Variance Estimation in Nonparametric Regression Via the Difference Sequence Method

by

Lawrence D. Brown
University of Pennsylvania

M. Levine
Purdue University

Technical Report #06-07

Department of Statistics
Purdue University
West Lafayette, IN USA

October 2006

◇

Variance estimation in nonparametric regression via the difference sequence method (short title: Variance estimation via difference sequences)

Lawrence D. Brown, Department of Statistics, University of Pennsylvania

M. Levine Department of Statistics, Purdue University *†‡§

Abstract

Consider the standard Gaussian nonparametric regression problem. The observations are (x_i, y_i) where

$$Y_i = g(x_i) + \sqrt{V(x_i)}\varepsilon_i$$

and where ε_i are iid with finite fourth moment $\mu_4 < \infty$. This article presents a class of difference-based kernel estimators for the variance

*AMS 2000 Subject Classification 62G08, 62G20

†Keywords and Phrases: Nonparametric regression, Variance estimation, Asymptotic minimaxity

‡The work of the first author was partially supported by the NSF grant # 0405716

§The work of the second author was partially supported by the 2004 Purdue Research Foundation Summer Faculty Grant

function V. Optimal convergence rates that are uniform over broad functional classes and bandwidths are fully characterized, and asymptotic normality is also established. We also show that for suitable asymptotic formulations our estimators achieve the minimax rate.

1 Introduction

Let us consider the following non-parametric regression problem

$$y_i = g(x_i) + \sqrt{V(x_i)}\epsilon_i,\ i = 1, \ldots, n \tag{1}$$

where $g(x)$ is an unknown mean function, the errors ϵ_i are iid with mean zero, variance 1 and the finite fourth moment $\mu_4 < \infty$ while the design is fixed. We assume that $\max\{x_{i+1} - x_i\} = O(n^{-1})$ for $\forall i = 0, \ldots, n$. Also, the usual convention $x_0 = 0$ and $x_{n+1} = 1$ applies. The problem we are interested in is estimating the variance $V(x)$ when the mean $g(x)$ is unknown. In other words, the mean $g(x)$ plays the role of a nuisance parameter. The problem of variance estimation in nonparametric regression was first seriously considered in the 1980's. The practical importance of this problem has been also amply illustrated. It is needed to construct a confidence band for any mean function estimate (see, e.g. Hart (1997), Chapter 4). It is of interest in confidence interval determination for turbulence modelling (Ruppert et al. 1997), financial time series (Härdle and Tsybakov (1997), Fan and Yao (1998)), covariance structure estimation of the nonstationary longitudinal data (see, for example, Diggle and Verbyla (1998)), estimating correlation structure of the heteroscedastic spatial data (Opsomer et al. 1999), nonparametric regression with lognormal errors as discussed in (Brown et al. 2005) and Shen and Brown (2006), and many other problems.

In what follows we describe in greater detail the history of a particular approach to the problem. von Neumann (1941), von Neumann (1942), and then Rice (1984) considered the special, homoscedastic situation in which $V(x) \equiv \sigma^2$ in the model (1). They proposed relatively simple estimators of

the following form:

$$\hat{V}(x) = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2. \tag{2}$$

The next logical step was made in Gasser et al. (1986) where three neighboring points were used to estimate the variance:

$$\hat{V}(x) = \frac{2}{3(n-2)} \sum_{i=1}^{n-2} \left(\frac{1}{2} y_i - y_{i+1} + \frac{1}{2} y_{i+2} \right)^2. \tag{3}$$

A further general step was made in Hall, Kay, and Titterington (1990). The following definition is needed first.

Definition 1.1. Let us consider a sequence of numbers $\{d_i\}_{i=0}^{r}$ such that

$$\sum_{i=0}^{r} d_i = 0 \tag{4}$$

while

$$\sum_{i=0}^{r} d_i^2 = 1. \tag{5}$$

Such a sequence is called a difference sequence of order r.

For example, when $r = 1$, we have $d_0 = \frac{1}{\sqrt{2}}, d_1 = -d_0$ which defines the first difference $\Delta Y = \frac{Y_i - Y_{i-1}}{\sqrt{2}}$. The estimator of Hall, Kay and Titterington (1990) estimator can be defined as

$$\hat{V}(x) = (n-r)^{-1} \sum_{i=1}^{n-r} \left(\sum_{j=0}^{r} d_j y_{j+i} \right)^2. \tag{6}$$

The conditions (4) and (5) are meant to insure the unbiasedness of the estimator (6) when g is constant and also the identifiability of the sequence $\{d_i\}$.

A different direction was taken in Hall and Carroll (1989) and Hall and Marron (1990) where the variance was estimated as an average of squared residuals from a fit to g; for other work on constant variance estimation, see also Buckley et al. (1988), Buckley and Eagleson (1989), and Carter and Eagleson (1992).

The difference sequence idea introduced by Hall, Kay and Titterington (see Hall, Kay, and Titterington (1990)) can be modified to the case of a nonconstant variance function $V(x)$. As a rule, the average of squared differences of observations has to be localized in one way or another - for example, by using the nearest neighbor average, spline approach or local polynomial regression. The first to try to generalize it in this way were probably Müller and Stadtmüller (1987). It was further developed in Hall, Kay, and Titterington (1990), Müller and Stadtmüller (1993), Seifert et al. (1993), Dette, Munk, and Wagner (1998), and many others. An interesting application of this type of variance function estimator for the purpose of testing the functional form of the given regression model is given in Dette (2002).

Another possible route to estimating the variance function $V(x)$ is to use the local average of the squared residuals from the estimation of $g(x)$. One of the first applications of this principle was in Hall and Carroll (1989). A closely related estimator was also considered earlier in Carroll (1982) and Matloff, Rose, and Tai (1984). This approach has also been considered in Fan and Yao (1998).

