Heiler, Siegfried; Feng, Yuanhua

Working Paper

A simple root n bandwidth selector for nonparametric regression

Diskussionsbeiträge - Serie II, No. 286

Provided in Cooperation with:
Department of Economics, University of Konstanz

Suggested Citation: Heiler, Siegfried; Feng, Yuanhua (1995) : A simple root n bandwidth selector for nonparametric regression, Diskussionsbeiträge - Serie II, No. 286, Universität Konstanz, Sonderforschungsbereich 178 - Internationalisierung der Wirtschaft, Konstanz

This Version is available at:
https://hdl.handle.net/10419/101751

Sonderforschungsbereich 178
„Internationalisierung der Wirtschaft“

Diskussionsbeiträge

Juristische
Fakultät

Fakultät für Wirtschafts-
wissenschaften und Statistik

Siegfried Heiler
Yuanhua Feng

A Simple Root n Bandwidth Selector for Nonparametric Regression

---

Postfach 5560
D-78434 Konstanz

Serie II — Nr. 286
Dezember 1995

A Simple Root n Bandwidth Selector For Nonparametric Regression

Siegfried Heiler

Yuanhua Feng

Serie II - Nr. 286

Dezember 1995

A Simple Root n Bandwidth Selector for Nonparametric Regression

Siegfried Heiler Yuanhua Feng

University of Konstanz

Abstract

The problem of selecting bandwidth for nonparametric regression is investigated. The methodology used here is a double-smoothing procedure with data-driven pilot bandwidths. After giving an extension of the asymptotic result of Härdle, Hall and Marron (1992) by transfering the ideas of Jones, Marron and Park (1991) into the context of nonparametric regression, some fast data-driven bandwidth selectors for nonparametric regression are proposed. One of them, h_{DS1}, is root n consistent. The performance of these bandwidth selectors is studied through simulation for local linear regression. They are also compared with the bandwidth selected by R criterion and the true ASE optimal bandwidth (h_{ASE}). Though all of them show a satisfactory performance, the root n bandwidth selector turns out to be the best.

Keywords: Bandwidth choice; Double-smoothing; Plug-in; Local linear regression.

1 Introduction and Motivation

Nonparametric regression has become a rapidly developing field as it is realized that parametric regression is not suitable for adequately fitting curves to many data sets that arise in practice. Many interesting examples of this may found in the monographs of Eubank (1988), Müller (1988), Härdle (1990) and Hastie and Tibshirani (1990). But effective use of these methods requires choice of the bandwidth or smoothing parameter. This is one of the most important aspects in nonparametric regression. In this paper we focus on the selection of a global bandwidth for univariate fixed design nonparametric regression. See Fan and Gijbels (1995) for a recent study on local bandwidth selection.

In the related field of nonparametric density estimation there has been major progress made in recent years in data-driven bandwidth selection. See the surveys of Jones, Marron and Sheather (1992, 1994) and the comparitive study of Cao, Cuevas and González-Manteiga (1994) for the progress in bandwidth selection for univariate density estimation. For the studies of multivariate density estimation see the monograph of Scott (1992). Jones, Marron and Sheather (1994) grouped the existing methods into "first generation" and "second generation" ones. For most first generation methods see the survey of Marron (1989). The second

generation methods, including various new plug-in methods (e.g. Park and Marron, 1990 and Jones and Sheather, 1991), smoothed cross-validation (Hall, Marron and Park, 1992), smoothed bootstrap (Marron, 1992, Cao, 1993 and Cao, Cuevas and González-Manteiga, 1994) and some root n convergent methods (Jones, Marron and Park, 1991 (JMP) and Marron, 1991), 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 (1983, 1984) and Härdle, Hall and Marron (1988). Developing of second generation methods in this field is still at the first steps. See Gasser, Kneip and Köhler (1991), Chiu (1991), Härdle, Hall and Marron (1992) (HHM) and Ruppert, Sheather and Wand (1995) for some proposed second generation bandwidth selectors. The proposal in HHM is a double-smoothing (DS) procedure. The consideration of this method is similar to those of the smoothed croos-validation and the smoothed bootstrap, both of them show a fairly satisfactory performance in the context of kernel density estimation (Cao, Cuevas and González-Manteiga, 1994). Under certain conditions the bandwidth selectors of HHM are root n consistent. Further, this proposal does not directly depend on asymptotic consideration 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 tried for bandwidth selection of time series decomposition with locally weighted regression (Heiler and Feng, 1995). We think that DS is a practically useful procedure for bandwidth selection. But there is a hurdle to actual use of this methodology, that is one has to choose a pilot bandwidth. This is an open question in HHM. 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 proposal of HHM following the ideas in JMP and give some special cases which provide a class of fast bandwidth selectors, some of them being root n consistent. It is shown that the best convergence rate $n^{-\frac{1}{2}}$ can even be achieved by kernel regression with nonnegative kernel functions in both pilot smoothing stage and main smoothing stage. This is a simple root n bandwidth selector, which involves the use of high order kernels only when one selects an unknown constant in the pilot bandwidth. 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 (1994). In section 3 the data-driven procedure for selecting the pilot bandwidth is described. It is shown that selecting the pilot bandwidth in a given case is equal to the selection of the unknown constant mentioned above. The data-driven procedure for selecting this constant is based on the results of Ruppert, Sheather and Wand (1994) and a first generation method. In this paper the R criterion of Rice (1983, 1984) is used. Section 4 gives the simulation results on the performances of the proposed bandwidth selectors for local linear regression. Some concluding remarks are given in section 5.

2 The DS Procedure and its Extension

The "double-smoothing" idea goes back at least to Müller (1985). Härdle, Hall and Marron (1992) studied DS bandwidth selectors and gave some important asymptotic properties. In the proposal of HHM a constant pilot bandwidth g is used. Jones, Marron and Park (1991) proposed the use of a pilot bandwidth of the form $g = Cn^{\nu}h^{\delta}$ in the smoothed cross-validation procedure discussed by Hall, Marron and Park (1992), where C, ν and δ are constants, which influence the performance of the bandwidth selector and must be chosen beforehand. The authors also allow for a so called nonstochastic term in the estimation of the Mean Integrated Squared Error (MISE), which was not taken into account by Hall, Marron and Park (1992). 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 *fixed design*

$$\mathbf{Y}_i = m(x_i) + \epsilon_i, \qquad 1 \le i \le n,$$

where each $x_i \in [0,1]$ and the errors are iid random variables with $E(\epsilon_i) = 0$ and $var(\epsilon_i) = \sigma^2$. Our goal is to estimate the mean function $m(\cdot)$ from these n observations. The nonparametric regression estimator considered here is a linear smoother, but for theorem 1 we only use the Nadraya-Watson estimator

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

where K is a kernel of order r (that is $\int u^p K(u)du = 0$ for $p < r$ and $\neq 0$ for $p = r$) and h is the bandwidth.

