On difference-based variance estimation in nonparametric regression when the covariate is high dimensional

Axel Munk, Nicolai Bissantz, Thorsten Wagner, Gudrun Freitag

Institut für Mathematische Stochastik, Georg-August-Universit ät Göttingen, Maschmühlenweg 8-10, D-37073 Göttingen, Germany

Summary. We consider the problem of estimating the noise variance in homoscedastic nonparametric regression models. For the case of low dimensional covariates $t \in \mathbb{R}^d$, $d = 1, 2$, difference-based estimators have been investigated in a series of papers. For a given length of such an estimator, difference schemes which minimize the asymptotic MSE can be computed for $d = 1$ and $d = 2$, respectively. However, from numerical studies it is known that for finite sample sizes the performance of these estimators may be deficient due to a large finite sample bias. In this paper, we provide theoretical support for these findings. In particular, we show that with increasing dimension d this becomes more drastic. If $d \geq 4$, these estimators even fail to be consistent. A different class of estimators is discussed which allow a better control of the bias and remain consistent when $d \geq 4$. These estimators are compared numerically with kernel type estimators (which are asymptotically efficient), and some guidance is given as to when their use becomes necessary.

1. Introduction

Recently, estimation of the error variance $\sigma^2 = E[\varepsilon_i^2]$ in a nonparametric regression model

$$Y_i = g(t_i) + \varepsilon_i \qquad t_i \in \mathbb{R}^d\,,\ i = 1, \ldots, N, \tag{1}$$

has received much interest. Knowledge of the variance is required for the purpose of signal estimation itself, for instance for the computation of confidence bands or the optimal choice of the bandwidth and other smoothing parameters. In addition, the variance or transforms of it are of direct interest in technical applications, image restoration, the analysis of financial time series and so on. For further applications we refer to Carroll and Ruppert (1988); Kay (1988); Härdle and Tsybakov (1997). In particular, the case of a one-dimensional predictor, $t \in \mathbb{R}^1$, has been treated extensively in the literature and various estimators have been suggested. The most popular of these estimators dates back to von Neumann (1941), a simple average of squared successive differences of the observations,

$$\hat{\sigma}^2 = \frac{1}{2(N-1)} \sum_{i=2}^{N} (Y_i - Y_{i-1})^2. \tag{2}$$

This estimator also has been used by Rice (1984) and was modified in various ways to so-called difference-based estimators (Gasser et al. (1986); Kay (1988); Hall et al. (1990, 1991); Thompson et al. (1991); among others). Difference estimators are only applicable when homogeneous noise is present, i.e. the error variance does not depend on the regressor t. For regression models with inhomogeneous variance, kernel-based estimators were suggested by

Müller and Stadtmüller (1987); Hall and Carroll (1989); Hall and Marron (1990); Neumann (1994), and more recently, local polynomial estimators were investigated independently by Ruppert et al. (1997); Härdle and Tsybakov (1997); Fan and Yao (1998).

Although variance estimation for $d = 1$ is extensively treated in the literature, the case $d = 2$ has been considered by relatively few authors, and to our knowledge $d \geq 3$ has never been explicitly treated. This might be founded in the tempting conjecture that, in principle, results for $d = 1, 2$ can be transferred in a straightforward manner to the higher dimensional case. We will see, however, this is not the case for difference-based estimators.

Difference-based estimators are very popular in practice, because they are easy to perform, in particular when the dimension of the covariate is larger than 1. In contrast, smoothing methods are computationally much more involved, and data driven selection of smoothing parameters is a difficult practical problem (Ruppert et al. (1997)). In particular, when the covariate is high dimensional additional difficulties occur due to the well known 'curse of dimensionality'. Nevertheless, as pointed out by Hall et al. (1990), difference estimators suffer from the fact that their asymptotic efficiency is less compared to smoothing methods, for instance kernel estimators, which achieve asymptotic minimax bounds (Hall and Marron (1990)). Therefore, within the class of difference estimators it is a reasonable goal to select a particular difference estimator which is obtained by minimizing the mean squared error (MSE). It turns out that this can be done under a constraint on the maximum number of observations (the *length of the difference scheme*) taken into account to compute a local residual required for such an estimator. This idea was exploited by Hall et al. (1990) for the case of a one-dimensional covariate and extended in Hall et al. (1991) to the case $d = 2$. In both cases it turns out that the bias contribution for the class of difference-based estimators with prescribed length is (asymptotically) negligible and hence it remains to minimize the variance as that part of the MSE which is asymptotically dominating. The resulting estimators will be called *optimal difference estimators*.

