of C for h_{DS0} and h_{DS1}, respectively; 6. $\hat{h}_{\mathrm{DS0}}$ and $\hat{h}_{\mathrm{DS1}}$ are then obtained with DS procedure by using local linear fitting in both pilot smoothing and main smoothing. The use of a simple pilot method in a fast bandwidth selection procedure was also proposed by Ruppert, Sheather and Wand [28] and Fan and Gijbels [5].

With $\hat{c}_0$ we obtain a direct plug-in estimator of h_0, $\hat{h}_{\mathrm{AM}} = \hat{c}_0 n^{-1/(2r+1)}$, denoted by $\hat{h}_{\mathrm{DPI}}$, as a by product of the procedure for h_{DS1}. $\hat{h}_{\mathrm{DPI}}$ in this paper is slightly different from that in the paper by Ruppert, Sheather and Wand [28] because we use $p=5$ instead of $p=3$ to estimate θ_{22}. The pilot method and the estimation of the variance are also different. The rate of convergence of $\hat{h}_{\mathrm{DPI}}$ is of order $n^{-2/5}$ because of the bias in h_{AM}. But the variance term of $(\hat{h}_{\mathrm{DPI}} - h_0)/h_0$ converges still faster.

SIMULATION RESULTS

To assess and compare each of the bandwidth selectors $\hat{h}_{\mathrm{DS0}}$, $\hat{h}_{\mathrm{DS1}}$ and $\hat{h}_{\mathrm{DPI}}$ we conducted a simulation study. In this paper we used the Quartic Kernel as a weight function for local linear regression in both pilot smoothing and main smoothing. The k-NN method was used to choose the bandwidth for estimating θ_{26} and θ_{46} because of the high

order of the polynomial. The following three functions are chosen as regressors:

$$m_1(x) = 2 - 5x + 5\exp[-100(x-0.5)^2],$$
$$m_2(x) = 2\sin(4\pi x),$$
$$m_3(x) = 10/(1+\exp(2-4\sin(2\pi(x+0.25)))).$$

The first two functions are r_1 and r_2 as used in Gasser, Kneip and Köhler [9]. The third one was chosen by us. As error terms i.i.d. standard normal variables were used. Observations were taken at $x_i = (i-0.5)/n$, for $n=50$ and $n=100$. The number of replications in the simulation was $T=300$. The true Averaged Squared Error (ASE) optimal bandwidths (h_{ASE}) for all samples were calculated. The bandwidth selected by the R criterion $\hat{h}_R$, was included in order to show a comparison between the first generation and the second generation methods. The results are summarized numerically in Tables II and III. The averages and the standard deviations of selected bandwidths are given in Table II. Table III gives the Averaged Squared Error to h_0, $\mathrm{ASE}(h_0) = T^{-1}\Sigma_i(\hat{h}_i - h_0)^2$, and the Averaged Squared Error to h_{ASE}, $\mathrm{ASE}(h_{ASE}) = T^{-1}\Sigma_i(h_i - h_{ASE,i})^2$ for each bandwidth selector as well as $\mathrm{ASE}(h_0)$ for h_{ASE}. The kernel density estimates of $\hat{h}_{DS0}$, $\hat{h}_{DS1}$, $\hat{h}_{DPI}$, $\hat{h}_R$ and h_{ASE} based on the values of $\log(\hat{h}) - \log(h_0)$ in 300 replications are given in Figures 1–3.

