Heiler, Siegfried; Feng, Yuanhua

Working Paper

A bootstrap bandwidth selector for local polynomial fitting

Diskussionsbeiträge - Serie II, No. 344

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

Suggested Citation: Heiler, Siegfried; Feng, Yuanhua (1997) : A bootstrap bandwidth selector for local polynomial fitting, Diskussionsbeiträge - Serie II, No. 344, Universität Konstanz, Sonderforschungsbereich 178 - Internationalisierung der Wirtschaft, Konstanz

This Version is available at:
https://hdl.handle.net/10419/101542

Sonderforschungsbereich 178
„Internationalisierung der Wirtschaft“

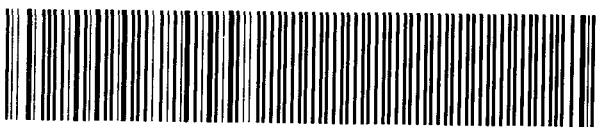
Diskussionsbeiträge

Juristische
Fakultät

Fakultät für Wirtschafts-
wissenschaften und Statistik

Siegfried Heiler
Yuanhua Feng

A Bootstrap Bandwidth Selector for Local Polynomial Fitting



D-78457 Konstanz

Serie II — Nr. 344
April 1997

A Bootstrap Bandwidth Selector for Local Polynomial Fitting

Siegfried Heiler

Yuanhua Feng

Serie II - Nr. 344

April 1997

A Bootstrap Bandwidth Selector for Local Polynomial Fitting

By

Siegfried Heiler Yuanhua Feng

University of Konstanz

Abstract

A bandwidth selector for local polynomial fitting is proposed following the bootstrap idea, which is just a double smoothing bandwidth selector with a bootstrap variance estimator, defined as the mean squared residuals of a pilot estimate. No simulated resampling is required in this context, since the needed expressions can be calculated explicitly. A simple, iterative data-driven procedure is proposed to estimate the variance and the bandwidth. A simulation study shows that this bandwidth selector performs very well, and it performs uniformly better than a double smoothing bandwidth selector using a difference-based variance estimator. The above mentioned bootstrap variance estimator is also a side result of this paper. It performs clearly better than the difference-based one. In a test example, the averaged squared error of this estimator in 500 replications achieved the theoretical lower bound already with a sample size of only $n = 200$.

Key Words: Local polynomial fitting, Bandwidth selection, Bootstrap, Double smoothing, Nonparametric variance estimation.

1 Introduction and motivation

Nonparametric regression using local polynomial fitting has proved to be a very attractive technique for functional estimation. The advantages of this approach include simplicity in terms of interpretability and mathematical analysis, ease of computation and of adaptation to various designs and a superior boundary behavior. See, for example, Cleveland and Devlin (1988), Fan (1992) and Ruppert and Wand (1994), and the monographs of Wand and Jones (1995) and Fan and Gijbels (1996) for recent contributions to the theory and practice in this setting. As with any nonparametric regression procedure, effective use of local polynomial regression requires the choice of some parameters, such as the polynomial degree p, the weight function (the kernel) $K(\cdot)$ and the bandwidth h. In this paper only the choice of the bandwidth h is investigated.

Bandwidth selection rules discussed in Rice (1984) and Härdle et al. (1988), the first generation methods following the terminology of Jones et al. (1996), have been shown to be subject to an unacceptably large amount of sample variability. Some more effective methods, the so-called second generation ones, have been proposed in recent years. Most of them may be found in Gasser et al. (1991), Chiu (1991), Ruppert et al. (1995) and Fan and Gijbels (1995). The latter two references are especially proposed for local polynomial fitting. Another bandwidth selection rule, proposed by Müller (1985) and Härdle et al. (1992) is the double smoothing (DS) criterion. A variate of their proposal was investigated by Heiler and Feng (1997) using a factorized pilot bandwidth. The second generation bandwidth selectors are far superior to the better known first generation ones (see the simulation studies in Chiu 1991, Herrmann 1994, and Heiler and Feng 1997).

An interesting idea for bandwidth selection in kernel density estimation is the smoothed bootstrap proposed by Taylor (1989) and Faraway and Jhun (1990). Developments in this direction may be found in Marron (1992) and Cao (1993). It was demonstrated in Cao, Cuevas and González-Manteiga (1994) that the smoothed bootstrap bandwidth selector shows a fairly satisfactory performance and could become one of the new standard methods for bandwidth selection in kernel density estimation (see also Jones et al. 1996). The goal of this paper is to develop a related bootstrap (BS) bandwidth selector for local polynomial fitting following the idea of bootstrap in nonparametric regression (Härdle and Bowman 1988). A curious feature of the bootstrapping in the context of bandwidth

selection is that no simulated resampling is required, since the needed functionals of the distribution can be calculated explicitly. The BS criterion is just a special case of the DS criterion, although the starting points are different. Hence this paper also provides a deeper understanding for the DS bandwidth selection rule.

The BS bandwidth selection rule requires pilot smoothings to estimate the variance and the bias. In this paper the R criterion (Rice 1984) is used as an initial method in order to obtain data-driven pilot estimates. The *residual squares criterion* (RSC) proposed by Fan and Gijbels (1995), for example, can also be used in this stage. The final proposal of these authors, the so-called *refined bandwidth selector* (RBS), is similar to the BS one. In this procedure the error criterion is also estimated using a pilot smoothing. A brief comparison between this proposal and the BS procedure can be found in section 8.

Some basic results for local polynomial fitting are described in section 2. The proposed bandwidth estimator is defined in section 3. Section 4 discusses the asymptotic properties of the bootstrap variance estimator. A simple data-driven DS procedure is proposed in section 5. Section 6 proposes an iterative data-driven procedure to estimate the variance and to select the BS bandwidth. Results of a simulation study are presented in section 7, while section 8 contains some conclusions.

2 Local polynomial fitting

Consider the regression model with a regular fixed design,

$$Y_i = m(x_i) + \epsilon_i, \qquad i = 1, ..., n,$$

where the ϵ_i's are i.i.d. random variables with $E(\epsilon_1) = 0$ and $\mathrm{var}\,(\epsilon_1) = \sigma^2$, the non-random design points x_1, ..., x_n are given according to a continuous design density $f>0$ on [0,1] through $x_i = F^{-1}\{(i - 0.5)/n\}$, where F^{-1} denotes the quantile function with respect to the density f.

Following Ruppert and Wand (1994) and Ruppert et al. (1995), the local polynomial estimator for $m(x)$ of degree p with bandwidth h and kernel $K(\cdot)$ is

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i,$$

with the weights

$$w_i(x) = \mathbf{e}_1'(\mathbf{X}_{p,x}'\mathbf{W}_x\mathbf{X}_{p,x})^{-1}\mathbf{X}_{p,x}'\mathbf{W}_x\mathbf{e}_i,$$

where

$$\mathbf{X}_{p,x} = \begin{bmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^p \end{bmatrix},$$

and $\mathbf{W}_x = diag\{K[(x_1-x)/h]/h, ..., K[(x_n-x)/h]/h\}$. $\mathbf{e}_j$ denotes a column vector having 1 as its jth entry and all other entries equal to zero. $K(\cdot)$ is assumed to be a symmetric density. It is also assumed that any polynomial degree used in this paper is odd to avoid the so-called "boundary effects". The mean averaged squared error (MASE)

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

is considered as a distance between $\hat{m}(x)$ and $m(x)$. It is well known that the MASE splits up into a variance part and a bias part, i.e. $M(h) = V + B$, where

$$V = V(h) = n^{-1}\sum_{i=1}^{n} \text{var}\,[\hat{m}(x_i)] = n^{-1}\sigma^2\sum_{i=1}^{n}\sum_{j=1}^{n} w_j(x_i)^2$$

and

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

Let h_0 denote the minimizer of MASE. The main point of this paper is to propose an effective data-driven estimator of h_0 following the bootstrap idea.

Useful notations for a function K are $\mu_k(K) = \int u^k K(u)du$ and $R(K) = \int K(u)^2 du$, assuming that the integrals converge. A well known approximation of h_0 is

$$h_{\text{AM}} = \left[\frac{(p+1)(p!)^2 R(K_p)\sigma^2}{2\mu_{p+1}(K_p)^2\theta_{p+1,p+1}n}\right]^{1/(2p+3)} =: c_0 n^{-1/(2p+3)}, \tag{2}$$

where $K_p(u) = \{|\mathbf{M}_p(u)|/|\mathbf{N}_p|\}K(u)$ is a kernel of order $p+1$, called an equivalent kernel of a local polynomial fitting, where $\mathbf{N}_p$ is the $(p+1)\times(p+1)$ matrix having (i,j) entry equal to $\int u^{i+j-2}K(u)du$ and $\mathbf{M}_p(u)$ is the same as $\mathbf{N}_p$, except that the first column is replaced by $(1, u, ..., u^p)'$. Further

$$\theta_{r,s} = \int_0^1 m^{(r)}(x)m^{(s)}(x)f(x)dx.$$

The next section is devoted to the bootstrap estimate of $\mathbf{M}(h)$.

3 Bootstrapping the MASE

In the sequel we define a bootstrap estimator of the MASE. To create a bootstrap estimate of the MASE we need pilot estimates for $m(\cdot)$. A bootstrap MASE estimator could be obtained from a single pilot estimate. However, in order to obtain a good estimate of it the degree of pilot smoothing for estimating the variance should not be at the same level as that for estimating the bias. Hence we define the bootstrap bias estimator with a pilot estimate and the bootstrap variance estimator with another pilot estimate. For bootstrap in nonparametric regression see Härdle and Bowman (1988). The weight functions and the degrees of polynomials are allowed to vary in different smoothing stages and will be given beforehand. Data-driven choices of the pilot bandwidths, g for estimating the bias, and g_v for estimating the variance, are discussed in section 5 and 6, respectively.

To obtain a bootstrap bias estimator we use the pilot estimate of the regression function, $\hat{m}_g(x) = \sum_{i=1}^{n} w_{ig}(x)Y_i$ with pilot bandwidth g, a polynomial of degree p_p and the kernel $K^p(\cdot)$. From this estimate we obtain the residuals $r_i = Y_i - \hat{m}_g(x_i)$. The residuals need not have mean 0, so, to let the resampled residuals reflect the behavior of the true observation errors, they should be first recentered as

$$\tilde{r}_i = Y_i - \hat{m}_g(x_i) - n^{-1}\sum_{j=1}^{n}\{Y_j - \hat{m}_g(x_j)\}.$$

The bootstrap observations are given by $Y_i^* = \hat{m}(x_i) + \epsilon_i^*$, where ϵ_i^* are the bootstrap residuals, which are created by sampling with replacement from $\{\tilde{r}_i\}$. A bootstrap estimator m^* of m, with h, p and K, is then obtained by smoothing $\{Y_i^*\}$ rather than $\{Y_i\}$. Now, the bias $b(x_i)$ could be estimated by $\hat{b}(x_i) = b^*(x_i)$, where $b^*(x_i)$ is the bootstrap bias (under the bootstrap distribution)

$$\begin{aligned} b^*(x_i) &= E^*[m^*(x_i)] - E[Y_i^*] \\ &= \sum_{j=1}^{n} w_j(x_i)\hat{m}_g(x_j) - \hat{m}_g(x_i), \end{aligned}$$

and the bias part of $M(h)$ could be estimated by

$$\begin{aligned} \hat{B} &= n^{-1}\sum_{i=1}^{n}\hat{b}(x_i)^2 \\ &= n^{-1}\sum_{i=1}^{n}\left\{\sum_{j=1}^{n} w_j(x_i)\hat{m}_g(x_j) - \hat{m}_g(x_i)\right\}^2. \end{aligned}$$

This is just the same as the bias part of the DS criterion with a fixed pilot bandwidth g (Müller 1985, Härdle et al. 1992).

