On the estimation of a monotone conditional variance in nonparametric regression

Holger Dette, Kay Pilz
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum, Germany
e-mail: holger.dette@ruhr-uni-bochum.de

July 22, 2004

Abstract

A monotone estimate of the conditional variance function in a heteroscedastic, nonparametric regression model is proposed. The method is based on the application of a kernel density estimate to an unconstrained estimate of the variance function and yields an estimate of the inverse variance function. The final monotone estimate of the variance function is obtained by an inversion of this function. The method is applicable to a broad class of nonparametric estimates of the conditional variance and particularly attractive to users of conventional kernel methods, because it does not require constrained optimization techniques. The approach is also illustrated by means of a simulation study.

AMS Subject Classification: 62G05
Keywords and phrases: nonparametric regression, heteroscedasticity, variance function, monotonicity, order restricted inference

1 Introduction

In regression analysis the assumption of homoscedasticity is often not satisfied and the efficiency of the statistical analysis can be improved substantially by taking heteroscedasticity into account. The classical example is the weighted least squares method, which requires estimates of the variance function. Other examples, where the estimation of the conditional variance is important include the choice of a local bandwidth in nonparametric regression [see Müller and Stadtmüller

(1987), Fan and Gijbels (1995)], the construction of confidence intervals for the conditional expectation [see Carroll (1987), Fan and Gijbels (1996)] and quality control [see Box (1988)]. In contrast to the problem of estimating the conditional mean much less effort has been spent on the construction of nonparametric variance function estimators. Carroll (1982) developed kernel estimators in the context of linear regression, Müller and Stadtmüller (1987) and Hall and Carroll (1989) analyzed kernel-type estimators without assuming a parametric form of the mean function, and Müller and Stadtmüller (1993) studied a broad class of estimators of the conditional variance, which are representable as quadratic forms. Local polynomial variance function estimators have been proposed by Fan and Gijbels (1995) and Ruppert, Wand, Holst and Hössjer (1997), where the latter authors also consider the problem of estimating derivatives of the variance function, a topic with applications in engineering. More recently the estimation of the conditional variance was considered by Fan and Yao (1998) in a time series context and by Yu and Jones (2004), who proposed a localized normal likelihood approach.

In many applications monotone estimates of the regression and variance function are required because of physical considerations. Such examples typically appear in growth curve models or in models, where the conditional variance is a function of the conditional mean, which depends monotonically on an explanatory variable. In contrast to the problem of estimating a monotone conditional expectation [see e.g. Brunk (1955), Mukerjee (1988), Mammen (1991), Hall and Huang (2001) among many others], the problem of estimating a monotone variance function has not been considered so far in the literature. In the present paper we propose a simple and efficient method for the estimation of a monotone conditional variance, which is based on the evaluation of a kernel density estimate from some (not necessarily monotone) estimated values of the variance function. This idea was introduced by Dette, Neumeyer and Pilz (2003) in the context of estimating a monotone regression function and will be adapted to the specific problem of statistical inference for the conditional variance. The method produces an estimate of the inverse of the monotone variance function and is applicable to any of the unconstrained variance function estimators mentioned in the previous paragraph.

In Section 2 we introduce some general notation and explain the basic idea of monotonizing a function by kernel density estimation. Because most work on unconstrained variance estimation suggests smoothing squared residuals from a nonparametric fit or pseudo-residuals by kernel smoothing we mainly restrict ourselves to this type of variance function estimators, but the results of the paper remain valid for other estimation methods. In Section 3 we prove asymptotic normality of the new estimate and show that it is first order asymptotically equivalent to the unconstrained variance function estimate. We also mention the corresponding statements for the local polynomial estimators of the conditional variance introduced by Fan and Gijbels (1995) and Ruppert, Wand, Holst and Hössjer (1997). For the sake of brevity we restrict ourselves to the case of a nonparametric regression model with a fixed design and independent errors, but extensions to more general models (random design, time series) are briefly mentioned in Section 3.3. Finally, in Section 4 a small simulation study is presented which illustrates the finite sample properties of the new monotone variance function estimates, while some of the more technical arguments are deferred to the appendix in Section 5.

2 Preliminaries: monotonizing by kernel density estimation

Consider the common nonparametric regression model

$$Y_{i,n} = m(x_{i,n}) + \sqrt{s(x_{i,n})}\varepsilon_{i,n}, \tag{2.1}$$

where $0 \le x_{1,n} < x_{2,n} < \ldots < x_{n,n} = 1$ are fixed design points satisfying

$$\int_0^{x_{i,n}} f(t)dt = \frac{i}{n}, \quad i = 1, \ldots, n \tag{2.2}$$

for a positive design density $f : [0,1] \to \mathbb{R}$ [see Sacks and Ylvisaker (1970)], $m : [0,1] \to \mathbb{R}$ denotes the regression and $s : [0,1] \to \mathbb{R}$ is a positive variance function. The errors $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ are assumed to be independent identically distributed with mean $E[\varepsilon_{i,n}] = 0$, variance $\mathrm{Var}(\varepsilon_{i,n}) = 1$ and fourth moment $E[\varepsilon_{i,n}^4] = m_4(x_{i,n})$, where $m_4 : [0,1] \to \mathbb{R}^+$ is a smooth function. For the sake of a simple notation we omit the index n, whenever it is clear from the context, i.e. we use the notation Y_i, x_i, ε_i instead of $Y_{i,n}, x_{i,n}, \varepsilon_{i,n}$ in such cases. We assume that the design density f and the variance function s are two times continuously differentiable and that the regression function m satisfies certain smoothness conditions which will be specified below. Moreover, the variance function is assumed to be strictly monotonic and we are interested in an estimate of this function, which also satisfies this restriction.

In order to fix ideas let s denote an arbitrary strictly increasing function on the interval $[0,1]$, then the inverse of s can be represented as

$$s^{-1}(t) = \int_0^1 I\{s(x) \le t\}dx. \tag{2.3}$$

Note that this function is not necessarily smooth, but smoothing can easily be accomplished by considering the function

$$s^{-1}(t, h_d) = \frac{1}{h_d}\int_0^1 \int_{-\infty}^t K_d\Big(\frac{s(x) - u}{h_d}\Big)dudx, \tag{2.4}$$

where h_d is a bandwidth satisfying $h_d \to 0$ with increasing sample size and K_d is a two times continuously differentiable, symmetric kernel with compact support, say $[-1,1]$. Note that for $h_d \to 0$ we have

$$\frac{1}{h_d}\int_0^1 \int_{-\infty}^t K_d\Big(\frac{s(x) - u}{h_d}\Big)dudx = \int_0^1 I\{s(x) \le t\}dx + o(1) \tag{2.5}$$

and that for a positive kernel K_d the function $s^{-1}(t, h_d)$ is always increasing, independently whether the original function s has this property, because

$$\frac{\partial}{\partial t} s^{-1}(t, h_d) = \frac{1}{h_d} \int_0^1 K_d\Big(\frac{s(x) - t}{h_d}\Big) dx \geq 0.$$

For more details discussing the role of the inverse of $s^{-1}(t, h_d)$ as a monotone approximation of the function s we refer to Section 2 in Dette, Neumeyer and Pilz (2003).
In the present context we will use this concept to obtain monotone estimates of the variance function. For the sake of transparency we restrict ourselves to the problem of estimating an increasing variance function, the corresponding case of a decreasing variance is briefly mentioned in Remark 2.1. Observing the discussion in the previous paragraph we only need an unconstrained estimate of the variance function, and for this purpose we will use

$$\hat{s}(x) = \frac{\sum_i K_r\left(\frac{x - x_i}{h_r}\right) \Delta_i^2}{\sum_i K_r\left(\frac{x - x_i}{h_r}\right)}, \tag{2.6}$$

where K_r and h_r denote a further kernel and bandwidth, respectively. We assume that the kernel K_r is symmetric and has also compact support contained in the interval $[-1, 1]$. In (2.6) the quantities Δ_i will denote residuals from a nonparametric fit [see e.g. Hall and Marron (1990)] or pseudo residuals [see e.g. Rice (1984) or Gasser, Sroka and Jennen-Steinmetz (1986)]. For the sake of brevity we concentrate on the Nadaraya-Watson estimate based on smoothing squared residuals, but other types of estimators could be considered as well [see Remark 3.3 for some examples]. Estimators of the form (2.6) have been considered by several authors, including Müller and Stadtmüller (1987, 1993), who mainly discussed pseudo residuals, Hall and Carroll (1989) and Akritas and van Keilegom (2001), who proposed to use residuals from a nonparametric fit. Different smoothing techniques in the context of estimating the conditional variance have been proposed by Ruppert, Wand, Host and Hössjer (1997), Fan and Yao (1998) and Yu and Jones (2004).
Following our general motivation for constructing an increasing variance function estimate we propose the statistic

$$\hat{s}_I^{-1}(t) = \frac{1}{N h_d} \sum_{i=1}^{N} \int_{-\infty}^{t} K_d\Big(\frac{\hat{s}(\frac{i}{N}) - u}{h_d}\Big) du \tag{2.7}$$

as an estimate of s^{-1}. The required monotone increasing estimate of the conditional variance is now obtained by a simple inversion of this function and will be denoted by $\hat{s}_I$ throughout this paper. The properties of this estimate depend on the particular method used for the unconstrained variance function estimate $\hat{s}$, but we prove below that in all important cases the monotone increasing estimate $\hat{s}_I$ is asymptotically normal distributed and first order equivalent to the corresponding unconstrained estimate. Note that the integral in (2.4) has been replaced by a simple quadrature formula with equidistant nodes. Moreover, the estimate $\hat{s}_I$ can be considered as a density estimate based on the "data" $\{(\frac{i}{N}, \hat{s}(\frac{i}{N}))|\ i = 1, \ldots, N\}$ and the number N used in this density estimator does not necessarily coincide with the sample size n used for the calculation of the unconstrained estimate. The indices "r" and "d" of the kernel functions K_r and K_d correspond to the phrase "regression" and "density", because we combine a regression with a density estimate to define the estimator in (2.7). In the following we will discuss the properties of the new estimate for two different types of residuals Δ_i separately.