In the following we will show, however, that for general $d \in \mathbb{N}$ an analogous result does not hold anymore. More precisely, we show that the bias of the variance minimizing difference estimator is of order $O(N^{-2/d})$. In contrast, the variance is of order $O(N^{-1})$, and hence for $d \geq 4$ the bias is not negligible for the asymptotic MSE. More drastically, if $d \geq 4$ this implies that these estimators are no longer $\sqrt{N}$-consistent. Note that this implies that the generalization of von Neumann's (1941) estimator (2) to regressors having dimension $d \geq 4$ leads to $\sqrt{N}$-inconsistent estimators. As a consequence the corresponding central limit theorems fail to hold.

It is well known (Hall et al. (1991); Thompson et al. (1991)) that already for the case $d = 1, 2$ optimal difference-based estimators may provide a rather large finite sample bias - although asymptotically not relevant - particularly for spiky or rapidly vaying signals g. Thompson et al. (1991) obtained improved performance using edge detecting algorithms which remove observations at points where rapid changes of the image occur. This will be explained in Section 3 by a second order expansion of the bias. Further, this shows that already for $d = 1, 2, 3$ difference-based estimators minimizing the *asymptotic* MSE have to be applied with greater caution, the more the dimension of the covariate increases.

To overcome these drawbacks of optimal difference schemes, in Section 3 a particular class of difference-based estimators with a polynomial weighted difference scheme is suggested. These estimators are characterized by estimating σ^2 unbiasedly for any d-dimensional polynomial g up to a specific degree. This extends ideas of Kay (1988); Thompson et al. (1991); Seifert et al. (1993); Dette et al. (1998). It is shown that the estimators have sufficiently small bias, even for large dimensions $d \geq 4$. In particular, these

estimators remain consistent, although they are asymptotically not as efficient as kernel estimators, say. Our findings are related to a minimax result by Spokoiny (2002) who showed for two-times differentiable regression functions that only for $d \leq 8$ the $\sqrt{N}$-rate is achievable, otherwise the optimal rate is $N^{-4/d}$. In fact, a modification of his proof yields for one-time differentiable function g the $\sqrt{N}$-rate for $d \leq 4$ as minimax rate and $N^{-2/d}$ if $d > 4$. Hence, our estimates are rate optimal for any $d \in \mathbb{N}$. In order to investigate this in more detail, in Section 4 we extend the kernel estimator of Hall and Marron (1990) to the case $d \geq 2$ and compare it to local polynomial and difference-based estimators, respectively, in a Monte-Carlo study. C++ code can be obtained from the authors on request. Our results can be summarized as follows. In practically all cases a kernel estimator with cross validated bandwidth outperforms any of the (asymptotically) optimal difference estimators. In addition to that, it turns out that the additional improvement by a local linear estimator is negligible. For fluctuating signals, kernel based estimators are outperformed by polynomial weighted estimators in general. This effect becomes more significant for increasing d. In addition, the computational effort of kernel based estimators increases drastically as d increases. In contrast, even for large d polynomial weighted estimators are easy to perform and computationally feasible. A good compromise between control of bias and a small variance is achieved for polynomial weighted estimators with length $r \leq 4$. Only for very smooth signals a significant improvement is obtained by optimal weighted estimators or kernel based estimators. In summary, difference estimators with polynomial weighting schemes of length $r \leq 4$ are a valid alternative to the more efficient but computationally intensive kernel estimators. They should always be used when it cannot be excluded a priori that the signal g is spiky or fluctuating.

