Statistica Sinica **6**(1996), 925-942

SHARP-OPTIMAL AND ADAPTIVE ESTIMATION FOR HETEROSCEDASTIC NONPARAMETRIC REGRESSION

Sam Efromovich and Mark Pinsker

University of New Mexico and Russian Academy of Sciences

Abstract: The problem is to estimate a smooth regression function for the case of heteroscedastic nonparametric regression. Fixed and random design models are studied simultaneously. Neither smoothness of the estimated regression function nor nuisance functions, that is variance of errors and design density of predictors, are supposed to be known. For this setting we suggest an asymptotically sharp data driven estimate which has minimal constant and maximal rate of local minimax Mean Integrated Squared Error convergence as sample size tends to infinity. The analysis is based on recent results on nonparametric local asymptotic normality and equivalence, in the sense of Le Cam's deficiency distance, between a heteroscedastic regression and corresponding signal-in-noise model. A simplified adaptive estimator is suggested for the case of small sample sizes. This estimator is analyzed via Monte Carlo simulations and compared with an optimal pseudo local linear estimator whose variable bandwidth is based on both underlying regression function and nuisance functions.

Key words and phrases: Heteroscedastic nonparametric regression, adaptation, sharp optimality, small samples.

1. Introduction

We consider simultaneously two models of heteroscedastic nonparametric regression: (i) the fixed design model,

$$Y_{ni} = f(x_{ni}) + \sigma(x_{ni})\xi_{ni}, \quad \int_{x_{n(i-1)}}^{x_{ni}} h(x)dx = (n+1)^{-1}, \ x_{n0} = 0, \quad i = 1, \ldots, n, \tag{1.1}$$

where observations are the pairs $\{(Y_{ni}, x_{ni}),\ i = 1, \ldots, n\}$, (ii) the random design model,

$$Y_{ni} = f(X_{ni}) + \sigma(X_{ni})\xi_{ni}, \ \ i = 1, \ldots, n, \tag{1.2}$$

where X_{ni} are i.i.d. according to design density $h(x)$ and observations are the pairs $\{(Y_{ni}, X_{ni}),\ i = 1, \ldots, n\}$. We refer to Y_{ni} as response and to x_{ni} or X_{ni} as predictor. Hereafter $h(x)$ is a positive design density supported on $[0, 1]$ and $\xi_{n1}, \ldots, \xi_{nn}$ are i.i.d. standard normal random variables (errors) which are independent of the predictors.

Brown and Low (1992) show that each of these rather complicated statistical models is asymptotically equivalent to a corresponding signal-in-white noise model. In particular, if we assume that:

A. An estimated function f belongs to a Sobolev space of periodic functions $\mathcal{F}(m,Q) = \{f : \int_0^1([f(x)]^2 + [f^{(m)}(x)]^2)dx \leq Q < \infty,\ f^{(s)}(0) = f^{(s)}(1)$ for $s = 0,\ldots,m-1\}$ where $f^{(s)}$ denotes the sth derivative and m is a positive integer,

B. Both $\sigma(x)$ and $h(x)$ are positive and have bounded first derivatives on $[0,1]$, and $\lambda^2(t) = \sigma^2(t)/h(t)$,

then due to Brown and Low (1992) the fixed design regression (1.1) is asymptotically equivalent to the following signal-in-noise model,

$$dY_n(t) = f(t)dt + n^{-1/2}\lambda(t)dw(t)\ , \quad 0 \leq t < 1, \tag{1.3}$$

where $w(t)$ denotes Brownian motion. We shall use this equivalence to assess both sharp lower bound for local minimax Mean Integrated Squared Error (MISE) convergence and adaptive estimation. More complicated equivalence holds for the case of the random design regression.

To this point the results on asymptotic sharp adaptive estimation, that is, the results on convergence of MISE with minimal constant and maximal rate, have been known only for normal noise with constant variance and equidistant fixed design model (see Nussbaum (1985), Speckman (1985), Efromovich (1986), Golubev and Nussbaum (1990, 1992)).

For the setting under consideration Fan (1992, 1993) shows that a local linear estimator, which is one of the most promising kernel type estimators, is asymptotically optimal over all linear estimators. We show in this paper that under some mild assumptions linear estimates are optimal among all possible procedures of estimation. Therefore, a pseudo local linear estimator whose variable bandwidth is based on underlying regression function and nuisance functions h and σ, may serve as an oracle for any adaptive estimate. We use this approach to explore our adaptive estimator for the case of small sample sizes.

In Section 2 a lower bound for local minimax MISE convergence is obtained. Section 3 is devoted to asymptotically sharp adaptive estimation. The case of small samples is considered in Section 4.

2. Lower Bound

In this section under assumptions A and B we investigate the asymptotics of a local minimax lower bound for MISE when $n \to \infty$. Let f_0 be a fixed function from $\mathcal{F}(m,Q)$ and restrict our attention to functions f such that the difference $f - f_0$ belongs to the intersection of $\mathcal{F}(m,Q)$ with shrinking L_2-balls

$B_n = \{g : \int_0^1 g^2(t)dt \leq (D_n/4)\delta_n^2\}$ where $\delta_n = n^{-m/(2m+1)}$ and $D_n \to \infty$ arbitrarily slowly as $n \to \infty$.

Due to Le Cam and Yang (1990), the Brown and Low's (1992) equivalence yields similar behavior of risks with bounded loss functions for the models (1.1) and (1.3). Therefore, in the discussion to follow, we use a sequence of bounded loss functions $\rho_n(\hat{f} - f) = \min(\delta_n^{-2} \int_0^1 (\hat{f}(t) - f(t))^2 dt, D_n)$. We also denote the expectation given an underlying function f as $E_f\{\cdot\}$, $o(1)$ and $O(1)$ as generic sequences in n which tend to zero as $n \to \infty$ and bounded, respectively, C's as positive constants, and $P(m) = [2m/(2\pi(m+1))]^{2m/(2m+1)}(2m+1)^{1/(2m+1)}$ as the Pinsker constant.

The following result is a generalization of the well known lower bound of Pinsker (1980) where the particular case of $\lambda(t) = \sigma$, squared loss function and global minimax was explored.

Theorem 2.1. *Let assumptions* A *and* B *hold for the signal-in-noise model* (1.3). *Then*

$$\inf \sup_{f - f_0 \in \mathcal{F}(m,Q) \cap B_n} E\{\rho_n(\tilde{f} - f)\} \geq P(m) Q^{1/(2m+1)} \Big[\int_0^1 \lambda^2(t)dt\Big]^{2m/(2m+1)} (1 + o(1)), \tag{2.1}$$

where the infimum is over all possible estimates $\tilde{f} = \tilde{f}_n(t, f_0, D_n, m, Q, \lambda)$ *based on parameters of smoothness* m *and* Q, *function* $\lambda(t)$, *the center of localization* $f_0(t)$ *and sequence* D_n.