Some of the latest work in the area of variance estimation includes attempts to derive methods that are suitable for the case where $X \in \mathcal{R}^d$ for $d > 1$; see, for example, Spokoiny (2002) for generalization of the residual-

based method and Munk, Bissantz, Wagner, and Freitag (2005) for generalization of the difference-based method.

The present research describes a class of nonparametric variance estimators based on difference sequences and local polynomial estimation, and investigates their asymptotic behavior. Section 2 introduces the estimator class and investigates its asymptotic rates of convergence as well as the choice of the optimal bandwidth. Section 3 establishes the asymptotic normality of these estimators. Section 4 investigates the question of asymptotic minimaxity for our estimator class among all possible variance estimators for the nonparametric regression.

2 Variance function estimators

Consider the model (1). We begin with the following formal definition.

Definition 2.1. A pseudoresidual of order r is

$$\Delta_i \equiv \Delta_{r,i} = \sum_{j=0}^{r} d_j y_{j+i-\lfloor r/2 \rfloor} \tag{7}$$

where $\{d_j\}$ is a difference sequence satisfying (4)-(5) and $i = \lfloor \frac{r}{2} \rfloor + 1, \ldots, n + \lfloor \frac{r}{2} \rfloor - r$.

Remark 2.2. The term "pseudoresidual" has been introduced for the first time, probably, in Gasser, Sroka, and Jennen-Steinmetz (1986) where it was used in a more restricted sense.

Remark 2.3. For $r = 1$ there is a unique difference sequence, up to multiplication by -1 and permutation of the coordinates. It is $(\pm\frac{1}{\sqrt{2}}, \mp\frac{1}{\sqrt{2}})$. For

larger r there can be many essentially different sequences. For example, for $r = 2$ the sequences $(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}})$ and $(\frac{1}{\sqrt{14}}, -\frac{3}{\sqrt{14}}, \frac{2}{\sqrt{14}})$ are each difference sequences satisfying (4) and (5).

Let $K(\cdot)$ be a real-valued function satisfying

1. $K(u) \geq 0$ and is not identically zero.
2. $K(u)$ is bounded: $\exists\, M > 0$ such that $K(u) \leq M$ for $\forall u$.
3. $K(u)$ is supported on $[-1, 1]$ and $\int K(u)\, du = 1$.

We use the notation $\sigma_K^2 = \int u^2 K(u)\, du$ and $R_K = \int K^2(u)\, du$. Then, based on $\Delta_{r,i}$, we define a variance estimator $\hat{V}_h(x)$ of order r as the local polynomial regression estimator based on $\Delta_{r,i}^2$:

$$\hat{V}_h(x) = \hat{a}_0 \tag{8}$$

where

$$(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p) = \arg \min_{a_0, a_1, \ldots, a_p} \sum_{i=\lfloor r/2 \rfloor + 1}^{n + \lfloor r/2 \rfloor - r} \Big[\Delta_{r,i}^2 - a_0 - a_1(x - x_i)$$

$$- \quad \ldots - a_p(x - x_i)^p\Big]^2 K\left(\frac{x - x_i}{h}\right).$$

The value h in (8) is called the bandwidth and K is the weight function.

It should be clear that these estimators are unbiased under the assumption of homoscedasticity $V(x) \equiv \sigma^2$ and constant mean $g(x) \equiv \mu$. We begin with the definition of the functional class that will be used in the asymptotic results to follow.

Definition 2.4. Let us define the functional class $\mathcal{C}_\gamma$ as follows. Let $C_1 > 0$, $C_2 > 0$. Let us denote $\gamma' = \gamma - \lfloor \gamma \rfloor$ where $\lfloor \gamma \rfloor$ denotes the greatest integer

less than γ. We say that the function $f(x)$ belongs to the class C_γ if for all $x, y \in (0,1)$

$$|f^{\lfloor\gamma\rfloor}(x) - f^{\lfloor\gamma\rfloor}(y)| \leq C_1|x-y|^{\gamma'} \tag{9}$$

$$|f^{(k)}(x)| \leq C_2 \tag{10}$$

for $k = 0, \ldots, \lfloor\gamma\rfloor - 1$. Note that C_γ depends on the choice of C_1, C_2, but for our convenience we omit this dependence from the notation. There are also similar types of dependence in the definitions that immediately follow.

Definition 2.5. Let $\delta > 0$. We say the function is in class C_γ^+ if it is in C_γ and in addition

$$f(x) \geq \delta \tag{11}$$

These classes of functions are familiar in the literature, as in (Fan 1992; Fan 1993) and are often referred to as Lipschitz balls.

Definition 2.6. We define the pointwise risk of the variance estimator $\hat{V}_h(x)$ (its mean squared error at a point x) as

$$R(V(x), \hat{V}_h(x)) = E\left[\hat{V}_h(x) - V(x)\right]^2.$$

Definition 2.7. Let us define the global mean squared risk of the variance estimator $\hat{V}_h(x)$ as

$$R(V, \hat{V}_h) = E\left(\int_0^1 (\hat{V}_h(x) - V(x))^2\, dx\right). \tag{12}$$

Then, the globally optimal in the minimax sense bandwidth h_{opt} is defined as

$$h_n = argmin\{\sup\{R(V, \hat{V}_h) : V \in C_\gamma,\, g \in C_\beta\} : h > 0\}.$$

Note that h_n depends on n as well as C_1, C_2, β and γ. A similar definition applies in the setting of Definition (2.6).

Remark 2.8. In a special case where $\gamma = 2$ and $\beta = 1$, the finite sample performance of this estimator has been investigated in Levine (2006) together with the possible choice of bandwidth. A version of K-fold crossvalidation has been recommended as the most suitable method. When utilized, it produces a variance estimator that in typical cases is not very sensitive to the choice of the mean function $g(x)$.