Remark 2.1. If the variance function s is supposed to be strictly decreasing the estimate can easily be modified as

$$\hat{s}_A^{-1}(t) := \frac{1}{Nh_d} \sum_{i=1}^{N} \int_t^{\infty} K_d\Big(\frac{\hat{s}(\frac{i}{N}) - u}{h_d}\Big) du \tag{2.8}$$

and the antitonic estimate is obtained by the inversion of this function.

3 Monotone variance function in action

3.1 Monotone variance function estimation with pseudo residuals

Following Hall, Kay and Titterington (1990) we define pseudo residuals by

$$\Delta_i = \Delta_{i,n} = \sum_{j=0}^{r} d_j Y_{i+j}, \tag{3.1}$$

where the quantities $d_0, \ldots, d_r$ are given weights satisfying

$$\sum_{j=0}^{r} d_j = 0, \quad \sum_{j=0}^{r} d_j^2 = 1. \tag{3.2}$$

In this case the preliminary estimator of the variance function is defined by

$$\hat{s}(x) = \frac{\sum_{i=1}^{n-r} K\Big(\frac{x-x_i}{h_r}\Big)\Delta_i^2}{\sum_{i=1}^{n-r} K\Big(\frac{x-x_i}{h_r}\Big)}. \tag{3.3}$$

Two special choices of pseudo residuals are very popular and have been considered by Rice (1984) $[r = 1, d_0 = -d_1 = 1/\sqrt{2}]$ and Gasser, Sroka and Jennen-Steinmetz (1986) $[r = 2, d_0 = d_2 = 1/\sqrt{6}, d_1 = -2/\sqrt{6}]$, while some general properties of variance estimates based on pseudo residuals are discussed in Dette, Munk and Wagner (1998) in the case of a homoscedastic regression model. Throughout this paragraph we assume that the regression function is Lipschitz continuous of order $\gamma > \frac{1}{4}$, which allows us to replace the quantities Δ_i in (3.1) by their unobservable counterparts

$$\Delta_i^{\varepsilon} = \Delta_{i,n}^{\varepsilon} = \sum_{j=0}^{r} d_j \sqrt{s(x_{i+j})}\varepsilon_{i+j} \tag{3.4}$$

with sufficiently accuracy [see the proofs in Section 5.1 of the Appendix]. The main properties of isotone variance function estimators using pseudo residuals are summarized in the following theorem, for which we require some assumptions regarding the bandwidths h_d, h_r and the number N used in the definition of the statistic $\hat{s}_I^{-1}$, that is

$$h_r \to 0 \ , \quad h_d \to 0, \tag{3.5}$$

$$nh_d \to \infty \ , \quad nh_r \to \infty \tag{3.6}$$

$$\lim_{h_d \to 0, h_r \to 0} h_r/h_d = \infty \tag{3.7}$$

$$nh_r^5 = O(1) \ , \quad n = O(N), \tag{3.8}$$

$$\frac{1}{nh_r h_d^2} = o(1). \tag{3.9}$$

Theorem 3.1. *Assume that the regression function m in the nonparametric regression model (2.1) is Lipschitz continuous of order $\gamma > 1/4$ and that the assumptions stated at the beginning of Section 2 and in (3.5) - (3.9) are satisfied. Let $\hat{s}_I$ denote the isotone estimate of the variance function s obtained as the inverse of the statistic (2.7) with the statistic (3.3) as preliminary estimate, then it follows that for every $t \in (0,1)$ with $s'(t) > 0$*

$$\sqrt{nh_r}\Big(\hat{s}_I(t) - s(t) - \Gamma(h_d, h_r, t)\Big) \overset{\mathcal{D}}{\Rightarrow} \mathcal{N}(0, \beta^2(t)), \tag{3.10}$$

where the asymptotic bias and variance are given by

$$\Gamma(h_d, h_r, t) \ = \ \kappa_2(K_d)\frac{s''(t)}{(s'(t))^2}h_d^2 + \kappa_2(K_r)\Big(\frac{s''f + 2s'f'}{f}\Big)(t)h_r^2 \ , \tag{3.11}$$

$$\beta^2(t) \ = \ \frac{s^2(t)\{m_4(t) - 1 + \delta_r\}}{f(t)} \int_{-1}^{1} K_r^2(u)du, \tag{3.12}$$

respectively, for a given kernel K the constant $\kappa_2(K)$ is defined as

$$\kappa_2(K) = \frac{1}{2}\int_{-1}^{1} v^2 K(v)dv, \tag{3.13}$$

and the quantitiy δ_r is given by

$$\delta_r = \sum_{k=1}^{r}(\sum_{j=0}^{r-k} d_j d_{j+k})^2 \quad (r \geq 1). \tag{3.14}$$

Remark 3.2. Note that the dominating term in the representation (3.11) for the bias is given by

$$\Gamma(h_d, h_r, t) = \kappa_2(K_r)\Big(\frac{s''f + 2s'f'}{f}\Big)(t)h_r^2 \ + o(h_r^2), \tag{3.15}$$

because $h_d = o(h_r)$ by assumption (3.7). It was observed by Dette, Neumeyer and Pilz (2003) in the context of estimating a monotone regression function that the choice of the bandwidth h_d in the density step is less critical compared to the choice of the bandwidth h_r in the regression step, and the same fact is true for the problem of estimating the conditional variance. Based on an extensive numerical study we recommend to choose h_d as $h_d = h_r^\alpha$ for some $\alpha \geq 1.5$ and the approximation (3.15) is well justified.

Remark 3.3. It follows from the proof of Theorem 3.1 that the choice of a different smoothing procedure in (2.6) does not change the asymptotic variance of the resulting monotone estimate of the variance function, but its asymptotic bias. For example, if a local linear estimate [see Fan and Gijbels (1996)] is applied to the squared pseudo residuals (3.1), then the resulting estimate $\hat{s}_I$ is asymptotically normal distributed, that is

$$\sqrt{nh_r}\Big(\hat{s}_I(t) - s(t) - \Gamma_{\text{loc}}(h_d, h_r, t)\Big) \overset{\mathcal{D}}{\Rightarrow} \mathcal{N}(0, \beta^2(t)), \tag{3.16}$$

where the asymptotic variance is given by (3.12) and the bias is defined by

$$\Gamma_{\text{loc}}(h_d, h_r, t) = \kappa_2(K_d)\frac{s''(t)}{(s'(t))^2}h_d^2 + \kappa_2(K_r)s''(t)h_r^2 = \kappa_2(K_r)s''(t)h_r^2 + o(h_r). \tag{3.17}$$

Other estimates for the regression step can be treated similarly. For example, if the local log-linear estimator proposed by Yu and Jones (2004) is used as preliminary unconstrained estimate of the conditional variance, the isotonized estimate $\hat{s}_I$ has still asymptotic variance $\beta^2(t)/nh_r$, asymptotic bias is given by

$$\Gamma_{YJ}(h_d, h_r, t) = \kappa_2(K_d)\frac{s''(t)}{(s'(t))^2}h_d^2 + \kappa_2(K_r)\Big(s''(x) - \frac{(s'(x))^2}{s(x)}\Big)h_r^2 + o(h_r^2)$$

and the appropriately standardized version of $\hat{s}_I$ is asymptotically normal distributed.

Remark 3.4. For the different estimates of the variance function considered in Theorem 3.1 and Remark 3.3 it follows from the results of Müller and Stadtmüller (1993), Yu and Jones (2004) and the proof of Theorem 3.1 that the isotone estimates of the variance function are first order asymptotically equivalent to the unconstrained estimates.

Remark 3.5. Note that the asymptotic variance in Theorem 3.1 depends on the constant δ_r defined in (3.14). For the estimator of Rice (1984) we have $r = 1$, $d_0 = -d_1 = 1/\sqrt{2}$, which yields $\delta_1 = d_0^4 = 1/4$ and

$$\beta_R^2(t) = \frac{s^2(t)}{f(t)}\Big(m_4(t) - \frac{3}{4}\Big)\int_{-1}^{1} K_r^2(u)du.$$

A different weighting scheme was suggested by Gasser, Sroka and Jennen-Steinmetz (1986), who used for a uniform design $(d_0, d_1, d_2) = \frac{1}{\sqrt{6}}(1, -2, 1)$ in the context of a nonparametric homoscedastic regression model, and argued that this sequence yields a smaller bias in the approximation of the pseudo residuals by the quantities defined in (3.4). For this choice we obtain in Theorem 3.1 $(r = 2)$ $\delta_2 = 17/36$,

$$\beta_G^2(t) = \frac{s^2(t)}{f(t)}\Big(m_4(t) - \frac{19}{36}\Big)\int_{-1}^{1} K_r^2(u)du. \tag{3.18}$$

Alternatively one could try to minimize the asymptotic variance (3.12) by an appropriate choice of the weights $d_0, \dots, d_r$. Hall, Kay and Titterington (1990) determined for a fixed order r optimal weights d_j such that the quantity δ_r in (3.14) becomes minimal [see Table 1 of their paper]. For this choice we have

$$\sum_{j=0}^{r-k} d_j d_{j+k} = -\frac{1}{2r},$$

the minimal value of δ_r is obtained as ${\delta_r}^{\text{opt}} = 1/4r$ and the resulting asymptotic variance is given by

$$\beta^2_{\text{opt}}(t) \;=\; \frac{s^2(t)}{f(t)}\Big(m_4(t) - \frac{4r-1}{4r}\Big)\int_{-1}^{1} K_r^2(t)dt. \tag{3.19}$$

Consequently the asymptotic variance in Theorem 3.1 can be decreased by using an optimal difference sequence in the sense of Hall, Kay and Titterington (1990) and an increasing order r. However, some care is appropriate in these asymptotic considerations. For realistic sample sizes it is also necessary to obtain a sufficiently small bias of the pseudo residuals Δ_i and optimal sequences usually produce a small variance but a large bias. The general choice of the weights in the definition of the pseudo residuals was carefully discussed by Dette, Munk and Wagner (1998) in the context of homoscedastic nonparametric regression. These authors give some data driven guidelines for choosing an appropriate order r and the corresponding weights $d_0, \dots, d_r$. In general difference sequences for $r = 1$ or $r = 2$ will be sufficient and the improvement in efficiency by using a larger order is negligible in most cases [compare also with the results of our simulation study in Section 4].