Note that the lower bound (2.1) yields the same lower bound for traditional minimax MISE.

Due to the equivalence result of Brown and Low (1992) the same lower bound holds for the fixed design regression (1.1). The same lower bound also holds for the case of the random design regression (1.2); the latter can be proved similarly to Efromovich (1992). Thus, for heteroscedastic nonparametric regression we obtain the following asymptotic lower bound for local minimax MISE convergence.

Corollary 2.1. *Let assumptions* A *and* B *hold. Then the lower bound* (2.1) *holds for the fixed* (1.1) *and the random* (1.2) *models of heteroscedastic nonparametric regression.*

Proof of Theorem 2.1. Following along the lines of proof in Pinsker (1980) one can easily check that if $\lambda(t)$ is constant and $D_n = \infty$ then the lower bound (2.1) holds for the considered local minimax approach.

To relax the latter assumption $D_n = \infty$ one can apply a projection operator P to an estimate $\tilde{f}$ where the operator P projects this estimate onto the class of considered functions f. Note that due to definitions of the truncated squared loss function and shrinking balls B_n this projection may only decrease the risk.

Moreover, the inequality $\delta_n^{-2}\int_0^1(P\tilde{f}(t)-f(t))^2dt \le D_n$ holds uniformly over all considered functions f.

Thus, the only essential difference between the considered model and Pinsker's (1980) setting is that the function $\lambda(t)$ is not constant. To explore this case we partition the unit interval $[0,1)$ into q subintervals and then approximate $\lambda(t)$ from below by a step function.

Let $T_i(q) = [(i-1)/q, i/q)$, $i = 1,\ldots,q$ be a partition of the unit interval into q subintervals and define the step function $\lambda_q(t) = \min_{(z\in T_i)}\lambda(z)$, $t \in T_i(q)$. Note that due to the definition of Brownian motion the stochastic differential Equation (1.3) can be written as $dY_n(t) = f(t)dt + n^{-1/2}\lambda_q(t)dw_1(t) + n^{-1/2}(\lambda^2(t) - \lambda_q^2(t))^{1/2}dw_2(t)$ where $w_1(t)$ and $w_2(t)$ are two independent standard Brownian motions. Therefore, every estimate based on observation (1.3) is dominated by an estimate based on observation (1.3) with the step function $\lambda_q(t)$ in place of $\lambda(t)$.

Set $\lambda_{qi} = \lambda_q((i-1)/q)$. Then

$$\begin{aligned}&\inf \sup_{f-f_0\in\mathcal{F}(m,Q)\cap B_n} E_f\{\rho_n(\tilde{f}-f)\}\\ &\ge \sum_{i=1}^q \inf \sup_{f\in\mathcal{F}_{ni}} E_f\Big\{\min(\delta_n^{-2}\int_{T_i}(\tilde{f}_n(t,f_0,D_n,m,Q,\lambda_q)-f(t))^2dt; D_n/q)\Big\}. \quad (2.2)\end{aligned}$$

Here $\mathcal{F}_{ni}$ is a class of functions f_0+g supported on T_i where $g^{(s)}((i-1)/q) = g^{(s)}(i/q) = 0$ for $s = 0,1,\ldots,m-1$, $\int_{T_i}[g^{(m)}(t)]^2dt \le Q_i = Q\lambda_{qi}^2/\sum_{s=1}^q\lambda_{qs}^2$ and $\int_{T_i}g^2(t)dt \le D_n\delta_n^2/4q$. The estimate $\tilde{f}_n(t,f_0,D_n,m,Q,\lambda_q)$ on the right-hand side of (2.2) is based on observation (1.3) with the function $\lambda_q(t)$ in place of $\lambda(t)$.

We have converted the original problem into q subproblems of filtering a function in white noise. The only difference between these subproblems and Pinsker's (1980) setting is that the subproblems consider a class of functions that vanish together with their $m-1$ derivatives at boundary points. Golubev and Nussbaum (1990) show that Pinsker's lower bound holds for this class of functions. Moreover, straightforward algebra shows that this lower bound holds for the considered shrinking neighborhoods $\mathcal{F}_{ni}$ as well. Therefore we can apply this lower bound to every term on the right-hand side of (2.2). We write for a fixed q,

$$\begin{aligned}&\sum_{i=1}^q \inf \sup_{f\in\mathcal{F}_{ni}} E_f\Big\{\min(\delta_n^{-2}\int_{T_i}(\tilde{f}_n(t,f_0,D_n,m,Q,\lambda_q)-f(t))^2dt; D_n/q)\Big\}\\ &\ge \sum_{i=1}^q P(m)\Big[q^{-2m}Q\lambda_{qi}^2/\sum_{s=1}^q\lambda_{qs}^2\Big]^{1/(2m+1)}[\lambda_{qi}^2]^{2m/(2m+1)}(1+o(1))\end{aligned}$$

$$\geq \sum_{i=1}^{q} P(m)Q^{1/(2m+1)}q^{-2m/(2m+1)}\lambda_{qi}^2\Big[\sum_{s=1}^{q}\lambda_{qs}^2\Big]^{-1/(2m+1)}(1+o(1))$$

$$= P(m)Q^{1/(2m+1)}\Big[q^{-1}\sum_{i=1}^{q}\lambda_{qi}^2\Big]^{2m/(2m+1)}(1+o(1)). \qquad (2.3)$$

Assumption B yields $q^{-1}\sum_{i=1}^{q}\lambda_{qi}^2 \geq \int_0^1 \lambda^2(t)dt(1-\gamma(q))$ where $\gamma(q)$ tends to zero as $q \to \infty$. Because q may be arbitrarily large, this proves the assertion of Theorem 2.1.

Remark 2.1. The Cauchy-Schwarz inequality implies $\int_0^1 \sigma^2(x)h^{-1}(x)dx \geq [\int_0^1 \sigma(x)dx]^2$ with the equality iff the design density h is equal to

$$h^*(x) = \sigma(x)/\int_0^1 \sigma(x)dx. \qquad (2.4)$$

Thus, h^* defines the optimal design of an experiment which minimizes asymptotic MISE for the heteroscedastic nonparametric regressions (1.1) and (1.2).

In the next section we show that the lower bound (2.1) is sharp, that is, there exists a data-driven estimator whose MISE attains this lower bound.