To obtain a bootstrap variance estimator we use another pilot smoothing, $\hat{m}_{g_v}(x) = \sum_{i=1}^{n} w_{ig_v}(x)Y_i$, with the pilot bandwidth g_v, a polynomial of degree p_v and kernel $K^v(\cdot)$. Let $\hat{\epsilon}_i = Y_i - \hat{m}_{g_v}(x_i)$ denote the residuals. In this case the residuals do not need to be recentered. Let $Y_i^{\#} = \hat{m}(x_i) + \epsilon_i^{\#}$ denote the bootstrap observations, where $\epsilon_i^{\#}$ are the bootstrap residuals, which are created by sampling with replacement from $\{\hat{\epsilon}_i\}$. A bootstrap estimator $m^{\#}$ of m in this case, with h, p and K, is then obtained by smoothing $\{Y_i^{\#}\}$. The variance of the bootstrap estimate at x_i is given by

$$\text{var}^{\#}[m^{\#}(x_i)] = \hat{\sigma}_{\#}^2 \sum_{j=1}^{n} w_j(x_i)^2,$$

where

$$\begin{aligned} \hat{\sigma}_{\#}^2 &= n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2 - \left\{n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i\right\}^2 \\ &\simeq n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2 =: \hat{\sigma}_{\text{B}}^2. \end{aligned}$$

The bootstrap estimate of V is defined by

$$\hat{V} = n^{-1}\hat{\sigma}_{\text{B}}^2 \sum_{i=1}^{n}\sum_{j=1}^{n} w_j(x_i)^2, \tag{3}$$

where

$$\hat{\sigma}_{\text{B}}^2 = n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

is called a bootstrap variance estimator, which is just the mean squared residuals of the pilot estimate. We use the variance estimator $\hat{\sigma}_{\text{B}}^2$ instead of $\hat{\sigma}_{\#}^2$ for simplicity. The final bootstrap estimate of the MASE is given by

$$\hat{M}_{\text{BS}}(h) = \hat{V} + \hat{B}. \tag{4}$$

The BS bandwidth selector $\hat{h}_{\text{BS}}$ is defined as the minimizer of (4). Observe that $\hat{h}_{\text{BS}}$ is just a special case of the DS bandwidth selection rule proposed by Müller (1985) (see the case of $\Delta = 1$ in Heiler and Feng 1997). The criterion of Härdle et al. (1992) includes an additional term, i.e. the case of $\Delta = 0$ in Heiler and Feng (1997). In $\hat{M}_{\text{BS}}(h)$ instead of a general root n consistent variance estimator a special one, $\hat{\sigma}_{\text{B}}^2$, obtained from the

residuals of a pilot smoothing is used. The BS criterion is more natural. The form of $\hat{\sigma}_B^2$ is very simple. We will see that $\hat{\sigma}_B^2$ has good asymptotic properties in common cases and that it can easily be obtained with a data-driven procedure. Since $\hat{\sigma}_B^2$ is also root n consistent, the BS bandwidth selector has the same asymptotic properties as given in theorem 1 of Heiler and Feng (1997). Now, the pilot bandwidth g is fixed. The finite sample performance of this banwidth selector is better than a DS bandwidth selector with the difference-based variance estimator of Gasser et al. (1986), $\hat{\sigma}_G^2$, say.

4 Asymptotic properties of $\hat{\sigma}_B^2$

A nonparametric variance estimator based on residuals or differences has the property $E(\hat{\sigma}_B^2 - \sigma^2)^2 = n^{-1}\alpha \operatorname{var}(\epsilon^2) + r_v$, where $1 \leq \alpha < \infty$ and $r_v = o(n^{-1})$ (Hall and Marron 1990 and Hall et al. 1990). It will turn out that the $\hat{\sigma}_B^2$ defined above achieves $\alpha = 1$.

At first sight it seems that the estimator $\hat{\sigma}_B^2$ is not reasonable, since

$$E\left\{\sum_{i=1}^{n}[Y_i - \hat{m}_{g_v}(x_i)]^2\right\} = \nu\sigma^2,$$

whenever m is a polynomial of degree less than or equal to p_v. Here $\nu = n - 2\sum_{i=1}^{n} w_{ig_v}(x_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} w_{jg_v}(x_i)^2$. Hence Ruppert et al. (1995) proposed a variance estimator for local polynomial fitting,

$$\tilde{\sigma}^2 = \nu^{-1}\sum_{i=1}^{n}[Y_i - \hat{m}_{g_v}(x_i)]^2 = n\nu^{-1}\hat{\sigma}_B^2.$$

See also Hall and Marron (1990) for the same proposal in the context of kernel regression. These authors showed that $\tilde{\sigma}^2$ has fair asymptotic properties. A disadvantage of it is that the calculation involving a new bandwidth selection problem. The evaluation of $\hat{\sigma}_B^2$ also requires a bandwidth. However this is not a problem, because the approximate optimal bandwidth for evaluating $\hat{\sigma}_B^2$ is just h_{AM} given in (2) with $p = p_v$ multiplied by a known constant.

Denote the equivalent kernel for evaluating $\hat{\sigma}_B^2$ as $K_{p_v}(u)$, which is of order $k_v = p_v + 1$ (Ruppert and Wand 1994). Assume that m has at least k_v continuous derivatives, $g_v \to 0$ and $ng_v \to \infty$ as $n \to \infty$, and put

$$C_1 = \{\mu_{k_v}(K_{p_v})/(k_v)!\}^2\theta_{k_v,k_v}, \quad C_2 = 2K_{p_v}(0) - R(K_{p_v})$$

and

$$C_3 = 2\sigma^4 \int (K_{p_v} * K_{p_v} - 2K_{p_v})^2 du - 2C_2 \mathrm{var}\,(\epsilon^2),$$

where $*$ denotes convolution. Then we have

$$E[\hat{\sigma}_{\mathrm{B}}^2(g_v) - \sigma^2] = C_1 g_v^{2k_v} - C_2\sigma^2(ng_v)^{-1} + o\{g_v^{2k_v} + (ng_v)^{-1}\} \tag{5}$$

and

$$\mathrm{var}\,[\hat{\sigma}_{\mathrm{B}}^2(g_v)] = n^{-1}\mathrm{var}\,(\epsilon^2) + C_3(n^2 g_v)^{-1} + \mathrm{O}(n^{-1}g_v^{2k_v}) + o\{(n^2 g)^{-1}\}. \tag{6}$$

The proof of (5) and (6) is given in the appendix.

The dominant part of $E(\hat{\sigma}_{\mathrm{B}}^2 - \sigma^2)^2 - n^{-1}\mathrm{var}\,(\epsilon^2)$ is the square of the bias given in (5). It can easily be shown that, for a second order kernel $K(u)$ with restriction $K(u) \leq K(0)$, $C_2 \geq R(K) > 0$. This also holds for many higher order kernels (see table 1). Hence with the additional assumption $\theta_{k_v,k_v} \neq 0$, the asymptotically MSE (mean squared error) optimal chioce of g_v comes from a trading off between the first term and the second term in the bias part, such that $C_1 g_v^{2k_v} - C_2\sigma^2(ng_v)^{-1} = 0$, that is

$$\begin{aligned} g_{v,opt} &= \left\{\frac{[2K_{p_v}(0) - R(K_{p_v})]\sigma^2}{[\mu_{k_v}(K_{p_v})/(k_v)!]^2\theta_{k_v,k_v}}\right\}^{1/(2k_v+1)} n^{-1/(2k_v+1)} \\ &=: \mathrm{CF}h_{\mathrm{AM}}(p_v), \end{aligned} \tag{7}$$

where

$$\mathrm{CF} = \{2k_v[2K_{p_v}(0)/R(K_{p_v}) - 1]\}^{1/(2k_v+1)},$$

which does not depend on any unknown quantity. We see, that the only difference between h_{AM} and $g_{v,opt}$ is given by the known constant CF, called a correction factor. This is a very important property of $\hat{\sigma}_{\mathrm{B}}^2$. It allows us to obtain a data-driven estimate $\hat{g}_{v,opt}$ by maens of a bandwidth $\hat{h}$ selected for m. The data-driven estimation of $\hat{\sigma}_{\mathrm{B}}^2$ will be discussed in section 5. The use of a correction factor was also proposed by Müller et al. (1987) for selecting bandwidths for derivatives.

The values of $K_{p,\mu}(0)$, $R(K_{p,\mu})$, $2K_{p,\mu}(0)/R(K_{p,\mu}) - 1$ and $\mathrm{CF}(p,\mu)$ for kernels $K_{p,\mu}$, $p = 1,3,5$, $\mu = 0,1,2,3$ are given in table 1. The $K_{p,\mu}$ are obtained from the formula $K_{p,\mu}(u) = \{|\mathbf{M}_{p,\mu}(u)|/|\mathbf{N}_{p,\mu}|\}K_\mu(u)$, where $K_\mu(u) = K_{1,\mu}(u) = K_{1,\mu}(0)(1-u^2)^\mu$. See table 5.7 in Müller (1988) for the explicit forms of $K_{p,\mu}$. The corresponding quantities for the normal kernel ($p = 1$, $\mu = \infty$) are also given in table 1. From table 1 we see that the values of $\mathrm{CF}(p,\mu)$ are always larger than 1, i.e. $\hat{\sigma}_{\mathrm{B}}^2$ demands a bandwidth larger than h_{AM}. This is different for $\tilde{\sigma}^2$. Its optimal bandwidth is smaller than h_{AM} (Ruppert et al. 1995).