3.2 Monotone variance function estimation with nonparametric residuals

Following Hall and Marron (1990) we consider residuals

$$\hat{\varepsilon}_i = Y_i - \hat{m}(x_i) \tag{3.20}$$

where

$$\hat{m}(x) = \frac{\sum_{i=1}^n K\Big(\frac{x-x_i}{h}\Big)Y_i}{\sum_{i=1}^n K\Big(\frac{x-x_i}{h}\Big)} \tag{3.21}$$

is the Nadaraya-Watson estimate of the regression function. The unconstrained estimate of the conditional variance is now given by

$$\hat{s}(x) = \frac{\sum_{i=1}^{n} K_r\left(\frac{x-x_i}{h_r}\right)\hat{\varepsilon}_i^2}{\sum_{i=1}^{n} K_r\left(\frac{x-x_i}{h_r}\right)}. \tag{3.22}$$

Note that different bandwidths are used for the estimation of the regression and variance function and that the kernels used in (3.21) and (3.22) do not necessarily coincide. The following result is an analogue of Theorem 3.1 for the case, where residuals from a nonparametric fit are used in the construction of a monotone estimate of the conditional variance. For its proof we require the following assumption regarding the bandwidth h in the Nadraya-Watson estimate (3.21)

$$h \to 0, \quad nh \to \infty, \quad h_r = O(h). \tag{3.23}$$

Theorem 3.6. *Assume that the regression function m in the nonparametric regression model (2.1) is two times continuously differentiable and that the assumptions stated at the beginning of Section 2, (3.5) - (3.9) and (3.23) are satisfied. Let $\hat{s}_I$ denote the isotone estimate of the variance function s obtained as the inverse of the statistic (2.7) with the statistic (3.22) as preliminary estimate, then it follows that for every $t \in (0,1)$ with $s'(t) > 0$*

$$\sqrt{nh_r}\Big(\hat{s}_I(t) - s(t) - \Gamma(h_d, h_r, t)\Big) \overset{\mathcal{D}}{\Rightarrow} \mathcal{N}(0, \delta^2(t)), \tag{3.24}$$

where the asymptotic bias is defined by (3.11) and the asymptotic variance is given by

$$\delta^2(t) = \frac{s^2(t)\{m_4(t) - 1\}}{f(t)} \int_{-1}^{1} K_r^2(u)du. \tag{3.25}$$

Note that the asymptotic bias of the monotone estimates based on (3.3) and (3.22) coincide, while there is a difference in the asymptotic variance. The asymptotic variance in (3.25) can be considered as a limit ($r \to \infty$) of the asymptotic variance of the monotone estimate using pseudo residuals with an optimal difference sequence. We note, however, that for realistic sample sizes these asymptotic differences are rarely observable.

Remark 3.7. A different choice of the estimator $\hat{m}$ (for example a local polynomial or the Gasser-Müller estimator) does not change the asymptotic result in Theorem 3.6. On the other hand, if a different estimator is used for the smoothing of the squared residuals in (3.22) the asymptotic bias has to be modified appropriately [compare with Remark 3.3]. Moreover, it can be shown by similar arguments as given in Fan and Yao (1998) that the estimates $\hat{s}_I$ considered in Theorem 3.6 and its corresponding preliminary estimate $\hat{s}$ defined in (3.22) are first order asymptotically equivalent.

3.3 Extension to other models

The results discussed so far remain valid (subject to an appropriate modification of the constants) for other nonparametric regression models. As an illustration consider the stochastic regression model

$$Y_i = m(X_i) + \sqrt{s(X_i)}\varepsilon_i, \tag{3.26}$$

where $(X_i, Y_i)_{i\in\mathbb{Z}}$ is a strictly stationary two dimensional process with $E[Y_i \mid X_i = x] = m(x)$, $\mathrm{Var}(Y_i \mid X_i = x) = s(x) \neq 0$, $E[\varepsilon_i^4 \mid X_i = x] = m_4(x)$. Fan and Yao (1998) proposed $\tilde{s}(x) = \hat{\alpha}$ as estimate of the conditional variance, where

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha,\beta}{\mathrm{argmin}} \sum_{i=1}^{n} \Big\{ \hat{r}_i - \alpha - \beta(X_i - x) \Big\}^2 K_r\Big(\frac{X_i - x}{h_r}\Big)$$

is the local linear estimate based on the nonparametric residuals $\hat{r}_1, \ldots, \hat{r}_n$. These quantities are defined by $\hat{r}_j = \hat{a}$, where

$$(\hat{a}, \hat{b}) = \underset{a,b}{\mathrm{argmin}} \sum_{i=1}^{n} \Big\{ Y_i - a - b(X_i - X_j) \Big\}^2 K\Big(\frac{X_i - X_j}{h}\Big)$$

is the local linear estimate of the regression function (and its derivative) at the point X_j. If $\tilde{s}_I$ denotes the isotonization of the conditional variance estimate obtained as the inverse of the statistic (2.7) with $\hat{s} = \tilde{s}$, the assumptions of Theorem 3.6 and the conditions 1-5 in Appendix 1 of Fan and Yao (1998) are satisfied, then the statistic

$$\sqrt{nh_r}\Big\{ \tilde{s}_I(x) - s(x) - \Gamma_{\mathrm{loc}}(h_d, h_r, x) \Big\}$$

is asymptotically normal with mean 0 and variance $\delta^2(x)$ defined in (3.25), where the quantity $\Gamma_{\mathrm{loc}}(h_d, h_r, x)$ is given by (3.17) and f is the marginal density of X. Again the monotonized estimate is first order asymptotically equivalent to the unconstrained estimate [see Fan and Yao (1998), Theorem 1].

4 Finite sample properties

In this section we illustrate the finite sample properties of the monotone estimates of the conditional variance by means of a small simulation study. We begin with a comparison of different estimates based on pseudo residuals [see Section 3.1] and then compare the best estimates in this class with the monotone variance estimates based on nonparametric residuals [see Section 3.2]. For the sake of brevity we restrict our study to two regression models, that is

$$Y_i = \sin(6x_i) + \sqrt{\frac{3}{2}x_i^2}\varepsilon_i; \quad i = 1, \ldots, n \tag{4.1}$$

$$Y_i = x_i + \sqrt{x_i}\varepsilon_i; \quad i = 1, \ldots, n \tag{4.2}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0,1)$ and the sample size is $n = 100$. As a design a uniform design ($f(x) = 1$) is considered, while the Epanechnikov kernel is used for the kernels K_d and K_r in the density and regression estimate. The bandwidth h_d for the density step is always given by $h_d = h_r^3$. We applied 2000 simulation runs to calculate the squared bias, variance and mean squared error in the interval $[0,1]$.

4.1 Finite sample properties of difference based estimates

In order to avoid boundary effects we use a local linear estimate based on the pseudo residuals (3.1) in the regression step [for a definition of this estimate see also Section 3.3], where different orders r and different sequences of weights are investigated. The choice of the bandwidth is important for the performance of the estimate and we use the following simple plug-in-rule

$$\hat{h}_r = \Big(\frac{\hat{A}}{n}\Big)^{1/5}, \tag{4.3}$$

where

$$\hat{A} = \frac{1}{n-r}\sum_{i=1}^{n-r}(\Delta_i^2 - \bar{\Delta}^2)^2 \tag{4.4}$$

is the empirical variance of the pseudo residuals $\Delta_1^2, \ldots, \Delta_n^2$ ($\bar{A}^2 = \frac{1}{n-r}\sum_{i=1}^{n-r}\Delta_i^2$). Because $\hat{A}$ is a consistent estimate of

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\Delta_i^2) \approx \int_0^1 s^2(x)\Big\{2 + (m_4(x)-3)\sum_{\ell=0}^{r} d_\ell^4\Big\} f(x)dx$$

the bandwidth (4.3) is (asymptotically) proportional to the global (with respect to the integrated mean squared error criterion) optimal bandwidth, if a local linear estimate is applied to the pseudo residuals $\Delta_1^2, \ldots, \Delta_n^2$. Smoothing parameters proportional to locally optimal bandwidths could be obtained similarly, but the bandwidth (4.3) yields reasonable results in all cases considered in our study.

Our first example investigates the optimal difference sequences introduced by Hall, Kay and Titterington (1990), which minimize the asymptotic variance of the monotone estimate $\hat{s}_I$. In Figure 4.1 we show the curves of the mean squared error, squared bias and variance with an optimal difference sequence of order $r = 1, 2, 3$. Variance estimates based on pseudo residuals with an optimal difference of larger order show a very similar picture and are therefore not depicted.