3. Asymptotically Sharp Adaptive Estimation

It is easy to check that for the signal-in-noise model (1.3) the estimates $\hat{\theta}_j = \int_0^1 \varphi_j(t)dY_n(t)$ of the Fourier coefficients $\theta_j = \int_0^1 \varphi_j(t)f(t)dt$ satisfy the equality

$$\sum_{i=-1}^{0} E_f\{(\hat{\theta}_{2j+i} - \theta_{2j+i})^2\} = 2n^{-1}\sigma^2,$$

where $\sigma^2 = \int_0^1 \lambda^2(t)dt$, $j = 1, 2, \ldots$, and $\{\varphi_s(t)\}$ is the classical trigonometric Fourier basis over the unit interval $[0,1]$, i.e., $\varphi_0(t) = 1$, $\varphi_{2j-1}(t) = \sqrt{2}\sin(2\pi jt)$ and $\varphi_{2j}(t) = \sqrt{2}\cos(2\pi jt)$, $j = 1, 2, \ldots$ If the parameter σ^2 is known, then an adaptive estimator of Efromovich and Pinsker (1984), which is based only on observation $Y_n(t)$, is asymptotically sharp minimax. For the case of unknown parameter σ^2 we note that $E\{\hat{\theta}_{2j-1}^2 + \hat{\theta}_{2j}^2\} = \theta_{2j-1}^2 + \theta_{2j}^2 + 2n^{-1}\sigma^2$. Thus, one may estimate the nuisance parameter σ^2 via statistics $\hat{\theta}_j^2$ where j are sufficiently large.

We shall use this promising idea, which is rather straightforward for the signal-in-noise model, to estimate a regression function for the heteroscedastic nonparametric regression models (1.1) and (1.2) when neither smoothness of estimated regression function nor noise variance $\sigma^2(x)$ nor design density $h(x)$ are known.

Set $S = \lfloor n^{1/6} \ln(n) \rfloor$, $L = 2\lfloor n^{4/9} \ln^6(n) \rfloor$ and $M = L + 1$. Here $\lfloor x \rfloor$ is the integer part of x. Set $s_n = 2$ for the fixed design model, $s_n = 2\lfloor \ln(\ln(n+3)+3) \rfloor$ for the random design model, and define the following sequence of subintervals of the unit interval, $A_{ni} = [x_{n\max(0,i-(1/2)s_n)}, x_{n\min(i+(1/2)s_n,n+1)})$. Recall that $x_{n0} = 0$ and $x_{n(n+1)} = 1$.

To analyze simultaneously the fixed and random design models we denote for the random design model (1.2) the ordered (according to ascending predictors) observations as $\{(Y_{ni}, x_{ni}), i = 1, \ldots, n\}$. From now on we use only these ordered observations.

First, we define the estimator for the Fourier coefficients θ_j,

$$\hat{\theta}_j = \sum_{i=1}^{n} Y_{ni} s_n^{-1} \int_{A_{ni}} \varphi_j(x)dx. \tag{3.1}$$

The reader might notice that (3.1) resembles the trapezoidal method of numerically computing the integral $\int_0^1 f(x)\varphi_j(x)dx$.

For the case when the parameter $\sigma^2 = \int_0^1 \sigma^2(x)h^{-1}(x)dx$ is unknown, we use the estimate

$$\hat{\sigma}^2 = nL^{-1} \sum_{j=M}^{M+L-1} \tilde{\theta}_j^2, \tag{3.2}$$

where $\tilde{\theta}_j$ is equal to $\hat{\theta}_j$ for the case of the fixed design model and it is defined by (3.1) with $s_n = s'_n = 2\lfloor n^{1/3} \ln^6(n) + 1 \rfloor$ for the random design model. From now on we set $\hat{\sigma}^2 = \sigma^2$ if the parameter σ^2 is known and $\hat{\sigma}^2$ is defined by (3.2) otherwise.

Finally we introduce the following statistics:

$$\hat{\Theta}(k,n) = (2k)^{-1} \sum_{j=(k-1)k+1}^{k(k+1)} (\hat{\theta}_j^2 - \hat{\sigma}^2 n^{-1}), \tag{3.3}$$

$$\hat{\Lambda}(k,n) = \hat{\Theta}(k,n)[\hat{\Theta}(k,n) + \hat{\sigma}^2 n^{-1}]^{-1}(\hat{\Theta}(k,n) - \ln^{-1}(k+3)\hat{\sigma}^2 n^{-1})_+, \tag{3.4}$$

where $k = 1, \ldots, S$ and $(x)_+ = \max(0, x)$.

The following statement is the main result of this section.

Theorem 3.1. *Let assumptions* A *and* B *hold. Then, the adaptive estimator*

$$\hat{f}_n(x) = \hat{\theta}_0 + \sum_{k=1}^{S} \hat{\Lambda}(k,n) \sum_{j=(k-1)k+1}^{k(k+1)} \hat{\theta}_j \varphi_j(x) \tag{3.5}$$

is asymptotically sharp minimax, that is,

$$\begin{aligned} &\sup_{f \in \mathcal{F}(m,Q)} E_f\Big\{ \int_0^1 (\hat{f}_n(t) - f(t))^2 dt \Big\} \\ &= P(m) Q^{1/(2m+1)} \Big[n^{-1} \int_0^1 \sigma^2(x) h^{-1}(x) dx \Big]^{2m/(2m+1)} (1 + o(1)). \end{aligned} \tag{3.6}$$

Remark 3.1. The estimator (3.5) with increasing s_n, chosen for the case of the random design model, is also asymptotically sharp optimal for the fixed design model. The inverse assertion does not hold, that is, the estimator (3.5) with $s_n = 2$ is not asymptotically sharp optimal for the random design model. However, this relatively simple estimator is rate optimal for the random design model.

Remark 3.2. There is a wide variety of groupings and shrinkings of the Fourier coefficients that yields sharp optimality (see Efromovich (1985)). In particular, one may set the cardinality of the first $J_n = O(1)\ln^\beta(n)$ groups to be equal to one. Here β is a positive real. We shall use such grouping in the following section. Also note that many promising adaptive estimators have been motivated by these smoothing procedures. The interested reader can find a comprehensive discussion of this issue and speculations on the future use in Birgé and Massart (1994) and Barron, Birgé and Massart (1995).

Remark 3.3. The assumption on Gaussian error in models (1.1) and (1.2) is not crucial for the validity of the upper bound. The interested reader can verify that this upper bound holds whenever $E\{\xi_{ni}^2\} = 1$ and $E\{\xi_{ni}^8\} < \infty$.

