The Annals of Statistics
1995, Vol. 23, No. 6, 1937–1959

AUTOMATIC BANDWIDTH CHOICE AND CONFIDENCE INTERVALS IN NONPARAMETRIC REGRESSION

BY MICHAEL H. NEUMANN

Weierstrass-Institut für Angewandte Analysis und Stochastik

In the present paper we combine the issues of bandwidth choice and construction of confidence intervals in nonparametric regression. Main emphasis is put on fully data-driven methods. We modify the $\sqrt{n}$-consistent bandwidth selector of Härdle, Hall and Marron such that it is appropriate for heteroscedastic data, and we show how one can optimally choose the bandwidth g of the pilot estimator $\hat{m}_g$. Then we consider classical confidence intervals based on kernel estimators with data-driven bandwidths and compare their coverage accuracy. We propose a method to put undersmoothing with a data-driven bandwidth into practice and show that this procedure outperforms explicit bias correction.

1. Introduction.

We assume observations

$$Y_i = m(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where the errors ε_i are independently, not necessarily identically distributed with zero mean and variance $v(x_i)$. The nonrandom design points x_i are assumed to be spaced on the unit interval $[0, 1]$, $x_1 < x_2 < \cdots < x_n$.

We aim at defining a confidence interval for the regression function m at some interior point x_0 of this interval.

There already exists a very developed theory for confidence intervals based on kernel estimators with nonrandom bandwidths. Under assumptions on the decay of the bandwidths it is shown that these methods are consistent and, moreover, there are rates for the errors in coverage probability calculated. Hall (1991, 1992a), for confidence intervals for a density, and Hall (1992b), for intervals in regression with i.i.d. errors, found optimal rates for the bandwidths involved in the confidence interval procedure by optimizing the coverage accuracy. On the other hand, the majority of the available literature does not take into account the bandwidth choice that is necessary for practical applications. Some exceptions we are aware of are papers of Faraway and Jhun (1990) for density estimation and Faraway (1990) for the regression case, where the bandwidth as well as the quantile for confidence bands are obtained on the basis of the same bootstrap sample. However, the authors do not provide any rigorous result on the real coverage probability in comparison to the prescribed level.

Received May 1992; revised April 1995.
AMS 1991 *subject classifications*. Primary 62G15; secondary 62G07, 62G20.
Key words and phrases. Nonparametric regression, bandwidth choice, confidence intervals, Edgeworth expansions.

Usually the first step in constructing asymptotic confidence intervals consists in the definition of an asymptotically normally distributed pivotal quantity. There are two commonly used methods to deal with the bias of the initial estimator $\hat{m}_h(x_0)$, undersmoothing and explicit bias correction on the basis of yet another kernel estimator. In Hall (1991, 1992a) it is shown that the undersmoothing method leads to a better coverage accuracy. An analogous result is proved in Neumann (1992b) for the regression case with not necessarily identical error distributions, again for kernel estimators with nonrandom bandwidths.

In the present paper we mainly focus on the practical bandwidth choice. Whereas we can apply the bias correction method with usual bandwidth selectors at all stages, we replace the pure undersmoothing method by a two-step procedure that yields the same rates for the coverage accuracy. As an estimator of the optimal global bandwidth we employ here, with some minor modifications, the $\sqrt{n}$-consistent bandwidth selector of Härdle, Hall and Marron [(1992), hereafter HHM]. It is based on plug-in estimates of the integrated variance and the integrated squared bias of $\hat{m}_h$. To make this method fully data-driven, we propose a method for how the bandwidth of the pilot estimate $\hat{m}_g$ needed for the bias estimation can be optimally chosen from the data.

The second step in getting confidence intervals is the recognition of the distributions of the aforementioned pivotal quantities. We restrict our considerations exclusively on a distribution recognition via Edgeworth expansions for the pivotal quantities. Recently, the application of bootstrap techniques in general, and in the context of heteroscedasticity the wild bootstrap proposed by Wu (1986) and by Härdle and Mammen (1993) in particular, has become quite popular. It is used by Härdle, Huet and Jolivet (1995) for the construction of asymptotic confidence intervals. However, in Neumann (1992b) it is shown that we obtain via Edgeworth expansions the same rate for the approximation of the cumulative distribution function. Since on the other hand the quantiles via Edgeworth expansions are explicitly given, there seems to be no need for the computationally more involved bootstrap.

The paper is organized as follows. In Section 2 we develop a completely data-driven, $\sqrt{n}$-consistent bandwidth selector. In Section 3 we introduce pivotal quantities used for the construction of confidence intervals. Estimates for the error in coverage probability of the intervals are derived in Section 4. Section 5 contains a discussion of the applied methods. A list of the assumptions, some technical lemmas and the proofs are deferred to the Appendix.

2. A fully data-driven bandwidth selector. As a starting point we take a kernel estimator, as proposed in Gasser and Müller (1979), which can be written as

$$\hat{m}_h(x) = \sum_{i=1}^{n} W(x, h)_i Y_i.$$

The weights are defined as $W(x,h)_i = \int_{s_{i-1}}^{s_i} h^{-1} w((z-x)/h)\,dz$, where w is some usual kernel of order r with support $[-1,1]$ if $h \le x \le 1-h$, and some boundary kernel otherwise. Explicit formulas for such kernels are given in Gasser, Müller and Mammitzsch (1985). Further we set $s_i = (x_i + x_{i+1})/2$ for $i = 1,\ldots,n-1$ and $s_0 = 0$, $s_n = 1$.

As already mentioned, all kernel estimators included in the procedure should be equipped with data-driven bandwidths. Although we adopt the coverage accuracy as our main criterion to evaluate the performance of our methods, we choose the bandwidths according to the risk behavior of the corresponding estimator, since a specific choice for confidence interval purposes seems to be difficult for some conceptual reasons discussed in Section 5. Since only the behavior of $\hat{m}_h$ at x_0 influences the properties of the confidence interval, it seems to be on first sight reasonable to seek an estimate of the locally optimal bandwidth. On the other hand, we agree with Härdle and Bowman (1988), who claim that the potential advantages of local adaptive bandwidth selection in the context of confidence intervals are not clear. Roughly speaking, the initial estimator $\hat{m}_h(x_0)$ will only serve as a vehicle to introduce a nondegenerate noise structure, whereas its bias will be corrected with the help of a second estimator $\hat{m}_g$. For a more detailed discussion of this subject we refer again to Section 5. On the other hand, since our methods are based on Edgeworth expansions of pivotal quantities with nonrandom bandwidths, we can expect a better coverage accuracy for intervals with random bandwidths that are very close to some fixed one. This is certainly an argument in favor of global bandwidth selectors, which are usually more stable than local ones.

For fixed h, the mean squared error of $\hat{m}_h(x_0) = \sum W(x_0,h)_i Y_i$ can be written as $\mathrm{MSE}(h) = V_h + B_h^2$, where $V_h = \sum_{i=1}^n W(x_0,h)_i^2 v(x_i)$ and $B_h = \sum_{i=1}^n W(x_0,h)_i m(x_i) - m(x_0)$.

Since we are not willing to impose restricting assumptions on the smoothness of v, we borrow the idea of the wild bootstrap and estimate V_h simply by

$$\hat{V}_h = \sum_{i=1}^n W(x_0,h)_i^2 \hat{v}_i, \tag{2.1}$$

where

$$\hat{v}_i = \hat{\varepsilon}_i^2 = \big(Y_i - \tilde{m}(x_i)\big)^2, \tag{2.2}$$

and where $\tilde{m}$ denotes yet another Gasser–Müller kernel estimator. Anticipating the following results, we remark that the consistency of the bandwidth selector considered in this section as well as of the confidence intervals in the next section require a higher degree of smoothness for m than the basic kernel estimator $\hat{m}_h$ can exploit. We assume throughout the present paper that m is $(r+s)$-times continuously differentiable, and therefore we take $\tilde{m}$ as a kernel estimator $\hat{m}_{\hat{f}}$ with an $(r+s)$th-order kernel and a bandwidth $\hat{f}$, which can for simplicity be chosen by cross-validation or by some other consistent bandwidth selector.

We opt here for these particular variance estimators (2.2) because they are close to the squared errors ε_i^2, which are easy to analyze. Another obvious possibility would be given by $\tilde{v}_i = (Y_i - Y_{i-1})(Y_i - Y_{i+1})$.

The bias B_h will be approximated by an estimator of the form

$$\hat{B}_{h,g} = \sum_{i=1}^{n} W(x_0, h)_i \hat{m}_g(x_i) - \hat{m}_g(x_0), \tag{2.3}$$

where $\hat{m}_g(x)$ is another Gasser–Müller kernel estimator with weights $\tilde{W}(x, g)_i$ based on an sth-order kernel $\tilde{w}$ and a bandwidth g.

Now we intend to modify the bandwidth selector of HHM such that it accounts for the possible heteroscedasticity of the data. An even more important issue is the data-dependent choice of the bandwidth of the auxiliary estimator $\hat{m}_g$, which is used for the estimation of the bias.