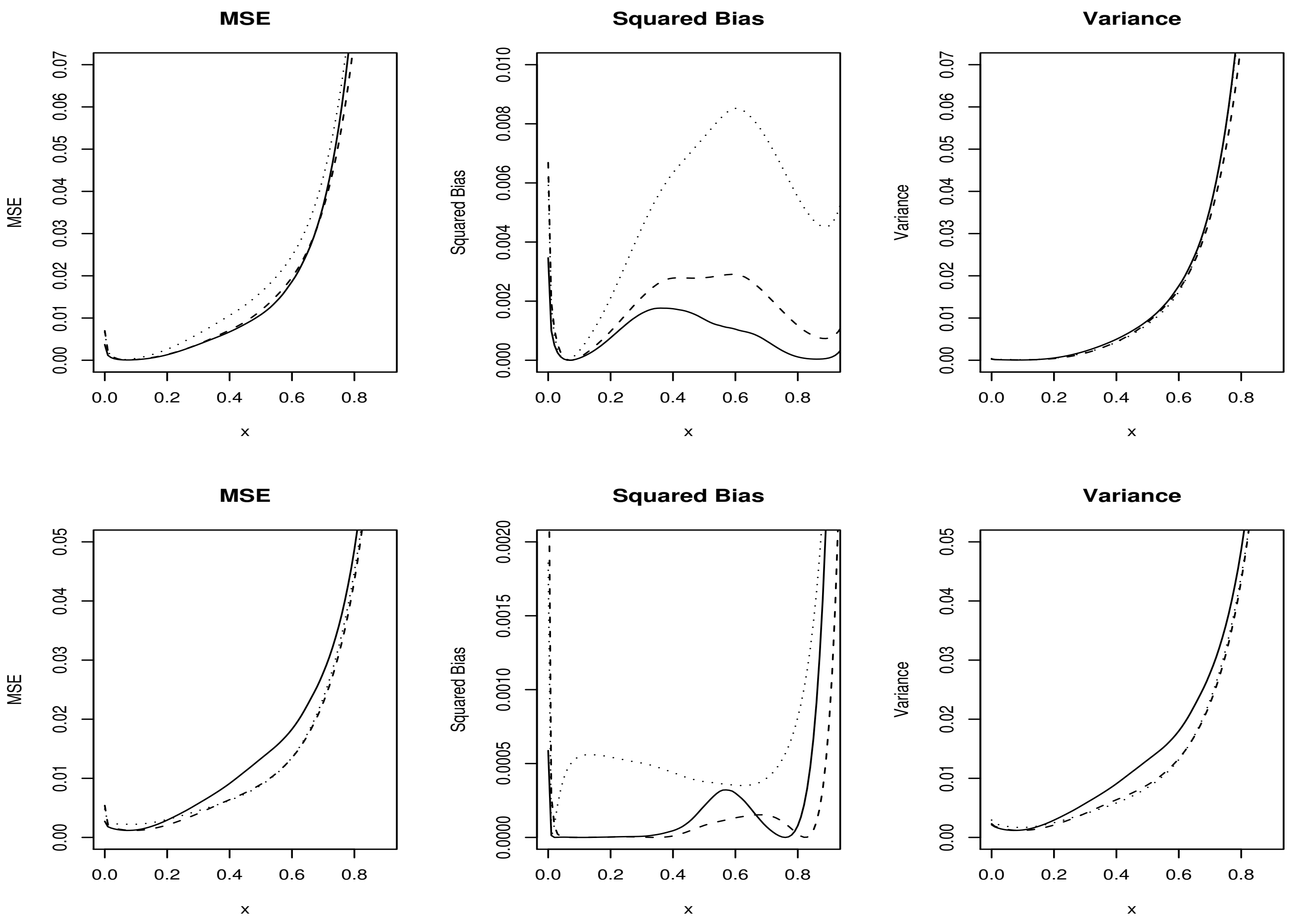

Figure 4.1. *Simulated mean squared error, squared bias and variance of the monotone variance estimate (2.7) based on pseudo residuals with an optimal difference sequence proposed by Hall, Kay and Titterington (1990); $r = 1$: solid line; $r = 2$: dashed line; $r = 3$: dotted line. The upper panel corresponds to model (4.1) and the lower panel to model (4.2).*

We observe that for model (4.1) all estimates behave very similary with respect to the variance criterion (with slight advantage for difference sequences of order $r = 2, 3$) and that the variance of the estimate $\hat{s}_I$ is strictly increasing. This reflects the asymptotic representation in Theorem 3.1, which shows that the variance must be proportional to

$$(nh_r)^{-1} \cdot (2 + \frac{1}{4r}) \cdot \frac{3}{2} \cdot t^2 \cdot 0.6$$

(recall that $f(x) \equiv 1$ and that for the Epanechnikov kernel $\int K^2(u)du = 3/5$). On the other hand there are advantages with respect to the squared bias criterion for the estimates using pseudo residuals with a lower order ($r = 1, 2$), while the monotone variance estimate based on pseudo residuals with an optimal difference sequence of order 3 has a substantial larger bias. A similar phenomenon was observed by Dette, Munk and Wagner (1998) in the context of variance estimation in a homoscedastic nonparametric regression model. These differences are also reflected in the mean squared error curves, where the estimates with pseudo residuals of order two and three have the best performance.

Note that for the regression model (4.2) the second derivative of the variance function vanishes, which results in a substantially smaller bias in Theorem 3.1. As a consequence the variance has a stronger impact on the mean squared error and we expect that variance estimates based on optimal difference sequences of larger order have a better performance. These asymptotic properties are clearly reflected in the squared bias and variance curve (see the lower panel of Figure 4.1). The variance estimates $\hat{s}_I$ based on pseudo residuals with an optimal difference sequence of order two and three have the best performance with respect to the mean squared error criterion and the differences between the three estimates are now mainly caused by the variance.

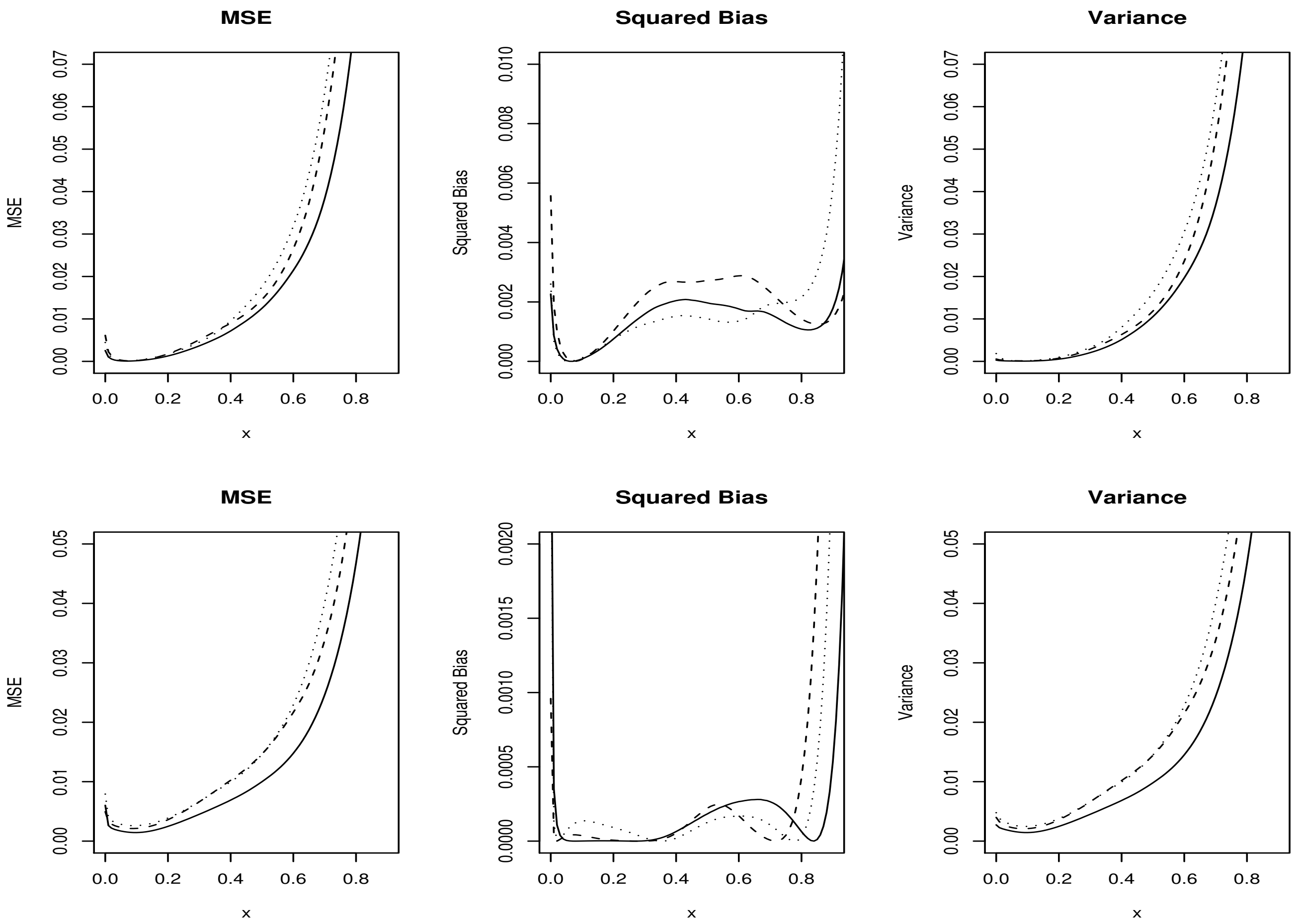

Figure 4.2. *Simulated mean squared error, squared bias and variance of the monotone variance estimate (2.7) based on pseudo residuals with a difference sequence of the form (4.5); $r = 1$: solid line; $r = 2$: dashed line; $r = 3$: dotted line. The upper panel corresponds to model (4.1) and the lower panel to model (4.2).*

Figure 4.2 shows the corresponding curves for model (4.1) and (4.2) if the difference sequence

$$d_i = (-1)^i \frac{\binom{r}{i}}{\binom{2r}{r}^{1/2}} \quad r = 1, 2, 3 \tag{4.5}$$

is used for the construction of the pseudo residuals Δ_i in (3.1). As pointed out by Dette, Munk and Wagner (1998) these difference sequences reduce the bias at the cost of a larger variance. Note that for $r = 1$ and $r = 2$ this choice yields the difference sequences proposed by Rice (1984) and Gasser, Sroka and Jennen-Steinmetz (1986), respectively. For order $r = 3$ this effect is clearly visible in model (4.1), where we observe a slightly smaller curve for the squared bias (compare also the upper panels in Figure 4.1 and 4.2), but a larger variance. For both models the difference sequence with $r = 1$ has the best performance in the class (4.5) and the decrease with respect to the bias does not compensate the increase in variance.
In model (4.1) the estimate with a difference sequence of order $r = 1$ produces the smallest mean squared error curve among the estimates using difference sequences of the form (4.5) [see the upper panel Figure 4.2], but the estimate with an optimal difference sequence of order $r = 2$ has a similar mean squared error [see Figure 4.1 and note that for $r = 1$ the optimal difference sequence and the difference sequence of the form (4.5) coincide]. In model (4.2) the best optimal difference sequence (obtained by using the order $r = 2$ or $r = 3$) yields a substantially smaller mean squared error than the best difference sequence from the class (4.5).