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 HHM). h_0, the minimizer of M, is taken as the optimal bandwidth. It is well known that the asymptotic MASE is given by

$$AM = n^{-1}h^{-1}\sigma^2(d-c)\int K^2(u)du + h^{2r}\left(\int u^r K(u)du/(r!)^2\right)^2 \int_c^d (m^{(r)}(x)dx)^2. \qquad (1)$$

From (1) one obtains $h_0 \approx h_{AM} = c_0 n^{-1/(2r+1)}$ (see also Herrmann and Gasser, 1994), where here and in the following,

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

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

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

with kernel K and bandwidth h, and a pilot smoothing

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

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, respectivly, and define

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

and

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

as in HHM. 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

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

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

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

for a general linear smoother or

$$\begin{aligned} A_j(x) &= n \textstyle\sum_k K[(x - x_k)/h] \{ L[(x_k - x_j)/g] \{ \sum_l L[(x_k - x_l)/g] \}^{-1} \\ &\quad - L[(x - x_j)/g] \{ \sum_l L[(x - x_l)/g] \}^{-1} \} \{ \textstyle\sum_l K[(x - x_l)/h] \}^{-1} \end{aligned}$$

in the special case of kernel regression, one obtains

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

The bias part of $M(h)$ is estimated by

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

There is a variance term, $n^{-3}\sigma^2 \sum_i^* \sum_{j=1}^n A_j(x_i)^2$, in this estimation, which was subtracted in HHM. This term does not depend on the data and plays a similar role as the nonstochastic term in Jones and Sheather (1991). These authors showed that this term should be taken into account in order to improve the performance of the plug-in bandwidth selector (Sheather and Jones, 1991). In JMP the same idea is also used to improve the performance of the smoothed cross-validation bandwidth selector proposed by Hall, Marron and Park (1992). It will turn out later that in the context of nonparametric regression the performance of DS bandwidth selector can also often be improved if this term is included in the estimation of the bias part. To handle this variance term we introduce an indicator variable Δ which takes the value 0 when this term is subtracted, as in HHM, and 1 when it is included. Hence the final estimation of $M(h)$ is

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

where $\hat{\sigma}^2$ is an estimation of σ^2 and $\hat{V} = n^{-1}\hat{\sigma}^2 \sum_i^* \sum_{j=1}^n w_{jh}(x_i)^2$. The DS estimator of h_0 is $\hat{h}$, the minimizer of (3). Note that the estimation of the variance part does not involve the pilot smoothing. In this point the DS and the smoothed bootstrap differ (Marron, 1992, Cao, 1993 and Cao, Cuevas and González-Manteiga, 1994).

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

Assumption 1. K and L are compactly supported kernels of orders r and s, respectively, K' and $L^{(r+1)}$ are bounded.

Asumpption 2. Let $r' = max(r, s)$. Assume that $m^{(r+r')}$ is continuous on (0, 1).

Assumption 3. $\hat{\sigma}^2$ is root n consistent for σ^2, that is, $\hat{\sigma}^2 = \sigma^2 + O_p(n^{-1/2})$.

Assumption 4. The pilot bandwidth is of the form $g = Cn^{\nu}h^{\delta}$.

Assumptions 1. - 3. are the same as in HHM. Assumption 4. is an additional assumption on the form of the pilot bandwidth.

It can be shown (see HHM and Härdle, Hall and Marron, 1988) that, under Assmptions 1-3, there exist positive constants c_1 and c_2 such that

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

Theorem 1. Under the assumptions 1. - 4.,

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

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

$$\gamma_1 = c_1^{-1}(d-c)\int K^2,$$

$$\gamma_2 = 2c_2^{-2}(d-c)[r-(r+\tfrac{1}{2})\delta]^2\kappa_r^4\sigma^4\int\left[\int L^{(r)}(y)L^{(r)}(y+z)dy\right]dz,$$

$$\gamma_3 = 4c_2^{-2}r^2\kappa_r^4\sigma^2\int_c^d (m^{(2r)})^2,$$

$$\gamma_4 = -2c_2^{-1}(2r+s\delta)\kappa_r^2\lambda_s\int_c^d m^{(r)}m^{(r+s)},$$

$$\gamma_5 = -2c_2^{-1}(r+s\delta)\kappa_r^2\lambda_s^2\int_c^d (m^{(r+s)})^2 \qquad \textit{and}$$

$$\gamma_6 = -2c_2^{-1}(d-c)[r-(r+\tfrac{1}{2})\delta]\sigma^2\kappa_r^2\int (L^{(r)})^2.$$

The constants γ_1 and γ_3 do not be affected by the selection of g and are the same as the ones in HHM. The proof of theorem 1 is similar to the proof of theorem 1 in HHM and is omitted here. If $\delta = 0$ and $\Delta = 0$ theroem 1 is the same as theorem 1 in HHM, 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 $\delta \neq -\frac{2r}{s}$. When $\delta = -\frac{2r}{s}$ we obtain $\gamma_4 = 0$ and the fourth term in theorem 1 vanishes.

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 so if one of them is given the other can easily be obtained. In the following we discuss some special cases. Because $\delta = -\frac{2r}{s}$ is a critical value we consider the cases of $\delta = -\frac{2r}{s}$ and $\delta \neq -\frac{2r}{s}$. Note that the second and third 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 $(\hat{h}-h_0)/h_0$,

$$\begin{aligned}AMSE &= \gamma_2 n^{-2}g_0^{-(4r+1)}+\gamma_3 n^{-1} \\ &+ [\gamma_1(\hat{\sigma}^2-\sigma^2)+\gamma_4 g_0^s+\gamma_5 g_0^{2s}+\Delta\gamma_6 n^{-1}g_0^{-(2r+1)}]^2.\end{aligned}$$

For given δ C has to be chosen so as to minimize AMSE. If $\Delta = 1$, the dominant term of AMSE is the bias part. Then C is chosen only to minimize the bias part of AMSE.