We are going to estimate the bandwidth h_0 that is optimal with respect to the mean integrated squared error (MISE) of $\hat{m}_h$, where the integration is because of boundary effects restricted to some interval $[c, d]$, $0 < c < d < 1$, that should include for our purposes the point x_0.

The MISE splits up into an integrated variance part,

$$\mathrm{IV}(h) = \int_c^d \sum_{i=1}^{n} W(x, h)_i^2 v(x_i)\, dx = \sum_{i=1}^{n} \int_c^d W(x, h)_i^2\, dx\, v(x_i),$$

and an integrated squared bias part,

$$\mathrm{ISB}(h) = \int_c^d \left(\sum_{i=1}^{n} W(x, h)_i m(x_i) - m(x) \right)^2 dx.$$

We estimate $\mathrm{IV}(h)$ analogously to (2.1) by

$$\widehat{\mathrm{IV}}(h) = \sum_{i=1}^{n} \int_c^d W(x, h)_i^2\, dx\, \hat{v}_i. \tag{2.4}$$

Provided an appropriate choice of g, the quantity

$$\begin{aligned} \overline{\mathrm{ISB}}(h, g) &= \int_c^d \left(\sum_{i=1}^{n} W(x, h)_i \hat{m}_g(x_i) - \hat{m}_g(x) \right)^2 dx \\ &= \sum_{i,j=1}^{n} A(h, g)_{ij} Y_i Y_j, \end{aligned} \tag{2.5}$$

with

$$\begin{aligned} A(h, g)_{ij} = \int_c^d &\left(\sum_k W(x, h)_k \tilde{W}(x_k, g)_i - \tilde{W}(x, g)_i \right) \\ &\times \left(\sum_l W(x, h)_l \tilde{W}(x_l, g)_j - \tilde{W}(x, g)_j \right) dx, \end{aligned}$$

could already serve as an estimator of $\mathrm{ISB}(h)$. A quantity similar to the sum of (2.4) and (2.5) was proposed by Müller (1988) as an estimator of the MISE. As remarked by HHM, $\overline{\mathrm{ISB}}(h, g)$ is biased due to the variance of the diagonal terms, and therefore it turns out to be better to estimate $\mathrm{ISB}(h)$ by

$$\widehat{\overline{\mathrm{ISB}}}(h, g) = \overline{\mathrm{ISB}}(h, g) - \sum_{i=1}^{n} A(h, g)_{ii} \hat{v}_i. \tag{2.6}$$

An alternative approach in the framework of density estimation is proposed by Sheather and Jones (1991). They recognize that the nonstochastic bias term has the opposite sign to the bias due to the smoothing, and they choose the auxiliary bandwidth such that these terms cancel. However, an appropriate choice of this bandwidth requires the estimation of higher derivatives. If enough smoothness is present to do this reasonably well, then we could alternatively use it to improve on the whole procedure at other stages. Now we have with $\widehat{\mathrm{IV}}(h) + \widehat{\overline{\mathrm{ISB}}}(h, g)$ a pattern to estimate $\mathrm{MISE}(h)$, but it remains to fix an appropriate value of g. This problem was not solved in an entirely satisfactory way in HHM, and the authors conjectured that there is no substitute for trying some number of different g's. Assuming slightly more regularity than in HHM, namely, $m \in C^{2r+s}[0, 1]$ instead of $m \in C^{2r \wedge r+s}[0, 1]$, we obtain that the term of order $O(n^{-1})$ of the mean squared error of $\widehat{\overline{\mathrm{ISB}}}(h, g)$ as an estimator of $\mathrm{ISB}(h)$ does not depend on g, whereas the largest two of the remaining terms do. These terms can be used to get a reasonable, asymptotically optimal choice for g.

The assumptions needed for the following lemma as well as for the assumptions in the sequel are given in the Appendix.

LEMMA 2.1. *Assume* (A_G) *and* (A_{BW}). *Then*

(i)
$$\begin{aligned} E\big(\widehat{\overline{\mathrm{ISB}}}(h, g) - \mathrm{ISB}(h)\big)^2 &= h^{4r} C(h) n^{-1} \\ &\quad + 4h^{4r} g^{2s} \kappa_r^4 \lambda_s^2 \left(\int_c^d m^{(r+s)}(x) m^{(r)}(x)\, dx \right)^2 \\ &\quad + 2h^{4r} n^{-2} g^{-(4r+1)} \kappa_r^4 \int_c^d \left(\frac{v(x)}{d(x)} \right)^2 dx \\ &\qquad \times \int \left(\int \tilde{w}^{(r)}(y) \tilde{w}^{(r)}(y+z)\, dy \right)^2 dz \\ &\quad + o\big(h^{4r}\big(g^{2s} + n^{-2} g^{-(4r+1)}\big)\big), \end{aligned}$$

where $C(h)$ *is bounded and* $\kappa_r = (-1)^r (r!)^{-1} \int z^r w(z)\, dz$, $\lambda_s =$

$(-1)^s(s!)^{-1}\int z^s\tilde{w}(z)\,dz$;

(ii)

$$g_{\text{opt}} = \left\{(4r+1)\int_c^d (v(x)/d(x))^2\,dx \times \int\left(\int \tilde{w}^{(r)}(y)\tilde{w}^{(r)}(y+z)\,dy\right)^2 dz \times \left[4s\lambda_s^2\left(\int_c^d m^{(r+s)}(x)m^{(r)}(x)\,dx\right)^2\right]^{-1} n^{-2}\right\}^{1/(2s+4r+1)} (1+o(1)).$$

Let $\bar{g}$ be any consistent estimator of g_{opt}, that is, $\bar{g} = g_{\text{opt}} + o_p(n^{-1/(s+2r+1/2)})$. Now we define $\hat{h}$ as a measurable minimizer of $\widehat{\text{MISE}}(h, \bar{g}) = \widehat{\text{IV}}(h) + \widehat{\text{ISB}}(h, \bar{g})$, whose existence is ensured by a lemma of Jennrich (1969).

REMARK. Analogously to Theorem 1 in HHM, one can prove that

$$\frac{\hat{h} - h_0}{h_0} = O_p(n^{-\Delta}),$$

where $\Delta = 1/2 \wedge s/(s+2r+1/2)$.

The next point concerns the appropriate choice of the bandwidth g for the local bias estimator $\hat{B}_{h,g}$. First we infer from Lemma A.2 that

$$\hat{B}_{h,g} - B_h = \int w(z)\int_0^{hz} \frac{(hz-y)^{r-1}}{(r-1)!}\left(\hat{m}_g^{(r)}(x_0+y) - m^{(r)}(x_0+y)\right)dy\,dz + O_P(n^{-2}g^{-1})$$

holds. Since $h \ll g$ holds for h and g of optimal orders, the task of estimating B_h is nearly equivalent to the estimation of $m^{(r)}(x_0)$ by $\hat{m}_g^{(r)}(x_0)$. To give a definite rule for the choice of g, we use the method proposed by Müller and Stadtmüller (1987) for the choice of the optimal global bandwidth $g_0^{(r,s)}$, which can be applied in the case of heteroscedasticity, too. They observed that

$$g_0^{(r,s)} = C_{r,s}(w)C(m,v)n^{-1/[2(r+s)+1]}(1+o(1))$$

holds with some constant $C_{r,s}$ that depends on the kernel function w but not on the unknown functions m and v. On the other hand, the optimal global bandwidth for an estimator of m itself with an $(r+s)$th-order kernel $\overline{w}$ has the form

$$g_0^{(0,r+s)} = C_{0,r+s}(\overline{w})C(m,v)n^{-1/[2(r+s)+1]}(1+o(1)).$$

Now we can estimate $g_0^{(0,r+s)}$ by some asymptotically optimal bandwidth $\widehat{g_0^{(0,r+s)}}$, and then we obtain a consistent estimate of $g_0^{(r,s)}$ by

$$\hat{g} = \frac{C_{r,s}(w)}{C_{0,r+s}(\overline{w})} \widehat{g_0^{(0,r+s)}}, \tag{2.7}$$

which spares us the more involved direct estimation of the constant $C(m, v)$.

3. Confidence intervals for $m(x_0)$.

3.1. *Construction principles for confidence intervals.* All commonly used methods to establish confidence intervals are based on the principle of first estimating $m(x_0)$ by an initial estimator $\hat{m}(x_0)$ and then estimating the distribution of $\hat{m}(x_0) - m(x_0)$. Usually a distinction is made between pivotal and nonpivotal methods. For the related problem of bootstrap confidence intervals, Hall (1992b) pointed out that pivotal methods, which are based on a quantity $\hat{V}^{-1/2}(\hat{\vartheta} - \vartheta)$ that contains an estimator $\hat{V}$ of the variance of $\hat{\vartheta}$, should be preferred to nonpivotal methods, which are simply based on an estimation of the distribution of $(\hat{\vartheta} - \vartheta)$. Hence, in the present paper we restrict ourselves to pivotal methods.

The main problem with confidence intervals in nonparametric regression rests on the fact that a consistent estimator of $m(x_0)$ is necessarily biased. Strictly speaking, MISE-optimal estimators have bias and standard deviation of the same order. There are two common methods to deal with this problem, undersmoothing and subsequent bias correction. Hall (1991, 1992a) shows in situations closely related to ours that the first method leads to a better asymptotic coverage accuracy, at least in the case of nonrandom bandwidths.