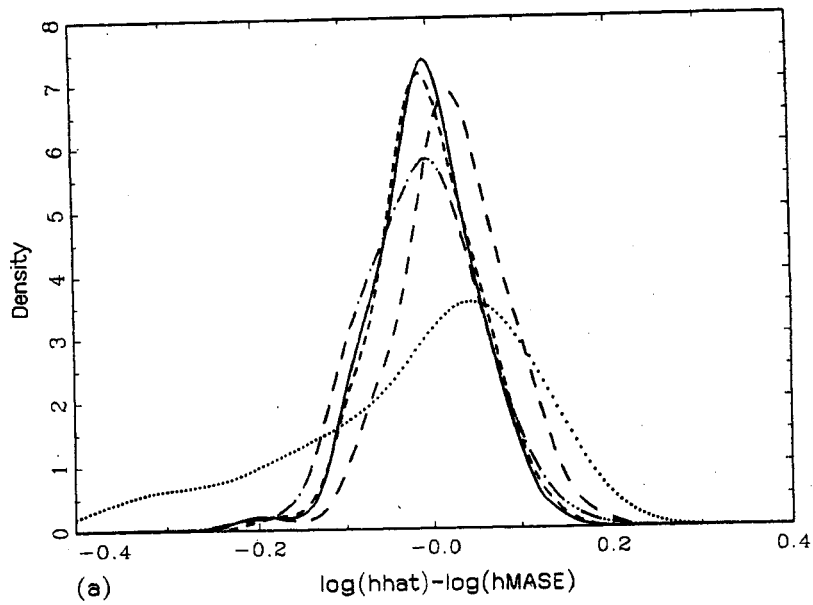
TABLE II *Average** and *Standard Deviation*** of Each Bandwidth Selector and of h_{ASE} in 300 Replications

Size	*1. n=50*			*2. n=100*		
Funct.	m_1	m_2	m_3	m_1	m_2	m_3
h_0	0.097	0.109	0.104	0.083	0.094	0.089
$\hat{h}_{DS0}$	0.103*	0.111	0.110	0.085	0.094	0.093
	1.41e-2**	1.79e-2	1.52e-2	8.67e-3	7.36e-3	9.98e-3
$\hat{h}_{DS1}$	0.096	0.102	0.100	0.081	0.089	0.087
	1.23e-2	1.19e-2	1.26e-2	7.70e-3	6.54e-3	8.84e-3
$\hat{h}_{DPI}$	0.096	0.101	0.099	0.080	0.087	0.087
	1.25e-2	1.19e-2	1.25e-2	7.85e-3	6.54e-3	9.13e-3
$\hat{h}_R$	0.096	0.105	0.103	0.081	0.090	0.086
	2.54e-2	2.85e-2	2.84e-2	1.87e-2	2.18e-2	2.64e-2
h_{ASE}	0.095	0.109	0.104	0.084	0.095	0.088
	1.45e-2	1.67e-2	1.73e-2	1.24e-2	1.64e-2	1.24e-2

TABLE III ASE(h_0) (first row), ASE(h_{ASE}) (second row) of each Bandwidth Selector and ASE(h_0) of h_{ASE} in 300 Replications, and their Rankings (numbers in parentheses)

Funct.	m_1	m_2	m_3
$n=50$			
$\hat{h}_{DS0}$	2.35e-4(3)	3.28e-4(3)	2.64e-4(3)
	6.34e-4(3)	8.85e-4(3)	7.35e-4(3)
$\hat{h}_{DS1}$	1.52e-4(1)	1.88e-4(1)	1.80e-4(2)
	5.04e-4(1.5)	6.70e-4(1)	6.05e-4(2)
$\hat{h}_{DPI}$	1.57e-4(2)	2.10e-4(2)	1.79e-4(1)
	5.04e-4(1.5)	6.71e-4(2)	5.79e-4(1)
$\hat{h}_R$	6.45e-4(4)	8.30e-4(4)	8.09e-4(4)
	1.08e-3(4)	1.53e-3(4)	1.38e-3(4)
h_{ASE}	2.13e-4	2.77e-4	3.01e-4
	–	–	–
$n=100$			
$\hat{h}_{DS0}$	8.02e-5(3)	5.42e-5(1)	1.18e-4(3)
	3.74e-4(3)	4.63e-4(1)	3.99e-4(3)
$\hat{h}_{DS1}$	6.49e-5(1)	6.73e-5(2)	8.16e-5(1)
	3.47e-4(1)	4.66e-4(2)	3.25e-4(2)
$\hat{h}_{DPI}$	6.91e-5(2)	9.11e-5(3)	8.84e-5(2)
	3.52e-4(2)	4.74e-4(3)	3.19e-4(1)
$\hat{h}_R$	3.52e-4(4)	4.92e-4(4)	5.24e-4(4)
	7.01e-4(4)	1.07e-3(4)	9.15e-4(4)
h_{ASE}	1.54e-4	2.69e-4	1.56e-4
	–	–	–