Theorem 2.9. *Consider the nonparametric regression problem described by* (1), *with estimator as described in* (8). *Fix* C_1, C_2, $\gamma > 0$ *and* $\beta > \gamma/(4\gamma+2)$ *to define functional classes* $\mathcal{C}_\gamma$ *and* $\mathcal{C}_\beta$ *according to the definition* (2.4). *Assume* $p > \lfloor\gamma\rfloor$. *Then the optimal bandwidth is* $h_n \asymp n^{-1/(2\gamma+1)}$. *Let* $0 < \underline{a} \leq \overline{a} < \infty$. *Then, there are constants* $\underline{B}$ *and* $\overline{B}$ *such that*

$$\begin{aligned} & \underline{B}n^{-2\gamma/(2\gamma+1)} + o\left(n^{-2\gamma/(2\gamma+1)}\right) \\ \leq \;& R(V, \hat{V}) \leq \overline{B}n^{-2\gamma/(2\gamma+1)} + o\left(n^{-2\gamma/(2\gamma+1)}\right) \end{aligned} \tag{13}$$

for all h *satisfying* $\underline{a} \leq n^{1/(2\gamma+1)}h \leq \overline{a}$, *uniformly for* $g \in \mathcal{C}_\beta$, $V \in \mathcal{C}_\gamma$.

Theorem (2.9) refers to properties of the integrated mean square error. Related results also hold for minimax risk at a point. The main results are stated in the following theorem.

Theorem 2.10. *Consider the setting of Theorem* (2.9). *Let* $x_0 \in (0, 1)$. *Assume* $p > \lfloor\gamma\rfloor$. *Then the optimal bandwidth is* $h_n(x) \asymp n^{-1/(2\gamma+1)}$. *Let*

$0 < \underline{a} \leq \overline{a} < \infty$. *Then, there are constants* $\underline{B}$ *and* $\overline{B}$ *such that*

$$\begin{aligned} & \underline{B}n^{-2\gamma/(2\gamma+1)} + o(n^{-2\gamma/(2\gamma+1)}) \leq R(V(x_0), \hat{V}_{h_n}(x_0)) \\ \leq \; & \overline{B}n^{-2\gamma/(2\gamma+1)} + o(n^{-2\gamma/(2\gamma+1)}) \end{aligned} \tag{14}$$

for all $h(x)$ *satisfying* $\underline{a} \leq n^{1/(2\gamma+1)}h \leq \overline{a}$, *uniformly for* $g \in \mathcal{C}_\beta$, $V \in \mathcal{C}_\gamma$.

The proof of these theorems can be found in the Appendix. At this point, the following remarks may be helpful.

Remark 2.11. The result stated in Theorem 2.10 is also valid when the bandwidth h depends on n as long as the inequality (49) in its proof remains valid.

Remark 2.12. If one assumes that $\beta = \gamma/(4\gamma + 2)$ in the definition of the functional class $\mathcal{C}_\beta$, the conclusions of the theorems (2.9) and (2.10) remain valid, but the constants $\underline{B}$ and $\overline{B}$ appearing in them become dependent on β. For more details see Appendix.

Remark 2.13. Müller and Stadtmüller (1993) made a very significant step forward in the nonparametric variance estimation area by considering the general quadratic form based estimator similar to our (8) and deriving convergence rates for its mean squared error. They also were the first to point out an error in the paper by Hall and Carroll (1989) (see Müller and Stadtmüller (1993), pp. 214 and 221). They use a slightly different (more restrictive) definition of the classes $\mathcal{C}_\gamma$ and $\mathcal{C}_\beta$ and only establish rates of convergence and error terms on those rates for fixed functions V and g within the classes $\mathcal{C}_\gamma$ and $\mathcal{C}_\beta$. Our results resemble these but we also establish the

rates of convergence *uniformly* over the functional classes $\mathcal{C}_\beta$ and $\mathcal{C}_\gamma$ and therefore our bounds are of the minimax type.

Remark 2.14. It is important to notice that the asymptotic mean square risks in Theorems (2.9) and (2.10) can be further reduced by the proper choice of the difference sequence $\{d_j\}$. The proof in the Appendix supplemented with material in Hall, Kay and Titterington (1990) shows that the asymptotic variance of our estimators will be affected by the choice of the difference sequence, but the choice of this sequence does not affect the bias in asymptotic calculations. The effect on the asymptotic variance is to multiply it by a constant proportional to

$$C = 2\left(1 + 2\sum_{k=1}^{r}\left(\sum_{j=0}^{r-1-k} d_j d_{j+k}\right)^2\right). \tag{15}$$

For any given value of r there is a difference sequence that minimizes this constant. A computational algorithm for these sequences is given in Hall, Kay and Titterington (1990). The resulting minimal constant as a function of r is

$$C_{min} = \frac{2r+1}{r}. \tag{16}$$

Hence, increasing r from 1 to ∞ achieves at most $\approx 33\%$ reduction in the asymptotic variance, and no change in the asymptotic bias term for the estimator. It follows that the improvement in the asymptotic risk is bounded by this same factor of 33%. Most of this reduction occurs via an increase of r from 1 to a moderate value such as $r = 3$ or 4.