An important goal of the present paper is to provide methods where all bandwidths are chosen by the data in a reasonable way. The only available guideline for a reasonable choice seems to be the risk behavior of the corresponding estimators; hence, the bandwidths we deal with are of MISE-optimal order, which means that bias and standard deviation of the estimator can be expected to decay at the same rate to zero. Therefore, we cannot apply the undersmoothing method in its pure form.

In contrast, it is possible to construct a bias-corrected pivotal quantity on the basis of MISE-optimal kernel estimators by a normalization of the initial estimator with estimates of its bias and variance. Now it seems to be more natural to estimate the bias first and then to divide the corrected quantity by an estimator of its standard deviation. It turns out that this method is equivalent to undersmoothing and, in accordance with the existing theory, we obtain a better coverage accuracy than by the first method.

3.2. *Definition of the pivotal quantities.* As already announced, we consider the bias-corrected pivotal quantity

$$T_{h,g} = \frac{\hat{m}_h(x_0) - \hat{B}_{h,g} - m(x_0)}{\hat{V}_h^{1/2}} = \frac{\Sigma \overline{W}_{h,g,i} \varepsilon_i + b_{h,g}}{\hat{V}_h^{1/2}}, \tag{3.1}$$

where $\hat{V}_h$ and $\hat{B}_{h,g}$ are defined by (2.1) and (2.3), respectively, and $b_{h,g} = \Sigma \overline{W}_{h,g,i} m(x_i) - m(x_0)$ denotes the remaining bias. Further, we obtain a method equivalent to undersmoothing by estimating the whole variance of the numerator of $T_{h,g}$ instead of that of $\hat{m}_h(x_0)$. In this case the usual condition $h \ll g$, which is introduced to keep the variance of the bias estimator of smaller order than that of $\hat{m}_h(x_0)$, is no longer necessary and we optimize g with respect to the asymptotic coverage accuracy by choosing it of the same order as h. Since there is no other guideline for doing this in practice, we simply set $g = h$, where h will be chosen later by some data-dependent rule.

With $\overline{W}_{h,i} = \overline{W}_{h,h,i}$ and $b_h = b_{h,h}$ we get the pivotal quantity

$$U_h = \frac{\Sigma \overline{W}_{h,i} \varepsilon_i + b_h}{\hat{\hat{V}}_h^{1/2}} \tag{3.2}$$

where $\hat{\hat{V}}_h = \Sigma \overline{W}_{h,i}^2 \hat{v}_i$.

Note that the roles of $\hat{B}_{h,g}$ and $\hat{B}_{h,h}$ are very different. Whereas the quantity $\hat{B}_{h,g}$ in $T_{h,g}$ *estimates* the bias, the term $\hat{B}_{h,h}$ in U_h reduces only the nonstochastic part of $\hat{m}_h(x_0)$, but contributes by a stochastic part of the same order as the initial estimator $\hat{m}_h(x_0)$. In both cases we get a new estimator, $\hat{m}_h(x_0) - \hat{B}_{h,g}$ and $\hat{m}_h(x_0) - \hat{B}_{h,h}$, respectively, with a squared bias of smaller order than its variance.

To obtain knowledge about the asymptotic distributions of the pivotal quantities, we intend to apply Edgeworth expansions as far as possible. For that we approximate the quantities of interest by certain smooth functions of random vectors. Using results of Skovgaard (1981, 1986), we can then prove the validity of these (formal) expansions. To draw conclusions from the size of the difference of two random variables to the difference of their cumulative distribution functions in a convenient way, we introduce the following notation.

DEFINITION 3.1. Let $\{Y_n\}$ and $\{Z_n\}$ ($Z_n \geq 0$ a.s.) be sequences of random variables, and let $\{\gamma_n\}$ be a sequence of positive reals. We write

$$Y_n = \tilde{O}(Z_n, \gamma_n)$$

if

$$P(|Y_n| > CZ_n) \leq C\gamma_n$$

holds for $n \geq 1$ and some $C < \infty$.

This notion differs obviously from the usual O_p, which would provide a similar property for an arbitrary constant γ instead of $C\gamma_n$ on the right-hand side. As a rule, for arbitrary $\delta, \lambda > 0$, we can conclude under sufficiently strong moment conditions on the distributions of the errors, by Markov's and Whittle's inequalities, that

$$(a_n)'\boldsymbol{\varepsilon} = \tilde{O}\big(n^{\delta} \|a_n\|, n^{-\lambda}\big) \tag{3.3}$$

and

$$\boldsymbol{\varepsilon}' A_n \boldsymbol{\varepsilon} - E\boldsymbol{\varepsilon}' A_n \boldsymbol{\varepsilon} = \tilde{O}\left(n^{\delta}\sqrt{\operatorname{tr}(A_n A_n')}, n^{-\lambda}\right) \tag{3.4}$$

hold uniformly over $a_n \in \mathbb{R}^n$ and arbitrary $(n \times n)$-matrices A_n, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. Furthermore, we obtain similar assertions for random quantities a_n and A_n, which is made rigorous by Lemma A.1 in the Appendix.

The following lemma shows that $\tilde{O}$ is an appropriate concept for the calculation of the cumulative distribution function of quantities that do not immediately admit an Edgeworth expansion.

LEMMA 3.1. *Let $\{X_n\}$ be a sequence of random variables that admit the Edgeworth expansion*

$$P(X_n < t) = \Phi(t) + p_n(t)\phi(t) + O(u_n),$$

with some polynomials p_n of bounded order with bounded coefficients. Further, we assume $Y_n = \tilde{O}(\gamma_{n1}, \gamma_{n2})$. Then

$$P(X_n + Y_n < t) = P(X_n < t) + O(u_n + \gamma_{n1} + \gamma_{n2}).$$

The proof of this lemma follows immediately from the inequalities

$$\begin{aligned} &P(X_n < t - C\gamma_{n1}) - P(|Y_n| > C\gamma_{n1}) \\ &\quad \le P(X_n + Y_n < t) \le P\left(X_n < t + C\gamma_{n_1}\right) + P(|Y_n| > C\gamma_{n1}) \end{aligned}$$

and the Lipschitz equicontinuity of the functions $\Phi(t) + p_n(t)\phi(t)$.

4. Coverage accuracy of the confidence intervals.

4.1. *Coverage accuracy in the case of nonrandom bandwidths.* First, we approximate the cumulative distribution functions of the pivotal quantities with *nonrandom* bandwidths via Edgeworth expansions. The following proposition serves then as a starting point to derive formulas for quantities with data-driven bandwidths.

PROPOSITION 4.1. *Assume (A_G), (A_E) and $h = h(n)$ and $g = g(n)$ to be nonrandom.*

(i) *If $nh \to \infty$, $g \to 0$ and $h/g \to 0$ as $n \to \infty$, then*

$$\begin{aligned} P(T_{h,g} < t) = \Phi(t) &- \frac{b_{h,g}}{V_h^{1/2}}\phi(t) + \rho_n \frac{2t^2 + 1}{6}\phi(t) \\ &- \frac{1}{2}\frac{\overline{V}_{h,g} - V_h}{V_h} t\phi(t) + O\left(g^{2s} + \left(\frac{h}{g}\right)^{2(r+1)} + (nh)^{-1}\right) \end{aligned}$$

holds uniformly in t, where $\rho_n = V_h^{-3/2} \sum_i W(x_0, h)_i^3 E\varepsilon_i^3$ and $\overline{V}_{h,g} = \sum_i \overline{W}_{h,g,i}^2 v(x_i)$.

(ii) *If* $nh \to \infty$ *and* $h \to 0$ *as* $n \to \infty$, *then*

$$P(U_h < t) = \Phi(t) - \frac{b_h}{\overline{V}_{h,h}^{1/2}}\phi(t) + \bar{\rho}_n \frac{2t^2 + 1}{6}\phi(t) + O\big(h^{2s} + (nh)^{-1}\big)$$

holds uniformly in t, *where* $\bar{\rho}_n = \overline{V}_{h,h}^{-3/2} \Sigma \overline{W}_{h,i}^3 E\varepsilon_i^3$.

The proof of this proposition is essentially the same as that of Proposition 3.2 in Neumann (1992b) and may be sketched, w.l.o.g. for (i), as follows. First we approximate $T_{h,g}$ by

$$\tilde{T}_{h,g} = \Big(\sum W(x_0, h)_i^2 \varepsilon_i^2\Big)^{-1/2}\Big(\sum \overline{W}_{h,g,i}\varepsilon_i + b_{h,g}\Big).$$

By Lemma 3.2 in Neumann (1992b) we have

$$T_{h,g} - \tilde{T}_{h,g} = \tilde{O}\big((nh)^{-1}, n^{-1}\big). \tag{4.1}$$

The vector $S_n = (\Sigma_i \overline{W}_{h,g,i}\varepsilon_i, \Sigma_i W(x_0, h)_i^2 \varepsilon_i^2)'$ is a sum of independent random vectors and admits, in accordance with results of Skovgaard (1986), an Edgeworth expansion with a residual term of order $O((nh)^{-1-\delta})$ for some $\delta > 0$. Since $\tilde{T}_{h,g}$ is a smooth function of S_n, we infer from Theorem 3.2 and Remark 3.4 in Skovgaard (1981) the validity of the formal Edgeworth expansion of $\tilde{T}_{h,g}$. To identify the expansion, we must calculate the cumulants of $\tilde{T}_{h,g}$, which has already been done in Neumann (1992b). By Lemma 3.1 we conclude from (4.1) that the expansions of $T_{h,g}$ and $\tilde{T}_{h,g}$ are identical up to a term of order $O((nh)^{-1})$, which completes the proof.