In order to keep the paper more readable we have deferred all proofs to an Appendix. We will start in the next section with a brief summary of results on the MSE of difference-based estimators which are available for the case $d = 1$. This will be helpful for a better understanding of the case $d \geq 2$.

2. Difference estimators for $d = 1$

Assume throughout this section that we observe independent data from model (1), where $d = 1$ and $E[\varepsilon_i] = 0$, $E[\varepsilon_i^2] = \sigma^2$, $\gamma_4 := \sigma^{-4} E[\varepsilon_i^4] < \infty$. Let $Y = (Y_1, \ldots, Y_N)'$, and let $tr\, D$ denote the trace of a matrix D. Throughout this paper, for triangular schemes of design points $(t_{1,N}, \ldots, t_{N,N})$ we will simply write $(t_1, \ldots, t_N)$.

Definition 1. *A difference (or weighting) scheme of order $r \in \mathbb{N}$ is a vector $d = (d_k)_{k=0,\ldots,r} \in \mathbb{R}^{r+1}$ such that*

$$\sum_{k=0}^{r} d_k = 0\,, \quad \sum_{k=0}^{r} d_k^2 = 1\,.$$

A difference estimator of order (or length) $r \in \mathbb{N}$ is a random quadratic form

$$\hat{\sigma}_D^2 = \frac{Y' D\, Y}{tr D}\,, \tag{3}$$

where $D = \tilde{D}'\tilde{D}$ *and*

$$\tilde{D} = \begin{pmatrix} d_0 & \dots & d_r & 0 & \dots & 0 \\ & \ddots & & \ddots & & \\ & & \ddots & & \ddots & \\ 0 & \dots & 0 & d_0 & \dots & d_r \end{pmatrix} \in \mathbb{R}^{(N-r)\times N} .$$

THEOREM 1. *(Hall et al. (1990)). Assume that the design points* $(t_i)_{i=1,\dots,N}$ *are located in the unit interval* $[0,1]$ *and fulfill the condition*

$$\int_0^{t_i} f(t)\,dt = i/N, \qquad i = 1,\dots,N, \tag{4}$$

for any $N \in \mathbb{N}$*, where* f *is a density on* $[0,1]$ *which is bounded away from zero. Assume further that* $g, f \in Lip_\gamma[0,1]^1$*,* $\gamma > 1/4$*, where*

$$Lip_\gamma[0,1]^d := \left\{ f : [0,1]^d \to \mathbb{R} \,:\, \exists\, c \in \mathbb{R} \text{ s.t. } |f(x) - f(y)| \le c\|x-y\|^\gamma ,\ x, y \in [0,1]^d \right\} .$$

Then, the asymptotic MSE of $\hat{\sigma}^2_D$ *is minimized among the class of difference-based estimators of order* r*, if and only if (iff)*

$$\sum_{i=\max(0,-k)}^{\min(r,r-k)} d_i\, d_{i+k} = -\frac{1}{2r} \tag{5}$$

for $1 \le |k| \le r$*. The corresponding weights* $(d_0^*, \dots, d_r^*)$ *are (up to the initial sign and reversal order) unique, and the* MSE *of the corresponding difference estimator* $\hat{\sigma}^2_{D^*}$ *has the following first order expansion*

$$MSE[\hat{\sigma}^2_{D^*}] = \frac{\sigma^4}{N}\left(\gamma_4 - 1 + \frac{1}{r}\right) + o(N^{-1}) . \tag{6}$$

A proof of Theorem 1 can be found in Hall et al. (1990). We mention that a simpler proof can be obtained when instead of the minimization problem in terms of the scheme $(d_k)_{k=0,\dots,r}$ the matrix D in (3) is minimized in an appropriate class (Munk (2002)). Note that from Theorem 1 it follows that for increasing r the asymptotic MSE of $\hat{\sigma}^2_{D^*}$ decreases. However, as stated in the introduction this can be in contrast to the finite sample MSE of $\hat{\sigma}^2_{D^*}$. This depends essentially on measures of curvature of g and the sample size N. Let $C^{(m)}[0,1]$ be the space of m times continuously differentiable functions on $[0,1]$.