3 Asymptotic Normality

As a next step, we establish that the estimator (8) is asymptotically normal. To do this, an additional insight into the nature of the estimator (8) is useful. Note that, in order to define (8), we define squared pseudoresiduals first and then smooth them locally to produce the kernel variance estimator. It is important to recall here that the local polynomial regression estimator $\hat{V}_h(x)$ can be represented as

$$\hat{V}_h(x) = \sum_{i=\lfloor r/2 \rfloor + 1}^{n+\lfloor r/2 \rfloor - r} K_{n;h,x}(x_i)\Delta_{r,i}^2. \tag{17}$$

where $K_{n;h,x}(x_i) = K_{n,x}\left(\frac{x-x_i}{h}\right)$. Here $K_{n,x}\left(\frac{x-x_i}{h}\right)$ can be thought of as a centered and rescaled local kernel function whose shape depends on the location of design points x_i, the point of estimation x and the number of observations n. Also, it is not usually non-negative the way the original kernel $K(\cdot)$ is. We know that $K_{n,x}\left(\frac{x-x_i}{h}\right)$ satisfies discrete moment conditions:

$$\sum_{i=\lfloor r/2 \rfloor + 1}^{n+\lfloor r/2 \rfloor - r} K_{n,x}\left(\frac{x-x_i}{h}\right) = 1 \tag{18}$$

$$\sum_{i=\lfloor r/2 \rfloor + 1}^{n+\lfloor r/2 \rfloor - r} (x-x_i)^q K_{n,x}\left(\frac{x-x_i}{h}\right) = 0 \tag{19}$$

for any $q = 1, \ldots, p$. We also need the fact that the support of $K^n(\cdot)$ is contained in the one of $K(\cdot)$; in other words, $K^n(\cdot) = 0$ whenever $|x_i - x| > h$. For more details see e.g. Fan and Gijbels (1995). Properties (18)-(19) are also needed to prove Theorems (2.9) and (2.10) (see Appendix). Now, we can state the following result.

Theorem 3.1. *Consider the nonparametric regression problem described by* (1), *with estimator as described in* (8). *We assume that the functions* $g(x)$ *and* $V(x)$ *are continuous for any* $x \in [0,1]$ *and* V *is bounded away from zero. Assume* $\mu_{4+\nu} = E(\varepsilon_i)^{4+\nu} < \infty$ *for a small* $\nu > 0$. *Then, as* $h \to 0$, $n \to \infty$ *and* $nh \to \infty$, *we find that*

$$\sqrt{nh}\left(\hat{V}_h(x) - V(x) - O(h^{2\gamma})\right) \tag{20}$$

is asymptotically normal with mean zero and variance σ^2 *where* $0 < \sigma^2 < \infty$.

Proof. To prove this result, we rely on CLT for partial sums of a generalized linear process

$$X_n = \sum_{i=1}^{n} a_{ni}\xi_i \tag{21}$$

where ξ_i is a mixing sequence. This and several similar results were established in Peligrad and Utev (1997). Recalling the representation (17), we observe that our estimator (8) can be easily represented in the form (21) with $K_{n;h,x}(x_i)$ as a_{ni}. Thus, we only need to verify the conditions of, say, their Theorem 2.2 (c). We will be checking these conditions one by one.

- The first condition is

$$\max_{1\le i\le n} |a_{ni}| \to 0 \tag{22}$$

as $n \to \infty$. In our notation, it can be expressed as

$$\max_{\lfloor r/2\rfloor+1\le i\le n+\lfloor r/2\rfloor-r} |K_{n;h,x}(x_i)| \to 0 \tag{23}$$

as $n \to \infty$, $h \to 0$ and $nh \to \infty$ uniformly for all $x \in [0,1]$. To achieve this, it is enough to notice that

$$K_{n;h,x}(x_i) = O((nh)^{-1}) \tag{24}$$

uniformly for all $x \in [0,1]$.

- The second condition is

$$\sup_n \sum_{i=1}^{n} a_{ni}^2 < \infty. \tag{25}$$

In our notation it is

$$\sup_n \sum_{i=\lfloor r/2 \rfloor + 1}^{n+\lfloor r/2 \rfloor - r} \left(K_{n;h,x}(x_i)\right)^2 < \infty. \tag{26}$$

From (24) and the Cauchy-Schwartz inequality we easily deduce that $\sum_{i=\lfloor r/2 \rfloor + 1}^{n+\lfloor r/2 \rfloor - r} \left(K_{n;h,x}(x_i)\right)^2 = O\left(\frac{1}{nh}\right)$ and therefore (25) is true.

- Now we need to establish uniform integrability of $\xi_i^2 \equiv \Delta_{r,i}^4$. To do this, we use a simple criterion mentioned in Shiryaev (1995) that requires existence of the non-negative, monotonically increasing function $G(t)$, defined for $t \geq 0$, such that

$$\lim_{t \to \infty} \frac{G(t)}{t} = \infty$$

and

$$\sup_i E\left[G(\Delta_{r,i}^4)\right] < \infty.$$

Indeed, choosing $G(t) = t^\nu$ we have $\lim_{t \to \infty} \frac{G(t)}{t} = \infty$ and $\sup_i E\left(\Delta_{r,i}^{4+\nu}\right) < \infty$; therefore, Shiryaev's conditions are true and the sequence $\{\Delta_{r,i}^4\}$ is uniformly integrable.

- The remaining three conditions of Peligrad and Utev (1997) are almost trivial. First, the sequence $\{\Delta_{r,i}^2\}$ is clearly strongly mixing as it is a measurable function of an iid sequence ε_i. Second, $\inf_i \operatorname{var} \Delta_{r,i}^2 > 0$ as

the variance function $V(\cdot)$ is bounded from below. Finally, the last condition requires that $\sum_i i^{2/\rho}\alpha(i) < \infty$ for some $\rho > 0$ where $\alpha(i)$ is a mixing coefficient of the sequence $\Delta_{r,i}^2$. Indeed, since $\Delta_{r,i}^2$ is an r-dependent sequence, the mixing coefficient $\alpha(i)$ is zero for any $n > r$. Thus, it is automatically true that $\sum_i i^{2/\delta}\alpha_i < 0$ for any $\delta > 0$.