For the rest of this subsection we assume that the nonrandom bandwidths h and g are chosen of the same order as $\hat{h}$ and $\hat{g}$ described above, namely, $h \asymp n^{-1/(2r+1)}$ and $g \asymp n^{-1/(2(r+s)+1)}$. Now it is easy to see that

$$b_{h,g} = O(h^r g^s), \qquad b_h = O(h^{r+s}),$$
$$V_h, \overline{V}_{h,h} \asymp (nh)^{-1/2},$$
$$\rho_n, \bar{\rho}_n = O\big((nh)^{-1/2}\big)$$

and

$$\frac{\overline{V}_{h,g} - V_h}{V_h} = O\left(\left(\frac{h}{g}\right)^{r+1}\right).$$

If $u_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the standard normal distribution, then we obtain

$$\begin{aligned} &P\Big(m(x_0) \in \Big(\hat{m}_h(x_0) - \hat{B}_{h,g} - u_{1-\alpha}\hat{V}_h^{1/2}, \infty\Big)\Big) \\ &\quad = P\big(T_{h,g} < u_{1-\alpha}\big) \\ &\quad = 1 - \alpha + O\big(g^s + (h/g)^{r+1} + (nh)^{-1/2}\big) \\ &\quad = 1 - \alpha + O\big(n^{-s/[2(r+s)+1]} + n^{-r/(2r+1)}\big) \end{aligned} \tag{4.2}$$

and

$$
\begin{aligned}
&P\Big(m(x_0) \in \big(\hat{m}_h(x_0) - \hat{B}_{h,h} - u_{1-\alpha}\hat{\hat{V}}_h^{1/2}, \infty\big)\Big) \\
&\qquad = 1 - \alpha + O\big(h^s + (nh)^{-1/2}\big) \\
&\qquad = 1 - \alpha + O\big(n^{-(s \wedge r)/(2r+1)}\big).
\end{aligned} \tag{4.3}
$$

Estimating ρ_n and $\bar{\rho}_n$ by

$$
\hat{\rho}_n = \hat{V}_h^{-3/2} \sum W(x_0, h)_i^3 \hat{v}_i^3 \quad \text{and} \quad \hat{\bar{\rho}}_n = \hat{\hat{V}}_h^{-3/2} \sum \overline{W}_{h,h,i}^3 \hat{v}_i^3,
$$

respectively, and inverting the expansions from Proposition 4.1, we obtain confidence intervals

$$
I_{h,g} = \left(\hat{m}_h(x_0) - \hat{B}_{h,g} - \left(1 + \frac{2u_{1-\alpha}^2 + 1}{6}\hat{\rho}_h\right) u_{1-\alpha}\hat{V}_h^{1/2}, \infty\right)
$$

and

$$
\tilde{I}_h = \left(\hat{m}_h(x_0) - \hat{B}_{h,h} - \left(1 + \frac{2u_{1-\alpha}^2 + 1}{6}\hat{\bar{\rho}}_h\right) u_{1-\alpha}\hat{\hat{V}}_h^{1/2}, \infty\right),
$$

with coverage probabilities

$$
\begin{aligned}
P\big(m(x_0) \in I_{h,g}\big) &= 1 - \alpha + O\big(g^s + (h/g)^{r+1} + (nh)^{-1}\big) \\
&= 1 - \alpha + O\big(n^{-s/[2(r+s)+1]}\big)
\end{aligned} \tag{4.4}
$$

and

$$
\begin{aligned}
P\big(m(x_0) \in \tilde{I}_h\big) &= 1 - \alpha + O\big(h^s + (nh)^{-1}\big) \\
&= 1 - \alpha + O\big(n^{-(s \wedge 2r)/(2r+1)}\big).
\end{aligned} \tag{4.5}
$$

Equations (4.4) and (4.5) are also proved in Neumann (1992b).

4.2. *The effect of the bandwidth choice on the coverage accuracy.* Now we are going to consider the performance of confidence intervals in practical situations, that is, intervals based on pivotal quantities $T_{\hat{h},\hat{g}}$ and $U_{\hat{h}}$ involving estimators with data-driven bandwidths. These quantities do not immediately admit Edgeworth expansions, because they cannot be written as smooth functions of a sum of independent random vectors. We will use the approximations to $T_{h,g}$ and U_h given by Proposition 4.1 and treat the differences between $T_{\hat{h},\hat{g}}$ and $T_{h,g}$ as well as between $U_{\hat{h}}$ and U_h by estimates based on $\tilde{O}$.

First, we consider the order of approximation of the optimal bandwidth by their estimates considered in Section 2. From here let $\delta > 0$ be an arbitrarily small quantity, whose occurrence is explained by application of Lemma A.1.

LEMMA 4.1. *Under* (A_G) *and* (A_{BW}) *we have*

$$\frac{\hat{h} - h_0}{h_0} = \tilde{O}(n^{\delta} n^{-\Delta}, n^{-1}),$$

where $\Delta = 1/2 \wedge s/(s + 2r + 1/2)$.

On the basis of Lemma A.1 it is easy to see that

$$\frac{d}{dg}\{T_{\hat{h},g}\} = \tilde{O}(n^{\delta} g^{s-1}, n^{-1}),$$

$$\frac{d}{dh}\{T_{h,g}\} = \tilde{O}(n^{\delta} h^{-1}, n^{-1})$$

and

$$\frac{d}{dh}\{\hat{\rho}_h\} = \tilde{O}\big(n^{\delta} h^{-1}(nh)^{-1/2}, n^{-1}\big).$$

From the decomposition

$$T_{\hat{h},\hat{g}} - \frac{2u_{1-\alpha}^2 + 1}{6}\hat{\rho}_{\hat{h}} = \left(T_{h,g} - \frac{2u_{1-\alpha}^2 + 1}{6}\hat{\rho}_h\right) + \frac{2u_{1-\alpha}^2 + 1}{6}\big(\hat{\rho}_h - \hat{\rho}_{\hat{h}}\big) + \big(T_{\hat{h},\hat{g}} - T_{\hat{h},g}\big) + \big(T_{\hat{h},g} - T_{h,g}\big)$$

we obtain, by Lemmas 3.1 and 4.1 and by (4.4), the following assertion.

THEOREM 4.1. *Assume* (A_G), (A_U), (A_{BW}), (A_E) *and* $|\hat{g} - g_0|/g_0 = \tilde{O}(n^{-\gamma}, n^{-1})$ *for some* $\gamma > 0$. *Then*

$$P\big(m(x_0) \in I_{\hat{h},\hat{g}}\big) = P\big(m(x_0) \in I_{h_0,g_0}\big) + O(n^{\delta} n^{-\Delta}) = 1 - \alpha + O(n^{-s/[2(r+s)+1]}).$$

For confidence intervals based on the pivotal statistic $U_{\hat{h}}$ we can derive in an analogous way estimates for the error in coverage probability. However, since U_h yields for nonrandom h better rates than $T_{h,g}$, the error due to the randomness of $\hat{h}$ is not automatically majorized by the error in coverage probability of the confidence interval with nonrandom bandwidths. Hence, we look for a better approximation to $\hat{h}$. The idea is quite simple. Neglecting the effect of the estimator $\hat{m}_{\hat{f}}$ involved in the $\hat{v}_i$'s, the pivotal statistic U_h depends only on $O(nh)$ of the n observations, whereas the bandwidth selector uses all of them to a certain amount. We define another bandwidth $\tilde{h}$ by a similar criterion, where only the observations in some neighborhood of x_0 of size $O(h_0)$ are excluded, such that the quantity $U_{\tilde{h}}$ is based on a set of observations disjoint from that used for the choice of $\tilde{h}$. Then the conditional distribution of $U_{\tilde{h}}$ under $\tilde{h}$ is the same as the unconditional distribution of U_h at the point $h = \tilde{h}$. Thus Proposition 4.1 remains valid for $U_{\tilde{h}}$ as well, and

because $\tilde{h}$ approximates $\hat{h}$ better than h_0, we obtain a better estimate for the error in coverage probability than via an approximation to $U_{\hat{h}}$ by U_{h_0}.