$\hat{h}_{DS0}$ performs much better than $\hat{h}_R$ but not as well as $\hat{h}_{DS1}$ or $\hat{h}_{DPI}$, since it is biased towards oversmoothing. Its standard deviation is always the largest amongst the proposed selectors. However for m_2, $n=100$, $\hat{h}_{DS0}$ happens to be the best following ASE(h_0) or ASE(h_{ASE}). This is due to the fact that the optimal bandwidth for m_2 with a polynomial of order 7 is $h_0(7)=0.5$, even when $n=100$. In this case the data-driven estimate of $h_0(7)$ is always smaller or equal to the true value. Sometimes $\hat{\theta}_{24}$ may have a sign which is different from that of θ_{24}. When this is the case, then $\hat{h}_{DS0}$ is much larger than its theoretical optimum. This occurred in the simulation study for m_2 eight times for $n=50$ and once for $n=100$. The average of these eight selected bandwidths for $n=50$ was 0.164. It was almost as large as the largest one selected by the R criterion (0.165) and the maximum that occurred was equal to 0.277. The value of the one outcome for $n=100$ was



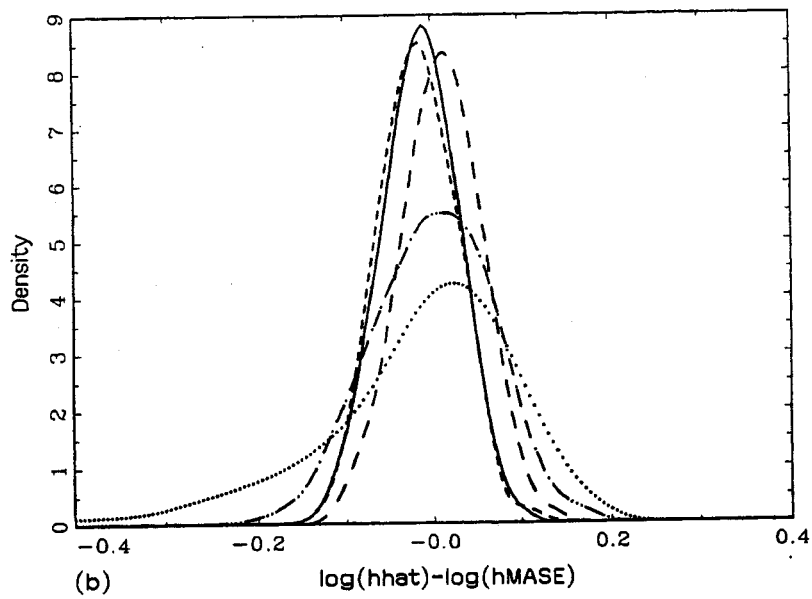
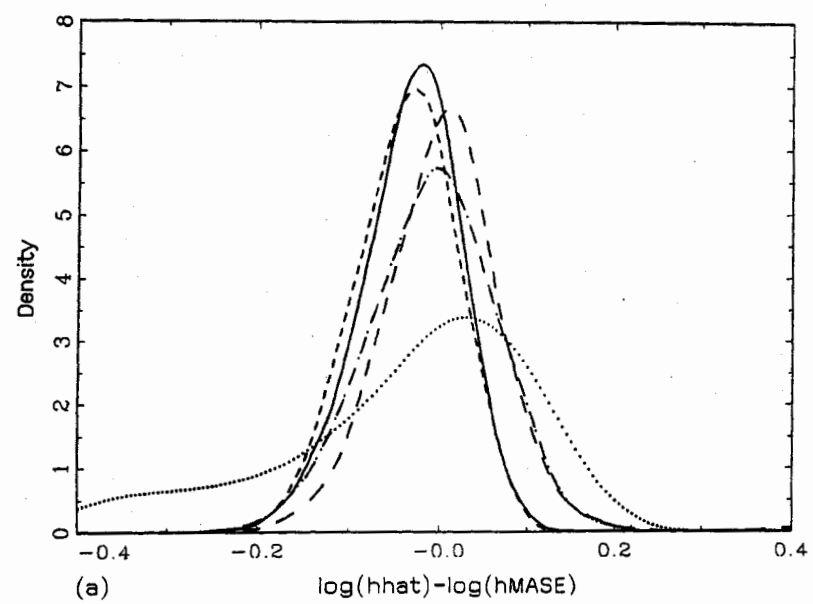