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

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

The resulting rate of convergence is

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

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

Case 2. $\delta \neq -\frac{2r}{s}$, $\Delta = 1$. We assume that $\delta < \frac{2r}{2r+1}$, that is $\gamma_6 < 0$. In this case the asymptotically best choices come from trading off or balancing the second and the fourth terms in the bias part of AMSE,

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

and

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

The resulting rate of convergence for the first subcase is $n^{-\frac{2s+1}{4r+2s+2}}$ if $s < 2r$ or $n^{-\frac{1}{2}}$ if $s \geq 2r$. And the resulting rate of convergence for the second subcase is $n^{-\frac{s}{2r+s+1}}$ if $s < 2r+1$ or $n^{-\frac{1}{2}}$ if $s \geq 2r+1$. We see that with $\Delta = 1$ we obtain a slightly higher rate of convergence when $\lambda_s \int_c^d m^{(r)} m^{(r+s)} < 0$, and a slightly slower one when $\lambda_s \int_c^d m^{(r)} m^{(r+s)} > 0$.

Case 3. $\delta = -\frac{2r}{s}$, $\Delta = 0$. The asymptotic best choices are

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

The resulting rate of convergence is

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

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

Case 4. $\delta = -\frac{2r}{s}$, $\Delta = 1$. Here, the asymptotically best choices come from trading off the third and the fourth terms in the bias part of AMSE,

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

The resulting rate of convergence is

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

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 nonparametric regression. Because s in DS procedure 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 $C \neq c_0^{\frac{2r}{s}}(-\gamma_6/\gamma_5)^{\frac{1}{2r+2s+1}}$ the rate of convergence is redused to $n^{-\frac{2s}{2r+2s+1}}$ if $s \leq r$ or $n^{-\frac{1}{2}}$ if $s > r$. Now the rate of convergence for $\Delta = 1$ is a little slower than that for $\Delta = 0$. However if $\delta = -\frac{2r}{s}$ and $s > r$ the DS bandwidth selector is always a root n estimator, 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 $\delta \neq -\frac{2r}{s}$, the rate of convergence does not depend on δ. Hence the choice of $\delta = 0$, as in HHM, is also favorable, because in this case the pilot bandwidth is a constant. For given r the larger s the higher is the rate of convergence. The choice of C is more difficult, which will be discussed in the next section.

3 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 rises another bandwidth selection problem. This is a hurdle to actual use of DS and was an open question in HHM. In this section we discuss the data-driven selection of the pilot bandwidth g for local polynomial fitting (Cleveland and Devlin, 1988 and Cleveland, Devlin and Grosse, 1988), especially for local linear regression. The results in section 2 are obtained for kernel regression, but they can be used directly for local polynomial fitting with so called asymptotically equvalent kernels (Ruppert and Wand, 1994). We consider here two special cases: (1) the bandwidth selector $\hat{h}_{DS0}$ in case 2 of section 2 with $r = s = 2$ and $\delta = 0$ and (2) the bandwidth selector $\hat{h}_{DS1}$ in case 4 of section 2 with $r = s = 2$ and $\delta = -\frac{2r}{s} = -2$. The pilot bandwidth for $\hat{h}_{DS0}$ does not depend on h, therefore the procedure for $\hat{h}_{DS0}$ is faster than that for $\hat{h}_{DS1}$. In both cases $\Delta = 1$.

It was shown in section 2 that the choice of ν in the form $g_0 = Cn^{\nu}h_0^{\delta}$ depends only on r and s, providing δ is given. Now the choice of g is equal to the choice of the constant C, in which the unkown term is of the form

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

For estimating $\hat{h}_{DS0}$ and $\hat{h}_{DS1}$ we need to estimate θ_{22}, θ_{24} and θ_{44}. The estimation of this expression is studied by Ruppert, Sheather and Wand (1994) 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, α_{kl} say. If we use local polynomials of order 5, according to (3.1) and (3.2) in Ruppert, Sheather and Wand (1994), the so called MSE-optimal bandwiths for estimating θ_{22}, θ_{24} and θ_{44} are:

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

where

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

and

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

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

where

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

and

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

and

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

where

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

and

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

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, as defined in Ruppert and Wand (1994). The values of C_{22}, C_{24} and C_{44} for some common kernels with support [-1, 1] are given in table 1.

We see that, in order to estimate α_{22}, α_{24} and α_{44} we have to estimate θ_{26} and θ_{46}. This leads to a new bandwidth selection problem. But at this stage the dependence of $\hat{\alpha}_{kl}$ on $\hat{\theta}$ is less important than the dependence of $\hat{C}$ on $\hat{\alpha}_{kl}$ or $\hat{h}$ on $\hat{C}$ at other stages. Therefore we can use a kernel estimator with bandwidth selected by a first generation method, for

Table 1: Kernel Dependent Constants

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

example the R criterion (Rice, 1983, 1984), in order to estimate the quantities of θ_{26} and θ_{46}. The proposal here is to use a local polynomial of order 7 to estimate θ_{24} and θ_{26} with the bandwidth selected by R criterion. This criterion is exactly the same as the M-Plot proposed by Cleveland and Devlin (1988) and Cleveland, Devlin and Grosse (1988) for locally weighted regression. The use of a simple method to estimate the pilot bandwidth was also proposed by Fan and Gijbels (1995) for a different procedure of bandwidth selection.

In this paper we use a simple difference-based estimator of the variance, σ^2, proposed by Gasser, Sroka and Jennen-Steinmetz (1986). This estimator, $\hat{\sigma}^2$, is in accordance with assumption 3 in theorem 1, because it is already a root n consistent estimator of σ^2. A more simple difference-based estimator of σ^2 can be found in Rice (1984). See Heiler and Feng (1995) for more references on this type of estimators. More complex estimators of σ^2 for locally weighted regression were proposed by Ruppert, Sheather and Wand (1994) and Fan and Gijbels (1995).