For appropriate C, let

$$\Delta_n = Cn^{-1/(2r+1)},$$
$$J_n = \{i \in \{1, \ldots, n\} \mid |x_i - x_0| \le \Delta_n\}.$$

We replace $\widehat{\mathrm{MISE}}(h, \bar{g})$ by

$$\tilde{M}(h) = \widetilde{\mathrm{IV}}(h) + \widetilde{\mathrm{ISB}}(h), \tag{4.6}$$

where

$$\widetilde{\mathrm{IV}}(h) = \sum_{i \notin J_n} \int_c^d W(x, h)_i^2 \, dx \, \varepsilon_i^2 + \sum_{i \in J_n} \int_c^d W(x, h)_i^2 \, dx \, v(x_i)$$

and

$$\widetilde{\mathrm{ISB}}(h) = \sum_{i, j \notin J_n} A(h, g_0)_{ij} Y_i Y_j + \sum_{(i,j):\, i \in J_n \text{ or } j \in J_n} A(h, g_0)_{ij} E Y_i Y_j - \sum A(h, g_0)_{ii} v(x_i),$$

and define $\tilde{h}$ as a measurable function with

$$\tilde{h} \in \operatorname*{arg\,min}_{h \in [h_0/2, 3h_0/2]} \tilde{M}(h).$$

Let the constant C be chosen so large that $U_{\tilde{h}}$ and $\tilde{M}(h)$ are based on disjoint sets of observations.

Now we can prove analogously to Lemma 4.1 that

$$\frac{|\hat{h} - \tilde{h}|}{h_0} = \tilde{O}(n^{\delta} n^{-\Delta'}, n^{-1}), \tag{4.7}$$

where $\Delta' = (1/2 + 1/(2(2r + 1))) \wedge s/(s + 2r + 1/2)$.

The additional factor $n^{-1/2(2r+1)}$ comes into play, because, roughly speaking, the number of the Y_i's included in $\tilde{M}(h) - \widehat{\mathrm{MISE}}(h, g_0)$ is $O(n\Delta_n)$ rather than $O(n)$ as in $\widehat{\mathrm{MISE}}(h, \bar{g})$.

Again by Lemma A.1 we obtain

$$\frac{d}{dh}\{U_h\} = \tilde{O}(n^{\delta} h^{-1}, n^{-1}) \quad \text{and} \quad \frac{d}{dh}\{\hat{\bar{\rho}}_h\} = \tilde{O}\big(n^{\delta} h^{-1} (nh)^{-1/2}, n^{-1}\big).$$

With the decomposition

$$U_{\hat{h}} - \frac{2u_{1-\alpha}^2 + 1}{6} \hat{\bar{\rho}}_h = \left(U_{\hat{h}} - \frac{2u_{1-\alpha}^2 + 1}{6} \hat{\bar{\rho}}_{\hat{h}}\right) + \frac{2u_{1-\alpha}^2 + 1}{6}\left(\hat{\bar{\rho}}_{\hat{h}} - \hat{\bar{\rho}}_{\hat{h}}\right) + \left(U_{\hat{h}} - U_{\hat{h}}\right)$$

we obtain, by Proposition 4.1(ii), (4.7) and Lemma 3.1, the following theorem.

THEOREM 4.2. *Under* (A_G), (A_U), (A_{BW}) *and* (A_E) *we have*

$$P\big(m(x_0) \in \tilde{I}_{\hat{h}}\big) = P\big(m(x_0) \in \tilde{I}_{\tilde{h}}\big) + O(n^{\delta} n^{-\Delta'})$$
$$= 1 - \alpha + O(n^{-(s \wedge 2r)/(2r+1)} + n^{\delta} n^{-\Delta'}).$$

Comparing the results of the Theorems 4.1 and 4.2 we see that the undersmoothing method retains its superiority to the explicit bias-correction also in the case of data-dependent bandwidths chosen by the above criteria.

5. Discussion.

5.1. Our estimates via the $\tilde{O}$-calculations seem to be, on first sight, somewhat rough, and there arises the question whether we would obtain better estimates by formal Edgeworth expansions of the pivotal quantities $T_{\hat{h},\hat{g}}$ and $U_{\hat{h}}$, respectively. Apart from the fact that the validity of these expansions is not immediately clear, it turns out that we would obtain the same rates as given by Theorems 4.1 and 4.2 with the exception of the factor n^{δ}. To see this, expand $U_{\hat{h}}$ in the Taylor series

$$U_{\hat{h}} = U_{\tilde{h}} + (\hat{h} - \tilde{h}) U'_h|_{h=\tilde{h}} + \frac{(\hat{h} - \tilde{h})^2}{2} U''_h\Big|_{h=h^*},$$

where h^* is between $\tilde{h}$ and $\hat{h}$. The third term on the right-hand side is of negligible order. All arguments can be conditioned on $\tilde{h}$, since the conditional distribution of $U_{\tilde{h}}$ is equal to the unconditional distribution of U_h at the point $h = \tilde{h}$. If we follow the proof of (4.7), we see that the leading term of order $h_0 n^{-\Delta'}$ of $\hat{h} - \tilde{h}$ is given by

$$\frac{\sum_{i \in J_n} W_i\big(\varepsilon_i^2 - v(x_i)\big) + \sum_{(i,j):\, i \in J_n \text{ or } j \in J_n} A(h, g_0)_{i,j}\big(m(x_i)\varepsilon_j + \varepsilon_i m(x_j)\big)}{-\tilde{M}''(\tilde{h})}.$$

On the other hand, we have

$$U'_h = \frac{\hat{m}'_h(x_0)}{\hat{V}_h^{1/2}} - \frac{\hat{m}_h(x_0)\hat{V}'_h}{2\hat{V}_h^{3/2}},$$

which depends mainly on Y_i's with $i \in J_n$ and has an order of magnitude of $O(h_0^{-1})$. Therefore, the second term of the above Taylor series contributes, by a term of order $n^{-\Delta'}$, to the first cumulant of $U_{\hat{h}}$, which leads to a difference of order at least $n^{-\Delta'}$ between the Edgeworth expansions of $U_{\tilde{h}}$ and $U_{\hat{h}}$.

5.2. One disappointing fact with confidence intervals in nonparametric regression is that we cannot obtain a size of the intervals that shrinks with the same rate as the standard deviation of optimal estimators. The reason is that we have actually two more or less separate problems, the estimation of

$m(x_0)$ by some estimator $\hat{m}(x_0)$ as well as the recognition of the distribution of $\hat{m}(x_0) - m(x_0)$, which essentially consists of the estimation of the bias $E\hat{m}(x_0) - m(x_0)$. To solve both problems satisfactorily, we have to apportion the smoothness assumption for both purposes, which requires the application of a suboptimal estimator $\hat{m}(x_0)$.

5.3. The methods used in the present paper can obviously be applied to kernel estimators with bandwidths chosen by other selectors. Neumann (1992a) shows that the cross-validation bandwidth $\hat{h}_{CV}$ can be approximated by some random bandwidth $\tilde{h}$, which is independent of those observations that enter into the estimator $\hat{m}_{\tilde{h}}(x_0)$, to an order of $\tilde{O}(h_0 n^{-1/(2r+1)} n^{\delta}, n^{-\lambda})$, which yields finally an error in coverage probability of $O(n^{\delta} n^{-1/(2r+1)})$. Another direction for an extension are alternative kernel estimators, as, for example, those of Nadaraya–Watson type.

5.4. A referee of this paper pointed out that a local bandwidth choice of h is perhaps more appropriate than a global one, and he gave some indication that the integral of the lengths of the intervals can be expected to be smaller for optimal local than for optimal global bandwidth choice. We agree that local bandwidth choice is a reasonable alternative to our proposal. However, we recall the trade-off between length of the confidence interval and bias of the estimator: to get a smaller length, we need a larger bandwidth, which results in a larger bias, also after the bias correction. Moreover, data-driven local bandwidths are usually less stable than global ones, which can influence the rate for the error in coverage probability, as shown in Section 4.2. Finally, we remark that neither a locally nor a globally optimal bandwidth for $\hat{m}_h$ is optimal in any sense for the ultimate estimator $\hat{\hat{m}}(x_0) = \hat{m}_h(x_0) - \hat{B}_h$ in our undersmoothing approach.

5.5. Another point which was raised by a referee is that our automatically chosen bandwidths are not optimal in any sense for the constructed confidence intervals. However, we think that a specific choice for confidence intervals is quite difficult in our completely nonparametric approach. Even from the theoretical point of view it is not clear which bandwidth should be considered to be optimal. There exist different reasonable aspects for the performance of confidence intervals, as, for example, coverage accuracy, length or a connection between length and miscentering of the interval as proposed by Beran (1986) in the discussion of Wu (1986), which would lead to different optimal bandwidths.

APPENDIX

A.1. Assumptions. Here we list the assumptions needed for the assertions in the previous sections.

GENERAL ASSUMPTIONS (A_G).

(i) The design points $x_i = x_i(n)$ are regularly spaced, that is, $\int_0^{x_i} d(t)\,dt = (i - 1/2)/n$, for some positive, continuous density d on $[0, 1]$;

(ii) $w \in C^0[-1, 1]$ is a kernel function of order $r \geq 2$, that is,

$$\int w(z) z^k\,dz = \begin{cases} 1, & \text{if } k = 0, \\ 0, & \text{if } k = 1, \ldots, r-1, \\ C \neq 0, & \text{if } k = r; \end{cases}$$

(iii) $\tilde{w} \in C^r[-1, 1]$ is a kernel function of order $s \geq 2$;

(iv) $m \in C^{r+s}[0, 1]$.