Table 1 Kernel based constants for selected kernels

p	μ	$K_{p,\mu}(0)$	$R(K_{p,\mu})$	$2K_{p,\mu}(0)/R(K_{p,\mu})-1$	$\mathrm{CF}(p,\mu)$
1	0	0.5000	0.5000	1.0000	1.3195
1	1	0.7500	0.6000	1.5000	1.4310
1	2	0.9375	0.7142	1.6250	1.4541
1	3	1.0938	0.8159	1.6812	1.4640
1	∞	$1/\sqrt{2\pi}$	$1/(2\sqrt{\pi})$	1.8284	1.4888
3	0	1.1250	1.1250	1.0000	1.2599
3	1	1.4063	1.2500	1.2500	1.2913
3	2	1.6406	1.4073	1.3315	1.3006
3	3	1.8457	1.5549	1.3740	1.3052
5	0	1.7578	1.7578	1.0000	1.2106
5	1	2.0508	1.8930	1.1667	1.2251
5	2	2.3071	2.0712	1.2278	1.2299
5	3	2.5378	2.2453	1.2624	1.2325

5 A practical DS procedure

The BS bandwidth selector is just a special case of the DS bandwidth selection rule. In this section we first propose a practical data-driven DS procedure for local polynomial fitting of odd order. Then we adapt it to the BS bandwidth selector in the next section. This procedure is very simple. It does not depend on asymptotic considerations and it does not involve estimation of derivatives. In this procedure a constant pilot bandwidth, $\hat{g}_{\mathrm{RGP}_p}$, which is selected by the R criterion (Rice 1984) using $\hat{\sigma}^2_{\mathrm{G}}$ (Gasser et al. 1986), is used. Here we use a slightly modified criterion, $\hat{R}(h) = \max\{\tilde{R}(h), \tilde{V}(h)\}$, also called an R criterion, where $\tilde{R}$ is the criterion of Rice (1984) and $\tilde{V}$ is as defined in (3), but with $\hat{\sigma}^2_{\mathrm{B}}$ replaced by $\hat{\sigma}^2_{\mathrm{G}}$, since a distance criterion should not be smaller than $\tilde{V}(h)$. The DS criterion in the following procedure means $M_{\mathrm{DS}} = \hat{B} + \tilde{V}$. The proposed method proceeds as follows (see also Müller 1985):

1. Estimate $\hat{\sigma}^2_{\mathrm{G}}$;
2. Estimate $m(\cdot)$ by fitting a local p_pth degree polynomial ($p_p > p$) with the bandwith, $\hat{g}_{\mathrm{RGP}_p}$ selected by the R criterion;

3. Minimize the DS criterion by means of the pilot estimate obtained in 2. w.r. to h, yielding the DS estimator $\hat{h}_{\mathrm{DSp}}$,

where p_p and p denote the polynomial degrees. Now, the equivalent kernels in the main stage and the pilot stage are K_p of order $r = p+1$ and K_{p_p} of order $s = p_p+1$, respectively. In fact, one can use any root n consistent variance estimator, $\hat{\sigma}^2$, in step 1. The rate of convergence of such a bandwidth selector depends only on r and s. Let $c_{0,s}$ denote the constant in the asymptotic optimal bandwidth in the pilot stage as given in (2), and let c_1, c_2 be the positive constants in the following approximation

$$\mathrm{MASE}''(h_0) \simeq c_1 n^{-1} h_0^{-3} \simeq c_2 h_0^{2p},$$

then the asymptotic properties of the DS bandwidth selector described above with any root n consistent variance estimator, $\hat{\sigma}^2$, are given by

Theorem 1: Under the assumptions

A1. K and L are compactly supported kernels, K' and $L^{(r+1)}$ are bounded.

A2. The design density f is continuous on $[0,1]$ and is bounded away from zero and bounded from above.

A3. Assume that $m^{(r+s)}$ is continuous on [0, 1].

Then

$$\begin{aligned}(\hat{h} - h_0)/h_0 &= \gamma_1(\hat{\sigma}^2 - \sigma^2) + (\gamma_2 n^{-(4s-4r+1)/(2s+1)} + \gamma_3 n^{-1})^{1/2} Z_n \\ &+ \gamma_4 n^{-s/(2s+1)}(1 + o(1)) + \gamma_5 n^{-2(s-r)/(2s+1)}(1 + o(1)),\end{aligned} \tag{8}$$

where Z_n is asymptotically normal $N(0,1)$, γ_1, ... ,γ_5 are constants, which are given by

$$\begin{aligned}\gamma_1 &= c_1^{-1} \int K^2(u)du, \\ \gamma_2 &= 4c_2^{-2} c_{0,s}^{-(4r+1)} r^2 \kappa_r^4 \sigma^4 \int \left[\int L^{(r)}(y) L^{(r)}(y+z)dy\right]^2 dz, \\ \gamma_3 &= 16 c_2^{-2} r^2 \kappa_r^4 \sigma^2 \theta_{r,r}, \\ \gamma_4 &= -4c_2^{-1} c_{0,s}^{s} r \kappa_r^2 \lambda_s \theta_{r,r+s}, \qquad \text{and} \\ \gamma_5 &= -2c_2^{-1} c_{0,s}^{-(2r+1)} r \sigma^2 \kappa_r^2 \int (L^{(r)}(u))^2 du.\end{aligned}$$

The proof of theorem 1 is omitted, since it is just a special case of theorem 1 in Heiler and Feng (1997), except that here a regular fixed design, not an equidistant one, is considered. We see that the dominant terms in the MSE of $(\hat{h} - h_0)/h_0$ are again the terms of the squared bias. The rate of convergence of such a bandwidth selector is the lower one of the two rates $n^{-s/(2s+1)}$ and $n^{-2(s-r)/(2s+1)}$. This rate is always lower than $n^{-1/2}$. Following case 1 in Heiler and Feng (1997), the pilot bandwidth $\hat{g}_{\mathrm{RGp}_p}$ is not optimal. However, the optimal chioce of the pilot bandwidth depends on asymptotic considerations and involves estimation of derivatives, but the proposal here does not.

The rates of convergence of the practical DS bandwidth selectors in common cases are:

1. $n^{-4/9}$ for $p = 1$, $p_p = 3$;
2. $n^{-6/13}$ for $p = 1$, $p_p = 5$;
3. $n^{-4/13}$ for $p = 3$, $p_p = 5$ and
4. $n^{-8/17}$ for $p = 3$, $p_p = 7$.

We see the rates of convergence of these simple bandwidth selectors are already very high. All the rates of convergence in 1., 2. and 4. are higher than $n^{-2/5}$, that is they are faster than a plug-in bandwidth selector based on h_{AM}. We are of the opinion that the DS (or BS) bandwidth selection rules are superior to the plug-in rule, because they aim at estimating the MASE directly and the data-driven procedure is very simple.

6 Data-driven BS bandwidth selection

In the sequel let RB stand for the R criterion with variance estimator $\hat{\sigma}^2_{\mathrm{B}}$ and BS stands for the bootstrap bandwidth selector. The key point to obtain a data-driven BS bandwidth selector is the problem of how to estimate $\hat{\sigma}^2_{\mathrm{B}}$. Hence we have only to develop a data-driven procedure for an estimator $\hat{\sigma}^2_{\mathrm{B}}$ with good finite sample properties. It is obvious that the performance of $\hat{\sigma}^2_{\mathrm{B}}$ depends strongly on the performance of $\hat{g}_v$. The dependence of $\hat{\sigma}^2_{\mathrm{B}}$ on the bandwidth is shown in figure 1(b). The curve shows the estimated variances against the bandwidths using the simulated data ($n = 200$) given in figure 1(a). They

were obtained from the regression function m_1 in the next section with $f(x) \equiv 1$ and standard normally distributed errors. Here and in the following $p_v = 1$ is used.

Insert figure 1 near here

The MSE of $\hat{\sigma}_B^2$ will be large if the MSE of $\hat{g}_v$ is large. $\hat{\sigma}_B^2$ will not be a good estimator if we simply use $\hat{g}_v = \mathrm{CF}\hat{h}_{\mathrm{RG1}}$ to calculate it, since the variation of $\hat{h}_{\mathrm{RG1}}$ is too large. The proposed procedure to calculate $\hat{\sigma}_B^2$ reads as follows: At first we obtain a DS bandwidth $\hat{h}_{\mathrm{DS1}}$ by using the procedure proposed in the last section with $p_p = 5$, $p = 1$ and $\hat{\sigma}_G^2$, which is then used for evaluating the bandwidth $\hat{g}_v = \mathrm{CF}\hat{h}_{\mathrm{DS1}}$ to calculate $\hat{\sigma}_B^2$. Now $\hat{\sigma}_B^2$ can be used in the DS procedure for $p = 1$ or $p = 3$ to obtain the BS bandwidth selector. This procedure can be iterated to obtain a more effective bandwidth selector.

In the simulation two BS bandwidth selectors $\hat{h}_{\mathrm{BS1}}$ and $\hat{h}_{\mathrm{BS3}}$ with $p_p = 5$ are considered. Other bandwidth selectors used for comparisons are $\hat{h}_{\mathrm{RGp}}$, $\hat{h}_{\mathrm{RBp}}$ and $\hat{h}_{\mathrm{DSp}}$, where p is equal to "1" or "3". Starting with "$i = 0$" the algorithm proceeds as follows:

1. If $i = 0$, calculate $\hat{\sigma}_G^2$ and put $\hat{\sigma}_{B,0}^2 = \hat{\sigma}_G^2$. Otherwise, let $\hat{g}_{v,i} = \mathrm{CF}\hat{h}_{1,i}$ and calculate $\hat{\sigma}_{B,i}^2$;

2. Select the bandwith with $p_p = 5$ by the R criterion with variance $\hat{\sigma}_{B,i}^2$ and fit $\hat{m}_p$;

3. Select the bandwidth $\hat{h}_{1,i}$ following the DS criterion with $\hat{m}_p$ and $\hat{\sigma}_{B,i}^2$. If $i > 0$ and $\hat{h}_{1,i} = \hat{h}_{1,i-1}$, put $\hat{h}_{\mathrm{DS1}} = \hat{h}_{1,0}$, $\hat{h}_{\mathrm{BS1}} = \hat{h}_{1,i}$, $\hat{\sigma}_B^2 = \hat{\sigma}_{B,i}^2$ and go to step 4. Otherwise, put $i := i + 1$ and go back to step 1;

4. Select the bandwidth $\hat{h}_{\mathrm{BS3}}$ following the DS criterion with $\hat{m}_p$ and $\hat{\sigma}_B^2$.

Other bandwidth estimators will also be calculated at the same time. This procedure does also not depend on any asymptotic considerations except for the constant CF.

The condition $\hat{h}_{1,i} = \hat{h}_{1,i-1}$ means that it is impossible to improve the variance estimator any more. Hence further iterations are not needed. If the number of iteritions (IS) is equal to 1, then $\hat{h}_{\mathrm{BS1}} = \hat{h}_{\mathrm{DS1}}$. In this case the gain of $\hat{\sigma}_B^2$ over $\hat{\sigma}_G^2$ is small, so that the selected bandwidth for $p = 1$ following the DS criterion cannot be improved by this procedure. However, even in this case there are also some gains for other bandwidth selectors calculated with $\hat{\sigma}_B^2$ over those with $\hat{\sigma}_G^2$.

7 Simulation results

In the simulations discussed here only an equidistant design, i.e. $f(x) \equiv 1$, with $x \in [0,1]$, was considered for simplicity. Hence the observations were taken at points $x_i = (i-0.5)/n$, $i = 1,2,...,n$. The bisquare kernel, $K(u) = 15(1-u^2)^2/16$, was used everywhere as weight function for local polynomial fitting. In this case CF=1.454. In this section the word "bandwidth" means the bandwidth of the observation indices. And we only consider integral bandwidths for simplicity. Now, the bandwidth is h, $h < n/2$, means that observations Y_i with $i \in [i_0 - h, i_0 + h]$, i.e. $x_i \in [x_{i_0} - h/n, x_{i_0} + h/n] \cap [0,1]$, are used to estimate the regression function at a point x_{i_0}.