Proof of Theorem 3.1. Let $\tilde{f}_n(x) = \sum_{j=0}^\infty \lambda_j \hat{\theta}_j \varphi_j(x)$ be an orthogonal series estimator. Then Parseval's identity yields $\int_0^1 (\tilde{f}_n(t) - f(t))^2 dt = \sum_{j=0}^\infty (\lambda_j \hat{\theta}_j - \theta_j)^2$. Thus, MISE of this estimator depends only on properties of λ_j and $\hat{\theta}_j$ but not on an underlying statistical model. Keeping this in mind, we see that Lemmas 1-3 in Efromovich (1985) yield the upper bound (3.6) whenever the following four relations hold,

$$E_f\{\hat{\theta}_j\} = \theta_j + r_{nj} \text{ where } \sum_{j=0}^J (1+j)^{-2} r_{nj}^2 = o(1)(Jn^{-1})^2 \text{ and } \sum_{j=0}^J r_{nj}^2 = o(1)Jn^{-1}, \tag{3.7}$$

$$E_f\{\hat{\theta}_j^2\} = \theta_j^2 + n^{-1}\int_0^1 \sigma^2(x)h^{-1}(x)\varphi_j^2(x)dx + R_{nj} \text{ where} \sum_{j=0}^J |R_{nj}| = o(1)Jn^{-1}, \tag{3.8}$$

$$E_f\{(\hat{\sigma}^2 - \sigma^2)^2\} \le (S\ln(n))^{-2}, \tag{3.9}$$

and

$$E_f\{(\hat{\Theta}(k,n) - \Theta(k,n))^4\} \le Ck^{-2}n^{-2}(\Theta(k) + n^{-1})^2. \tag{3.10}$$

The relations (3.7) and (3.8) are to be held for all natural $J \in \{\lfloor \ln(n+3) \rfloor, \ldots, S(S+1)\}$. Hereafter $\Theta(k) = (2k)^{-1}\sum_{j=(k-1)k+1}^{k(k+1)} \theta_j^2$, $0 < k \le S$ and $0 \le j \le S(S+1)$.

Thus, it remains to verify (3.7)-(3.10).

Proof of (3.7). Assume that the following two inequalities hold,

$$E_f\Big\{\sum_{i=1}^n s_n^{-1}\int_{A_{ni}}|(f(x_{ni})-f(x))\varphi_j(x)|dx\Big\}\le Cs_nn^{-1} \tag{3.11}$$

and

$$\Big|E_f\Big\{\sum_{i=1}^n s_n^{-1}\int_{A_{ni}}f(x)\varphi_j(x)dx\Big\}-\theta_j\Big|\le Cs_nn^{-1}. \tag{3.12}$$

Then we may write

$$\begin{aligned}&E_f\{\hat\theta_j\}=E_f\Big\{\sum_{i=1}^n f(x_{ni})s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx\Big\}\\&=E_f\Big\{\sum_{i=1}^n s_n^{-1}\int_{A_{ni}}f(x)\varphi_j(x)dx\Big\}+\sum_{i=1}^n E_f\Big\{s_n^{-1}\int_{A_{ni}}(f(x_{ni})-f(x))\varphi_j(x)dx\Big\}\\&\triangleq\theta_j+r_{nj}.\end{aligned}$$

Note that (3.11)-(3.12) yield the inequality

$$|r_{nj}|\le Cs_nn^{-1} \tag{3.13}$$

which implies (3.7).

Now we are in a position to verify (3.11) and (3.12). Using assumption A and the Cauchy-Schwarz inequality we obtain $\int_{A_{ni}}|(f(x_{ni})-f(x))\varphi_j(x)|dx\le C\int_{A_{ni}}[\int_{A_{ni}}|f^{(1)}(u)|du]dx\le Cl_{ni}^{3/2}[\int_{A_{ni}}|f^{(1)}(u)|^2du]^{1/2}$. Hereafter $l_{ni}=\int_{A_{ni}}dx$ denotes the length of the corresponding subinterval A_{ni}. Then, again using the Cauchy-Schwarz inequality we write

$$\begin{aligned}&E_f\Big\{\sum_{i=1}^n s_n^{-1}\int_{A_{ni}}|(f(x_{ni})-f(x))\varphi_j(x)|dx\Big\}\\&\le CE_f\Big\{\sum_{i=1}^n s_n^{-1}l_{ni}^{3/2}[\int_{A_{ni}}|f^{(1)}(u)|^2du]^{1/2}\Big\}\\&\le Cs_n^{-1}E_f\Big\{[\sum_{i=1}^n l_{ni}^3]^{1/2}[\sum_{i=1}^n\int_{A_{ni}}|f^{(1)}(u)|^2du]^{1/2}\Big\}.\end{aligned}$$

Note that $\sum_{i=1}^n\int_{A_{ni}}|f^{(1)}(u)|^2du\le s_n\int_0^1|f^{(1)}(u)|^2du\le Cs_n$ where the former inequality follows at once from definition of the subintervals A_{ni} and the latter from assumption A for the case $m=1$ and from assumption A and Parseval's identity for $m>1$. On the other hand, the inequality

$$E\{l_{ni}^k\}\le C(s_nn^{-1})^k \tag{3.14}$$

holds for any natural k. Indeed, for the case of the fixed design predictors this inequality follows at once from the fact that due to assumption B the design density $h(x)$ is bounded below from zero. For the case of the random design predictors with a design density $h(x)$ bounded from below the inequality (3.14) is a well known property of ordered statistics (see, e.g., Wilks (1962)).

Combining the obtained inequalities, we get

$$E_f\Big\{\sum_{i=1}^n s_n^{-1}\int_{A_{ni}}|(f(x_{ni})-f(x))\varphi_j(x)|dx\Big\} \le Cs_n^{-1}\Big[E_f\sum_{i=1}^n l_{ni}^3\Big]^{1/2}s_n^{1/2} \le Cs_n n^{-1}.$$

Thus, the inequality (3.11) is proved.

The inequality (3.12) is proved similarly. First, it is easy to see that

$$\begin{aligned}&\Big|\sum_{i=1}^n s_n^{-1}\int_{A_{ni}} f(x)\varphi_j(x)dx - \int_0^1 f(x)\varphi_j(x)dx\Big|\\ &< s_n^{-1}\Big[s_n\int_{A_{n(s_n/2)}\cup A_{n(n-s_n/2)}}|f(x)\varphi_j(x)|dx\Big].\end{aligned}$$

Note that due to assumption A the considered regression functions f are uniformly bounded, that is, $\sup_{f\in\mathcal{F}(m,Q)}\max_x|f(x)|<C$. Also recalling that $\theta_j=\int_0^1 f(x)\varphi_j(x)dx$ we get the inequality $|\sum_{i=1}^n s_n^{-1}\int_{A_{ni}} f(x)\varphi_j(x)dx-\theta_j|<C(l_{n(s_n/2)}+l_{n(n-s_n/2)})$. This inequality at once implies (3.12) for the case of fixed design regression; for the random design regression the right-hand side of the last inequality may be estimated via the inequality (3.14) and this immediately yields (3.12).

Thus, line (3.7) is proved.