ASSUMPTION FOR UNIFORM APPROXIMATIONS (A_U). All moments of the ε_i's are uniformly bounded by corresponding constants, that is, $E|\varepsilon_i|^M \leq C(M)$ for some $C(M) < \infty$. (If we assume instead that only a finite number of moments are bounded, then we have to choose δ in dependence on this number and on the entropy of the families of vectors and matrices, as can be seen in the proof of Lemma A.1.)

ASSUMPTIONS ESPECIALLY FOR THE CHOICE OF THE OPTIMAL GLOBAL BANDWIDTH (A_{BW}).

(i) $m \in C^{2r+s}[0, 1]$, $\int_c^d (m^{(r)}(x))^2\,dx \neq 0$,

$$\int_c^d \left[\int \frac{1}{h} w_{x,h}\left(\frac{z - x}{h}\right) m(z)\,dz - m(x) \right]^2 dx \neq 0 \quad \text{for all } h > 0;$$

(ii) $v \in C^0[0, 1]$ is bounded from zero.

ASSUMPTIONS FOR EDGEWORTH EXPANSIONS (A_E).

(i) A sufficiently large number of moments of the ε_i's are uniformly bounded;

(ii) Cramér's condition is uniformly satisfied by the random vectors $\alpha_i = (\varepsilon_i, \varepsilon_i^2, \varepsilon_i^3)'$ in some neighborhood of x_0, that is,

$$\sup_{i:\,|x_i - x_0| \leq C} \sup_{\|t\| > b} |E \exp\{it'\alpha_i\}| < 1,$$

for some $C > 0$ and all $b > 0$.

A.2. Some technical lemmas.

LEMMA A.1 (Uniform $\tilde{O}$-approximation). *Let $\mathscr{A}^n = \{a_\theta^{(n)}\}_{\theta \in \Theta}$ and $\mathscr{A}^{n \times n} = \{A_\theta^{(n)}\}_{\theta \in \Theta}$ be families of n-vectors and $(n \times n)$-matrices, respectively. Further, define the ε-entropy $E_\varepsilon(\mathscr{A}^{n \times n})$ of $\mathscr{A}^{n \times n}$ as the minimal number of $(n \times n)$-matrices A_i with the property that each $A \in \mathscr{A}^{n \times n}$ can be approximated by some A_i with $\|A - A_i\| \leq \varepsilon$. Analogously, we define the ε-entropy $E_\varepsilon(\mathscr{A}^n)$ of $\mathscr{A}^n$. Assume (A_U), $E_{n^{-1/2-\beta}}(\mathscr{A}^n) = O(n^\lambda)$ and $E_{n^{-1-\beta}}(\mathscr{A}^{n \times n}) = O(n^\lambda)$ for some $\beta > 0$ and $\lambda < \infty$. Then*:

(i) $$\sup_{\theta \in \Theta} \left\{ \left(\|a_\theta^{(n)}\| + n^{-\beta} \right)^{-1} |a_\theta^{(n)'} \boldsymbol{\varepsilon}| \right\} = \tilde{O}(n^\delta, n^{-\gamma}),$$

$$\text{(ii)}\quad \sup_{\theta\in\Theta}\left\{\left(\sqrt{\operatorname{tr}\left(A_\theta^{(n)}A_\theta^{(n)'}\right)}+n^{-\beta}\right)^{-1}\left|\boldsymbol{\varepsilon}'A_\theta^{(n)}\boldsymbol{\varepsilon}-E\boldsymbol{\varepsilon}'A_\theta^{(n)}\boldsymbol{\varepsilon}\right|\right\}=\tilde{O}(n^{\delta},n^{-\gamma})$$

hold for $\delta>0$ *and* $\gamma<\infty$, *which can be chosen arbitrarily small and large, respectively.*

PROOF. For a one-element set $\Theta=\{\theta_0\}$ we obtain (i) and (ii) by Markov's and Whittle's inequalities [see Whittle (1960)]. For general Θ we derive (i) and (ii) on the basis of that set of vectors and matrices just given by the definition of the $n^{-1/2-\beta}$-entropy and $n^{-1-\beta}$-entropy, respectively. Let $\hat{\theta}$ denote this parameter from the approximating grid with $\|a_\theta^{(n)}-a_{\hat{\theta}}^{(n)}\|\le n^{-1/2-\beta}$. By Markov's, Whittle's and Bonferroni's inequalities we obtain that, for arbitrary positive δ and γ,

$$\begin{aligned}\|(a_\theta^{(n)})'\boldsymbol{\varepsilon}\| &\le \|(a_{\hat{\theta}}^{(n)})'\boldsymbol{\varepsilon}\|+\|a_\theta^{(n)}-a_{\hat{\theta}}^{(n)}\|\,\|\boldsymbol{\varepsilon}\|\\ &= O\left(n^{\delta}\|a_{\hat{\theta}}^{(n)}\|+n^{-1/2-\beta}n^{1/2+\delta}\right)\\ &= O\left(n^{\delta}\|a_\theta^{(n)}\|+n^{\delta}n^{-\beta}\right)\end{aligned}$$

holds uniformly over $\theta\in\Theta$ with a probability exceeding $1-O(n^{-\gamma})$, which implies (i). Statement (ii) can be proved analogously. □

The next lemma improves the remainder term of order n^{-1} given in Gasser and Müller (1979) for the expectation of their kernel estimator.

LEMMA A.2. *Let* $w_{x,h}$ *be uniformly* (*in* x *and* h) *Lipschitz continuous of order* 1, *and let* $\{g_n\}$ *be a sequence of twice-differentiable functions. Further, assume that the design satisfies the condition given in* (A_G). *Then*

$$\begin{aligned}\sum_{j=1}^{n}W(x,h)_j g_n(x_j) &= \int_0^1\frac{1}{h}w_{x,h}\left(\frac{z-x}{h}\right)g_n(z)\,dz\\ &\quad+O\left(n^{-2}h^{-1}\sup_{0\le z\le 1}\{|g_n'(z)|\}+n^{-2}\sup_{0\le z\le 1}\{|g_n''(z)|\}\right).\end{aligned}$$

The proof of this lemma is straightforward and can be found in Neumann (1992b).

A.3. Proofs.

PROOF OF LEMMA 2.1. The calculations are very similar to those in the proof of Theorem 1 in HHM. Therefore we indicate only the sources of the terms in (i). Some of these formulas will be used in the course of the proof of

Lemma 4.1. First, we approximate the entries of the matrix $A(h,g)$ by

$$A_{ij} = \frac{h^{2r}}{g^{2r+1}}(s_i - s_{i-1})(s_j - s_{j-1})\kappa_r^2 \int \tilde{w}^{(r)}(x)\tilde{w}^{(r)}\left(x + \frac{x_i - x_j}{g}\right) dx + o\left(\frac{h^{2r}}{g^{2r+1}}n^{-2}\right) + O(n^{-4}h^{-4}g) \quad \text{if } |x_i - x_j| \le Cg, \tag{A.1}$$

whereas $A_{ij} = 0$ holds if $|x_i - x_j| > Cg$. Further, we have, by $m \in C^{2r+s}[0,1]$,

$$(A(h,g)m)_i = C(h,i)n^{-1}h^{2r} + O(n^{-1}h^{2r}g^s) + O(n^{-3}h^{r-1}g^{-r-1} + n^{-3}h^{r-3}). \tag{A.2}$$

Now we split up

$$\begin{aligned}\operatorname{Var}(\overline{\mathrm{ISB}}(h,g)) &= \operatorname{Var}(\boldsymbol{\varepsilon}'A(h,g)\boldsymbol{\varepsilon}) + 4\operatorname{Var}(\mathbf{m}'A(h,g)\boldsymbol{\varepsilon}) \\ &\quad + 4\operatorname{Cov}(\boldsymbol{\varepsilon}'A(h,g)\boldsymbol{\varepsilon}, \mathbf{m}'A(h,g)\boldsymbol{\varepsilon}),\end{aligned} \tag{A.3}$$

where $\mathbf{m} = (m(x_1), \dots, m(x_n))'$, and we estimate the terms on the right-hand side separately. Those terms which enter into formula (i) are underlined. We have

$$\begin{aligned}&\operatorname{Var}(\boldsymbol{\varepsilon}'A(h,g)\boldsymbol{\varepsilon}) \\ &\quad = 2\sum_{i,j} A_{ij}^2 v(x_i)v(x_j)(1 + o(1)) \\ &\quad = 2\frac{h^{4r}}{g^{4r+1}}n^{-2}\kappa_r^4 \underline{\int_0^1 \left(\frac{v(x)}{d(x)}\right)^2 dx \int \left(\int \tilde{w}^{(r)}(y)\tilde{w}^{(r)}(y+z)\,dy\right)^2 dz} \\ &\qquad \times (1 + o(1)).\end{aligned} \tag{A.4}$$