FIGURE 1 Kernel density estimates of $\log(\hat{h})-\log(h_0)$ for m_1 in 300 replications. (a) $n=50$, (b) $n=100$. The curves are for $\hat{h}_{DS1}$ (solid line), h_{DPI} (short dashes) $\hat{h}_{DS0}$ (long dashes), h_{ASE} (dashes and dots) and $\hat{h}_R$ (dots), respectively.



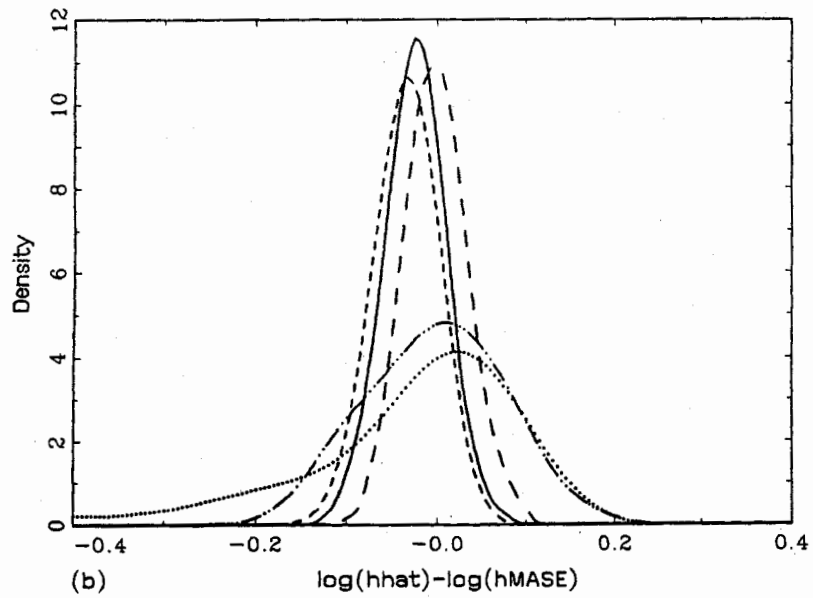
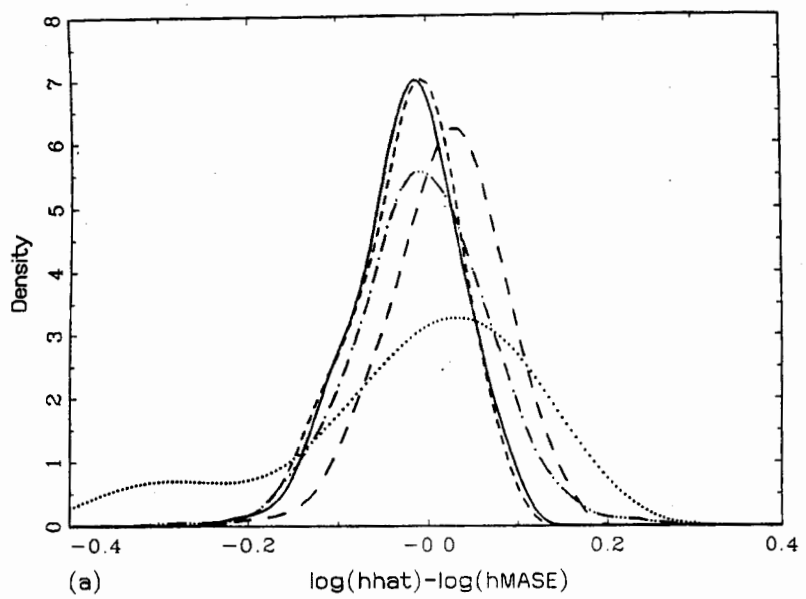