Since we have to estimate the very important constant c_0 in order to estimate the constant C for h_{DS1}, we obtain a so called direct plug-in estimator of h_0, $\hat{h}_{AM} = \hat{c}_0 n^{-1/(2r+1)}$, written as $\hat{h}_{DPI}$ as in Ruppert, Sheather and Wand (1994), as a by-product of the procedure for $\hat{h}_{DS1}$. $\hat{h}_{DPI}$ in this paper is different from the direct plug-in estimator in Ruppert, Sheather and Wand (1994) in three points: 1. Here we use $p = 5$ instead of $p = 3$ in Ruppert, Sheather and Wand (1994) to estimate θ_{22}; 2. The bandwidth used to estimate θ_{26} is selected by R criterion and 3. The estimator of variance is also different. The procedure for selecting $\hat{h}_{DPI}$ is more simple than that for $\hat{h}_{DS1}$. The rate of convergence of $\hat{h}_{DPI}$ is higher than the one in Ruppert, Sheather and Wand (1994), where $\hat{h}_{DPI}$ is an $O_p(n^{-2/7})$ bandwidth estimator. Here the rate of convergence of $\hat{h}_{DPI}$ is of order $n^{-2/5}$ because of the bias in h_{AM}. But the variance term of $(\hat{h}_{DPI} - h_0)/h_0$ converges still faster than the bias term.

4 Simulation Results

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

$$\begin{aligned} m_1(x) &= 2 - 5x + 5exp[-100(x-0.5)^2], \\ m_2(x) &= 2\sin(4\pi x) \qquad\qquad\qquad\qquad\qquad\qquad and \\ m_3(x) &= 10/(1 + exp(2 - 4sin(2\pi(x+0.25)))). \end{aligned}$$

The first two functions are r_1 and r_2 used in Gasser, Kneip and Köhler (1991). Independent standard normally distributed errors 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 by R criterion, $\hat{h}_R$, was included in order to give a comparison between the first generation methods and the second generation methods. The numerical results are summarized in table 2 and table 3. The kernel density estimations of $\hat{h}_{DS0}$, $\hat{h}_{DS1}$, $\hat{h}_{DPI}$, $\hat{h}_R$ and h_{ASE} in 300 replications are given in figure 1-3.

From table 2 and table 3 we can see that both, $\hat{h}_{DPI}$ and $\hat{h}_{DS1}$ perform very well, but are slightly biased towards undersmoothing. This situation is a little more serious for function 2. This is due to the fact that the optimal bandwidth for function 2 with a polynomial of order 7 is $h_0(7) = 0.5$, even when $n = 100$. This is very high and comes already close to a global (not local) model. In this case the data-driven estimation of $h_0(7)$ is always smaller or equal to the true value. This situation is improved when n changes from 50 to 100. For $n = 50$ $\hat{h}_{DPI}$ and $\hat{h}_{DS1}$ perform quite similar. For $n = 100$ $\hat{h}_{DS1}$ is better than $\hat{h}_{DPI}$ for all three regressions, following the criterion of Averaged Squared Error to h_0, $ASE(h_0) =: \frac{1}{T}\sum_{j=1}^{T}(\hat{h}_j - h_0)^2$. Now both, the bias and the standard deviation of $\hat{h}_{DS1}$ are smaller than the ones of $\hat{h}_{DPI}$. This conforms with the theoretical results, because the rate of convergence of $\hat{h}_{DS1}$ is higher. We think the difference will be more evident if a simulation with larger n is done.

$\hat{h}_{DS0}$ often also performs well. It is biased towards oversmoothing with larger standard deviation. For function 2, $n = 100$, $\hat{h}_{DS0}$ happens to be the best one due to the same reason mentioned above. But sometimes, $\hat{\theta}_{24}$ may have a sign different from θ_{24}. When this happens to be the case, $\hat{h}_{DS0}$ is much larger than its theoretical optimum. This occurred in the simulation study for function 2 8 times for $n = 50$ and once for $n = 100$. The average of these 8 selected bandwidths for $n = 50$ was 0.164. It was almost as large as the maximal one selected by the R criterion (0.165) and the maximum that occurred was 0.277. The

one for $n = 100$ was 0.114. Hence we think that $\hat{h}_{DS0}$ is not a good bandwidth selector, especially when n is small.

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 $\hat{h}_R$. But the Averaged Squared Error to h_{ASE}, $ASE(h_{ASE}) =: \frac{1}{T}\sum_{j=1}^{T}(\hat{h}_j - h_{ASE,j})^2$, is much larger than $ASE(h_0)$. Following $ASE(h_{ASE})$, $\hat{h}_{DPI}$ is sometimes better than $\hat{h}_{DS1}$. $\hat{h}_{DS1}$ 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 results conform 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 $n^{-1/10}$ (Härdle, Hall and Marron, 1988).

5 Concluding Remarks

We think that the most important lesson to be learned from this study is that the DS procedure provides an interesting alternative to the plug-in method or the consideration in Chiu (1991) to obtain very fast data-driven bandwidth selectors. This study shows that $\hat{h}_{DS1}$ has not only very good theoretical performance but yields also 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 and to compare $\hat{h}_{DS1}$ with other proposals. 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 enough smoothed $\hat{h}_{DS1}$ might not be a suitable bandwidth selector.

The very good performance of $\hat{h}_{DS1}$ and $\hat{h}_{DPI}$ is due to their very small sample variability. The bias part does not often play an important role. Our experiment shows that the bias of the final bandwidth selector depends on the bandwidth selector at the first stage. In this paper the R criterion was used in the simulation study, following similar considerations as DS in Heiler and Feng (1995). But a bandwidth selector which is biased towards slightly oversmoothing should be better in order to reduce the negative bias in $\hat{h}_{DS1}$ and $\hat{h}_{DPI}$. For example the biased cross-validation of Scott and Terrell (1987) could possibly be adapted to be used in the first stage. For the estimation of the variance one can use other root-n estimators of σ^2, for example the estimator used by Ruppert, Sheather and Wand (1994).

If we put $\delta = 0$, another simple data-driven procedure is that one selects at first a bandwidth $\hat{g}_s$ for $s > r$ with the R criterion or other methods, then one selects a bandwidth following DS by using $\hat{g}_s$ (or $\hat{g}_s$ multiplied by a factor) as pilot bandwidth. This method is

used in Heiler and Feng (1995). A simulation study should still be done to investigate its performance.

References

[1] CAO, R. (1993): Bootstrapping the Mean Integrated Squared Error. *Journal of Multivariate Analysis*, **45**, 137-60.

[2] CAO, R., CUEVAS, A. and GONZáLEZ-MANTEIGA, W. (1994): A Comparative Study of Several Smoothing Methods in Density Estimation. *Computational Statistics & Data Analysis*, **17**, 153-76.

[3] CHIU, S-T. (1991): Some Stabilized Bandwidth Selectors for Nonparametric Regression. *The Annals of Statistics*, **19**, 1528-46.