4 Asymptotic minimaxity and related issues

Lower bounds on the asymptotic minimax rate for estimating a nonparametric variance in formulations related to that in (1) have occasionally been studied in earlier literature. Two papers seem particularly relevant. Munk and Ruymgaart (2002) study a different, but related problem. Their paper contains a lower bound on the asymptotic minimax risk for their setting. In particular, their setting involves a problem with random design, rather than the fixed design case in (1). Their proof uses van Trees inequality and relies heavily on the fact that their (X_i, Y_i) pairs are independent and identically distributed. While it may well be possible to do so, it is not immediately evident how to modify their argument to apply to the setting (1).

Hall and Carroll (1989) consider a setting similar to ours. Their equation (2.13) claims (in our notation) that there is a constant $K > 0$, possibly depending on C_1, C_2, β such that for any estimator $\tilde{V}$

$$\sup\{R(V(x_0), \tilde{V}(x_0)) : V \in \mathcal{C}_\gamma, g \in \mathcal{C}_\beta\} \geq K \max\{n^{\frac{-2\gamma}{2\gamma+1}}, n^{\frac{-4\beta}{2\beta+1}}\}. \qquad (27)$$

Note that $n^{-2\gamma/(2\gamma+1)} = o(n^{-4\beta/(2\beta+1)})$ for $\beta < \gamma/(2\gamma+2)$. It thus follows from (14) in our Theorem (2.10) that for any $\gamma/(4\gamma+2) < \beta < \gamma/(2\gamma+2)$

and n sufficiently large

$$\sup\{R(V(x_0), \hat{V}_{h_n}(x_0)) : V \in \mathcal{C}_\gamma, g \in \mathcal{C}_\beta\} \lll K \max\{n^{\frac{-2\gamma}{2\gamma+1}}, n^{\frac{-4\beta}{2\beta+1}}\} \quad (28)$$

where h_n is yet again the optimal bandwidth. This contradicts the assertion in Hall and Carroll (1989), and shows that their assertion (2.13) is an error - as is the argument supporting it that follows (C.3) of their article. For a similar commentary see also Müller and Stadtmüller (1993). Because of this contradiction it is necessary to give an independent statement and proof of a lower bound for the minimax risk. That is the goal of this section, where we treat the case in which $\beta \geq \gamma/(4\gamma + 2)$. The minimax lower bound for the case in which $\beta < \gamma/(4\gamma + 2)$ requires different methods which are more sophisticated. That case, as well as some further generalizations have been treated in Wang, Brown, Cai and Levine (2006) as a sequel to the present paper. That paper proves rate-wise sharp lower and upper bounds for the case where $\beta < \gamma/(4\gamma + 2)$.

We have treated both mean squared error at a point (in theorem (2.10)) and integrated mean squared error (in theorem (2.9)). Correspondingly, we provide statements of lower bounds on the minimax rate for each of these cases. Both results are obtained under the assumption of normality of errors ε_i which enables us to use Hellinger distance-based results. See Section 2 for the definition of R, and other quantities that appear in the following statements.

Theorem 4.1. *Consider the nonparametric regression problem described by* (1). *Fix* C_1, C_2, β *and* γ *to define functional classes* $\mathcal{C}_\gamma$, $\mathcal{C}_\beta$ *according to* (2.4). *Also assume that* $\varepsilon_i \sim N(0, 1)$ *and independent. Then there is a*

constant $K > 0$ *such that*

$$\inf\{\sup\{R(V,\tilde{V}) : V \in \mathcal{C}_\gamma^+, g \in \mathcal{C}_\beta\} : \tilde{V}\} \geq Kn^{-2\gamma/(2\gamma+1)} \qquad (29)$$

where the inf *is taken over all possible estimators of the variance function* V.

Our argument relies on the so-called "two-point" argument, introduced and extensively analyzed in Donoho and Liu (1990,1991).

Theorem 4.2. *Consider the nonparametric regression problem described by* (1). *Fix* C_1, C_2, β *and* γ *to define functional classes* $\mathcal{C}_\gamma$, $\mathcal{C}_\beta$ *according to* (2.4). *Also assume that* $\varepsilon_i \sim N(0,1)$ *and independent. Then there is a constant* $K > 0$ *such that*

$$\inf\{\sup\{R(V(x_0)), \tilde{V}(x_0)) : V \in \mathcal{C}_\gamma, g \in \mathcal{C}_\beta\} : \tilde{V}\} \geq Kn^{-2\gamma/(2\gamma+1)} \qquad (30)$$

where the inf *is taken over all possible estimators of the variance function* V.

Proof of Theorem 4.2 It is easier to begin with the proof of theorem (4.2) and then proceed to the proof of theorem (4.1). We will use a two-point modulus-of-continuity argument to establish the lower bound. Such an argument was pioneered by Donoho and Liu (1990, 1991) for a different, though related, problem. See also Hall and Carroll (1989) and Fan (1993).

Define the function

$$h(t) = \begin{cases} 2 - |t|^\gamma & \text{if} \quad 0 \leq |t| \leq 1 \\ (2 - |t|)^\gamma & \text{if} \quad 1 < |t| \leq 2 \\ 0 & \text{if} \quad |t| > 2. \end{cases} \qquad (31)$$

Assume (for convenience only) that $C_1 > 2$. Let d be a constant satisfying $0 < d < C_2$ and let

$$f_{\delta,l}(x) = d + l\delta h\left(\frac{x - x_0}{\delta^{1/\gamma}}\right). \tag{32}$$

Then $f_{\varepsilon,\pm 1} \in \mathcal{C}_\gamma$ for $\delta > 0$ sufficiently small. Let H denote the Hellinger distance between densities: that is, for any two probability densities m_1,m_2 dominated by a measure $\mu(dz)$

$$H^2(m_1, m_2) = \int (\sqrt{m_1(z)} - \sqrt{m_2(z)})^2 \mu(dz). \tag{33}$$

Here are two basic facts about this metric that will be used below. If $Z = \{Z_j : j - 1, \ldots, n\}$ where the Z_j are independent with densities $\{m_{kj} : j = 1, \ldots, n\}$, $k = 1, 2$ and $m_k = \Pi_j m_{kj}$ denotes the product density then

$$H^2(m_1, m_2) \leq \sum_j H^2(m_{1j}, m_{2j}); \tag{34}$$

and if m_i are univariate normal densities with mean 0 and variance σ_i^2, $i = 1, 2$, then

$$H^2(m_1, m_2) \leq 2\left(\frac{\sigma_1^2}{\sigma_2^2} - 1\right)^2 \tag{35}$$

For more details see Brown and Low (1996) and Brown et al. (2002).