FIGURE 2 Kernel density estimates of $\log(\hat{h})-\log(h_0)$ for m_2 in 300 replications. (a) $n = 50$, (b) $n = 100$. The curves are for $\hat{h}_{DS1}$ (solid line), $\hat{h}_{DPI}$ (short dashes), $\hat{h}_{DS0}$ (long dashes), h_{ASE} (dashes and dots) and $\hat{h}_R$ (dots), respectively.



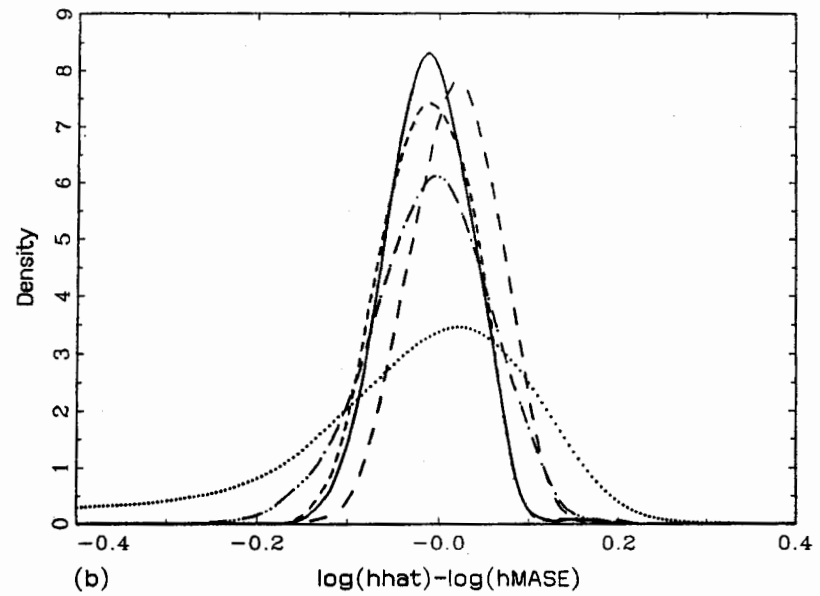


FIGURE 3 Kernel density estimates of $\log(\hat{h})-\log(h_0)$ for m_1 in 300 replications. (a) $n = 50$, (b) $n = 100$. The curves are for $\hat{h}_{DS1}$ (solid line), $\hat{h}_{DPI}$ (short dashes), $\hat{h}_{DS0}$ (long dashes), h_{ASE} (dashes and dots) and $\hat{h}_R$ (dots), respectively.