[4] CLEVELAND, W.S., DELVIN, S.J. (1988): Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J. Amer. Statist. Assoc.*, 83, 596–610.

[5] CLEVELAND, W.S., DELVIN, S.J., GROSSE, E. (1988): Regression by Local Fitting. Methods, Properties and Computational Algorithmus. *Journal of Econometrics*, 37, 87–114.

[6] EUBANK, R. L. (1988): *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

[7] FAN, J. and GIJBELS, I. (1995): Data-driven Bandwidth Selection in Local Polynomial Fitting: Variable bandwidth and Spatial Adaptation. *Journal of the Royal Statistical Society*, series B, 57, 371-394.

[8] GASSER, T., KNEIP, A. and KÖHLER, W. (1991): A Flexible and Fast Method for Automatic Smoothing. *J. Amer. Statist. Assoc.*, 86, 643–52.

[9] GASSER, T., SROKA, L. and JENNEN-STEINMETZ, C. (1986): Residual Variance and Residual Pattern in Nonparametric Regression. *Biometrica*, 73, 625–33.

[10] HÄRDLE, W. (1990): *Applied Nonparametric Regression.* Cambridge University Press, New York.

[11] HÄRDLE, W., HALL, P. and MARRON, J.S. (1988): How Far Are Automa-tically Chocen Regression Smoothing Parameters from Their Optimum?. *J. Amer. Statist. Assoc.*, 83, 86–95.

[12] HÄRDLE, W., HALL, P. and MARRON, J.S. (1992): Regression Smoothing Parameters That Are not Far from Their Optimum. *J. Amer. Statist. Assoc.*, 87, 227–33.

[13] HALL, P. MARRON, J.S. and PARK, U. (1992): Smoothed Cross-Validation. *Probability Theory*, **92**, 1-20.

[14] HASTIE, T. and TIBSHIRANI, R. (1990): *Generalized Additive Models.* Chapman and Hall, New York.

[15] HEILER, S. and FENG, Y. (1995): Data-Driven Optimal Decomposition of Time Series. *Discussion paper, University of Konstanz.*

[16] HERRMANN, E. and GASSER, T. (1994): Iterative Plug-in Algorithm for Bandwidth Selection in Kernel Regression Estimation. *Discussion Paper.*

[17] JONES, M.C., MARRON, J.S. and PARK, B.U. (1991): A Simple Root n Bandwidth Selector. *Ann. Statist.*, 19, 1919–32.

[18] JONES, M.C., MARRON, J.S. and SHEATHER, S.J. (1992): Progress in Data Based Bandwidth Selection for Kernel Density Estimation. Submitted to *Computational Statistics.*

[19] JONES, M.C., MARRON, J.S. and SHEATHER, S.J. (1994): A Brief Survey of Bandwidth Selection for Density Estimation. *J. Amer. Statist. Assoc.*, to appear.

[20] JONES, M.C. and SHEATHER, S.J. (1991): Using Nonstochastic Terms to Advanstage in Kernel-based Estimation of Integrated Squared Density Derivatives. *Statist. Probab. Lett.*, **11**, 511-514.

[21] MARRON, J.S. (1989): Automatic Smoothing Parameter Selection. In ULLAH, A. (Ed.): *Semiparametric and Nonparametric Econometrics*, 65-86. Physica-Verlag, Heidelberg.

[22] MARRON, J.S. (1991): Root N Bandwidth Selection. In ROUSSAS, G. (Ed.): *Nonparametric Functional Estimation and Related Topics*, 251-260. Kluwer Academic Publishers, Dordrecht.

[23] MARRON, J.S. (1992): Bootstrap Bandwidth Selection. In LePAGE, R. and BILARD, L. (Eds): *Exploring the Limits of Bootstrap*, 249-62. John Wiley & Sons, New York.

[24] MÜLLER, H.-G. (1985): Empirical Bandwidths Choice for Nonparametric Kernel Regression by Means of Pilot Estimators. *Statistical and Decisions*, Supplement Issue 2, 193-206.

[25] MÜLLER, H.-G. (1988): *Nonparametric Analysis of Longitudinal Data.* Springer-Verlag, Berlin.

[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. (1983): Methods for Bandwidth Choice in Nonparametric Kernel Regression. *Computer Science and Statistics.* 1983, 186–90.

[28] RICE, J. (1984): Bandwidth Choice for Nonparametric Regression. *Annals Statistics*, **12**, 1215–30.

[29] RUPPERT, D., SHEATHER, S.J. and WAND, M.P. (1994): An Effective Bandwidth Selector for Local Least Squares Regression. *J. Amer. Statist. Assoc.*, to appear.

[30] RUPPERT, D. and WAND, M.P. (1994): Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics*, **22**, 1346-70.

[31] SCOTT, D.W. (1992): *Miltivariate Density Estimation - Theory, Practice, and Visualization.* John Wiley & Sons, New York.

[32] SCOTT, D.W. and TERRELL, G.R. (1987): Biased and Unbiased Cross-Validation in Density Estimation. *J. Amer. Statist. Assoc.*, **82**, 1131-46.

[33] SHEATHER, S.J. and JONES, M.C. (1991): A Reliable Data-based Bandwidth Selection Method for Kernel Density. *J. R. Statist. Soc.* B, **53**, 683–90.

Address of the authors:

Department of Economics and Statistics

Universität Konstanz

Universitätstrasse 10

78434 Konstanz, Germany

Fax: (+49) 7531 - 883324

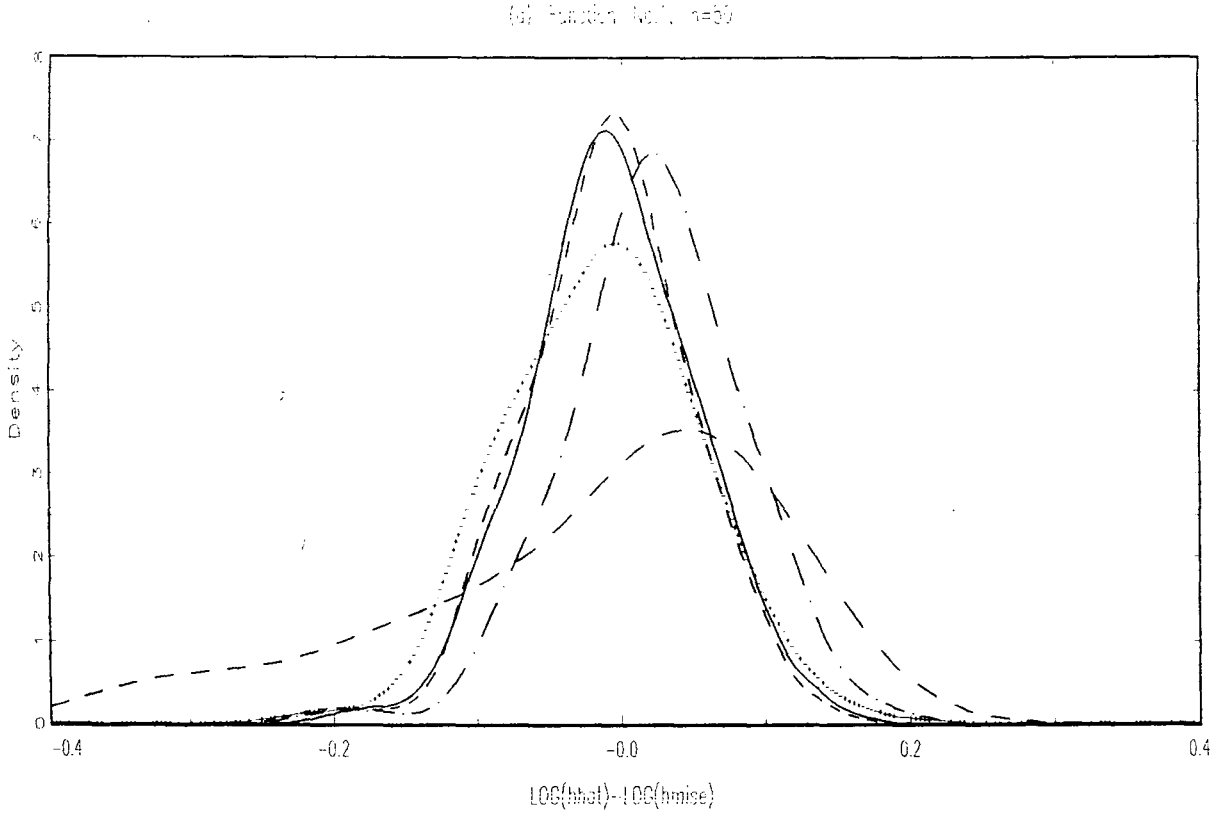
E-mail: siegfried.heiler@uni-konstanz.de

Table 2: *Average** **and** *Standard deviation*** **of Bandwidth Estimators in 300 Replications**

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

Table 3: $ASE(h_0)^*$ and $ASE(h_{ASE})^{**}$ of Bandwidth Estimators in 300 Replications

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



(b) Function No.1, n=100

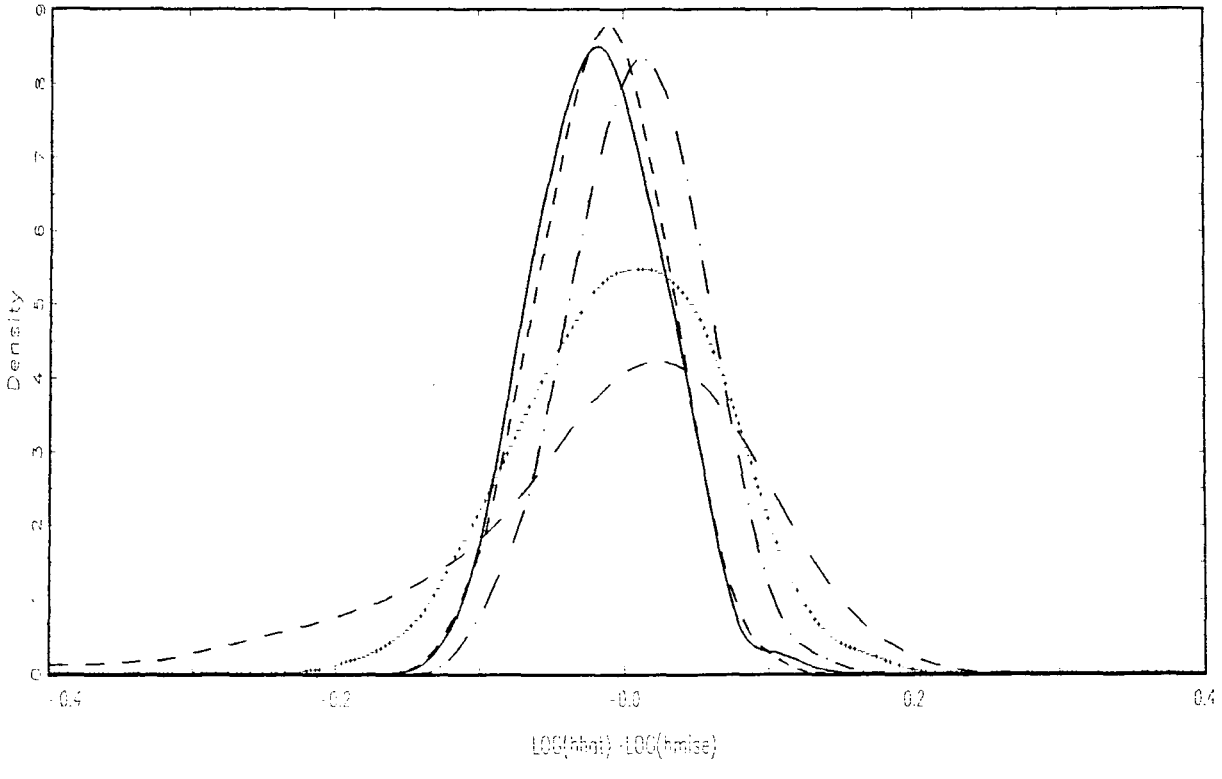
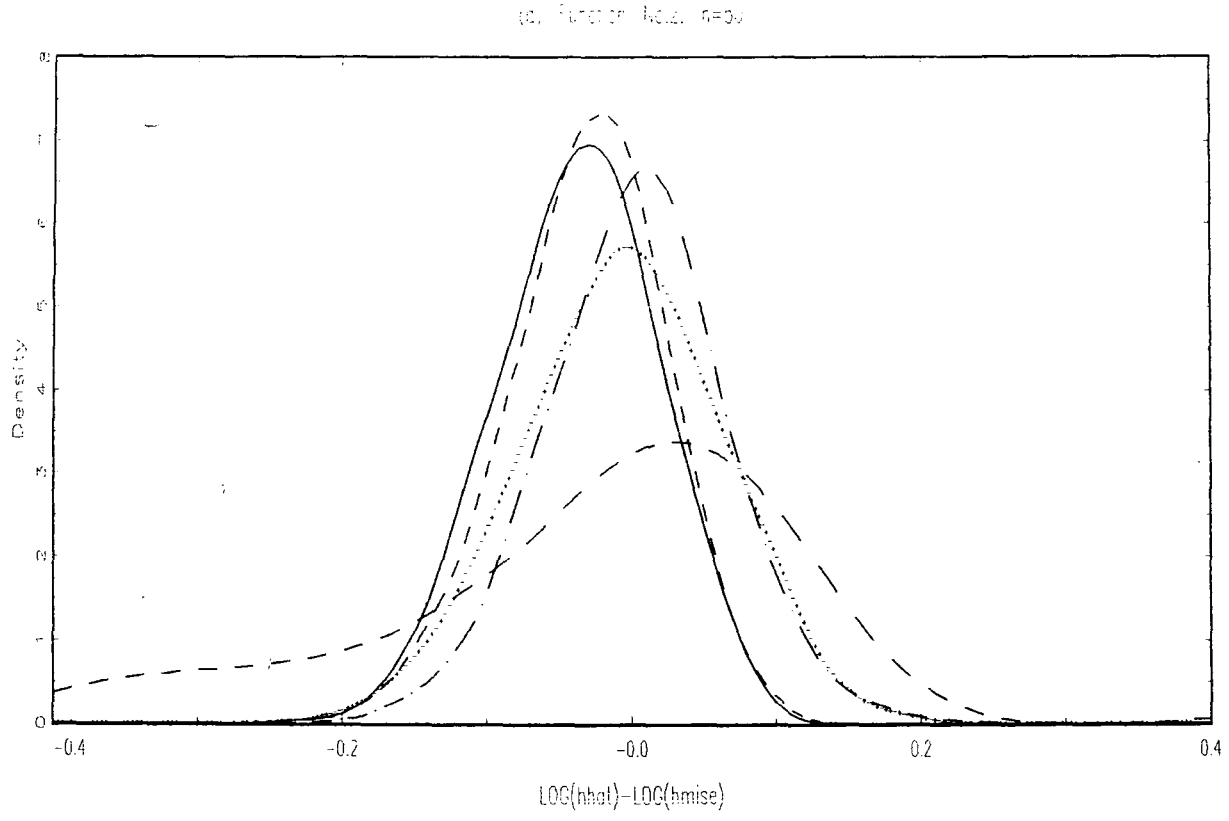