THEOREM 2. *(Dette et al. (1998)). Let* $\gamma_3 := \sigma^{-3}E[\varepsilon_i^3] = 0$ *and assume that the design is equidistantly spaced, i.e.* $t_i = i/N$*,* $i = 1,\dots,N$*. Then we have for* $g \in C^{(2)}[0,1]$ *with* $\|g\|_2^2 = \int_0^1 g^2(t)\,dt$*,*

$$MSE[\hat{\sigma}^2_{D^*}] = \frac{\sigma^4}{N}\left(\gamma_4 - 1 + \frac{1}{r}\right) + \frac{(2r+1)^2\,(r+1)^2}{144N^4}\left(\|g'\|_2^4 + \frac{4\sigma^2}{N}\|g''\|_2^2\right) + o(N^{-5}) .$$

From this result it becomes apparent that for large values of $\|g'\|_2$ and $\|g''\|_2$, respectively, the finite sample MSE becomes large. Moreover, the MSE increases as r increases. Recall that this is in contrast to the first order expansion (6), which suggests to choose r

as large as possible. For some illustrating examples and numerical investigations we refer to Thompson et al. (1991) or Dette et al. (1998). A class of difference-based estimators which allows the reduction of the bias to any order $O(N^{-m})$, $m \in \mathbb{N}$, was introduced by Kay (1988); Thompson et al. (1991); Seifert et al. (1993). It is given by the polynomial difference scheme

$$d_k^m = \frac{\binom{m}{k}(-1)^k}{\binom{2m}{m}^{1/2}}, \qquad k = 0, \ldots, m. \tag{7}$$

If we denote the corresponding estimator by $\hat{\sigma}^2_{D^m}$ it can be shown that for $g \in C^{(2m)}[0,1]$,

$$MSE[\hat{\sigma}^2_{D^m}] \approx \frac{\sigma^4}{N}\left(\gamma_4 - 1 + 2\binom{4m}{2m}\binom{2m}{m}^{-2}\right) + \binom{2m}{m}^{-2}\left(\frac{\|g^{(m)}\|_2^4}{N^{4m}} + \frac{4\sigma^2\,\|g^{(2m)}\|_2^2}{N^{4m+1}}\right).$$

In particular, the bias is of order $O(N^{-2m})$. Note that the polynomial difference scheme (7) provides an unbiased estimator of the variance whenever the signal is a polynomial of order $m-1$ (Thompson et al. (1991)).

Remark 1. We indicate briefly how the above discussion can be transferred to non-equidistant or random designs, respectively. For more details we refer to Wagner (1999). To reduce the bias up to order $O(N^{-2m})$ in any non-equidistant design model there are in general two options. One is that the matrix D is constructed according to the method of divided differences, analogously to the Newton interpolation formula (see Stoer (1979)). In this case, the resulting difference scheme additionally depends on the design points t_i, and it yields an unbiased estimator for all polynomials g of a certain degree, exactly as in the equidistant case (see Seifert et al. (1993)).

The other possibility is to pose the restriction $f \in C^{(r-1)}[0,1]$ on the density f in (4) generating the design points. In this case, it can be shown that the same difference scheme as in the equidistant design model can be used in order to achieve asymptotically the same order of bias. Note, however, that the resulting estimator is no longer unbiased for the class of regression polynomials.

To reduce the bias in a random design model up to order $O(N^{-2m})$, difference schemes not depending on the design points are no longer valid. To see this, consider the simple difference estimator for $r = 2$ and $(d_0, d_1, d_2) = (1, -2, 1)/\sqrt{6}$. Let $g(x) = x$ and assume independent, identically distributed (i.i.d.) design points $X_i \sim U(0,1)$ independent of the error ϵ_i. In this case a simple calculation shows that the bias of the variance estimator $\hat{\sigma}^2_D$ equals $E[(X_{(i+2)} - X_{(i+1)}) - (X_{(i+1)} - X_{(i)})]^2/\sqrt{6} = 2/(\sqrt{6}(N+1)(N+2))$ and hence is of order $O(N^{-2})$, instead of order $O(N^{-4})$ as for the fixed design. However, again a valid possibility for a bias reduction is to constitute the matrix D as a function of the design points with the above mentioned method of divided differences.