The pattern regression functions were:

$$\begin{aligned} m_1(x) &= 2 - 5x + 5exp[-100(x-0.5)^2], \\ m_2(x) &= \sin(6\pi x) \qquad\qquad\qquad\qquad\qquad\qquad and \\ m_3(x) &= 8x + 8exp[-65(x-0.5)^2] + 6sin(3\pi x). \end{aligned}$$

m_1 is a linear regression function with Gaussian peak (Gasser et al. 1991). m_2 is a simple sine function. m_3 is a combination of the sine function and the linear regression function with Gaussian peak. Independently, normally distributed errors, $\epsilon(i) = \sigma_i \epsilon$, $i = 1,2,3$, were used, where ϵ is standard normally distributed and $\sigma_1 = 1$ for m_1, $\sigma_2 = 0.5$ for m_2, $\sigma_3 = 2$ for m_3, respectively. We used the sample sizes $n = 100$ and $n = 200$. The bandwidth for calculating $\hat{\sigma}_B^2$ in the ith iteration is $\hat{g}_{v,i} = [1.454\hat{h}_{1,i} + 0.5]$, where $[x]$ denotes the largest integer less than or equal to x.

Recall that IS=1 means $\hat{h}_{DS1} = \hat{h}_{BS1}$. We divide all 500 replications in each case into two groups: group A: replications with IS=1 and group B: replications with IS> 1. Besides the numerical results of the 500 replications, the same results for these two groups are also given. We will see that the performances of $\hat{\sigma}_G^2$, $\hat{h}_{RGp}$ and $\hat{h}_{DSp}$ in group A and group B are often very different, but the performances of $\hat{\sigma}_B^2$, $\hat{h}_{RBp}$ and $\hat{h}_{BSp}$ in group A and group B are quite similar.

We compare at first the performances of the two variance estimators, $\hat{\sigma}_G^2$ and $\hat{\sigma}_B^2$. The results are exhibited in table 2. In each case the variance of $\hat{\sigma}_B^2$ is smaller than that of $\hat{\sigma}_G^2$. $\hat{\sigma}_B^2$ tends slightly towards under estimation but performs better than $\hat{\sigma}_G^2$ in both groups. The performance of $\hat{\sigma}_B^2$ in group B is quite similar to, and often a little better than that in

Table 2

The numerical results of $\hat{\sigma}_G^2$ and $\hat{\sigma}_B^2$ in 500 replications

	n	Results of all repl.			Results of group A			Results of group B		
		Aver.	Var.	ASE	Aver.	Var.	ASE	Aver.	Var.	ASE
m_1										
$\hat{\sigma}_G^2$	100	1.00	4.44e-2	4.44e-2	0.93	2.76e-2	3.28e-2	1.08	5.02e-2	5.63e-2
	200	1.00	1.81e-2	1.81e-2	0.95	1.26e-2	1.52e-2	1.03	1.95e-2	2.03e-2
$\hat{\sigma}_B^2$	100	0.96	2.51e-2	2.65e-2	0.95	2.46e-2	2.73e-2	0.98	2.51e-2	2.56e-2
	200	0.98	1.06e-2	1.11e-2	0.98	1.25e-2	1.29e-2	0.98	9.14e-3	9.76e-3
m_2										
$\hat{\sigma}_G^2$	100	.254	2.45e-3	2.46e-3	.240	1.80e-3	1.90e-3	.280	2.64e-3	3.57e-3
	200	.250	1.14e-3	1.14e-3	.240	7.60e-4	8.62e-4	.261	1.33e-3	1.46e-3
$\hat{\sigma}_B^2$	100	.240	1.29e-3	1.39e-3	.239	1.35e-3	1.47e-3	.242	1.16e-3	1.23e-3
	200	.244	5.87e-4	6.24e-4	.244	6.17e-4	6.53e-4	.244	5.53e-4	5.89e-4
m_3										
$\hat{\sigma}_G^2$	100	3.97	6.15e-1	6.16e-1	3.84	5.34e-1	5.60e-1	4.15	6.68e-1	6.89e-1
	200	4.01	3.12e-1	3.12e-1	3.81	2.20e-1	2.57e-1	4.13	3.29e-1	3.46e-1
$\hat{\sigma}_B^2$	100	3.82	3.54e-1	3.86e-1	3.85	3.76e-1	4.00e-1	3.79	3.22e-1	3.68e-1
	200	3.91	1.60e-1	1.69e-1	3.89	1.72e-1	1.83e-1	3.92	1.53e-1	1.60e-1

ASE = averaged squared error

group A. However, the performance of $\hat{\sigma}_G^2$ in group B is clearly worse than that in group A. The variance of $\hat{\sigma}_G^2$ in group B was very large, and so was the ASE (averaged squared error) of it. The bias of $\hat{\sigma}_G^2$ in group A was always negative and it was always positive in group B, although $\hat{\sigma}_G^2$ itself was almost unbiased. $\hat{\sigma}_B^2$ is a very stable variance estimator. When $\hat{\sigma}_G^2$ performs well, i.e. in group A, $\hat{\sigma}_B^2$ performs still a little better than $\hat{\sigma}_G^2$. When $\hat{\sigma}_G^2$ performs bad, i.e. in group B, the performance of $\hat{\sigma}_B^2$ did not worsen but also improved slightly. Figure 2 shows kernel density estimates for $\hat{\sigma}_G^2$ and $\hat{\sigma}_B^2$ in 500 replications, while the kernel density estimates for $\hat{\sigma}_G^2$ and $\hat{\sigma}_B^2$ in group A and group B for m_1 with $n = 100$ are presented in figure 3.

Table 3

The empirical efficiencies of $\hat{\sigma}^2_G$, $\hat{\sigma}^2_B$, gain of $\hat{\sigma}^2_B$ over $\hat{\sigma}^2_G$ (%)

Function	m_1		m_2		m_3	
n	100	200	100	200	100	200
Eff($\hat{\sigma}^2_G$)	45	55	51	55	52	51
Eff($\hat{\sigma}^2_B$)	76	90	91	100	83	95
Gain($\hat{\sigma}^2_B$)	40	39	44	45	37	46

Insert figure 2 and figure 3 near here

The asymptotic lower bound of the MSE for a variance estimator is $n^{-1}\text{var}(\epsilon^2)$, i.e. $2\sigma^4/n$ in the case of normally distributed ϵ's. The empirical efficiencies, Eff($\hat{\sigma}^2$):= nASE($\hat{\sigma}^2$)/ $(2\sigma^4)\times 100$, of these two estimators and the empirical gain of $\hat{\sigma}^2_B$ over $\hat{\sigma}^2_G$, (1-ASE($\hat{\sigma}^2_B$)/ASE($\hat{\sigma}^2_G$))$\times 100$, are listed in table 3. The asymptotic efficiency of $\hat{\sigma}^2_G$ is 51% (Hall et al. 1990) and it is 100% for $\hat{\sigma}^2_B$. However, following the exact finite sample analysis of Seifert et al. (1993), the efficiency of $\hat{\sigma}^2_G$ in finite samples is sometimes larger than its asymptotic efficiency. In the simulation one half of the empirical efficiencies of $\hat{\sigma}^2_G$ were larger than 51%. The largest one was 55% for m_1 and m_2 with $n = 200$. $\hat{\sigma}^2_B$ performs not only very well in theory, but also in practice. In any case the practical performance of $\hat{\sigma}^2_B$ was clearly better than that of $\hat{\sigma}^2_G$. The smallest empirical efficiency of $\hat{\sigma}^2_B$ was 76% in the case of m_1 with $n = 100$. This is much larger than the largest value for $\hat{\sigma}^2_G$. The asymptotic gain of $\hat{\sigma}^2_B$ over $\hat{\sigma}^2_G$ should be 49%. Although the empirical gain in each case was smaller than the asymptotic one, it was already very large. The smallest value was 37%. The empirical efficiency of $\hat{\sigma}^2_B$ increased as n increased, and they also depended on the regression function. In the case of m_2 with $n = 200$, the empirical efficiency of $\hat{\sigma}^2_B$ already achieved 100%. Despite its worse performance, $\hat{\sigma}^2_G$ plays an important role in our procedure, since it is the initial quantity for calculating $\hat{\sigma}^2_B$.

Now, let us look at the practical performances of the bandwidth selectors. Although our proposal is to use only the BS bandwidth selector, we will discuss all the bandwidth selectors mentioned in section 6 in order to understand, how the variance estimators and the error criteria influence the performances of the bandwidth selectors. The true optimal

Table 4 The true bandwidths h_0 of all cases

	m_1				m_2				m_3			
n	100		200		100		200		100		200	
p	1	3	1	3	1	3	1	3	1	3	1	3
h_0	8	17	14	31	6	16	12	30	10	22	17	39

bandwidths in all cases are given in table 4. The numerical results for the bandwidth selectors in 500 replications, the results in group A and group B are exhibited in tables 5–7, for m_1, m_2 and m_3, respectively. Kernel density estimates of these bandwidth selectors in 500 replications are presented in figures 4–6, respectively.

The bandwidth selectors following the R criterion with $\hat{\sigma}_{\mathrm{G}}^2$, $\hat{h}_{\mathrm{RGp}}$, were the worst ones in all cases. Like $\hat{\sigma}_{\mathrm{G}}^2$, the performance of $\hat{h}_{\mathrm{RGp}}$ in group B was much worse than that in group A. All of the variances and the biases of $\hat{h}_{\mathrm{RGp}}$ in group B were much larger than those in group A. The biases of $\hat{h}_{\mathrm{RGp}}$ in group A were small. However their biases in group B were positive and very large. The positive bias of $\hat{\sigma}_{\mathrm{G}}^2$ in group B was often inflated due to the nonlinear relationship between $\hat{\sigma}^2$ and $\hat{h}$. Although $\hat{h}_{\mathrm{RGp}}$ has disadvantages, the simulation shows that it is a good starting point of the DS as well as BS procedures.

By using the DS criterion, the performance of the bandwidth selectors was clearly improved. The gains of $\hat{h}_{\mathrm{DSp}}$ over $\hat{h}_{\mathrm{RGp}}$ following the criterion of ASE in 500 replications were very large (see table 8). One half of them were larger than 90%. The smallest one was 78%. The gain of $\hat{h}_{\mathrm{DS1}}$ over $\hat{h}_{\mathrm{RG1}}$ was often larger than that of $\hat{h}_{\mathrm{DS3}}$ over $\hat{h}_{\mathrm{RG3}}$, since the rate of convergence of $\hat{h}_{\mathrm{DS1}}$ is higher. Sometimes the DS bandwidth selectors using $\hat{\sigma}_{\mathrm{G}}^2$ also shared the drawback of $\hat{\sigma}_{\mathrm{G}}^2$. That is, the performances of these bandwidth selectors in group B were often much worse than those in group A, and their biases in group B were sometimes very large, which were not large in group A. The biases of $\hat{h}_{\mathrm{DSp}}$ mainly come from replications in group B.