Proof of (3.8). First we verify (3.8) for the fixed design model. At the same time, we shall derive general assertions whenever this is possible.

Assumption B, definition (1.1) of the design density $h(x)$ and the Taylor expansion give the following asymptotic equality:

$$ns_n^{-1}l_{ni} = h^{-1}(x_{ni}) + O(1)n^{-1}. \tag{3.15}$$

Also note that $s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx = s_n^{-1}l_{ni}\varphi_j(x_{ni})+s_n^{-1}\int_{A_{ni}}(\varphi_j(x)-\varphi_j(x_{ni}))dx$. This equality together with assumption B and (3.15) yields

$$\sum_{i=1}^n\sigma^2(x_{ni})\Big[s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx\Big]^2 = n^{-1}\int_0^1\sigma^2(x)h^{-1}(x)\varphi_j^2(x)dx + O(1)js_nn^{-2}\,. \tag{3.16}$$

Now we may write, using our assumption that $\xi_{n1},\dots,\xi_{nn}$ are i.i.d. standard normal random variables, that

$$\begin{aligned}E_f\{\hat\theta_j^2\} &= E_f\Big\{\sum_{i,l=1}^n Y_{ni}Y_{nl}s_n^{-2}\int_{A_{ni}}\varphi_j(x)dx\int_{A_{nl}}\varphi_j(x)dx\Big\}\\ &= E_f\Big\{\sum_{i,l=1}^n f(x_{ni})f(x_{nl})s_n^{-2}\int_{A_{ni}}\varphi_j(x)dx\int_{A_{nl}}\varphi_j(x)dx\\ &\qquad +\sum_{i=1}^n\sigma^2(x_{ni})[s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx]^2\Big\} \qquad (3.17)\end{aligned}$$

and using (3.11)-(3.16) we obtain

$$E_f\{\hat\theta_j^2\} = \theta_j^2 + n^{-1}\int_0^1\sigma^2(x)h^{-1}(x)\varphi_j^2(x)dx + R_{nj},$$

where it is easy to see that $|R_{nj}|\le C[js_nn^{-2}+|\theta_j||r_{nj}|+r_{nj}^2]$. This together with (3.13) and the Cauchy-Schwarz inequality $\sum_{j=0}^J|\theta_j||r_{nj}|\le[\sum_{j=0}^J(1+j)^2\theta_j^2]^{1/2}$ $[\sum_{j=0}^J(1+j)^{-2}r_{nj}^2]^{1/2}$ yields (3.8) for the fixed design model.

To prove (3.8) for the random design model, we use, instead of (3.15), the well known expansion,

$$ns_n^{-1}l_{ni} = h^{-1}(x_{ni})+\eta_{ni}, \qquad (3.18)$$

where, under assumption B, the random variables η_{ni} satisfy the inequality $E\{\eta_{ni}^4\}\le Cs_n^{-2}$ (see Wilks (1962)). This expansion explains why we should use an increasing sequence s_n for sharp optimal estimating whenever predictors are random.

Using (3.18) we write

$$\begin{aligned}s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx &= s_n^{-1}l_{ni}\varphi_j(x_{ni}) + s_n^{-1}\int_{A_{ni}}(\varphi_j(x)-\varphi_j(x_{ni}))dx\\ &= n^{-1}h^{-1}(x_{ni})\varphi_j(x_{ni}) + n^{-1}\eta_{ni}\varphi_j(x_{ni})\\ &\qquad + s_nn^{-2}j\Big[s_n^{-1}\int_{A_{ni}}(\varphi_j(x)-\varphi_j(x_{ni}))dx/(s_nn^{-2}j)\Big]. \qquad (3.19)\end{aligned}$$

Then, using (3.19), the inequality $E\{\eta_{ni}^4\}\le Cs_n^{-2}$ mentioned above, assumption B and the Cauchy-Schwarz inequality, similarly to (3.16) we get

$$\begin{aligned}&E_f\Big\{\sum_{i=1}^n\sigma^2(x_{ni})[s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx]^2\Big\}\\ &= E_f\Big\{\sum_{i=1}^n\sigma^2(x_{ni})s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx[n^{-1}h^{-1}(x_{ni})\varphi_j(x_{ni})\end{aligned}$$

$$+n^{-1}\eta_{ni}\varphi_j(x_{ni}) + s_n n^{-2} j[s_n^{-1}\int_{A_{ni}}(\varphi_j(x)-\varphi_j(x_{ni}))dx/(s_n n^{-2}j)]]\Big\}$$
$$= n^{-1}\int_0^1 \sigma^2(x)h^{-1}(x)\varphi_j^2(x)dx + O(1)n^{-2}\sum_{i=1}^n E^{1/2}\{\eta_{ni}^2\} + O(1)js_n n^{-2}. \quad (3.20)$$

Thus, using (3.20) for estimating the right-hand side of (3.17), we can easily see that for all natural j the following inequality holds,

$$|R_{nj}| \le C[n^{-1}s_n^{-1/2} + js_n n^{-2} + |\theta_j|s_n n^{-1} + s_n^2 n^{-2}]. \quad (3.21)$$

The validity of (3.8) follows immediately from (3.21). Note that the inequality (3.21) holds for the fixed design model as well, so we can always use it.

Proof of (3.9). We restrict our attention to the more complex case of the random design model. The fixed design model is analyzed similarly and yet simpler, so we leave this case to the interested reader.

First of all, we recall that here we use $s_n = s_n' = 2\lfloor n^{1/3}\ln^6(n)+1\rfloor$. Using (3.8) we write

$$E_f\{\hat\sigma^2\} = nL^{-1}\sum_{j=M}^{M+L-1} E_f\{\hat\theta_j^2\} = nL^{-1}\sum_{j=M}^{M+L-1}\theta_j^2 + \sigma^2 + nL^{-1}\sum_{j=M}^{M+L-1} R_{nj}.$$

Recall that estimated regression functions belong to $\mathcal{F}(m,Q)$ with $m \ge 1$ and therefore

$$\sum_{j\ge M}\theta_j^2 \le QM^{-2}. \quad (3.22)$$

Also, using (3.21) and the Cauchy-Schwarz inequality we see that

$$nL^{-1}\sum_{j=M}^{M+L-1}|R_{nj}| \le C\Big[s_n^{-1/2} + s_n n^{-1}(L+M) + nL^{-1}\sum_{j=M}^{M+L-1}\theta_j^2 + s_n^2 n^{-1}\Big].$$

Thus, we obtain the following estimate for the squared bias,

$$(E_f\{\hat\sigma^2\}-\sigma^2)^2 \le C[n^2L^{-2}M^{-4} + s_n^{-1} + s_n^2 n^{-2}(L+M)^2 + s_n^4 n^{-2}] = o(1)(S\ln(n))^{-2}. \quad (3.23)$$