Further, we obtain, by (A.2),

$$\begin{aligned}\operatorname{Var}(\mathbf{m}'A(h,g)\boldsymbol{\varepsilon}) &= \mathbf{m}'A(h,g)\operatorname{Diag}[v(x_1), \dots, v(x_n)]A(h,g)\mathbf{m} \\ &= \underline{h^{4r}C_3(h)n^{-1}} + O(h^{4r}n^{-1}g^s)\end{aligned} \tag{A.5}$$

and, by (A.1),

$$\operatorname{Cov}(\boldsymbol{\varepsilon}'A(h,g)\boldsymbol{\varepsilon}, \mathbf{m}'A(h,g)\boldsymbol{\varepsilon}) = O(h^{4r}n^{-2}g^{-2r-1}), \tag{A.6}$$

where the residual terms both will be majorized by the underlined term in (A.4). By

$$\operatorname{Var}\left(\widehat{\overline{\mathrm{ISB}}}(h,g)\right) = \operatorname{Var}(\overline{\mathrm{ISB}}(h,g)) + O\left(\sqrt{\operatorname{Var}(\overline{\mathrm{ISB}}(h,g))}\sqrt{\operatorname{Var}\left(\sum A_{ii}\hat{v}_i\right)} + \operatorname{Var}\left(\sum A_{ii}\hat{v}_i\right)\right)$$

and

$$\operatorname{Var}\left(\sum A_{ii}\hat{v}_i\right) \le E\left(\sum A_{ii}(\hat{v}_i - v_i)\right)^2 = O(h^{4r}n^{-3}g^{-4r-2}),$$

we see that the remaining residual terms do not enter into the asymptotic formula. Finally, we have

$$
\begin{aligned}
&E\widehat{\mathrm{ISB}}(h,g) - \mathrm{ISB}(h) \\
&\quad = \mathbf{m}'A(h,g)\mathbf{m} - \mathrm{ISB}(h) + E\sum A_{ii}(v_i - \hat{v}_i) \\
&\quad = 2h^{2r}g^{s}\kappa_r^2\lambda_s\int m^{(r+s)}(x)m^{(r)}(x)\,dx(1+o(1)) \\
&\qquad + O(n^{-2}) + O\bigl(h^{2r}n^{-3/2}g^{-2r-1}\bigr),
\end{aligned}
\tag{A.7}
$$

which completes the calculations needed for the proof of (i). □

PROOF OF LEMMA 4.1. First, we investigate how well MISE(h) is approximated by its estimate $\widehat{\mathrm{MISE}}(h,\bar{g})$. Let $W_i = \int_c^d W(x,h)_i^2\,dx$. We split up

$$
\begin{aligned}
\widehat{\mathrm{IV}}(h) - \mathrm{IV}(h) &= \sum_i W_i\bigl(\varepsilon_i^2 - v(x_i)\bigr) \\
&\quad + \sum_i W_i\bigl(m(x_i) - \hat{m}_{\hat{f}}(x_i)\bigr)^2 \\
&\quad + 2\sum_i W_i\varepsilon_i\Bigl(m(x_i) - \sum_j \overline{W}\bigl(x_i,\hat{f}\bigr)_j m(x_j)\Bigr) \\
&\quad - 2\sum_i W_i\varepsilon_i\sum_j \overline{W}\bigl(x_i,\hat{f}\bigr)_j\varepsilon_j \\
&= T_1 + \cdots + T_4.
\end{aligned}
$$

By means of Lemma A.1 we can easily estimate the terms T_1 through T_4. For convenience, let h first be restricted to the interval $[n^{-1},1/2]$. Using $W_i = O(n^{-2}h^{-1})$, we get

$$
T_1 = \tilde{O}\bigl((nh)^{-1}n^{\delta-1/2}, n^{-\lambda}\bigr).
$$

By $n^{-1}\Sigma(m(x_i) - \hat{m}_{\hat{f}}(x_i))^2 = \tilde{O}((n\hat{f})^{-1} + \hat{f}^{2(r+s)}, n^{-\lambda})$ and $\hat{f} = f_0 + \tilde{O}(f_0^{3/2}n^{\delta}, n^{-\lambda})$, with some $f_0 \asymp n^{-1/(2(r+s)+1)}$, we obtain

$$
T_2 = \tilde{O}\bigl((nh)^{-1}n^{-2(r+s)/[2(r+s)+1]}, n^{-\lambda}\bigr)
$$

and

$$
\begin{aligned}
T_3 &= \tilde{O}\left(\sqrt{\sum_i\left(W_i\Bigl[m(x_i) - \sum_j \overline{W}\bigl(x_i,\hat{f}\bigr)_j m(x_j)\Bigr]\right)^2}\,n^{\delta}, n^{-\lambda}\right) \\
&= \tilde{O}\bigl(n^{-2}h^{-1}n^{1/2}\hat{f}^{r+s}n^{\delta}, n^{-\lambda}\bigr) \\
&= \tilde{O}\bigl((nh)^{-1}n^{-1/2}n^{-(r+s)/[2(r+s)+1]}n^{\delta}, n^{-\lambda}\bigr).
\end{aligned}
$$

If we write T_4 in the form $\boldsymbol{\varepsilon}'M(h,\hat{f})\boldsymbol{\varepsilon}$, we obtain, by the relation $\mathrm{tr}(M(h,f)'M(h,f)) = O(n^{-1}(nh)^{-2}(nf)^{-1})$ and Whittle's inequality for

quadratic forms, the following estimate:

$$\begin{aligned} T_4 &\le \left|\boldsymbol{\varepsilon}' M\left(h, \hat{f}\right)\boldsymbol{\varepsilon} - E\boldsymbol{\varepsilon}' M(h, f)\boldsymbol{\varepsilon}|_{f=\hat{f}}\right| + E\boldsymbol{\varepsilon}' M(h, f)\boldsymbol{\varepsilon}|_{f=\hat{f}} \\ &= \tilde{O}\left(\sqrt{\operatorname{tr}\left(M\left(h, \hat{f}\right)' M\left(h, \hat{f}\right)\right)}\, n^{\delta}, n^{-\lambda}\right) + O\left(\sum_j W_j \overline{W}\left(x_j, \hat{f}\right)_j\right) \\ &= \tilde{O}\left((nh)^{-1} n^{-1/2} n^{-(r+s)/[2(r+s)+1]} n^{\delta}, n^{-\lambda}\right) \\ &\quad + \tilde{O}\left((nh)^{-1} n^{-2(r+s)/[2(r+s)+1]}, n^{-\lambda}\right). \end{aligned}$$

Next, we decompose

$$\begin{aligned} \widehat{\mathrm{ISB}}(h, \bar{g}) - \mathrm{ISB}(h) &= \boldsymbol{\varepsilon}' A(h, \bar{g})\boldsymbol{\varepsilon} - \sum A(h, \bar{g})_{ii} v(x_i) \\ &\quad + \sum_i A(h, \bar{g})_{ii}\left(v(x_i) - \hat{v}_i\right) \\ &\quad + 2\mathbf{m}' A(h, \bar{g})\boldsymbol{\varepsilon} \\ &\quad + \mathbf{m}' A(h, \bar{g})\mathbf{m} - \mathrm{ISB}(h) \\ &= T_5 + \cdots + T_8. \end{aligned}$$

By (A.1) we see

$$\begin{aligned} T_5 &= \tilde{O}\left(\sqrt{\sum A(h, \bar{g})_{ij}^2}\, n^{\delta}, n^{-\lambda}\right) \\ &= \tilde{O}\left(h^{2r} n^{-1} \bar{g}^{-2r-1/2} n^{\delta}, n^{-\lambda}\right). \end{aligned}$$

Analogously to the estimation of $\widehat{\mathrm{IV}}(h) - \mathrm{IV}(h)$, we obtain

$$\begin{aligned} T_6 &= \tilde{O}\left(h^{2r} n^{-2} \bar{g}^{-2r-1} n^{1/2} n^{\delta}, n^{-\lambda}\right) \\ &= \tilde{O}\left(h^{2r} n^{-1} \bar{g}^{-2r-1/2} (n\bar{g})^{-1/2} n^{\delta}, n^{-\lambda}\right), \\ T_7 &= \tilde{O}\left(\|A(h, \bar{g})\mathbf{m}\| n^{\delta}, n^{-\lambda}\right) \\ &= \tilde{O}\left(h^{2r} n^{-1/2} n^{\delta}, n^{-\lambda}\right) \end{aligned}$$

and

$$T_8 = \tilde{O}\left(h^{2r} \bar{g}^{s}, n^{-\lambda}\right).$$

Analogous estimates can be derived for $h \in [0, n^{-1})$, where $O(1)$-terms take the place of the $(nh)^{-1}$-terms. It is known that

$$\mathrm{MISE}(h) \ge C\left(h^{2r} + \left((nh)^{-1} \wedge 1\right)\right), \tag{A.8}$$

which implies in conjunction with the above calculations that

$$\frac{\widehat{\mathrm{MISE}}(h, \bar{g}) - \mathrm{MISE}(h)}{\mathrm{MISE}(h)} = \tilde{O}\left(n^{-\Delta} n^{\delta}, n^{-\lambda}\right).$$