It follows that if m_k, $k = 1, 2$, are the joint densities of the observations $\{x_i, Y_i,\ i = 1, \ldots, n\}$ of (1) with $g \equiv 0$ and $f_k = f_{\delta,(-1)^k}$ then

$$\begin{aligned} H^2(m_1, m_2) &\leq \sum_i 2\left(\frac{f_{\delta,-1}(x_i)}{f_{\delta,1}(x_i)} - 1\right)^2 \\ &\leq 8\sum_i \delta^2 h^2\left(\frac{x_i - x_0}{\delta^{1/\gamma}}\right) = O(n\gamma^{(2\gamma+1)/\gamma}). \end{aligned} \tag{36}$$

For this setting the Hellinger modulus-of-continuity, $\omega(\cdot)$, (Donoho and Liu (1991), equation (1.1)) is defined as the inverse function corresponding to the value

$H(m_1, m_2)$. Hence it satisfies

$$\omega^{-1}(\gamma) = O(n^{1/2}\gamma^{(2\gamma+1)/2\gamma}). \tag{37}$$

Equation (30) (for $g \equiv 0$) then follows, as established in Donoho and Liu (1991). Although this completes the proof of theorem (4.2), we provide a sketch of the argument based on (37) since a few details will be needed in the proof of theorem (4.1). See Donoho and Liu (1991) and references cited there for more details.

Let L denote the L_1 distance between m_1 and m_2; that is,

$$L(m_1, m_2) = \int |m_1(z) - m_2(z)| \, \mu_{dz}. \tag{38}$$

Note that

$$L(m_1, m_2) \leq 2H(m_1, m_2). \tag{39}$$

The Bayes risk, $\mathcal{B}$, for the prior giving mass $1/2$ for each of m_1 and m_2 satisfies

$$\begin{aligned} \mathcal{B} &\doteq \inf \Big\{ \frac{1}{2} \Big[\int (\tilde{\gamma}(z) - \gamma(-1))^2 m_1(z) \, \mu(dz) \\ &+ \int (\tilde{\gamma}(z) - \gamma)^2 m_2(z) \, \mu(dz) \Big] : \tilde{\gamma} \Big\} \\ &\geq \gamma^2 \int \min\{m_1(z), m_2(z)\} \, \mu(dz) = \gamma^2 \left(1 - \frac{L_1(m_1, m_2)}{2} \right). \end{aligned} \tag{40}$$

Note that because of (36), for small enough $\varepsilon > 0$ the choice

$$\gamma \sim \varepsilon n^{-\gamma/(2\gamma+1)} \tag{41}$$

yields

$$\mathcal{B} \geq \frac{\varepsilon^2 n^{-2\gamma/(2\gamma+1)}}{2} \left(1 - K_0 n (\varepsilon n^{-\gamma/(2\gamma+1)})^{(2\gamma+1)/\gamma} \right) \tag{42}$$

for some constant $K_0 \leq \infty$. Hence, choosing small enough $\varepsilon > 0$ yields

$$\inf\{\sup\{R(V(x_0), \tilde{V}(x_0)) : V \in \mathcal{C}_\gamma, g \in \mathcal{C}_\beta\} : \tilde{V}\}$$

$$\inf\{\sup\{R(V(x_0), \tilde{V}(x_0)) : V \in \mathcal{C}_\gamma, g \equiv 0\} : \tilde{V}\} \geq \mathcal{B} \geq Kn^{-2\gamma/(2\gamma+1)} \quad (43)$$

Proof of Theorem 4.1 Let $1 \geq \varepsilon > 0$ be small enough so that (43) holds. Let

$$n^* = \frac{n^{1/(2\gamma+1)}}{4\sqrt{\varepsilon}} - 1$$

and

$$x_j^* = 4n^{-1/(2\gamma+1)} j\sqrt{\varepsilon},\ j = 1, \ldots, n$$

Let

$$L = \{l_j : j = 1, \ldots, n^*\} \text{ with each } l_j = 1 \text{ or } 2$$

Let

$$f_L = d + \sum_j \left(l_j \delta_n h \left(\frac{x - x_j^*}{\delta_n^{1/\gamma}} \right) \right)$$

with $\delta_n = \varepsilon n^{-\gamma/(2\gamma+1)}$. Note that the components $l_j \delta_n h \left(\frac{x - x_j^*}{\delta_n^{1/\gamma}} \right)$ that appear in the definition of f_L have disjoint support. Note also that

$$\int \left(\left(l_j \delta_n h \left(\frac{x - x_j^*}{\delta_n^{1/\gamma}} \right) \right)^2 \right) dx = K_1 \varepsilon^{(2\gamma+1)/\gamma} n^{-1}$$

where $K_1 = \int h^2\, dt$.