Now we estimate the variance of $\hat\sigma^2$. Write

$$\hat\sigma^2 = nL^{-1}\sum_{j=M}^{M+L-1}\Big[\sum_{i=1}^n Y_{ni}s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx\Big]^2$$
$$= nL^{-1}\sum_{j=M}^{M+L-1}\Big[\sum_{i=1}^n f(x_{ni})s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx\Big]^2$$

$$+ 2nL^{-1}\sum_{j=M}^{M+L-1}\Big[\sum_{i=1}^{n} f(x_{ni})s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx\Big]\Big[\sum_{l=1}^{n}\sigma(x_{nl})\xi_{nl}s_n^{-1}\int_{A_{nl}}\varphi_j(x)dx\Big]$$

$$+ nL^{-1}\sum_{j=M}^{M+L-1}\sum_{i=1}^{n}\sigma^2(x_{ni})\xi_{ni}^2\Big[s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx\Big]^2$$

$$+ nL^{-1}\sum_{1\le i,l\le n; i\ne l}\sigma(x_{ni})\sigma(x_{nl})\xi_{ni}\xi_{nl}\sum_{j=M}^{M+L-1}\Big[s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx\Big]\Big[s_n^{-1}\int_{A_{nl}}\varphi_j(x)dx\Big]$$

$$\triangleq D_1 + D_2 + D_3 + D_4. \tag{3.24}$$

Obviously

$$E_f\{(\hat\sigma^2 - E_f\{\hat\sigma^2\})^2\} \le C\sum_{r=1}^{4} E_f\{(D_r - E_f\{D_r\})^2\} \tag{3.25}$$

and therefore it suffices to estimate the variances of D_1-D_4.

To estimate the variance of D_1 we first deduce similarly to (3.11)-(3.12) the following expansion,

$$\sum_{i=1}^{n} f(x_{ni})s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx = \theta_j + \kappa_{nj}, \text{ where } \quad E_f\{\kappa_{nj}^4\} \le C(s_n n^{-1})^4. \tag{3.26}$$

Using (3.26) write

$$D_1 - E_f\{D_1\} = nL^{-1}\sum_{j=M}^{M+L-1}[(\theta_j + \kappa_i)^2 - (\theta_j^2 + 2\theta_j E_f\{\kappa_{nj}\} + E_f\{\hat\kappa_{nj}^2\})]$$
$$= nL^{-1}\sum_{j=M}^{M+L-1}[2\theta_j(\kappa_{nj} - E_f\{\kappa_{nj}\}) + (\kappa_{nj}^2 - E_f\{\kappa_{nj}^2\})].$$

Using the Cauchy-Schwarz inequality we get

$$E_f\{(D_1 - E_f\{D_1\})^2\}$$
$$\le Cn^2L^{-2}\Big[\sum_{j=M}^{M+L-1}\theta_j^2\sum_{r=M}^{M+L-1}E_f\{(\hat\kappa_{nr} - E_f\{\hat\kappa_{nr}\})^2\} + (M+L)\sum_{j=M}^{M+L-1}E_f\{\kappa_{nj}^4\}\Big].$$

Finally, applying (3.22) and (3.26) to the right-hand side of the last inequality, we see that

$$E_f\{(D_1 - E_f\{D_1\})^2\} \le Cn^2L^{-2}[M^{-2}(M+L)s_n^2n^{-2} + (M+L)^2s_n^4n^{-4}]$$
$$= o(1)(S\ln(n))^{-2}. \tag{3.27}$$

To analyze D_2 we note that its expectation is zero and that via (3.26)

$$D_2 = 2nL^{-1} \sum_{j=M}^{M+L-1} \Big[(\theta_j + \kappa_{nj}) \sum_{i=1}^{n} \sigma(x_{ni})\xi_{ni} s_n^{-1} \int_{A_{ni}} \varphi_j(x)dx\Big].$$

Then, applying the Cauchy-Schwarz inequality we get

$$E_f\{D_2^2\} \leq Cn^2L^{-2}\Big[\sum_{j=M}^{M+L-1} \theta_j^2 \sum_{r=M}^{M+L-1} E_f\{(\sum_{i=1}^{n}\sigma(x_{ni})\xi_{ni}s_n^{-1}\int_{A_{ni}} \varphi_r(x)dx)^2\} \\ + E_f\{(\sum_{j=M}^{M+L-1} \kappa_{nj} \sum_{i=1}^{n}\sigma(x_{ni})\xi_{ni}s_n^{-1}\int_{A_{ni}} \varphi_j(x)dx)^2\}\Big].$$

Using our assumption that the errors $\xi_{n1}, \ldots, \xi_{nn}$ are i.i.d. standard normal random variables which are independent of predictors, (3.22) and then applying the Cauchy-Schwarz inequality we write

$$E_f\{D_2^2\} \leq Cn^2L^{-2}\Big[M^{-2}(M+L)n^{-1} + E_f^{1/2}\{\sum_{j,r=M}^{M+L-1} \kappa_{nj}^2\kappa_{nr}^2\} \\ \times E_f^{1/2}\{\sum_{j,r=M}^{M+L-1} (\sum_{i=1}^{n} \sigma^2(x_{ni})[s_n^{-1}\int_{A_{ni}} \varphi_j(x)dx][s_n^{-1}\int_{A_{ni}} \varphi_r(x)dx])^2\}\Big].$$

Now, using (3.19), write $\sum_{i=1}^{n} \sigma^2(x_{ni})[s_n^{-1}\int_{A_{ni}} \varphi_j(x)dx][s_n^{-1}\int_{A_{ni}} \varphi_r(x)dx]$ as $n^{-1}\int_0^1 \sigma^2(x)h^{-1}(x)\varphi_j(x)\varphi_r(x)dx$ plus some smaller terms. Then via (3.26) we obtain

$$E_f\{D_2^2\} \leq o(1)(S\ln(n))^{-2} + Cn^2L^{-2}[(M+L)(s_nn^{-1})^2] \\ \times\Big[\{\sum_{j,r=M}^{M+L-1} [n^{-1}\int_0^1 \sigma^2(x)h^{-1}(x)\varphi_j(x)\varphi_r(x)dx]^2\}^{1/2} \\ + (M+L)n^{-1}s_n^{-1/2} + (M+L)^{3/2}s_nn^{-2}\Big].$$

Note that the Bessel inequality and assumption B yield

$$\sum_{j=M}^{M+L-1} \Big[\int_0^1 \sigma^2(x)h^{-1}\varphi_j(x)\varphi_r(x)dx\Big]^2 \leq \int_0^1 \sigma^4(x)h^{-2}\varphi_r^2(x)dx \leq C \qquad (3.28)$$

and therefore, after some straightforward simplifications, we obtain the desired asymptotic bound

$$E_f\{D_2^2\} \leq o(1)(S\ln(n))^{-2} + Cs_n^2L^{-2}(M+L)[(M+L)^{1/2}n^{-1} + (M+L)n^{-1}s_n^{-1/2} \\ + (M+L)^{3/2}s_nn^{-2}] = o(1)(S\ln(n))^{-2}. \qquad (3.29)$$