The performance of a DS bandwidth estimator was further improved by using the BS bandwidth selection rule, that is, $\hat{\sigma}_{\mathrm{G}}^2$ is replaced by $\hat{\sigma}_{\mathrm{B}}^2$ in the DS criterion. This improvement comes from the improvement of the variance estimation. Note that only the finite sample performance of a bandwidth selector can be improved by using a more

Table 5 The numerical results of the bandwidth selectors for m_1

n	$\hat{h}$	Results of all repl.			Results of group A			Results of group B		
		Aver.	Var.	ASE	Aver.	Var.	ASE	Aver.	Var.	ASE
100	$\hat{h}_{\mathrm{RG1}}$	9.7	15.7	18.5	7.8	3.97	4.04	11.6	20.2	33.3
	$\hat{h}_{\mathrm{RB1}}$	8.1	2.54	2.56	7.9	2.75	2.77	8.4	2.16	2.35
	$\hat{h}_{\mathrm{DS1}}$	8.6	2.14	2.55	7.9	.932	.941	9.4	2.26	4.20
	$\hat{h}_{\mathrm{BS1}}$	8.1	.917	.920	7.9	.932	.941	8.2	.858	.899
	$\hat{h}_{\mathrm{RG3}}$	19.7	48.8	56.3	16.2	14.5	15.1	23.3	58.5	98.5
	$\hat{h}_{\mathrm{RB3}}$	16.8	9.13	9.17	16.3	9.75	10.3	17.3	7.93	8.04
	$\hat{h}_{\mathrm{DS3}}$	18.7	9.41	12.5	17.2	4.12	4.14	20.4	9.56	21.0
	$\hat{h}_{\mathrm{BS3}}$	17.3	4.31	4.42	17.1	4.11	4.12	17.6	4.35	4.73
200	$\hat{h}_{\mathrm{RG1}}$	17.1	40.1	49.9	13.4	7.31	7.69	19.9	46.3	81.4
	$\hat{h}_{\mathrm{RB1}}$	14.1	5.02	5.03	13.8	4.95	4.97	14.3	4.99	5.08
	$\hat{h}_{\mathrm{DS1}}$	15.0	3.72	4.67	13.9	1.32	1.33	15.8	4.01	7.16
	$\hat{h}_{\mathrm{BS1}}$	14.0	1.34	1.34	13.9	1.32	1.33	14.1	1.35	1.35
	$\hat{h}_{\mathrm{RG3}}$	35.9	130	154	29.2	33.8	37.0	40.9	145	242
	$\hat{h}_{\mathrm{RB3}}$	30.1	20.2	20.9	29.8	23.1	24.6	30.4	17.8	18.1
	$\hat{h}_{\mathrm{DS3}}$	33.3	19.5	24.8	30.6	7.76	7.90	35.3	18.9	37.5
	$\hat{h}_{\mathrm{BS3}}$	30.9	7.34	7.37	30.7	7.43	7.50	30.9	7.27	7.27

efficient variance estimator. Neither the rate of convergence nor the constants of the dominant terms could be improved, if the bandwidth selector is not a root n one, such as $\hat{h}_{\mathrm{DSp}}$. Hence $\hat{h}_{\mathrm{BSp}}$ and $\hat{h}_{\mathrm{DSp}}$ have the same rate of convergence for given p. The drawback of $\hat{h}_{\mathrm{DSp}}$ could be overcome by using $\hat{\sigma}_{\mathrm{B}}^2$. The performances of $\hat{h}_{\mathrm{BSp}}$ in group B were better than those in group A, except for two cases of m_1 with $n = 100$, $p = 3$ and $n = 200$, $p = 1$, in which the performances of $\hat{h}_{\mathrm{BSp}}$ in group B were slightly worse than those in group A. The variances of $\hat{h}_{\mathrm{BSp}}$ were clearly smaller than those of $\hat{h}_{\mathrm{DSp}}$. The biases of them were sometimes larger than the biases of $\hat{h}_{\mathrm{DSp}}$. Following the criterion ASE in 500 replications $\hat{h}_{\mathrm{BSp}}$ perform uniformly better than $\hat{h}_{\mathrm{DSp}}$ (see table 8). The gains of $\hat{h}_{\mathrm{BSp}}$ over $\hat{h}_{\mathrm{DSp}}$ depended strongly on the regression function and changed from case to case. The largest one was 71%, but the smallest one was only 6%. When $\hat{h}_{\mathrm{DSp}}$ performed very well, the gains were often small, since now only a small gap remained for further improvement.

Table 6 The numerical results of the bandwidth selectors for m_2

n	$\hat{h}$	Results of all repl.			Results of group A			Results of group B		
		Aver.	Var.	ASE	Aver.	Var.	ASE	Aver.	Var.	ASE
100	$\hat{h}_{\mathrm{RG1}}$	7.2	6.03	7.51	6.3	3.20	3.27	9.1	6.15	15.8
	$\hat{h}_{\mathrm{RB1}}$	6.3	.995	1.07	6.1	1.12	1.13	6.6	.625	.941
	$\hat{h}_{\mathrm{DS1}}$	6.6	.539	.848	6.3	.356	.438	7.1	.478	1.65
	$\hat{h}_{\mathrm{BS1}}$	6.3	.304	.372	6.3	.356	.438	6.2	.200	.243
	$\hat{h}_{\mathrm{RG3}}$	17.6	24.4	27.0	15.8	15.3	15.3	21.2	23.1	49.7
	$\hat{h}_{\mathrm{RB3}}$	15.3	4.82	5.27	15.2	5.39	5.99	15.5	3.64	3.86
	$\hat{h}_{\mathrm{DS3}}$	16.2	2.35	2.38	15.7	2.01	2.11	17.2	1.56	2.92
	$\hat{h}_{\mathrm{BS3}}$	15.5	1.35	1.59	15.5	1.57	1.86	15.6	.889	1.04
200	$\hat{h}_{\mathrm{RG1}}$	13.2	21.1	22.6	11.0	5.72	6.74	15.8	26.6	40.9
	$\hat{h}_{\mathrm{RB1}}$	11.3	2.36	2.87	11.1	2.46	3.21	11.5	2.18	2.47
	$\hat{h}_{\mathrm{DS1}}$	11.6	.880	1.03	11.2	.362	1.00	12.1	1.05	1.06
	$\hat{h}_{\mathrm{BS1}}$	11.2	.348	.920	11.2	.362	1.00	11.3	.328	.828
	$\hat{h}_{\mathrm{RG3}}$	33.2	89.7	100	28.5	38.6	40.9	38.7	92.7	169
	$\hat{h}_{\mathrm{RB3}}$	28.7	13.8	15.6	28.4	14.6	17.0	28.9	12.8	14.0
	$\hat{h}_{\mathrm{DS3}}$	29.8	4.68	4.73	28.8	2.70	4.05	30.9	4.69	5.50
	$\hat{h}_{\mathrm{BS3}}$	28.8	2.18	3.74	28.8	2.21	3.75	28.7	2.15	3.73

insert figures 4–6 near here

From the asymptotic point of view the $\hat{h}_{\mathrm{RBp}}$ are not good bandwidth selectors, since they converge with the same rate as $\hat{h}_{\mathrm{RGp}}$. They are given here in order to show the influences of variance estimators on bandwidth selection. The finite sample performances of $\hat{h}_{\mathrm{RBp}}$ were much better than those of $\hat{h}_{\mathrm{RGp}}$. The gains of $\hat{h}_{\mathrm{RBp}}$ over $\hat{h}_{\mathrm{RGp}}$ were large. The smallest one was 80%. We see that the bad performances of $\hat{h}_{\mathrm{RGp}}$ were largely due to the variance estimator. In any case the gains of $\hat{h}_{\mathrm{RBp}}$ over $\hat{h}_{\mathrm{RGp}}$ were larger than the gains of $\hat{h}_{\mathrm{BSp}}$ over $\hat{h}_{\mathrm{DSp}}$. This shows that, if the bandwidth selection rule is not good, the influence of the variance estimator on the bandwidth selection is large. For m_1 and m_3, $\hat{h}_{\mathrm{RB3}}$ performs even better than $\hat{h}_{\mathrm{DS3}}$, despite its lower rate of convergence.

$\hat{h}_{\mathrm{BSp}}$ and $\hat{h}_{\mathrm{RBp}}$ are bandwidth selectors using the same variance estimator but different

Table 7 The numerical results of the bandwidth selectors for m_3

n	$\hat{h}$	Results of all repl.			Results of group A			Results of group B		
		Aver.	Var.	ASE	Aver.	Var.	ASE	Aver.	Var.	ASE
100	$\hat{h}_{\mathrm{RG1}}$	11.8	34.5	37.7	9.9	17.6	17.6	14.2	45.7	63.6
	$\hat{h}_{\mathrm{RB1}}$	9.5	4.04	4.34	9.3	4.57	5.07	9.7	3.27	3.39
	$\hat{h}_{\mathrm{DS1}}$	9.7	1.93	2.01	9.4	1.66	2.04	10.1	1.95	1.97
	$\hat{h}_{\mathrm{BS1}}$	9.4	1.47	1.88	9.4	1.66	2.04	9.3	1.21	1.69
	$\hat{h}_{\mathrm{RG3}}$	27.2	149	177	24.1	108	112	31.3	175	260
	$\hat{h}_{\mathrm{RB3}}$	21.8	27.4	27.5	22.3	31.7	31.8	21.2	21.2	21.9
	$\hat{h}_{\mathrm{DS3}}$	24.5	30.8	37.3	23.3	29.0	30.6	26.2	28.4	46.0
	$\hat{h}_{\mathrm{BS3}}$	22.6	21.5	21.8	23.1	24.7	25.9	21.9	16.5	16.5
200	$\hat{h}_{\mathrm{RG1}}$	22.4	88.7	118	17.1	20.6	20.6	25.7	102	179
	$\hat{h}_{\mathrm{RB1}}$	17.2	8.18	8.21	17.2	8.55	8.61	17.1	7.94	7.95
	$\hat{h}_{\mathrm{DS1}}$	17.5	3.62	3.82	16.6	2.64	2.77	18.0	3.56	4.48
	$\hat{h}_{\mathrm{BS1}}$	16.6	2.30	2.47	16.6	2.64	2.77	16.6	2.09	2.29
	$\hat{h}_{\mathrm{RG3}}$	51.7	481	644	39.5	138	138	59.3	544	958
	$\hat{h}_{\mathrm{RB3}}$	38.5	50.3	50.6	38.8	62.6	62.6	38.3	42.6	43.1
	$\hat{h}_{\mathrm{DS3}}$	43.8	69.7	92.4	39.3	43.0	43.1	46.5	66.3	123
	$\hat{h}_{\mathrm{BS3}}$	38.9	31.0	31.0	39.3	37.2	37.4	38.6	26.9	27.0

Table 8 Empirical gains (%) of bandwidth selectors following ASE

Function	m_1				m_2				m_3			
n	100		200		100		200		100		200	
p	1	3	1	3	1	3	1	3	1	3	1	3
$\hat{h}_{\mathrm{DSp}}$ over $\hat{h}_{\mathrm{RGp}}$	86	78	91	84	89	91	95	95	95	79	97	86
$\hat{h}_{\mathrm{BSp}}$ over $\hat{h}_{\mathrm{DSp}}$	64	65	71	70	56	33	10	21	6	42	35	66
$\hat{h}_{\mathrm{RBp}}$ over $\hat{h}_{\mathrm{RGp}}$	86	84	90	86	86	80	87	84	88	84	93	92
$\hat{h}_{\mathrm{BSp}}$ over $\hat{h}_{\mathrm{RBp}}$	64	52	73	65	65	70	68	76	57	21	70	39
$\hat{h}_{\mathrm{BSp}}$ over $\hat{h}_{\mathrm{RGp}}$	95	92	97	95	95	94	96	96	95	88	98	95