Now, assume that the values of l_j, $j = 1, \ldots, n$ are independent random variables each taking the value $l_j = 1$ or 2 with probability 1/2. Let $\mathcal{B}^*$ denote the Bayes risk and let $\mathcal{B}_j$ denote the Bayes risk in each component problem. Because of the disjoint support property noted above $\mathcal{B}^* = \sum_{j=1}^n \mathcal{B}_j$. Hence

$$\mathcal{B}^* = \sum_{j=1}^{n^*} \mathcal{B}_j \geq n^* (K_1 \varepsilon^{(2\gamma+1)/\gamma} n^{-1}) K_2$$

where

$$K_2 = \frac{1}{2}\left(1 - K_0 n(\varepsilon n^{-\gamma/(2\gamma+1)})^{(2\gamma+1)/\gamma}\right)$$

by (42) and the reasoning leading to it. It follows that $\varepsilon > 0$ can be chosen sufficiently small so that $\mathcal{B}^* \geq Kn^{-2\gamma/(2\gamma+1)}$ for some $K > 0$. This proves the assertion of the theorem.

5 Acknowledgement

We wish to thank T. Cai and L. Wang for pointing out the significance of the article of Hall and Carroll (1989) and its relation to our (14).

Appendix

Proof.

Fix r and functional classes $\mathcal{C}_\gamma$ and $\mathcal{C}_\beta$. For the sake of brevity, we write $\Delta_i \equiv \Delta_{r,i}$. Our main tools in this proof are the representation (17) of the variance estimator $\hat{V}_h(x)$ and the properties (18)-(19). We also use the following property:

$$\sum_{i=\lfloor r/2 \rfloor - 1}^{n+\lfloor r/2 \rfloor - r} (K_{n;h,x}(x_i))^2 = O\left(\frac{1}{nh}\right). \tag{44}$$

(44) follows from (24) and the Cauchy-Schwarz inequality, as was already noted when discussing condition (25) in the proof of theorem (3.1). Here and later, O is uniform for all $V \in \mathcal{C}_\gamma$, $g \in \mathcal{C}_\beta$ and $\{h\} = \{h_n\}$.

Now,

$$E(\Delta_i^2) = Var\,(\Delta_i) + (E(\Delta_i))^2 \tag{45}$$

where

$$Var\left(\Delta_i\right) = \sum d_j^2 Var(y_{j+i-\lfloor r/2 \rfloor}) = V(x_i) + O\left(\left(\frac{1}{n}\right)^{\gamma}\right) \tag{46}$$

and

$$E(\Delta_i) = O\left(\left(\frac{1}{n}\right)^{\beta}\right) \tag{47}$$

since $\sum d_j = 0$, $\sum d_j^2 = 1$ and $x_{i+r-\lfloor r/2 \rfloor} - x_{i-\lfloor r/2 \rfloor} = O\left(\frac{1}{n}\right)$. This provides an asymptotic bound on the bias as

$$\begin{aligned} & Bias\, \hat{V}_h(x) \equiv E\left(\hat{V}_h(x) - V(x)\right) \qquad (48) \\ = & \sum_{i=\lfloor r/2 \rfloor + 1}^{n+\lfloor r/2 \rfloor - r} \left(V(x_i) - V(x)\right) K_{n;h,x}(x_i) \\ + & \; O(n^{-\gamma}) + O(n^{-\beta}) = O(h^{\gamma}) + O(n^{-\gamma}) + O(n^{-\beta}). \end{aligned}$$

The last step in (48) uses the fact that $V \in \mathcal{C}_{\gamma}$, the standard technique for bounding the summation based on properties (18)-(19) and the Taylor expansion of the function V up to the order $\lfloor \gamma \rfloor$. It is a very minor variation of the technique employed in Wang, Brown, Cai and Levine (2006) (see pp. 10-11).

Next, we need to use the fact that Δ_i and Δ_j are independent if $|i-j| \geq$

$r+1$. Hence,

$$
\begin{aligned}
Var\,\hat{V}_h(x) &= Var\left(\sum_{i=\lfloor r/2\rfloor+1}^{n+\lfloor r/2\rfloor-r} K_{n;h,x}(x_i)\Delta_i^2\right) \\
&= \sum_{i=\lfloor r/2\rfloor+1}^{n+\lfloor r/2\rfloor-r}\ \sum_{j=\lfloor r/2\rfloor+1}^{n+\lfloor r/2\rfloor-r} K_{n;h,x}(x_i)K_{n;h,x}(x_j)Cov\left(\Delta_i^2,\Delta_j^2\right) \\
&= \sum_{i=\lfloor r/2\rfloor+1}^{n+\lfloor r/2\rfloor-r}\ \sum_{j=i-r}^{i+r} K_{n;h,x}(x_i)K_{n;h,x}(x_j)Cov\left(\Delta_i^2,\Delta_j^2\right) \\
&\leq \sum_{i=\lfloor r/2\rfloor+1}^{n+\lfloor r/2\rfloor-r}\ \sum_{j=i-r}^{i+r} 4^{-1}\left((K_{n;h,x}(x_i))^2+(K_{n;h,x}(x_j))^2\right)\left(Var\,\Delta_i^2+Var\,\Delta_j^2\right)
\end{aligned}
$$

It is easy to see that

$$
\begin{aligned}
\Delta_i^2 &= \left(\sum_{j=0}^{r} d_j y_{j+i-\lfloor r/2\rfloor}\right)^2 \\
&= \left(\sum_{j=0}^{r} d_j\sqrt{V(x_{j+i-\lfloor r/2\rfloor})}\varepsilon_{i+j-\lfloor r/2\rfloor}+O(n^{-\beta})\right)^2,
\end{aligned}
$$

and this means, in turn, that

$$
\begin{aligned}
Var\,\Delta_i^2 &\leq C_2^2 Var\left(\sum_{j=0}^{r} d_j\varepsilon_{i+j-\lfloor r/2\rfloor}+O(n^{-\beta})\right)^2 \\
&\leq C_2^2(r+1)\mu_4+O(n^{-2\beta})+O(n^{-4\beta})=O(1).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
Var\,\hat{V}_h(x) &\leq O(1)\sum_{i=\lfloor r/2\rfloor+1}^{n+\lfloor r/2\rfloor-r}\ \sum_{j=i-r}^{i+r}\left((K_{n;h,x}(x_i))^2+(K_{n;h,x}(x_j))^2\right) \\
&= O\left(\frac{1}{nh}\right). \qquad (49)
\end{aligned}
$$

Combining the bounds in (48) and (49) yields the assertion of the theorem since $2\beta > \gamma/(2\gamma+1)$.