Variance estimates based on pseudo residuals with optimal difference sequences produce a substantially smaller variance and mean squared error compared to the estimators using the difference sequences of the form (4.5). Because other simulation results (which are not depicted here for the sake of brevity) show a similar picture we recommend the use of the optimal difference sequences if pseudo residuals are used in the construction of the monotone estimate $\hat{s}_I$ of the conditional variance. We now compare these estimates with the monotone variance estimates based on nonparametric residuals introduced in Section 2.2.

4.2 Pseudo or nonparametric residuals?

For the construction of the nonparametric residuals $\hat{\varepsilon}_i = Y_i - \hat{m}(x_i)$ we use a local linear estimate $\hat{m}$ with bandwidth

$$h = \Big(\frac{\hat{\sigma}^2}{n}\Big)^{1/5}, \tag{4.6}$$

where $\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=2}^n (Y_i - Y_{i-1})^2$ is the nonparametric estimate of Rice (1984) for the integrated variance. Again a local linear estimate based on the nonparametric residuals $\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2$ is used in the preliminary regression step. The bandwidth h_r was chosen according to the plug-in rule (4.3) where the pseudo residuals Δ_i^2 in (4.4) are now replaced by the nonparametric residuals $\hat{\varepsilon}_i^2$.

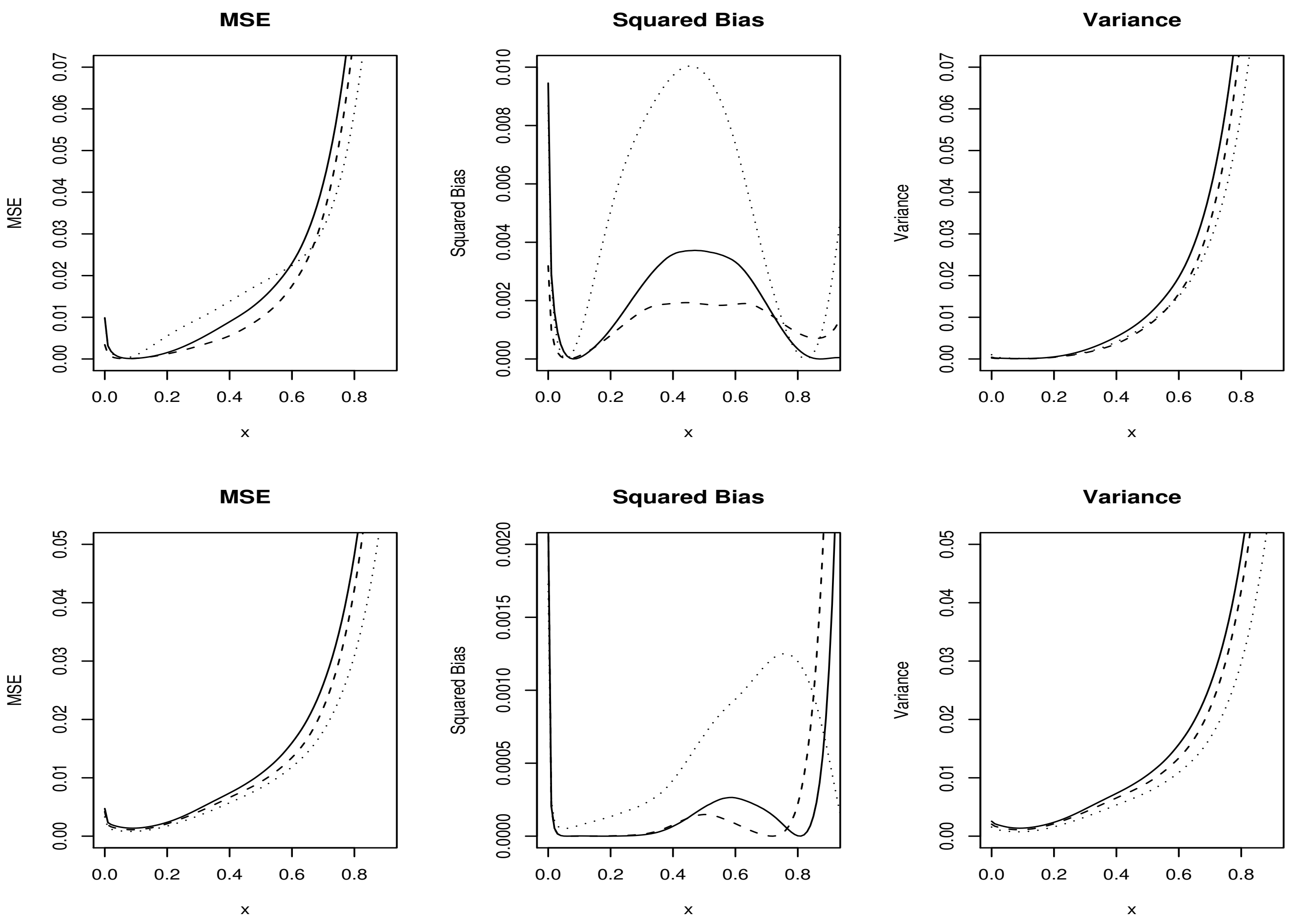

Figure 4.3. *Simulated mean squared error, squared bias and variance of the monotone variance estimate (2.7) based on pseudo residuals with an optimal difference sequence order $r = 1$ (solid line), with an optimal difference sequence of order $r = 2$ (dashed line) and based on nonparametric residuals (dotted line). The upper panel corresponds to model (4.1) and the lower panel to model (4.2).*

Throughout this section monotone variance estimators obtained from the nonparametric residuals $\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2$ will be denoted by $\hat{s}_I^N$, while the estimates obtained from pseudo residuals with the best optimal variance sequence ($r = 2$) and the best sequence of the form (4.5) ($r = 1$) are denoted by $\hat{s}_I^{D_2}$ and $\hat{s}_I^{D_1}$, respectively. Note that in the case $r = 1$ the optimal difference sequence and the difference sequence of the form (4.5) coincide. For both models (4.1) and (4.2) we observe in Figure 4.3 that the estimate $\hat{s}_I^N$ has the smallest variance followed by $\hat{s}_I^{D_2}$ and $\hat{s}_I^{D_1}$. This corresponds to asymptotic theory, which shows that the asymptotic variance of the statistics $\hat{s}_I^N, \hat{s}_I^{D_1}, \hat{s}_I^{D_2}$ is given by

$$\frac{6}{5}\frac{s^2(t)}{nh_r}, \ \frac{51}{40}\frac{s^2(t)}{nh_r}, \ \frac{27}{20}\frac{s^2(t)}{nh_r},$$

respectively. However, Figure 4.3 also shows that there are differences in the behaviour with respect to the squared bias criterion. In both models the estimate $\hat{s}_I^N$ produces the largest bias (but this is negligible in the model (4.2)). The estimate $\hat{s}_I^{D_2}$ has a smaller (squared) bias in both models than $\hat{s}_I^{D_1}$. In model (4.1) the estimates based on pseudo residuals have a smaller mean squared error than $\hat{s}_I^N$ over a broad range of the interval $[0,1]$. Only at the right boundary of the interval $[0,1]$ the smaller variances of $\hat{s}_I^N$ compensate its larger bias, such that it becomes the best estimate in our comparison. On the other hand in model (4.2) the bias can be neglected and the mean squared error is dominated by the variance. As a consequence the monotone variance estimate $\hat{s}_I^N$ based on nonparametric residuals yields the smallest mean squared error for the complete interval $[0,1]$.

5 Proofs

5.1 Proof of Theorem 3.1.

The proof is performed in several steps. At first we calculate the asymptotic bias and variance of the statistic $\hat{s}_I^{-1}$ defined in (2.7), secondly, we establish asymptotic normality of this estimate and finally we use this result to obtain the assertion of Theorem 3.1. For the sake of transparency we assume that $N = n$; the general case is obtained by exactly the same arguments with an additional amount of notation.

For the calculation of the asymptotic bias we first note that it follows from Lemma 2.1 in Dette, Neumeyer and Pilz (2003)

$$\hat{s}_I^{-1}(t) = s^{-1}(t) + \kappa_2(K_d)h_d^2(s^{-1})''(t) + \Delta_n(t) + o(h_d^2) + O\Big(\frac{1}{nh_d}\Big), \tag{5.1}$$

where the term $\Delta_n(t)$ is given by

$$\Delta_n(t) = \frac{1}{nh_d}\sum_{i=1}^n \int_{-\infty}^t \Big\{K_d\Big(\frac{\hat{s}(\frac{i}{n}) - u}{h_d}\Big) - K_d\Big(\frac{s(\frac{i}{n}) - u}{h_d}\Big)\Big\}du = \Delta_n^{(1)}(t) + \frac{1}{2}\Delta_n^{(2)}(t), \tag{5.2}$$

and the quantities $\Delta_n^{(j)}(t)$ $(j = 1, 2)$ in this decomposition are defined by

$$\Delta_n^{(1)}(t) = \frac{-1}{nh_d}\sum_{i=1}^n K_d\Big(\frac{s(\frac{i}{n}) - t}{h_d}\Big)\Big\{\hat{s}(\frac{i}{n}) - s(\frac{i}{n})\Big\}, \tag{5.3}$$

$$\Delta_n^{(2)}(t) = \frac{1}{nh_d^3}\sum_{i=1}^n \int_{-\infty}^t K_d''\Big(\frac{\xi_i - u}{h_d}\Big)\Big\{\hat{s}(\frac{i}{n}) - s(\frac{i}{n})\Big\}^2 du, \tag{5.4}$$

with $|\xi_i - s(\frac{i}{n})| < |\hat{s}(\frac{i}{n}) - s(\frac{i}{n})|$ $(i = 1, \dots, n)$. With an appropriate modification at the boundary it follows by similar arguments as in Müller and Stadtmüller (1993) for the second term

$$\Delta_n^{(2)}(t) = O\Big(\frac{1}{h_d}\Big(h_r^4 + \frac{1}{nh_r}\Big)\Big). \tag{5.5}$$