Figure 1. Kernel density estimates based on $log(\hat{h})$-$log(h_0)$ values for: $\hat{h}_{ASE}$(dotted), $\hat{h}_R$(dashed), $\hat{h}_{DS0}$(dots and dashes), $\hat{h}_{DS1}$(short dashes) and $\hat{h}_{DPI}$(solid line).



(b) Function No.2, n=100

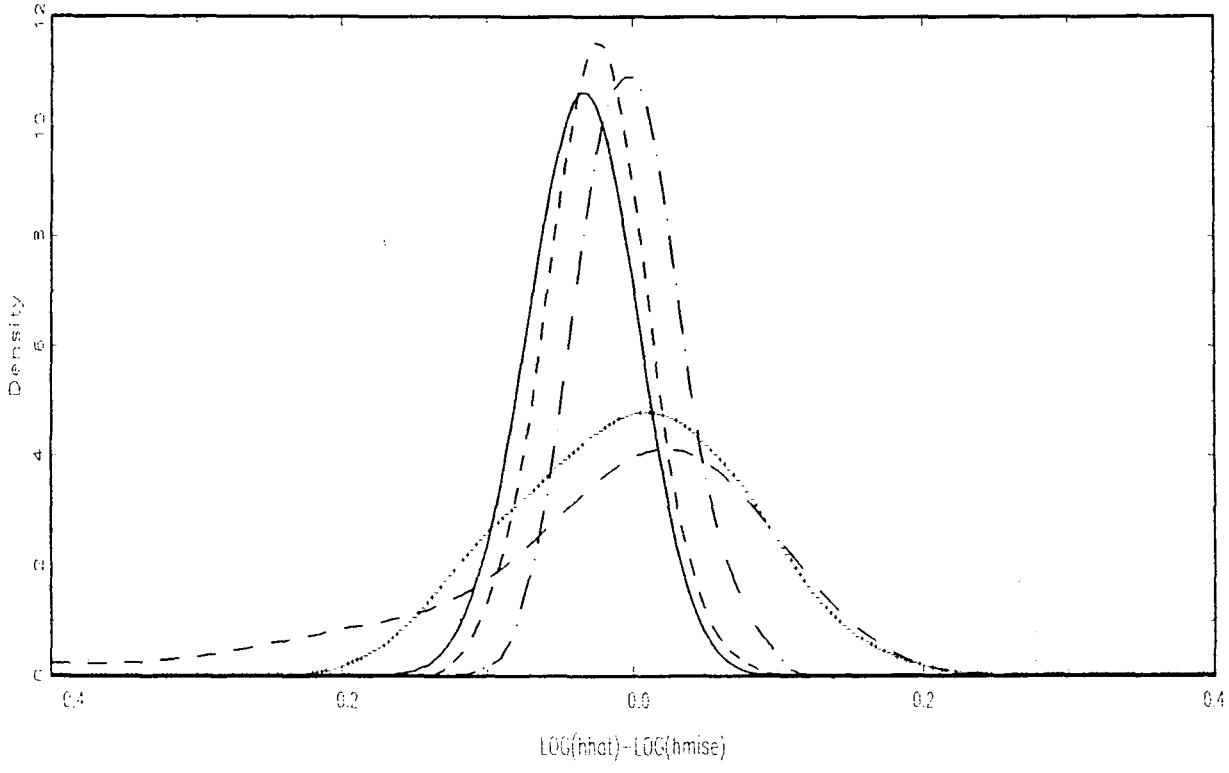
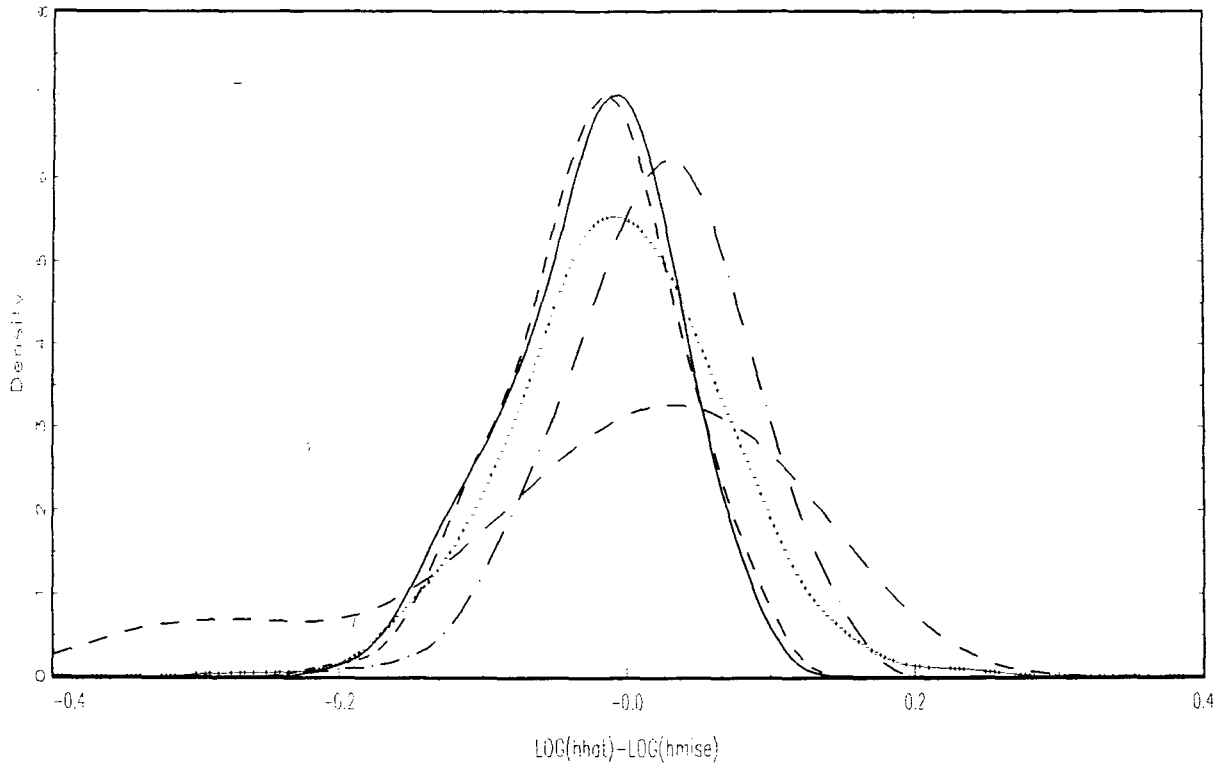