References

Brown, L. D., T. Cai, M. Low, and C.-H. Zhang (2002). Asymptotic equivalence theory for nonparametric regression with random design. *The Annals of Statistics 30*, 688–707.

Brown, L. D., N. Gans, N. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association 100*, 36–50.

Brown, L. D. and M. Low (1996). A constrained risk inequality with applications to nonparametric function estimation. *The Annals of Statistics 24*, 2524–2535.

Buckley, M. J. et al. (1988). The estimation of the residual variance in non-parametric regression. *Biometrika 75*, 189–199.

Buckley, M. J. and G. Eagleson (1989). A graphical method for estimating the residual variance in non-parametric regression. *Biometrika 76*, 203–210.

Carroll, R. and D. Ruppert (1988). *Transformation and Weighting in Regression.* Chapman& Hall.

Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *The Annals of Statistics 10*, 1224–1233.

Carroll, R. J. (1987). The effect of variance function estimation on prediction and calibration: an example. In J. Berger and S. Gupta (Eds.), *Statistical Decision Theory and Related Topics IV.* Heidelberg:

Springer.

Carter, C. K. and G. Eagleson (1992). A comparison of variance estimators in nonparametric regression. *Journal of the Royal Statistical Society B 54*, 773–780.

Dette, H. (2002). A consistent test for heteroscedasticity in nonparametric regression based on the kernel method. *Journal of Statistical Planning and Inference 103 (1-2)*, 311–329.

Dette, H., A. Munk, and T. Wagner (1998). Estimating the variance in nonparametric regression-what is a reasonable choice? *Journal of the Royal Statistical Society B 60*, 751–764.

Diggle, P. J. and A. Verbyla (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics 54*, 401–415.

Donoho, D. and R. Liu (1990). Minimax risk over hyperrectangles, and implications. *The Annals of Statistics 18*, 1416–1437.

Donoho, D. and R. Liu (1991). Geometrizing the rates of convergence, ii. *The Annals of Statistics 19*, 633–667.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association 87*, 998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics 21*, 196–216.

Fan, J. and I. Gijbels (1995). *Local polynomial modelling and its applications*. Chapman& Hall.

Fan, J. and Q. Yao (1998). Efficient estimation of conditional variance

functions in stochastic regression. *Biometrika 85*, 645–660.

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

Graybill, F. (1983). *Matrices with applications in Statistics.* Duxbury Press.

Hall, P. and R. Carroll (1989). Variance function estimation in regression: the effect of estimating the mean. *Journal of the Royal Statistical Society B 51*, 3–14.

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

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

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press.

Härdle, W. and A. Tsybakov (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics 81*, 223–242.

Hart, J. (1997). *Nonparametric Smoothing and Lack of Fit Tests.* Springer: New York.

Levine, M. (2003). *Variance estimation for nonparametric regression and its applications.* Ph. D. thesis, University of Pennsylvania.

Levine, M. (2006). Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: A possible approach. *Computational Statistics and Data Analysis 50*, 3405–3431.

Matloff, N., R. Rose, and R. Tai (1984). A comparison of two methods for estimating optimal weights in regression analysis. *Journal of Statistical Computation and Simulation 19*, 265–274.

Müller, H.-G. and U. Stadtmüller (1987). Estimation of heteroskedasticity in regression analysis. *The Annals of Statistics 15*, 610–625.

Müller, H.-G. and U. Stadtmüller (1993). On variance function estimation with quadratic forms. *Journal of Statistical Planning Inference 35*, 213–231.

Munk, A., N. Bissantz, T. Wagner, and G. Freitag (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *Journal of the Royal Statistical Society B 2005*, 19–41.

Munk, A. and F. Ruymgaart (2002). Minimax rates for estimating the variance and its derivatives in non-parametric regression. *Australian and New Zealand Journal of Statistics 44(4)*, 479–488.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications 9*, 141–142.

Opsomer, J., D. Ruppert, M. Wand, U. Holst, and O. Hössjer (1999). Kriging with nonparametric variance function estimation. *Biometrics 55*, 704–710.

Peligrad, M. and S. Utev (1997). Central limit theorem for linear processes. *The Annals of Probability 25*, 443–456.

Rice, J. (1984). Bandwidth choice for nonparametric kernel regression. *Annals of Statistics 12*, 1215–1230.

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

Sacks, J. and D. Ylvisaker (1981). Asymptotically optimum kernels for density estimation at a point. *The Annals of Statistics 9*, 334–346.

Seifert, B. et al. (1993). Nonparametric estimation of the residual variance revisited. *Biometrika 80*, 373–383.

Shen, H. and L. Brown (2006). Nonparametric modelling of time-varying customer service times at a ban call-center. *Applied Stochastic Models in Business and Industry*.

Shiryaev, A. N. (1995). *Probability*. Springer.

Spokoiny, V. (2002). Variance estimation for high-dimensional regression models. *Journal of Multivariate Analysis 82*, 111–133.

von Neumann, J. (1941). Distribution of the ratio of the mean squared successive difference to the variance. *The Annals of Mathematical Statistics 12*, 367–395.

von Neumann, J. (1942). A further remark concerning the distribution of the ratio of the mean squared successive difference to the variance. *The Annals of the Mathematical Statistics 13*, 86–88.

Wang, L., L. Brown, T. Cai, and M. Levine (2006). Effect of

mean on variance function estimation in nonparametric regression. Technical report, University of Pennsylvania. Available at www-stat.wharton.upenn.edu/~tcai.