Estimating the variance of D_3 is rather simple. Using the easy verified inequality $\mathrm{Var}_f\{\sigma^2(x_{ni})\xi_{ni}^2[s_n^{-1}l_{ni}]^2\} \le Cn^{-4}$ we obtain

$$\mathrm{Var}_f\{D_3\} \le C(nL^{-1})^2(M+L)^2 \sum_{i=1}^n \mathrm{Var}_f\Big\{\sigma^2(x_{ni})\xi_{ni}^2[s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx]^2\Big\}$$
$$= o(1)(S\ln(n))^{-2}. \tag{3.30}$$

Now we estimate the variance of D_4. The expectation of D_4 is equal to zero and therefore we are studying the second moment of D_4. We can write $D_4^2 = n^2L^{-2}\sum_{j,r=M}^{M+L-1} a_j a_r$ where

$$a_j = \sum \sigma(x_{ni})\sigma(x_{nl})\xi_{ni}\xi_{nl}\Big[s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx\Big]\Big[s_n^{-1}\int_{A_{nl}}\varphi_j(x)dx\Big]$$

and hereafter the summation is over $1 \le i, l \le n$ except for $i = l$. Denote $E'\{\cdot\}$ as the expectation with respect to the distribution of errors given the predictors. Recall that $\xi_{n1},\dots,\xi_{nn}$ are i.i.d. standard normal random variables and therefore $E'\{a_j a_r\} = 2\sum\sigma^2(x_{ni})\sigma^2(x_{nl})[s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx]\ [s_n^{-1}\int_{A_{ni}}\varphi_r(x)dx]$ $[s_n^{-1}\int_{A_{nl}}\varphi_j(x)dx][s_n^{-1}\int_{A_{nl}}\varphi_r(x)dx]$. The last equation shows that

$$E_f\{D_4^2\} = n^2L^{-2}\sum_{j,r=M}^{M+L-1} E_f\{E'\{a_j a_r\}\} \le Cn^2L^{-2}\sum_{j,r=M}^{M+L-1} E_f\{a_{jr}\},$$

where $a_{jr} = (\sum_{i=1}^n \sigma^2(x_{ni})[s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx][s_n^{-1}\int_{A_{ni}}\varphi_r(x)dx])^2$.

Now we repeat the same technical step which was used for estimating the second moment of D_2. Namely, using (3.19) we are writing $E_f\{a_{jr}\}$ as $(n^{-1}\int_0^1\sigma^2(x)$ $h^{-1}(x)\varphi_j(x)\varphi_r(x)dx)^2$ plus some smaller terms. We get

$$E_f\{a_{jr}\} = E_f\Big\{(\sum_{i=1}^n \sigma^2(x_{ni})[s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx][n^{-1}h^{-1}(x_{ni})\varphi_r(x_{ni}) + n^{-1}\eta_{ni}\varphi_r(x_{ni})$$
$$+ s_n n^{-2} r[s_n^{-1}\int_{A_{ni}}(\varphi_r(x)-\varphi_r(x_{ni})dx/(s_n n^{-2}r)]]])^2\Big\}$$
$$\le C\Big[E_f\{[n^{-1}\sum_{i=1}^n \sigma^2(x_{ni})h^{-1}(x_{ni})\varphi_r(x_{ni})s_n^{-1}\int_{A_{ni}}\varphi_j(x)dx]^2\}$$
$$+E_f\{[n^{-1}\sum_{i=1}^n|\eta_{ni}|s_n^{-1}l_{ni}]^2\} + (s_n n^{-2}r)^2\Big].$$

The first term in the right hand side of the last inequality is not greater than $C[n^{-1}\int_0^1\sigma^2(x)h^{-1}(x)\varphi_j(x)\varphi_r(x)dx]^2 + Cn^{-2}(s_n n^{-1}r)^2$; the second one is not greater than $Cn^{-2}s_n^{-1}$. Then, using (3.28) we finally obtain

$$E_f\{D_4^2\} \le Cn^2L^{-2}\sum_{j,r=M}^{M+L-1} E_f\{a_{jr}\} \le C[L^{-2}L + L^{-2}n^{-2}s_n^2(M+L)^4 + s_n^{-1}]$$
$$= o(1)(S\ln(n))^{-2}.$$

Using the last inequality, (3.27), (3.29) and (3.30) to estimate the right-hand side of (3.25) and then combining the result with (3.23) we establish (3.9).

Proof of (3.10). This proof follows straightforwardly along the lines of the proof of Lemma 1 in Efromovich (1985) and we leave it to the interested reader.

Theorem 3.1 is now proved.

4. Adaptive Estimation for Small Samples

In this section we discuss a modified (for small sample sizes) adaptive estimator which mimics the proposed asymptotically sharp adaptive estimator.

Below we give a step-by-step explanation of the recommended procedure and then illustrate it via comparison with a pseudo local linear estimator of Fan (1992). This local linear estimator is one of the most promising kernel type estimators whose variable bandwidth is based on both underlying regression function and nuisance functions $\sigma(x)$ and $h(x)$. We employ this estimator as an oracle whose MISE is used as lower bound for MISE of an adaptive estimator. We also would like to note that a promising procedure of data-driven bandwidth selection in local polynomial fitting is suggested by Fan and Gijbels (1995).

4.1. Procedure for estimation

There are three steps in the procedure which we recommend for use when the sample size is in the range from twenty to several hundred observations. Hereafter we use the basis $\psi_1(x) = 1$ and $\psi_j(x) = \sqrt{2}\cos((j-1)\pi x)$ for $j = 2, 3, \ldots, J$ which allows us to approximate aperiodic functions.

Step 1. Let J be the minimal integer which is greater than $2 + \ln(n)$. For $j = 1, \ldots, J$, estimate the Fourier coefficients θ_j by the estimator

$$\hat{\theta}_j = \sum_{i=1}^{n} w_i(j) Y_i, \tag{4.1}$$

where $w_1(j) = x_{n2}\varphi_j(x_{n1})/2$, $w_n(j) = (1 - x_{n(n-1)})\varphi_j(x_{nn})/2$ and $w_i(j) = (x_{n(i+1)} - x_{n(i-1)})\varphi_j(x_i)/2$ for $i = 2, 3, \ldots, n-1$. The reader might realize that $\hat{\theta}_j$ mimics the estimator (3.1) with $s_n = 2$.