Replacing the density estimate in the denominator of $\hat{s}(\frac{i}{n})$ by $nh_r f(\frac{i}{n})$ we obtain for the first term of the decomposition (5.2)

$$\Delta_n^{(1)}(t) = \Big(\Delta_n^{(1.1)}(t) + \Delta_n^{(1.2)}(t) + \Delta_n^{(1.3)}(t)\Big)(1 + o_p(1)), \tag{5.6}$$

with

$$\Delta_n^{(1.1)}(t) = \frac{-1}{n^2 h_d h_r} \sum_{i,j=1}^{n} K_d\Big(\frac{s(\frac{i}{n}) - t}{h_d}\Big) K_r\Big(\frac{x_j - \frac{i}{n}}{h_r}\Big) \frac{(\Delta_j^\varepsilon)^2 - s(\frac{i}{n})}{f(\frac{i}{n})}, \tag{5.7}$$

$$\Delta_n^{(1.2)}(t) = \frac{-1}{n^2 h_d h_r} \sum_{i,j=1}^{n} K_d\Big(\frac{s(\frac{i}{n}) - t}{h_d}\Big) K_r\Big(\frac{x_j - \frac{i}{n}}{h_r}\Big) \frac{(\Delta_j^m)^2}{f(\frac{i}{n})}, \tag{5.8}$$

$$\Delta_n^{(1.3)}(t) = \frac{-2}{n^2 h_d h_r} \sum_{i,j=1}^{n} K_d\Big(\frac{s(\frac{i}{n}) - t}{h_d}\Big) K_r\Big(\frac{x_j - \frac{i}{n}}{h_r}\Big) \frac{\Delta_j^m \Delta_j^\varepsilon}{f(\frac{i}{n})}, \tag{5.9}$$

where for $j = 1, \ldots, n-r$ the quantities $\Delta_j^\varepsilon, \Delta_j^m$ are defined by

$$\Delta_j^m = \sum_{\ell=0}^{r} d_\ell m(x_{j+\ell}) \tag{5.10}$$

$$\Delta_j^\varepsilon = \sum_{\ell=0}^{r} d_\ell \sqrt{s(x_{j+\ell})}\varepsilon_{j+\ell}, \tag{5.11}$$

respectively, and we use the notation $\Delta_j^\varepsilon = \Delta_j^m = 0$, whenever $j \in \{n-r+1, \ldots, n\}$. A straightforward calculation and the assumption of Lipschitz continuity for the regression function show that

$$\Delta_j^m = \sum_{\ell=0}^{r} d_\ell m(x_{j+\ell}) = \sum_{\ell=0}^{r-1}\Big(\sum_{k=0}^{\ell} d_k\Big)\Big(m(x_{j+\ell}) - m(x_{j+\ell+1})\Big) = O\Big(\frac{1}{n^\gamma}\Big)$$

(uniformly with respect to $j = 1, \ldots, n$), and it follows that

$$\Delta_n^{(1.2)}(t) = O\Big(\frac{1}{n^{2\gamma}}\Big). \tag{5.12}$$

Next, consider the first term in (5.7), which has expectation

$$\begin{aligned} E[\Delta_n^{(1.1)}(t)] &= \frac{-1}{n^2 h_d h_r} \sum_{i,j} K_d\Big(\frac{s(\frac{i}{n}) - t}{h_d}\Big) K_r\Big(\frac{x_j - \frac{i}{n}}{h_r}\Big) \frac{\sum_{\ell=0}^{r} d_\ell^2 s(x_{j+\ell}) - s(\frac{i}{n})}{f(\frac{i}{n})} \\ &= -\frac{1}{h_r h_d}\int_0^1\int_0^1 K_d\Big(\frac{s(x) - t}{h_d}\Big) K_r\Big(\frac{y - x}{h_r}\Big) f(y)\frac{s(y) - s(x)}{f(x)} dy dx \cdot (1 + o(1)) \\ &= -h_r^2 \kappa_2(K_r) \int_0^1 \frac{1}{h_d} K_d\Big(\frac{s(x) - t}{h_d}\Big)\Big\{s''(x) + \frac{2s'(x)f'(x)}{f(x)}\Big\} dx \cdot (1 + o(1)) \\ &= -h_r^2 \kappa_2(K_r)\Big(\frac{s''f + 2s'f'}{fs'}\Big)(s^{-1}(t)) \cdot (1 + o(1)), \end{aligned} \tag{5.13}$$

The remaining third term has obviously expectation $E[\Delta_n^{(1.3)}(t)] = 0$, while the second moment can be estimated similarly as in the previous paragraph, that is

$$
\begin{aligned}
E[(\Delta_n^{(1.3)}(t))^2] &= \frac{4}{n^4 h_d^2 h_r^2} \sum_{i,i',j,j'} K_d\Big(\frac{s(\frac{i}{n})-t}{h_d}\Big) K_r\Big(\frac{x_j-\frac{i}{n}}{h_r}\Big) \\
&\quad \times K_d\Big(\frac{s(\frac{i'}{n})-t}{h_d}\Big) K_r\Big(\frac{x_{j'}-\frac{i'}{n}}{h_r}\Big) \frac{\Delta_j^m \Delta_{j'}^m E[\Delta_j^\varepsilon \Delta_{j'}^\varepsilon]}{f(\frac{i}{n}) f(\frac{i'}{n})} \qquad (5.14) \\
&= O\Big(\frac{1}{n^{1+2\gamma} h_r}\Big),
\end{aligned}
$$

where we used the fact that Δ_i^ε and Δ_j^ε are uncorrelated, whenever $|i-j| > r$. Therefore Markov's inequality yields

$$
\Delta_n^{(1.3)}(t) = O_p\Big(\frac{1}{n^{1/2+\gamma} h_r^{1/2}}\Big) = o_p\Big(\frac{1}{\sqrt{nh_r}}\Big), \qquad (5.15)
$$

and a combination with (5.1), (5.2), (5.5), (5.13), (5.15) shows that

$$
\begin{aligned}
&\sqrt{nh_r}\Big\{\hat{s}_I^{-1}(t) - s^{-1}(t) - \kappa_2(K_d) h_d^2 (s^{-1})''(t) + h_r^2 \kappa_2(K_r) \Big(\frac{s'' f + 2s' f}{f s'}\Big)(s^{-1}(t))\Big\}, \qquad (5.16) \\
&= Z_n + o_p(1),
\end{aligned}
$$

where the random variable Z_n is defined as

$$
Z_n = \frac{-1}{n^{3/2} h_d \sqrt{h_r}} \sum_{i,j=1}^{n} K_d\Big(\frac{s(\frac{i}{n})-t}{h_d}\Big) K_r\Big(\frac{x_j-\frac{i}{n}}{h_r}\Big) \frac{(\Delta_j^\varepsilon)^2 - E[(\Delta_j^\varepsilon)^2]}{f(\frac{i}{n})} \qquad (5.17)
$$

For the variance of Z_n we obtain

$$
\begin{aligned}
\mathrm{Var}(Z_n) &= \frac{1}{n^3 h_d^2 h_r} \sum_{i,i',j,j'} K_d\Big(\frac{s(\frac{i}{n})-t}{h_d}\Big) K_d\Big(\frac{s(\frac{i'}{n})-t}{h_d}\Big) K_r\Big(\frac{x_j-\frac{i}{n}}{h_r}\Big) K_r\Big(\frac{x_{j'}-\frac{i'}{n}}{h_r}\Big) \frac{L_{j,j'}}{f(\frac{i}{n}) f(\frac{i'}{n})} \\
&= \frac{(1+o(1))}{n^3 h_d^2 h_r} \sum_{i,i',j} K_d\Big(\frac{s(\frac{i}{n})-t}{h_d}\Big) K_d\Big(\frac{s(\frac{i'}{n})-t}{h_d}\Big) K_r\Big(\frac{x_j-\frac{i}{n}}{h_r}\Big) K_r\Big(\frac{x_j-\frac{i'}{n}}{h_r}\Big) \sum_{k=-r}^{r} \frac{L_{j,j+k}}{f(\frac{i}{n}) f(\frac{i'}{n})}, \qquad (5.18)
\end{aligned}
$$

where the quantities $L_{j,j'}$ are defined by

$$
L_{j,j'} = E[(\Delta_j^\varepsilon)^2 (\Delta_{j'}^\varepsilon)^2] - E[(\Delta_j^\varepsilon)^2] E[(\Delta_{j'}^\varepsilon)^2]. \qquad (5.19)
$$