Table 9

ASE of bandwidth selectors in replications, when $|\hat{\sigma}_B^2 - \sigma^2| > |\hat{\sigma}_G^2 - \sigma^2|$

	m_1				m_2				m_3			
n	100		200		100		200		100		200	
p	1	3	1	3	1	3	1	3	1	3	1	3
$\hat{h}_{RGp}$	9.30	30.4	31.1	98.5	4.71	17.4	12.7	73.2	20.4	132	87.2	514
$\hat{h}_{RBp}$	2.24	8.62	4.70	22.0	.995	5.62	2.63	14.3	3.51	20.2	7.55	49.0
$\hat{h}_{DSp}$	1.53	7.19	2.78	16.2	.608	1.66	.747	3.48	1.40	27.7	2.63	76.8
$\hat{h}_{BSp}$	1.02	3.93	1.46	8.04	.344	1.84	1.02	4.40	2.01	16.5	2.43	30.6

error criteria. The gains of $\hat{h}_{BSp}$ over $\hat{h}_{RBp}$, i.e. the gains from using the DS criterion, when $\hat{\sigma}_B^2$ is used, were always smaller than those of $\hat{h}_{DSp}$ over $\hat{h}_{RGp}$. That is, if the variance estimator is improved, the gains by using the DS criterion will be reduced. In any case $\hat{h}_{BSp}$ turned out to be the best ones. The gains of $\hat{h}_{BSp}$ over $\hat{h}_{RGp}$ were all larger than 90%, except for the case of m_3 with $n = 100$, $p = 3$, in which it was 88%. The gains increased as n increased. Both, the use of DS criterion and the use of $\hat{\sigma}_B^2$ contributed to these gains. Although the use of the DS criterion plays a more important role, sometimes also the use of $\hat{\sigma}_B^2$ is important.

Furthermore, the bandwidth used to evaluate $\hat{\sigma}_B^2$ depends on the variation in a sample. But the order of the difference-based variance estimator (Hall et al. 1990) is given beforehand. It is equal to 2 in the simulation study. If the variation in a special sample is small, then the selected bandwidth tends to be small, hence $\hat{\sigma}_B^2$ tends to be smaller than σ^2. Conversely, if the variation in a special sample is large, $\hat{\sigma}_B^2$ tends to be larger than σ^2. Whence $\hat{\sigma}_B^2$ adapts automatically to the structure of a special sample and tends to simulate the "true" variation in the sample. The bandwidth selected by using $\hat{\sigma}_B^2$ is not too far from h_0 even when $\hat{\sigma}_B^2$ is far from σ^2. To show this table 9 gives the ASE's of the bandwidth selectors in replications, in which $|\hat{\sigma}_B^2 - \sigma^2| > |\hat{\sigma}_G^2 - \sigma^2|$. We see that in this case $\hat{h}_{RBp}$ still performs much better than $\hat{h}_{RGp}$, and in two thirds of the cases, $\hat{h}_{BSp}$ outperforms $\hat{h}_{DSp}$. And even in this case there are no large changes in the performances of $\hat{h}_{RBp}$ and $\hat{h}_{BSp}$ (compare with tables 5–7). This shows that bandwidth selectors using $\hat{\sigma}_B^2$ are very stable.

8 Conclusions

In this paper we have proposed a simple variance estimator, $\hat{\sigma}_B^2$, and a BS bandwidth selector. $\hat{\sigma}_B^2$ performs clearly better than $\hat{\sigma}_G^2$. And the BS bandwidth selector performs uniformly better than the DS one using $\hat{\sigma}_G^2$. The performances of bandwidth selectors following the same error criterion but with different variance estimators are compared through simulation. So far as we know, this is the first paper to do this. It is shown that the gain of a bandwidth selector from an improved variance estimator depends strongly on the bandwidth selection rule. It is interesting to see how large the influence of a variance estimator on a plug-in bandwidth selector is. From (2) we know that the influence of $\hat{\sigma}^2$ will be reduced substantially by taking the $(2p+3)$th root. However, in this case $\hat{\sigma}^2$ will also influence the estimate of the functional $\int\{m^{(p+1)}(x)\}^2dx$. This problem has not been discussed in detail in this paper.

Fan and Gijbels (1995) discussed bandwidth selection for estimation of regression functions and their derivatives by local polynomial fitting. Their suggestion, RBS, is also a method using a pilot local polynomial fit of degree $p+a$, e.g. $a=2$. The main differences between the RBS and the DS procedure are: 1. In the RBS the bias is estimated with the $p+1$, ..., $p+a$ terms of the pilot estimate, while the BS criterion uses a bootstrap bias estimator; 2. In the RBS the bias and the variance are estimated using a single pilot estimate, while the bias and variance in the BS criterion are estimated with separate pilot estimates. This characteristic of the BS procedure allows us to estimate the bias using a factorized pilot bandwidth in order to reduce the bias in $\hat{h}_{BS}$ (Heiler and Feng 1997). Like the RBS, the DS procedure can also be used to select bandwidth for estimating derivatives (Müller 1985). It is not difficult to extend the results of Heiler and Feng (1997) to this case. It would be worthwhile to carry out a simulation study in order to compare the practical performance of the RBS with that of the BS procedure.

Another point of view is to take $\tilde{h}_0$, the minimizer of the ASE between $\hat{m}$ and m, as the true optimal bandwidth (Härdle et al. 1988). $\hat{h}_{BSp}$ should also be very close to $\tilde{h}_0$, since $\hat{\sigma}_B^2$ adapts automatically to the structure of a special sample.

Appendix: Proof of (5) and (6)

Results (5) and (6) can easily be derived by adapting the proofs of Hall and Marron (1990) and Ruppert et al. (1995). Note that

$$\hat{\sigma}_{\mathrm{B}}^2 = n^{-1}\sum_{i=1}^{n}[Y_i - \hat{m}_{g_v}(x_i)]^2,$$

and if m is a polynomial of degree less than or equal to p_v, $E(\hat{\sigma}_{\mathrm{B}}^2) = \nu n^{-1}\sigma^2$. Following Hall and Marron (1990),

$$E(\hat{\sigma}_{\mathrm{B}}^2 - \sigma^2) = n^{-1}\sum_{i=1}^{n}\delta_i^2 - (1 - \nu n^{-1})\sigma^2,$$

and

$$\mathrm{var}\,(\hat{\sigma}_{\mathrm{B}}^2) = n^{-2}\left\{\sum_{j=1}^{n}E(\Delta_j^2) + 2\sigma^4\sum\sum_{j\neq k}t_{jk}^2\right\},$$

where

$$\delta_i = m(x_i) - \sum_{j=1}^{n}w_j(x_i)m(x_j),$$

$$t_{jk} = \sum_{i=1}^{n}w_j(x_i)w_k(x_i) - 2w_k(x_j),$$

and

$$\Delta_j = \left(\delta_j - \sum_{i=1}^{n}w_j(x_i)\delta_i\right)\epsilon_j + \left(1 - 2w_{jj} + \sum_{i=1}^{n}w_j(x_i)\right)(\epsilon^2 - \sigma^2).$$

Using theorem 4.1 of Ruppert and Wand (1994) and noting that $\nu n^{-1} \simeq (1 - C_2(ng_v)^{-1})$, we obtain

$$n^{-1}\sum_{i=1}^{n}\delta_i^2 \simeq \{\mu_{k_v}(K_{p_v})/(k_v)!\}^2\theta_{k_v,k_v}g_v^{2p_v+2} = C_1 g_v^{2p_v+2}$$

and

$$E(\hat{\sigma}_{\mathrm{B}}^2(g_v) - \sigma^2) \simeq C_1 g_v^{2p_v+2} - C_2\sigma^2(ng_v)^{-1}.$$

Noting that

$$w_{ij} \simeq K_{p_v}\{(x_i - x_j)/g_v\}\{ng_v f(x_i)\}^{-1}$$

we have

$$t_{jk} \simeq (K_{p_v} * K_{p_v} - 2K_{p_v})\{(x_j - x_k)/g_v\}\{ng_v f(x_j)\}^{-1}$$

and

$$\sum\sum_{j\neq k}t_{jk}^2 \simeq g_v^{-1}R(K_{p_v} * K_{p_v} - 2K_{p_v}).$$

From Hall and Marron (1990),

$$\sum_{j=1}^{n} E(\Delta_j^2) = \mathrm{var}\,(\epsilon^2) n\{1 - 2C_2(ng_v)^{-1}\} + O(n^{-1}g_v^{2k_v}) + o(g_v^{-1}),$$

whence

$$\mathrm{var}\,(\hat{\sigma}_{\mathrm{B}}^2(g_v)) \simeq n^{-1}\mathrm{var}\,(\epsilon^2) + C_3(n^2 g_v)^{-1} + \mathrm{O}(n^{-1}g_v^{2k_v}).$$

$\diamond$

References

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

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

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

[4] W.S. Cleveland, and S.J. Devlin (1988), Locally weighted regression: An approach to regression analysis by local fitting, *J. Amer. Statist. Assoc.* **83**, 596–610.

[5] J. Fan (1992), Design-adaptive nonparametric regression, *J. Amer. Statist. Assoc.* **87**, 998–1004.

[6] J. Fan and I. Gijbels (1995), Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation, *J. Roy. Statist. Soc. Ser. B* **57**, 371–394.

[7] —— (1996), *Local Polynomial Modeling and its Applications*, Chapman & Hall, London.

[8] J.J. Faraway and M. Jhun (1990), Bootstrap choice of bandwidth for density estimation, *J. Amer. Statist. Assoc.* **85**, 1119–1122.

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

[10] T. Gasser, L. Sroka, and C. Jennen-Steinmetz (1986), Residual variance and residual pattern in nonlinear regression, *Biometrika* **73**, 625–633.

[11] W. Härdle and A.W. Bowman (1988), Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands, *J. Amer. Statist. Assoc.* **83**, 102–110.

[12] W. Härdle, P. Hall, and J.S. Marron (1988), How far are automatically chosen regression smoothing parameters from their optimum? (with discussion), *J. Amer. Statist. Assoc.* **83**, 86–99.

[13] — (1992), Regression smoothing parameters that are not far from their optimum, *J. Amer. Statist. Assoc.* **87**, 227–233.

[14] P. Hall, J.W. Kay, and D.M. Titterington (1990), Asymptotically optimal difference-based estimation of variance in nonparametric regression, *Biometrika* **77**, 521–528.

[15] P. Hall and J.S. Marron (1990), On variance estimation in nonparametric regression, *Biometrika* **77**, 415–419.

[16] S. Heiler and Y. Feng (1997), A simple root n bandwidth selector for nonparametric regression, to appear in *Jurnal of Nonparametric Statistics.*

[17] E. Herrmann (1994), Asymptotic distribution of bandwidth selectors in kernel regression estimation. *Statistical Papers* **35**, 17–26.