Step 2. Calculate for $1 \le j \le J$ the statistics $\hat{\Theta}_j = \max((\hat{\theta}_j^2 - \hat{\sigma}^2), 0)$. Here $\hat{\sigma}^2 = (2J)^{-1}\sum_{j=J+1}^{3J}\hat{\theta}_j^2$ is an estimator for the integrated squared error $n^{-1}\int_0^1 \sigma^2(x)\, h^{-1}(x)dx$.

Step 3. Find the optimal cutoff $\hat{J}$ which minimizes the empirical risk, $\hat{J}\hat{\sigma}^2 + \sum_{j=\hat{J}+1}^{J}\hat{\Theta}_j$, over $1 \le \hat{J} \le J$ and then compute the modified adaptive estimator,

$$\tilde{f}_n(x) = \sum_{j=1}^{\hat{J}} [\hat{\Theta}_j/(\hat{\Theta}_j + \hat{\sigma}^2)]\hat{\theta}_j\psi_j(x). \tag{4.2}$$

4.2. Comparison with a local linear estimator via Monte Carlo simulations

So far we have discussed an orthogonal series approach. It is of interest to compare the suggested adaptive orthogonal series estimator with a kernel type estimator. One of the most promising kernel estimators is a local linear estimator (see e.g., Fan (1992, 1993)). Here we compare our adaptive estimator with the pseudo local linear estimator of Fan (1992) which utilizes a variable bandwidth based on the underlying regression function, design density and variance of errors. The latter allows us to compare our data driven estimator with this ideal kernel type estimator.

We compare these estimators using Monte Carlo simulations for the following two families of underlying regression functions: $\mathcal{F}_1 = \{f : f(x) = .5 + \pi^{-1}\arctan(a(x-b)),\ .5 \le a \le 10,\ .2 \le b \le .7\}$ and $\mathcal{F}_2 = \{f : f(x) = .5(1+\sin(cx)),\ .5 \le c \le 5\}$. The first family mimics a wide variety of monotone functions, the second is suggested by Fan (1992) and mimics differently oscillated functions. The predictors are i.i.d. normal (μ, d^2) random variables, where $0 \le \mu \le 1$ and $.1 \le d^2 \le 4$, which have been truncated on the interval $[0,1]$.

The variance $\sigma^2(x)$ of errors is assumed to belong to the family $\{\sigma^2 : \sigma^2(x) = 1 + kx^2,\ -.8 \le k \le .8\}$. For each Monte Carlo simulation the parameters a, b, c, μ, d and k are chosen uniformly and independently from their ranges.

For each family of regression functions and for sample sizes 25, 50, 100, 250 and 500, ten thousand independent Monte Carlo simulations were performed. Recall that each simulation is based on randomly chosen functions $f(x)$, $h(x)$ and $\sigma^2(x)$. The basis for comparison of the adaptive estimator (4.2) with the pseudo local linear estimator is the integrated squared error of an estimate f_n, that is, $\int_0^1 (f_n(x) - f(x))^2 dx$. For each set of 10000 simulations, the sample mean and sample median of integrated squared errors were computed for each estimator. Table 1 displays the ratios of the sample mean and sample median of integrated squared errors of the adaptive estimate (4.2) with those of the pseudo local linear estimator.

Table 1. Ratios of sample means and sample medians

Sample	Ratios			
	$\mathcal{F}_1$		$\mathcal{F}_2$	
Size	Mean	Median	Mean	Median
25	3.7	1.8	2.5	1.3
50	3.9	2.0	2.7	1.5
100	4.0	2.0	2.9	1.7
250	3.6	1.8	2.8	1.7
500	3.6	1.7	2.5	1.6

Table 1 shows that even in comparison with the pseudo local linear estimator, that is one of the best kernel type estimator whose variable bandwidth is based on both underlying regression function and nuisance functions, our adaptive estimator performs reasonably well for the small sample sizes. It is not surprising that the orthogonal series estimator performs better for the family $\mathcal{F}_2$ of oscillated functions.

The fact that the sample means of the studied ratios are greater than the corresponding sample medians is due to extreme Monte Carlo samples. In particular, we observed samples, that is combination of an underlying regression function, nuisance functions, random predictors and errors, when the ratios exceeded twelve hundred. That is, some samples are extremely favorable to the pseudo estimator which is based on underlying regression function and nuisance functions. However, it follows from Table 1 that the probability of the extreme samples is extremely low and overall our data-driven estimator performs relatively well in comparison with the pseudo local linear estimator.

Acknowledgements

The work presented in this article was supported in part by a National Science Foundation (NSF) Grant DMS-9123956 and Sandia National Laboratories Grant AE-1679. We thank the referee and associate editor for helpful comments and suggestions.

References

Barron, A., Birgé, L. and Massart, P. (1995). Risk bounds for model selection via penalization. Technical Report 95.54, Universite de Paris-Sud.

Birgé, L. and Massart, P. (1994). From model selection to adaptive estimation. Technical Report 94.8, Universite de Paris-Sud.

Brown, L. and Low, M. (1992). Asymptotic equivalence of nonparametric regression and white noise. Technical Report Cornell University (to appear in *Ann. Statist.*).

Efromovich, S. and Pinsker, M. S. (1984). An adaptive algorithm of nonparametric filtering. *Automat. Remote Control* **45**, 58-65.

Efromovich, S. (1985). Nonparametric estimation of a density with unknown smoothness. *Theory Probab. Appl.* **30**, 557-568.

Efromovich, S. (1986). The adaptive algorithm of nonparametric regression. In *Abstracts of Second IFAC Symposium on Stochastic Control.* Vilnuis **2**, 112-114.

Efromovich, S. (1992). On nonparametric regression for i.i.d. observations in general setting. Technical Report, UNM (to be published in *Ann. Statist.* **24**, 1996).

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Dekker, New York.

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

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.* **21**, 196-216.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser.B* **2**, 371-394.

Golubev, G. K. and Nussbaum, M. (1990). A risk bound in Sobolev class regression. *Ann. Statist.* **18**, 758-778.

Golubev, G. K. and Nussbaum, M. (1992). Adaptive spline estimates for nonparametric regression models. *Theory Probab. Appl.* **37**, 521-529.

Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics: Some Basic Concepts.* Springer-Verlag, New York.

Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13**, 984-997.

Pinsker, M. S. (1980). Optimal filtration of functions from L_2 in gaussian noise. *Problems Information Transmission* **16**, 52-68.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13**, 970-983.

Wilks, S. (1962). *Mathematical Statistics.* John Wiley & Sons.

Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131, U.S.A.

Institute for Problems of Information Transmission, Russian Academy of Sciences, 19 Ermolova, Moscow 110051, Russia.

(Received February 1993; accepted December 1995)