We now calculate these expectations separately, that is

$$
\begin{aligned}
\sum_{k=-r}^{r} E[(\Delta_j^\varepsilon)^2] E[(\Delta_{j+k}^\varepsilon)^2] &= \sum_{k=-r}^{r} \Big(\sum_{\ell=0}^{r} d_\ell^2 s(x_{j+\ell})\Big)\Big(\sum_{\ell'=0}^{r} d_{\ell'}^2 s(x_{j+k+\ell'})\Big) \qquad (5.20) \\
&= (2r+1) s^2(x_j)(1+o(1)),
\end{aligned}
$$

uniformly with respect to $j = 1, \ldots, n$, where we used the convention $s(x_i) = 0$, whenever $i \notin \{1, \ldots, n\}$. The investigation of the first term in (5.19) is more difficult, but a straightforward calculation gives

$$\begin{aligned}
\sum_{k=-r}^{r} E[(\Delta_j^\varepsilon)^2(\Delta_{j+k}^\varepsilon)^2] &= \sum_{\ell=0}^{r} d_\ell^4 s^2(x_{j+\ell}) m_4(x_{j+\ell}) + 3 \sum_{\substack{\ell,\ell'=0\\ \ell\neq\ell'}}^{r} d_\ell^2 d_{\ell'}^2 s(x_{j+\ell}) s(x_{j+\ell'}) \\
&+ 2\sum_{k=1}^{r} \sum_{\ell,\ell',p,p'=0}^{r} d_\ell d_{\ell'} d_p d_{p'} \sqrt{s(x_{j+\ell}) s(x_{j+\ell'}) s(x_{j+k+p}) s(x_{j+k+p'})} \\
&\qquad\qquad \times E[\varepsilon_{j+\ell}\varepsilon_{j+\ell'}\varepsilon_{j+k+p'}\varepsilon_{j+k+p}] \\
&= s^2(x_j)\Big\{(m_4(x_j)-3)\sum_{\ell=0}^{r} d_\ell^4 + 3 + 2m_4(x_j)\sum_{k=1}^{r}\sum_{\ell=0}^{r-k} d_\ell^2 d_{\ell+k}^2 \\
&+ 2\sum_{k=1}^{r}\Big(\sum_{\substack{\ell,s=0\\ \ell\neq s+k}}^{r} d_\ell^2 d_s^2 + 2\sum_{\substack{\ell,s=0\\ \ell\neq s}}^{r} d_\ell d_{\ell+k} d_s d_{s+k}\Big)\Big\}(1+o(1)) \\
&= s^2(x_j)\Big\{(m_4(x_j)-3)\Big(\sum_{\ell=0}^{r} d_\ell^4 + 2\sum_{k=1}^{r}\sum_{\ell=0}^{r-k} d_\ell^2 d_{\ell+k}^2\Big) + 3 \\
&+ 2\sum_{k=1}^{r}\Big(\Big[\sum_{\ell=0}^{r} d_\ell^2\Big]^2 + 2\Big[\sum_{\ell=0}^{r-k} d_\ell d_{\ell+k}\Big]^2\Big)\Big\}(1+o(1)) \\
&= s^2(x_j)\Big\{m_4(x_j) + 2r + 4\sum_{k=1}^{r}\Big[\sum_{\ell=0}^{r-k} d_\ell d_{\ell+k}\Big]^2\Big\}(1+o(1)), \qquad (5.21)
\end{aligned}$$

uniformly with respect to $j = 1, \ldots, n$. Combining (5.18) - (5.21) and observing the definition of δ_r in (3.14) we thus obtain

$$\begin{aligned}
\mathrm{Var}(Z_n) &= \frac{(1+o(1))}{n^3 h_d^2 h_r} \sum_{i,i',j=1}^{n} K_d\Big(\frac{s(\frac{i}{n})-t}{h_d}\Big) K_d\Big(\frac{s(\frac{i'}{n})-t}{h_d}\Big) \\
&\quad \times K_r\Big(\frac{x_j-\frac{i}{n}}{h_r}\Big) K_r\Big(\frac{x_j-\frac{i'}{n}}{h_r}\Big) \frac{(m_4(x_j)-1+\delta_r)s^2(x_j)}{f(\frac{i}{n})f(\frac{i'}{n})} \qquad (5.22) \\
&= \frac{1}{h_d^2 h_r}\int_0^1 K_d\Big(\frac{s(z)-t}{h_d}\Big)\int_0^1 K_d\Big(\frac{s(y)-t}{h_d}\Big) \\
&\quad \times \int_0^1 \frac{s^2(x)(m_4(x)-1+\delta_r)}{f(y)f(z)} K_r\Big(\frac{x-y}{h_r}\Big) K_r\Big(\frac{x-z}{h_r}\Big) f(x)\,dx\,dy\,dz \cdot (1+o(1)) \\
&= \frac{s^2(s^{-1}(t))(m_4(s^{-1}(t))-1+\delta_r)}{(s'(s^{-1}(t))^2 f(s^{-1}(t))} \int\int\int K_d(w)K_d(v)K_r(u) \\
&\quad \times K_r\Big(\frac{s^{-1}(t+h_d v)-s^{-1}(t+h_d w)}{h_r} + u\Big)\,du\,dv\,dw \cdot (1+o(1)) \\
&= \frac{t^2[m_4(s^{-1}(t))-1+\delta_r]}{(s'(s^{-1}(t))^2 f(s^{-1}(t))} \int_{-1}^{1} K_r^2(u)\,du \cdot (1+o(1)).
\end{aligned}$$

A similar calculation and an application of Orey's (1958) central limit theorem for arrays of m-

dependent random variables finally shows that Z_n is asymptotically normal distributed, that is

$$Z_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \xi^2(t)), \tag{5.23}$$

where the asymptotic variance $\xi^2(t)$ is defined by

$$\xi^2(t) = \frac{t^2\{m_4(s^{-1}(t)) - 1 + \delta_r\}}{(s'(s^{-1}(t))^2 f(s^{-1}(t))} \int_{-1}^{1} K_r^2(u)du,$$

and from (5.16) we have

$$\sqrt{nh_r}\Big\{\hat{s}_I^{-1}(t) - s^{-1}(t) - \kappa_2(K_d)h_d^2(s^{-1})''(t) + h_r^2\kappa_2(K_r)\Big(\frac{s''f + 2s'f'}{fs'}\Big)(s^{-1}(t))\Big\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \xi^2(t)). \tag{5.24}$$

The final assertion regarding the asymptotic normality of the estiamte $\hat{s}_I$ is now obtained by similar arguments as presented in the proof of Theorem 3.2 in Dette, Neumeyer and Pilz (2003), and for the sake of self-consistency we indicate the main steps in this derivation. By a second order Taylor expansion we obtain [see Dette, Neumeyer and Pilz (2003), Lemma A.1]

$$\hat{s}_I(t) - s(t) = -\frac{(\hat{s}_I^{-1} - s^{-1})}{(s^{-1})'}(s(t)) + o_p\Big(\frac{1}{\sqrt{nh_r}}\Big),$$

which yields

$$\begin{aligned}
&\sqrt{nh_r}\Big\{\hat{s}_I(t) - s(t) - \Gamma(h_d, h_r, t)\Big\} \\
&\quad = -\sqrt{nh_r}\Big\{\frac{(\hat{s}_I^{-1} - s^{-1})}{(s^{-1})'}(s(t)) + \Gamma(h_d, h_r, t)\Big\} + o_p(1) \\
&\quad = -\sqrt{nh_r}s'(t)\Big\{(\hat{s}_I^{-1} - s^{-1}) \circ s(t) + \kappa_2(K_d)\frac{s''(t)}{(s'(t))^3}h_d^2 + \kappa_2(K_r)\Big(\frac{s''f + 2s'f}{fs'}\Big)(t)h_r^2\Big\} \\
&\quad \xrightarrow{\mathcal{D}} \mathcal{N}(0, (s'(t))^2\xi^2(s(t))),
\end{aligned}$$

where we used (5.24) and the fact that $s''/(s')^3 = -(s^{-1})''$. Finally, a straightforward calculation shows that

$$(s'(t))^2\xi^2(s(t)) = \frac{s^2(t)\{m_4(t) - 1 + \delta_r\}}{f(t)} \int_{-1}^{1} K_r^2(u)du = \beta^2(t),$$

where $\beta^2(t)$ is the asymptotic variance defined in (3.12).

□

5.2 Proof of Theorem 3.6.

The proof of Theorem 3.6 is performed by similar arguments as the proof of Theorem 3.1 and for this reason we will only indicate the main differences. First we note that the arguments given at the beginning of the proof of Theorem 3.1 remain valid. This follows by some standard calculations using the differentiability of the regression function and some basic properties of the Nadaraya-Watson estimate. Therefore we obtain

$$\sqrt{nh_r}\Big\{\hat{s}_I^{-1}(t) - s^{-1}(t) - \kappa_2(K_d)h_d^2(s^{-1})''(t) + h_r^2\kappa_2(K_r)\Big(\frac{s''f + 2s'f}{fs'}\Big)(s^{-1}(t))\Big\} = W_n + o_p(1), \tag{5.25}$$

where the statistic W_n is defined by

$$W_n = \frac{-1}{n^{3/2}h_d\sqrt{h_r}}\sum_{i,j=1}^{n} K_d\Big(\frac{s(\frac{i}{n}) - t}{h_d}\Big)K_r\Big(\frac{x_j - \frac{i}{n}}{h_r}\Big)\frac{\tilde{\varepsilon}_j^2 - E[\tilde{\varepsilon}_j^2]}{f(\frac{i}{n})}, \tag{5.26}$$

the quantities $\tilde{\varepsilon}_j$ are given by

$$\tilde{\varepsilon}_j = \sqrt{s(x_j)}\varepsilon_j - \sum_{\ell=1}^{n} w_{j\ell}\sqrt{s(x_\ell)}\varepsilon_\ell = \sum_{\ell=1}^{n} w_{j\ell}(\sqrt{s(x_j)}\varepsilon_j - \sqrt{s(x_\ell)}\varepsilon_\ell), \tag{5.27}$$

and

$$w_{j\ell} = \frac{K\Big(\frac{x_\ell - x_j}{h}\Big)}{\sum_{q=1}^{n} K\Big(\frac{x_q - x_j}{h}\Big)} \tag{5.28}$$

denote the weights of the Nadaraya-Watson estimate. In the following we will make use of the estimate

$$W_n = V_n + o_p(1), \tag{5.29}$$

where the statistic V_n is defined by

$$V_n = \frac{-1}{n^{3/2}h_d\sqrt{h_r}}\sum_{i,j=1}^{n} K_d\Big(\frac{s(\frac{i}{n}) - t}{h_d}\Big)K_r\Big(\frac{x_j - \frac{i}{n}}{h_r}\Big)\frac{s(x_j)(\varepsilon_j^2 - 1)}{f(\frac{i}{n})}. \tag{5.30}$$