[18] M.C. Jones, J.S. Marron, and S.J. Sheather (1996), A brief survey of bandwidth selection for density estimation, *J. Amer. Statist. Assoc.* **91**, 401–407.

[19] J.S. Marron (1992), Bootstrap bandwidth selection, in *Exploring the Limits of Bootstrap*, eds. R. LePage and L. Billard, John Wiley, New York, pp. 249–262.

[20] H.-G. Müller (1985), Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators, *Statist. Decisions*, Supp. Issue **2**, 193–206.

[21] H.-G. Müller (1988), *Nonparametric Analysis of Longitudinal Data*, Springer-Verlag, Berlin.

[22] H.-G. Müller, U. Stadtmüller, and T. Schmitt (1987), Bandwidth choice and confidence intervals for derivatives of noisy data, *Biometrika* **74**, 743–750.

[23] J. Rice (1984), Bandwidth choice for nonparametric regression, *Ann. Statist.* **12**, 1215–1230.

[24] D. Ruppert, S.J. Sheather, and M.P. Wand (1995), An effective bandwidth selector for local least squares regression, *J. Amer. Statist. Assoc.* **90**, 1257–1270.

[25] D. Ruppert and M.P. Wand (1994), Multivariate locally weighted least squares regression, *Ann. Statist.* **22**, 1346–1370.

[26] B. Seifert, T. Gasser, and A. Wolf (1993), Nonparametric estimation of residual variance revisited, *Biometrika* **80**, 373–383.

[27] C.C. Taylor (1989), Bootstrap chioce of the smoothing parameter in kernel density estimation, *Biometrika* **76**, 705–712.

[28] M.P. Wand and M.C. Jones (1995), *Kernel Smoothing*, Chapman & Hall, London.

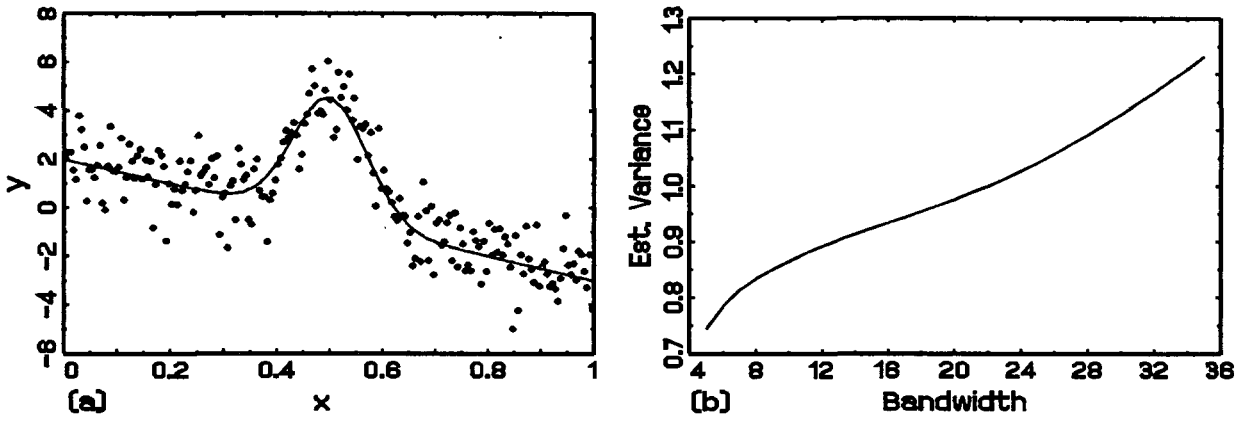


Figure 1: (a) the regression function m_1 and a simulated data ($n = 200$), (b) the estimated variances for this data with different bandwidths (The scale of the abscissa is $200h$).

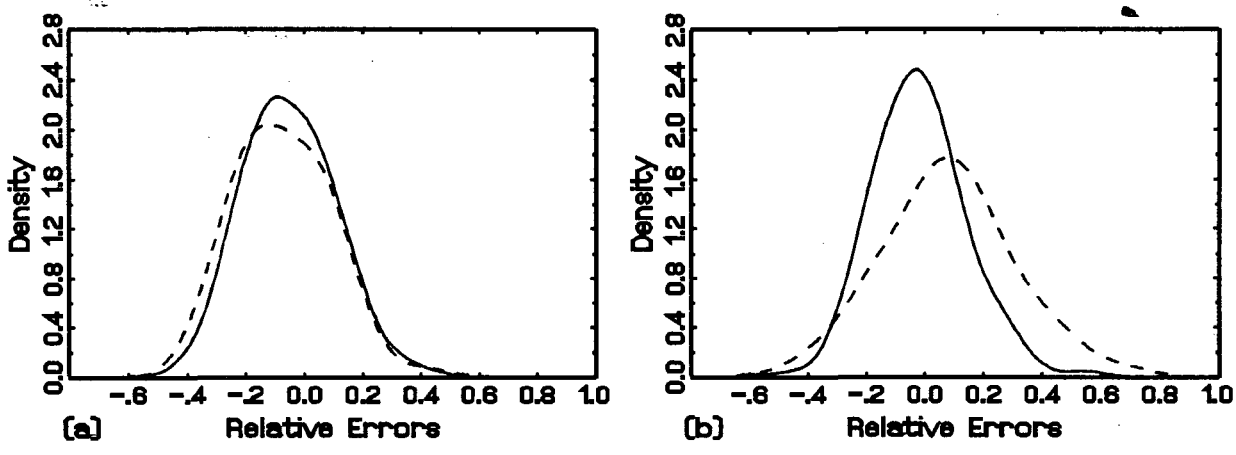


Figure 3: Kernel density estimates for the relative errors of $\hat{\sigma}^2_{\mathrm{G}}$ (dashed line) and $\hat{\sigma}^2_{\mathrm{B}}$ (solid line) for m_1 with $n = 100$. (a) for replications in group A, (b) for replications in group B.

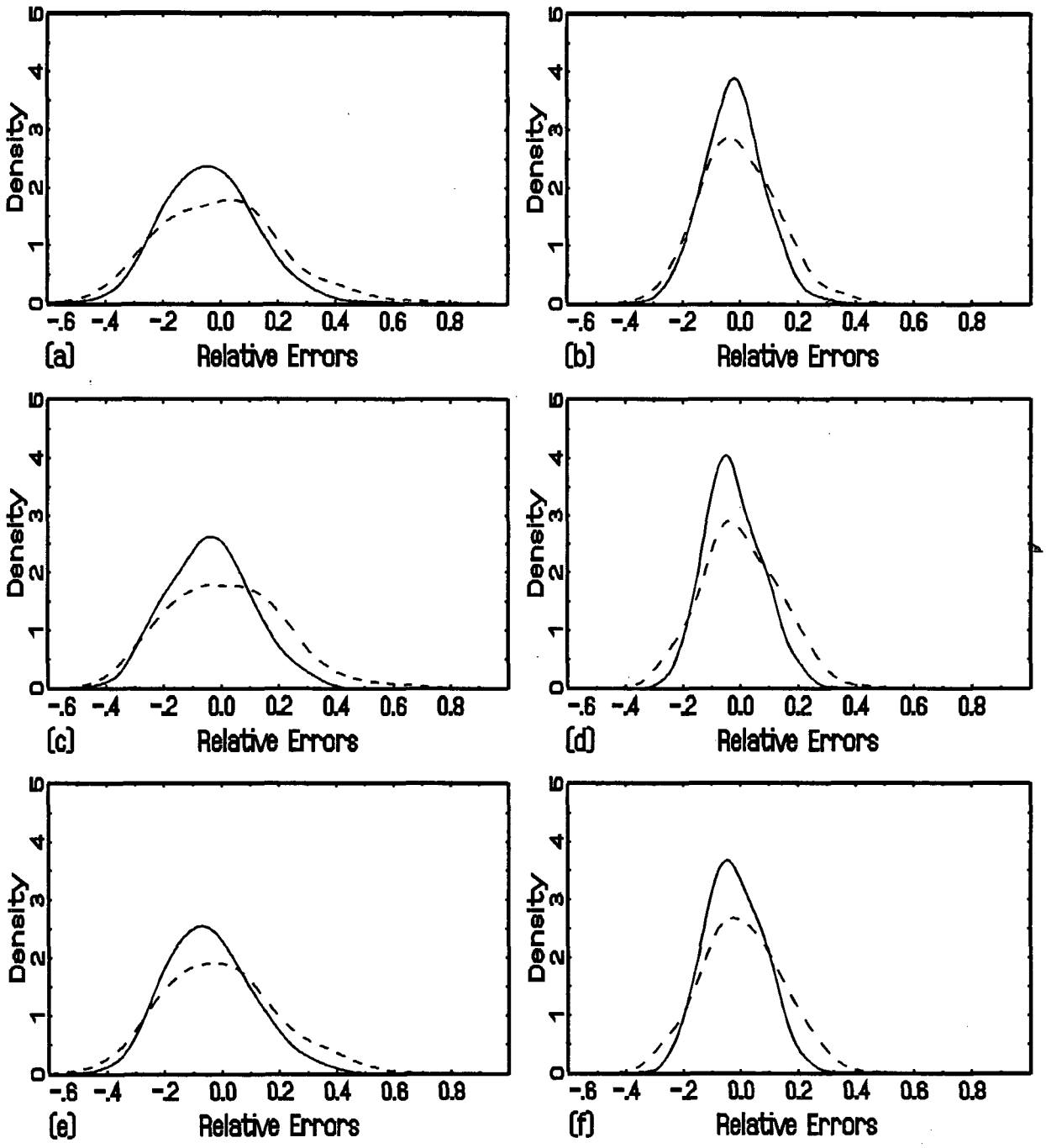


Figure 2: Kernel density estimates for the relative errors of $\hat{\sigma}^2_{\mathrm{G}}$ (dashed line) and $\hat{\sigma}^2_{\mathrm{B}}$ (solid line) for m_1 (upper), m_2 (middle) and m_3 (lower), respectively, with $n = 100$ (left) and $n = 200$ (right).