Figure 2. Kernel density estimates based on $log(\hat{h})$-$log(h_0)$ values for: h_{ASE}(dotted), h_R(dashed), h_{DS0}(dots and dashes), h_{DS1}(short dashes) and h_{DPI}(solid line).



(b) Function No.3, n=100

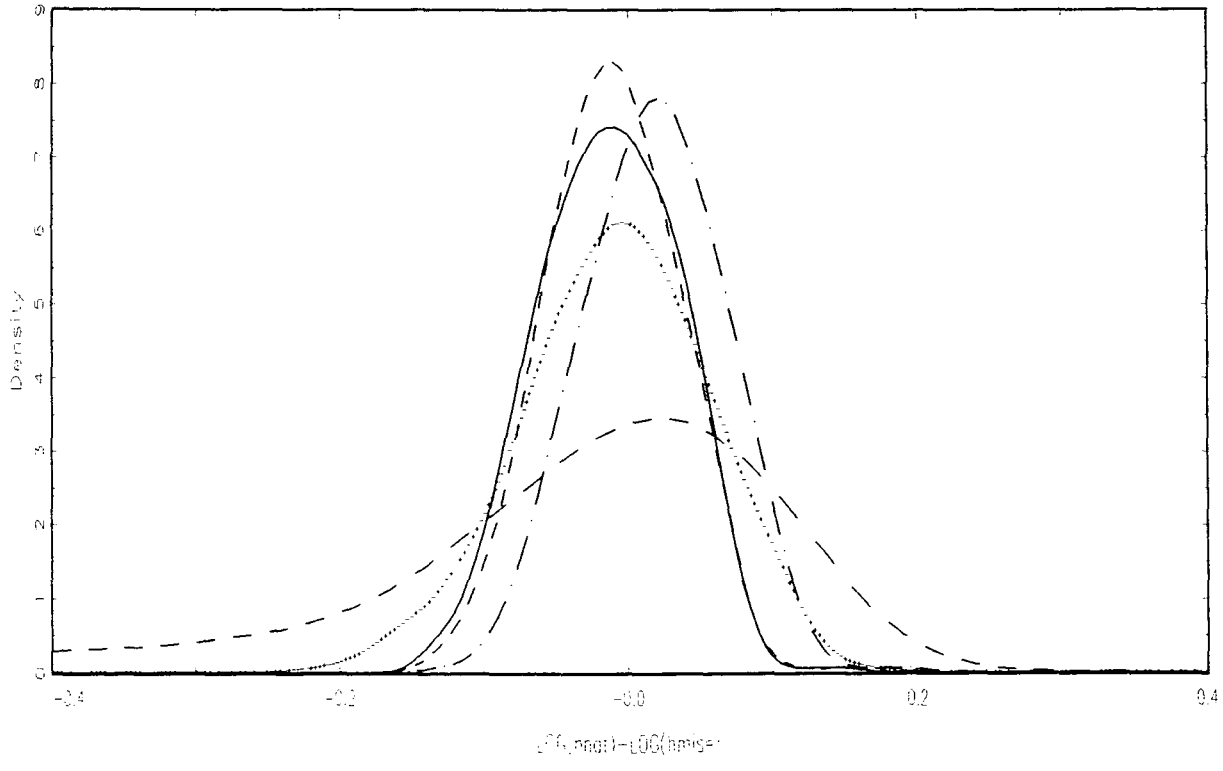


Figure 3. Kernel density estimates based on $log(\hat{h})$-$log(h_0)$ values for: h_{ASE}(dotted), h_R(dashed), h_{DS0}(dots and dashes), h_{DS1}(short dashes) and h_{DPI}(solid line).