On the other hand, we have $\mathrm{MISE}(h_0) = O(n^{-2r/(2r+1)})$, which implies by (A.8) that

$$P\left(|\hat{h} - h_0| > \frac{h_0}{2}\right) = O\left(n^{-\lambda}\right). \tag{A.9}$$

For brevity we set $\hat{M}(h) = \widehat{\mathrm{MISE}}(h, \bar{g})$ and $M(h) = \mathrm{MISE}(h)$. Because $\hat{M}'(h)|_{h=\hat{h}} = M'(h)|_{h=h_0} = 0$, we obtain

$$\begin{aligned} 0 &= (\hat{M} - M)'(\hat{h}) + \big(M'(\hat{h}) - M'(h_0)\big) \\ &= (\hat{M} - M)'(\hat{h}) + \big(\hat{h} - h_0\big)M''(h^*) \end{aligned}$$

for some h^* between h_0 and $\hat{h}$, which implies

$$\hat{h} - h_0 = \frac{(\hat{M} - M)'(\hat{h})}{-M''(h^*)}.$$

By straightforward calculations one obtains that

$$M''(h) \asymp n^{-1}h^{-3} \tag{A.10}$$

holds for $h \asymp h_0$. The term $(\hat{M} - M)'(h)$ can be decomposed in the same way as $\hat{M}(h) - M(h)$. It turns out that T_1' through T_8' are of the same order as T_1 through T_8, respectively, with an additional factor of order h_0^{-1}. Hence, we have

$$(\hat{M} - M)'(\hat{h}) = \tilde{O}\big(n^{-1}h_0^{-2}n^{-\Delta}n^{\delta}, n^{-\lambda}\big),$$

which implies

$$\hat{h} - h_0 = \tilde{O}\big(h_0 n^{-\Delta}n^{\delta}, n^{-\lambda}\big). \qquad \square$$

PROOF OF (4.7). The proof of this equation is very similar to that of Lemma 4.1.

First, one can derive, analogously to (A.9), that

$$P\left(|\tilde{h} - h_0| > \frac{h_0}{2}\right) = O(n^{-\lambda}) \tag{A.11}$$

holds. By $\hat{M}'(h)|_{h=\hat{h}} = \tilde{M}'(h)|_{h=\tilde{h}} = 0$, we obtain

$$\begin{aligned} 0 &= (\hat{M} - \tilde{M})'(\hat{h}) + \big(\tilde{M}'(\hat{h}) - \tilde{M}'(\tilde{h})\big) \\ &= (\hat{M} - \tilde{M})'(\hat{h}) + (\hat{h} - \tilde{h})\tilde{M}''(h^{**}) \end{aligned}$$

for some h^{**} between $\tilde{h}$ and $\hat{h}$, which implies

$$\hat{h} - \tilde{h} = \frac{(\hat{M} - \tilde{M})'(\hat{h})}{-\tilde{M}''(h^{**})}.$$

By (A.10) we can prove, due to Lemma A.1, that $\tilde{M}''(h^{**})^{-1} = \tilde{O}(nh_0^3, n^{-\lambda})$. We have

$$\begin{aligned} \widehat{\mathrm{IV}}(h) - \widetilde{\mathrm{IV}}(h) &= \sum_{i \in J_n} W_i\big(\varepsilon_i^2 - v(x_i)\big) + T_2 + T_3 + T_4 \\ &= T_1^* + T_2 + T_3 + T_4. \end{aligned}$$

Let $h \asymp h_0$. By $\#J_n = O(n\Delta_n)$, we obtain

$$(T_1^*)' = \tilde{O}\big((nh)^{-1}h^{-1}n^{-1/2}\Delta_n^{1/2}n^{\delta}, n^{-\lambda}\big),$$

which differs from T_1' by the factor $\Delta_n^{1/2}$. It can be seen that the terms T_2' through T_4' are all majorized by $(T_1^*)'$. Next, we decompose

$$\begin{aligned}
&\widehat{\mathrm{ISB}}(h,\bar g) - \widetilde{\mathrm{ISB}}(h,g_0)\\
&\quad = T_5 - \sum_{i,j\notin J_n} A(h,g_0)\varepsilon_i\varepsilon_j + \sum_{i\notin J_n} A(h,g_0)_{ii}v(x_i)\\
&\qquad + T_6\\
&\qquad + \mathbf{m}'A(h,\bar g)\mathbf{m} - \mathrm{ISB}(h) - \mathbf{m}'A(h,g_0)\mathbf{m} + \mathrm{ISB}(h)\\
&\qquad + 2\mathbf{m}'\big(A(h,\bar g) - A(h,g_0)\big)\boldsymbol{\varepsilon}\\
&\qquad + 2\sum_{(i,j):\, i\in J_n \text{ or } j\in J_n} A(h,g_0)_{ij}m(x_i)\varepsilon_j\\
&\quad = U_1 + \cdots + U_5.
\end{aligned}$$

The terms U_1', U_2' and U_3' are of the same order as T_5', T_6' and T_8', respectively.

By (A.2) we conclude that $\|(A(h,\bar g) - A(h,g_0))\mathbf{m}\| = \tilde O(h^{2r}n^{-1/2}g_0^s n^\delta, n^{-\lambda})$ holds, which implies that

$$U_4' = \tilde O\big(h^{2r}n^{-1/2}g_0^s n^\delta, n^{-\lambda}\big).$$

Finally, we have

$$U_5' = \tilde O\big(h^{2r}n^{-1/2}h^{-1}\Delta_n^{1/2}n^\delta, n^{-\lambda}\big).$$

Collecting the upper estimates for $(T_1^*)'$, T_2' through T_4' and U_1' through U_5', we obtain

$$(\hat M' - \tilde M')(\hat h) = \tilde O\big(n^{-1}h_0^{-2}n^{-\Delta'}n^\delta, n^{-\lambda}\big),$$

which yields (4.7). □

Acknowledgments. Some part of the research was carried out during a visit of the author to the INRA, Jouy-en-Josas, France. I would like to thank Emmanuel Jolivet for suggesting this interesting subject and for helpful discussions. Further, I thank Wolfgang Härdle and Steve Marron for their interest and valuable comments. Finally, I thank the referees and an Associate Editor for their critical remarks that led to a considerable improvement of this paper.

REFERENCES

BERAN, R. (1986). Discussion of "Jackknife, bootstrap and other resampling methods in nonparametric regression analysis," by C. F. J. Wu. *Ann. Statist.* **14** 1295–1298.

FARAWAY, J. (1990). Bootstrap selection of bandwidth and confidence bands for nonparametric regression. *J. Statist. Comput. Simulation* **37** 37–44.

FARAWAY, J. and JHUN, M. (1990). Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc.* **85** 1119–1122.

GASSER, T. and MÜLLER, H. G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 23–68. Springer, New York.

GASSER, T., MÜLLER, H.-G. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238–252.

HALL, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics* **22** 215–232.

HALL, P. (1992a). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Statist.* **20** 675–694.

HALL, P. (1992b). On bootstrap confidence intervals in nonparametric regression. *Ann. Statist.* **20** 695–711.

HÄRDLE, W. and BOWMAN, A. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83** 102–110.

HÄRDLE, W., HALL, P. and MARRON, J. S. (1992). Regression smoothing parameters that are not far from their optimum. *J. Amer. Statist. Assoc.* **87** 227–233.

HÄRDLE, W., HUET, S. and JOLIVET, E. (1995). Better bootstrap confidence intervals for regression curve estimation. *Statistics*. To appear.

HÄRDLE, W. and MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21** 1926–1947.

JENNRICH, R. I. (1969). Asymptotic properties of nonlinear least squares estimators. *Ann. Math. Statist.* **40** 633–643.

JONES, M. C. and SHEATHER, S. J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **11** 511–514.

MÜLLER, H. G. (1988). *Nonparametric Regression Analysis of Longitudinal Data. Lecture Notes in Statist.* **46**. Springer, Berlin.

MÜLLER, H. G. and STADTMÜLLER, U. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.* **15** 182–201.

NEUMANN, M. H. (1992a). On completely data-driven pointwise confidence intervals in nonparametric regression. Rapport Technique 92-02, INRA, Dépt. Biométrie, Jouy-en-Josas, France.

NEUMANN, M. H. (1992b). Pointwise confidence intervals in nonparametric regression with heteroscedastic error structure. Preprint No. 34, Institut für Angewandte Analysis und Stochastik, Berlin.

SHEATHER, S. J. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53** 683–690.

SKOVGAARD, I. M. (1981). Transformations of an Edgeworth expansion by a sequence of smooth functions. *Scand. J. Statist.* **8** 207–217.

SKOVGAARD, I. M. (1986). On multivariate Edgeworth expansions. *Internat. Statist. Rev.* **54** 169–186.

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

WU, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in nonparametric regression analysis (with discussion). *Ann. Statist.* **14** 1261–1344.

WEIERSTRASS-INSTITUT FÜR ANGEWANDTE
ANALYSIS UND STOCHASTIK
MOHRENSTRASSE 39
D-10117 BERLIN
GERMANY