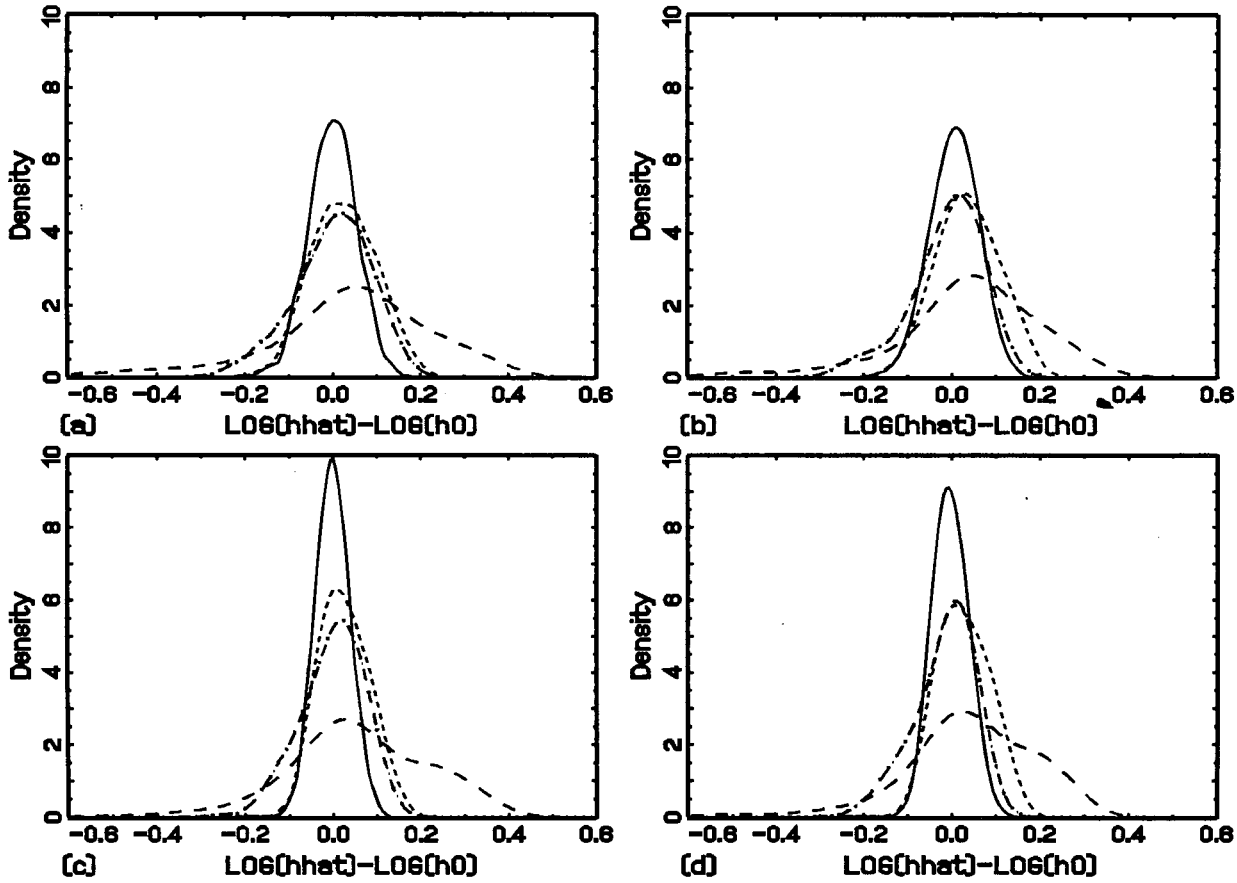


Figure 4: Kernel density estimates based on $log(\hat{h}) - log(h_0)$ values for m_1. (a) $n = 100$, $p = 1$, (b) $n = 100$, $p = 3$, (c) $n = 200$, $p = 1$ and (d) $n = 200$, $p = 3$. The curves are for $\hat{h}_{\mathrm{RGp}}$ (dashed line), $\hat{h}_{\mathrm{DSp}}$ (short dashes), $\hat{h}_{\mathrm{RBp}}$ (dashes and dots) and $\hat{h}_{\mathrm{BSp}}$ (solid line).

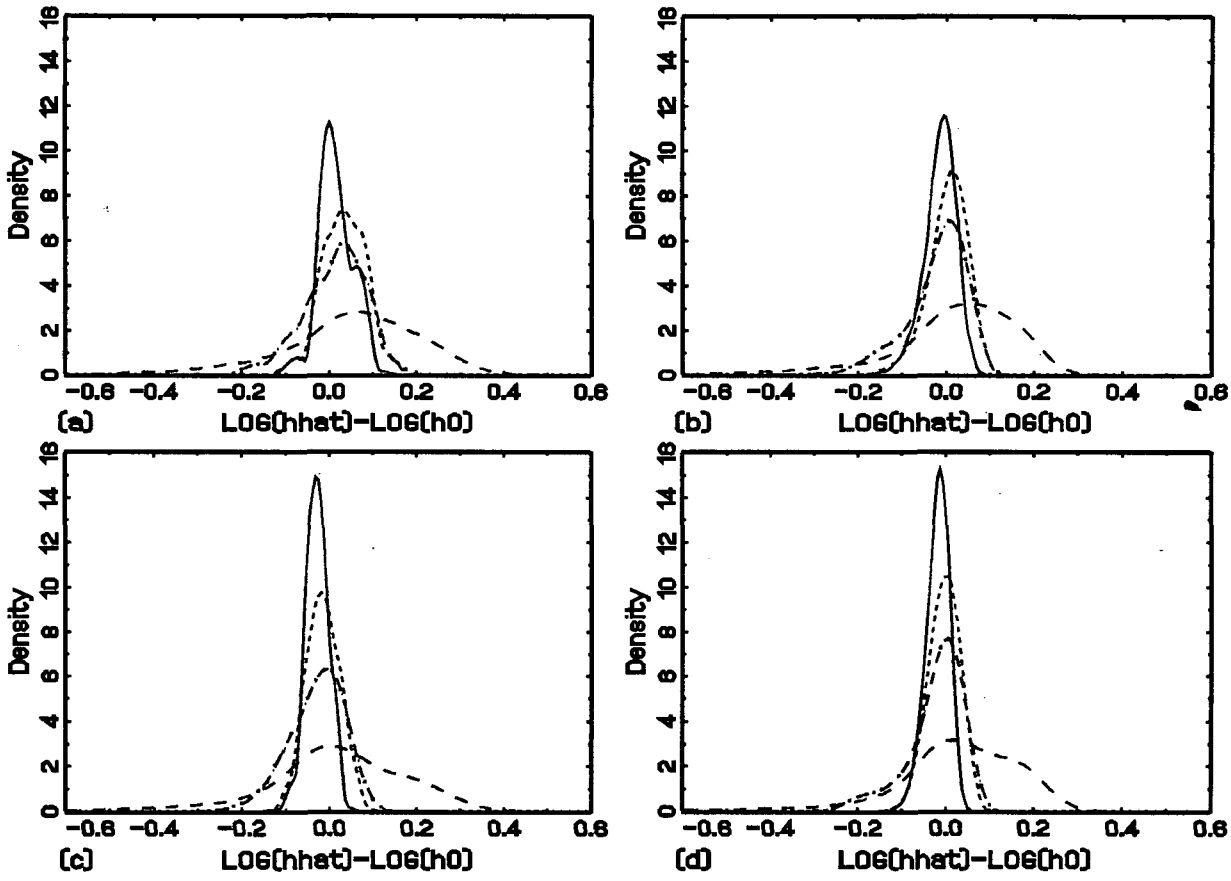


Figure 5: Kernel density estimates based on $log(\hat{h}) - log(h_0)$ values for m_2. (a) $n = 100$, $p = 1$, (b) $n = 100$, $p = 3$, (c) $n = 200$, $p = 1$ and (d) $n = 200$, $p = 3$. The curves are for $\hat{h}_{\mathrm{RGp}}$ (dashed line), $\hat{h}_{\mathrm{DSp}}$ (short dashes), $\hat{h}_{\mathrm{RBp}}$ (dashes and dots) and $\hat{h}_{\mathrm{BSp}}$ (solid line).

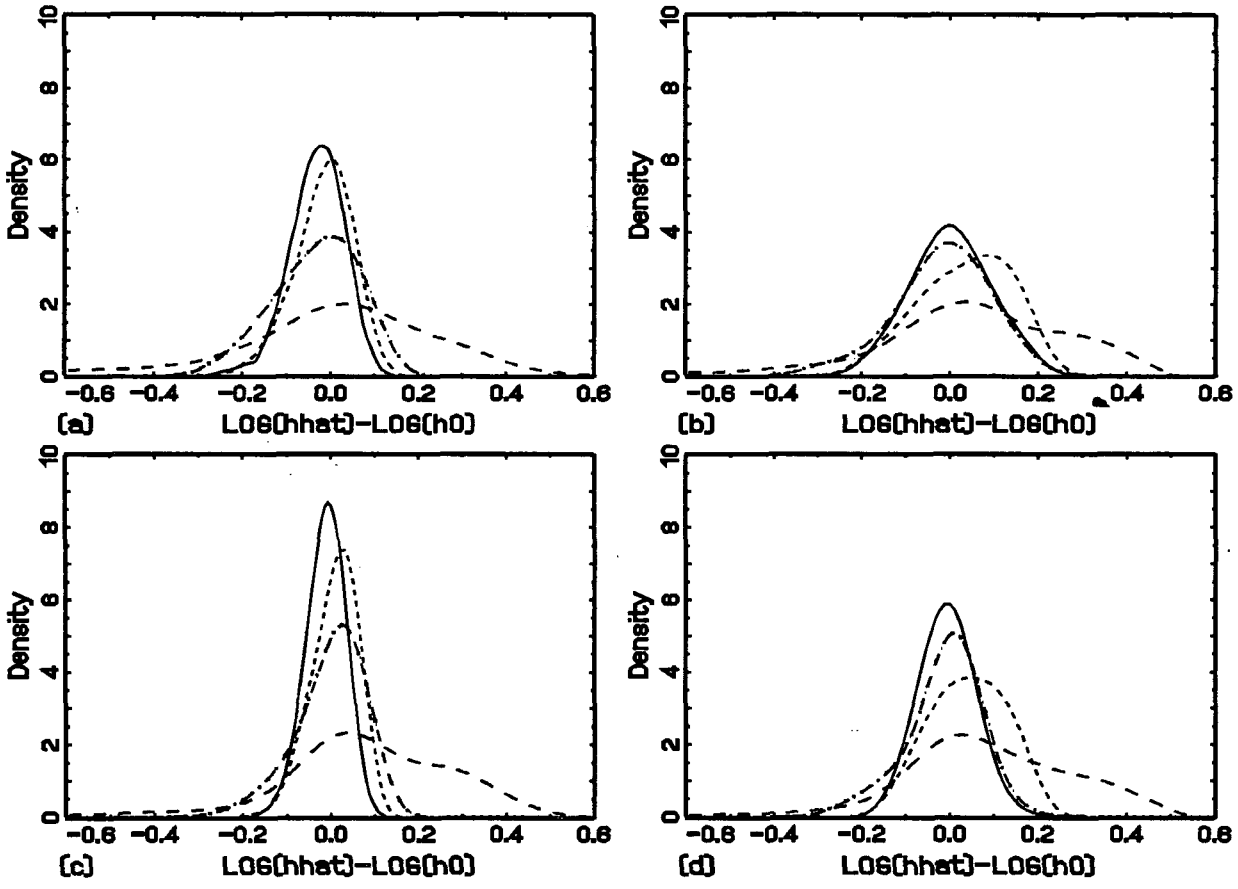


Figure 6: Kernel density estimates based on $log(\hat{h}) - log(h_0)$ values for m_3. (a) $n = 100$, $p = 1$, (b) $n = 100$, $p = 3$, (c) $n = 200$, $p = 1$ and (d) $n = 200$, $p = 3$. The curves are for $\hat{h}_{\mathrm{RGp}}$ (dashed line), $\hat{h}_{\mathrm{DSp}}$ (short dashes), $\hat{h}_{\mathrm{RBp}}$ (dashes and dots) and $\hat{h}_{\mathrm{BSp}}$ (solid line).