0.114. Hence we are of the opinion that $\hat{h}_{DS0}$ is not a good bandwidth selector, especially when n is small.

Both, $\hat{h}_{DPI}$ and $\hat{h}_{DS1}$ perform very well but are slightly biased towards undersmoothing. The situation is a little more serious for m_2. This is due to the same reason as mentioned above. This situation is improved when n increases from 50 to 100. Following ASE(h_0), $\hat{h}_{DS1}$ always perform better than $\hat{h}_{DPI}$ except for m_3 with $n=50$. When n changes from 50 to 100 the difference between ASE(h_0) of h_{DS1} and ASE(h_0) of h_{DPI} is clearly increased. For $n=100$ both the bias and the standard deviation of $\hat{h}_{DS1}$ are not larger than the ones of $\hat{h}_{DPI}$. Hence $\hat{h}_{DS1}$ turns out to be the best choice. The simulation results conform with theorem 1, which leads us to think that the difference would be more evident if a simulation with a larger n were to be carried out.

All of these three bandwidth selectors are not only much closer to the MASE optimal bandwidth h_0 but also much closer to the true ASE optimal bandwidth h_{ASE} than $\hat{h}_R$. ASE(h_{ASE}) is much larger than ASE(h_0). Following ASE(h_{ASE}), $\hat{h}_{DPI}$ is sometimes better than $\hat{h}_{DS1}$, however $\hat{h}_{DS1}$ seems to give a more stable performance. $\hat{h}_{DS1}$ and $\hat{h}_{DPI}$ are even much closer to h_0 than the true optimal bandwidth h_{ASE}. When $n=100$, $\hat{h}_{DS0}$ is also closer to h_0 than h_{ASE}. These confirm with the theoretical results because the best rate of convergence is only $n^{-1/10}$ if h_{ASE} is taken to be the optimal bandwidth, and the difference between h_{ASE} and h_0 is also of order $n^{-1/10}$.

CONCLUDING REMARKS

We are of the opinion that the most important lesson to be learnt from this analysis is that the DS procedure provides an interesting alternative to the plug-in method or the consideration in Chiu [3], to obtain very fast data-driven bandwidth selectors. This study shows that $\hat{h}_{DS1}$ not only yields a very good theroretical performance but also yields very good practical results. Therefore we suggest the use of $\hat{h}_{DS1}$ for bandwidth selection of nonparametric regression in practice, especially when n is large, although a larger simulation study would be required to confirm this suggestion. The drawbacks of $\hat{h}_{DS1}$ are its computational complexity and the necessity of using polynomials of order 7 at the first stage. When n is small or when the underlining

function is not very smooth, $\hat{h}_{\mathrm{DS1}}$ might not be a suitable bandwidth selector.

The excellent performance of $\hat{h}_{\mathrm{DS1}}$ and $\hat{h}_{\mathrm{DPI}}$ is due to their very small sample variability. The bias part often does not play an important role. Our experiment shows that the bias of the final bandwidth selector depends on the bandwidth selector used at the first stage. Other methods, e.g., the rule-of-thumb (Ruppert, Sheather and Wand [28]) or the biased cross-validation (Scott and Terrell [30]), could also be used at this stage. A pilot bandwidth selector which is biased towards slight oversmoothing might be better in order to reduce the negative bias in $\hat{h}_{\mathrm{DS1}}$ and $\hat{h}_{\mathrm{DPI}}$. One can also use other root n consistent methods to estimate the variance, e.g., the estimator used by Ruppert, Sheather and Wand [28].

One can construct a data-driven DS procedure in different ways. The simplest way is that one obtains a pilot estimate by using a simple method, and then selects a bandwidth following DS by means of this pilot estimate (Müller [23]). Heiler and Feng [15] used this method to select bandwidths for time series decomposition. When the pilot method is consistent, the asymptotic property of such a bandwidth selector can be obtained from theorem 1 with $\delta=0$, $\Delta=1$ and $g=O(n^{-1/(2s+1)})$. A simulation study should still be done to investigate its sample performance.

References

[1] Cao, R. (1993). Bootstrapping the mean integrated squared error, *J. Multivariate Anal.*, **45**, 137–160.