With this representation it now follows by a similar calculation as given in the proof of Theorem 3.1 that

$$\begin{aligned} \mathrm{Var}(V_n) &= \frac{(1+o(1))}{h_d^2 h_r}\int\int\int K_d\Big(\frac{s(x_1) - t}{h_d}\Big)K_d\Big(\frac{s(x_2) - t}{h_d}\Big) \\ &\quad \times K_r\Big(\frac{x_3 - x_1}{h_r}\Big)K_r\Big(\frac{x_3 - x_2}{h_r}\Big)\frac{f(x_3)s^2(x_3)(m_4(x_3) - 1)}{f(x_1)f(x_2)}dx_1dx_2dx_3 \\ &= \frac{t^2(m_4(s^{-1}(t)) - 1)}{f(s^{-1}(t))(s'(s^{-1}(t))^2}\int K_r^2(u)du\cdot(1+o(1)) , \end{aligned} \tag{5.31}$$

and a straightforward application of Ljapunoff's Theorem yields

$$V_n \stackrel{\mathcal{D}}{\Rightarrow} \mathcal{N}(0, \tilde{\delta}^2(t)) \tag{5.32}$$

where the asymptotic variance $\tilde{\delta}^2(t)$ is defined as

$$\tilde{\delta}^2(t) = \frac{t^2\{m_4(s^{-1}(t)) - 1\}}{(s'(s^{-1}(t))^2 f(s^{-1}(t))}\int K_r^2(u)du.$$

The assertion of Theorem 3.6 now follows by exactly the same arguments as given at the end of the proof of Theorem 3.1.
We finally prove the remaining estimate (5.29) noting that

$$W_n - V_n = 2A_n - B_n, \tag{5.33}$$

where

$$\begin{aligned} A_n &= \frac{1}{n^{3/2} h_d \sqrt{h_r}} \sum_{i,j,\ell=1}^{n} K_d\Big(\frac{s(\frac{i}{n}) - t}{h_d}\Big) K_r\Big(\frac{x_j - \frac{i}{n}}{h_r}\Big) \frac{w_{j\ell}\varepsilon_j\varepsilon_\ell - E[w_{j\ell}\varepsilon_j\varepsilon_\ell]}{f(\frac{i}{n})} \sqrt{s(x_j)s(x_\ell)} \\ B_n &= \frac{1}{n^{3/2} h_d \sqrt{h_r}} \sum_{i,j,\ell,\ell'=1}^{n} K_d\Big(\frac{s(\frac{i}{n}) - t}{h_d}\Big) K_r\Big(\frac{x_j - \frac{i}{n}}{h_r}\Big) \frac{w_{j\ell}w_{j\ell'}\varepsilon_\ell\varepsilon_{\ell'} - E[w_{j\ell}w_{j\ell'}\varepsilon_\ell\varepsilon_{\ell'}]}{f(\frac{i}{n})} \sqrt{s(x_\ell)s(x_{\ell'})}\,. \end{aligned}$$

Obviously, we have $E[A_n] = E[B_n] = 0$, while we obtain for the variance of A_n

$$\begin{aligned} \mathrm{Var}(A_n) &= E[A_n^2] \\ &= \frac{1}{n^3 h_d^2 h_r} \sum_{i,i',j,j',\ell,\ell'} K_d\Big(\frac{s(\frac{i}{n}) - t}{h_d}\Big) K_r\Big(\frac{x_j - \frac{i}{n}}{h_r}\Big) \qquad (5.34) \\ &\qquad \times K_d\Big(\frac{s(\frac{i'}{n}) - t}{h_d}\Big) K_r\Big(\frac{x_{j'} - \frac{i'}{n}}{h_r}\Big) \frac{E[w_{j\ell}w_{j'\ell'}\varepsilon_j\varepsilon_\ell\varepsilon_{j'}\varepsilon_{\ell'}]}{f(\frac{i}{n})f(\frac{i'}{n})} \sqrt{s(x_j)s(x_{j'})s(x_\ell)s(x_{\ell'})} \\ &= \frac{(1+o(1))}{n^3 h_d^2 h_r} \Big\{ c_1 \sum_{i,i',j,\ell} K_d\Big(\frac{s(\frac{i}{n}) - t}{h_d}\Big) K_r\Big(\frac{x_j - \frac{i}{n}}{h_r}\Big) K_d\Big(\frac{s(\frac{i'}{n}) - t}{h_d}\Big) K_r\Big(\frac{x_j - \frac{i'}{n}}{h_r}\Big) \frac{w_{j\ell}^2 s(x_j)s(x_\ell)}{f(\frac{i}{n})f(\frac{i'}{n})} \\ &\qquad + c_2 \sum_{i,i',j,\ell} K_d\Big(\frac{s(\frac{i}{n}) - t}{h_d}\Big) K_r\Big(\frac{x_j - \frac{i}{n}}{h_r}\Big) K_d\Big(\frac{s(\frac{i'}{n}) - t}{h_d}\Big) K_r\Big(\frac{x_\ell - \frac{i'}{n}}{h_r}\Big) \frac{w_{j\ell}w_{\ell j} s(x_j)s(x_\ell)}{f(\frac{i}{n})f(\frac{i'}{n})} \Big\} \end{aligned}$$

for some constants $c_1, c_2 > 0$. Observing the definition of w_{ij} in (5.28) it therefore follows

$$\begin{aligned} \mathrm{Var}(A_n) &= \frac{(1+o(1))}{n h_d^2 h_r h^2} \Big\{ c_1 \int\!\!\int\!\!\int\!\!\int K_d\Big(\frac{s(x_1) - t}{h_d}\Big) K_r\Big(\frac{x_2 - x_1}{h_r}\Big) \\ &\qquad \times K_d\Big(\frac{s(x_3) - t}{h_d}\Big) K_r\Big(\frac{x_2 - x_3}{h_r}\Big) K^2\Big(\frac{x_2 - x_4}{h}\Big) \frac{s(x_2)s(x_4)f(x_4)dx_1dx_2dx_3dx_4}{f(x_1)f(x_2)f(x_3)} \\ &\qquad + c_2 \int\!\!\int\!\!\int\!\!\int K_d\Big(\frac{s(x_1) - t}{h_d}\Big) K_r\Big(\frac{x_2 - x_1}{h_r}\Big) \\ &\qquad \times K_d\Big(\frac{s(x_3) - t}{h_d}\Big) K_r\Big(\frac{x_4 - x_3}{h_r}\Big) K^2\Big(\frac{x_2 - x_4}{h}\Big) \frac{s(x_2)s(x_4)dx_1dx_2dx_3dx_4}{f(x_1)f(x_3)} \Big\} \\ &= O\Big(\frac{1}{nh}\Big). \end{aligned}$$

A similar but tedious calculation shows that

$$\mathrm{Var}(B_n) = O\Big(\frac{h_r}{nh^2}\Big), \tag{5.35}$$

and from (5.33) the estimate (5.29) follows, which completes the proof of Theorem 3.6.

□

Acknowledgements. The authors are grateful to Isolde Gottschlich who typed numerous versions of this paper with considerable technical expertise and to M. Birke, E. Mammen and W. Polonik for useful discussions and some help with the references. The work of the authors was supported by the Sonderforschungsbereich 475, Komplexitätsreduktion in multivariaten Datenstrukturen.

References

M. Akritas, I. van Keilegom (2001). Nonparametric estimation of the residual distribution. Scand. J. Statist., 28, 549-567.

G. E. P. Box (1988). Signal to noise ratios, performance criteria and transformation. Technometrics (with discussions) 30, 1-40.

H. D. Brunk (1955). Maximum likelihood estimates of monotone parameters. Ann. Math. Statist., 26, 607-616.

R. J. Carroll (1982). Adapting for heteroscedasticity in linear models. Ann. Statist., 10, 1224-1233.

R. J. Carroll (1987). The effect of variance function estimation on prediction-intervals. In: Proc. 4th Purdue Symp. Statistical Decision Theory and Related Topics (eds. J. O. Berger and S. S. Gupta), Vol. II, Springer Heidelberg.

H. Dette, A. Munk, T. Wagner (1998). Estimating the variance in nonparametric regression - what is a reasonable choice? J. Roy. Statist. Soc., Ser. B, 60, 751-764.

H. Dette, N. Neumeyer, K. F. Pilz (2003). A simple nonparametric estimator of a monotone regression function. Technical report, Department of Mathematics. http://www.ruhr-uni-bochum.de/mathematik3/preprint.htm

J. Fan, I. Gijbels (1995). Data driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaption. J. Roy. Statist. Soc., Ser. B, 57, 371-394.

J. Fan, I. Gijbels (1996). Local polynomial modelling and its applications. Chapman and Hall, London.

J. Fan, W. Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85, 645-660.

T. Gasser, L. Sroka, G. Jennen-Steinmetz (1986). Residual variance and residual pattern in nonlinear regression. Biometrika, 73, 626-633.

P. Hall, R. J. Carroll (1989). Variance estimation in regression: the effect of estimating the mean. J. Roy. Statist. Soc., Ser. B, 51, 3-14.

P. Hall, L. S. Huang (2001). Nonparametric kernel regression subject to monotonicity constraints. Ann. Statist., 29, 624-647.

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

P. Hall, J.S. Marron (1990). On variance estimation in nonparametric regression. Biometrika 77, 415-19.

E. Mammen (1991). Estimating a smooth monotone regression function. Ann. Statist., 19, 724-740.

R. Mukerjee (1988). Monotone nonparametric regression. Ann. Statist., 16, 741-750.

H. G. Müller, U. Stadtmüller (1987). Estimation of heteroscedasticity in regression analysis. Ann. Statist. 15, 610-625.

H. G. Müller, U. Stadtmüller (1993). On variance function estimation with quadratic forms. J. Statist. Plann. Inf. 35, 213-231.

J. Rice (1984). Bandwidth choice for nonparametric regression. Ann. Statist., 12, 1215-1230.

D. Ruppert, M. P. Wand, U. Holst, O. Hössjer (1997). Local polynomial variance-function estimation. Technometrics 39, 262-273.

J. Sacks, D. Ylvisaker (1970). Designs for regression problems for correlated errors. Ann. Math. Statist., 41, 2057-2074.

K. Yu, M.C. Jones (2004). Likelihood-based local linear estimation of the conditional variance function. J. Americ. Statist. Assoc. 99, 139-155.