[2] Cao, R., Cuevas, A. and González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation, *Comput. Statist. Data Anal.*, **17**, 153–176.

[3] Chiu, S.-T. (1991). Some stabilized bandwidth selectors for nonparametric regression, *Ann. Statist.*, **19**, 1528–1546.

[4] Cleveland, W. S. (1979). Robust locally regression and smoothing scatter-plots, *J. Amer. Statist. Assoc.*, **74**, 829–836.

[5] Fan, J. and 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.

[6] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and its Applications*, Chapman and Hall, London.

[7] Gasser, T., Kneip, A. and Köhler, W. (1991). A flexible and fast method for automatic smoothing, *J. Amer. Statist. Assoc.*, **86**, 643–652.

[8] Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions, *Lecture Notes in Math.*, **757**, Springer, New York, pp. 23–68.

[9] Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1996). Residual variance and residual pattern in nonlinear regression, *Biometrika*, **73**, 625–633.
[10] Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.
[11] Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion), *J. Amer. Statist. Assoc.*, **83**, 86–99.
[12] Regression smoothing parameters that are not far from their optimum, *J. Amer. Statist. Assoc.*, **87**, 227–233 (1992).
[13] Hall, P. and Marron, J. S. (1990). On variance estimation in nonparametric regression, *Biometrika*, **77**, 415–419.
[14] Hall, P., Marron, J. S. and Park, B. U. (1992). Smoothed cross-validation, *Probab. Theory Related Fields*, **92**, 1–20.
[15] Heiler, S. and Feng, Y. (1995). Data-driven optimal decomposition of time series, SFB II-287, University of Konstanz.
[16] Jones, M. C., Marron, J. S. and Park, B. U. (1991). A Simple root n bandwidth selector, *Ann. Statist.*, **19**, 1919–1932.
[17] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation, *J. Amer. Statist. Assoc.*, **91**, 401–407.
[18] Progress in data based bandwidth selection for kernel density estimation, *Comput. Statist.*, to appear.
[19] Jones, M. C. and Sheather, S. J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives, *Statist. Probab. Lett.*, **11**, 511–514.
[20] Marron, J. S. (1989). Automatic smoothing parameter selection: A survey, in *Semi-parametric and Non-parametric Econometrics*, ed. Ullah, A. Physica-Verlag, Heidelberg, pp. 65–86.
[21] Root n bandwidth selection, in *Nonparametric Functional Estimation and Related Topics*, ed. Roussas, G. Kluwer Academic Publishers, Dordrecht, 1991, pp. 251–260.
[22] Bootstrap bandwidth selection, in *Exploring the Limits of Bootstrap*, eds. LePage, R. and Billard, L. John Wiley, New York, 1992, pp. 249–262.
[23] Müller, H.-G. (1985). Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators, *Statist. Decisions*, Supp. Issue **2**, 193–206.
[24] Müller, H.-G. (1988). Nonparametric Regression Analysis of Longitudinal Data, *Lecture Notes in Statistics*, **46**, Springer-Verlag, Berlin.
[25] Nadaraya, E. A. (1964). On estimating regression, *Theory Probab. Appl.*, **9**, 141–142.
[26] Park, B. U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors, *J. Amer. Statist. Assoc.*, **85**, 66–72.
[27] Rice, J. (1984). Bandwidth choice for nonparametric regression, *Ann. Statist.*, **12**, 1215–1230.
[28] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression, *J. Amer. Statist. Assoc.*, **90**, 1257–1270.
[29] Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression, *Ann. Statist.*, **22**, 1346–1370.
[30] Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation, *J. Amer. Statist. Assoc.*, **82**, 1131–1346.
[31] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation, *J. Roy. Statist. Soc. Ser. B*, **53**, 683–690.
[32] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman Hall, London.
[33] Watson, G. S. (1964). Smooth regression analysis, *Sankhyā Ser. A*, **26**, 359–372.