3. Higher dimensions

Now we turn to the investigation of difference-based estimators in higher dimensions. In particular, the case $d = 2$ occurs in various applications (see Bissantz and Munk (2002) for an application in astrophysics) and is of particular interest in imaging, because here model (1) is a standard model for digital image processing where a noisy version of an image g has to be recovered from the data. Here the knowledge of the variance is important for the choice of smoothing parameters, and as a global measure of quality of the resulting image. Early

references are Lee (1981); Kay (1988); further references can be found in Thompson et al. (1991) or Hall et al. (1991). Herrmann et al. (1995) considered a two-dimensional difference-based estimator taking into account the edges of all possible Delaunay triangulations (Ripley (1981)).

Consider model (1) where $t_i = (t_{i_1}, \dots, t_{i_d})' \in \mathbb{R}^d$. We assume that observations are drawn from a d-dimensional grid so that $i_k = 1, \dots, n_k$, $k = 1, \dots, d$. Throughout the following, let

$$n := \min_{k=1,\dots,d} n_k \ , \quad N := \prod_{k=1}^{d} n_k \, , \tag{8}$$

and let

$$\frac{i_k}{n_k} = \int_0^{t_{i_k}} f_k(s)\, ds, \qquad k = 1, \dots, d, \tag{9}$$

where f_k, $k = 1, \dots, d$, are design densities bounded away from zero. The errors are assumed to fulfill

$$E[\varepsilon_i] = 0 \ , \quad E[\varepsilon_i^2] = \sigma^2, \quad \text{and } \gamma_p := \sigma^{-p} E[\varepsilon_i^p] < \infty, \, p = 3, 4 \, . \tag{10}$$

Generalizing ideas of Kay (1988); Thompson et al. (1991) and Hall et al. (1991), we introduce a class of difference-based estimators for arbitrary dimension $d \in \mathbb{N}$ as follows.

Definition 2. *A generalized difference scheme estimator for the variance σ^2 in the regression model (1) is defined as*

$$\hat\sigma^2 = \sum_{l=1}^{L} \mu_l \, n_{R_l}^{-1} \sum_{i \in R_l} \left(\sum_{j \in J_l} d_j^{(l)} Y_{i+j} \right)^2 =: \sum_{l=1}^{L} \mu_l \, \hat\sigma_l^2, \tag{11}$$

where $\hat\sigma_l^2$ is defined by the last identity. Here the sum of the weights μ_l equals 1, and the so-called generalized difference scheme $(d_j^{(l)})_{j \in J_l}$, $l = 1, \dots, L$, satisfies $\sum_{j \in J_l} d_j^{(l)} = 0$ and $\sum_{j \in J_l} \left(d_j^{(l)} \right)^2 = 1$.

Further, $J_l \subset \mathbb{Z}^d$ denotes some index set, the set R_l is given by

$$R_l := \left\{ i \in \times_{k=1}^{d} \{1, \dots, n_k\} \,|\, \forall j \in J_l : \, (i + j) \in \times_{k=1}^{d} \{1, \dots, n_k\} \right\} ,$$

where $A \times B$ denotes the cartesian product of two sets A and B, and n_{R_l} denotes the cardinality $\# R_l$ of R_l.

In order to illustrate this definition and notation we will consider throughout the following various special cases.

Example 1. (The case $d = 2$). *Let $n_1 = n_2 = n$. We consider the following configurations.*
a) $J_l = \{(0,0), (1,1), (2,2), (3,3)\}$, $R_l = \{(i_1, i_2) : i_1, i_2 = 1, \dots, (n-3)\}$, $n_{R_l} = (n-3)^2$.
b) $J_l = \{(0,0), (0,1), (0,2)\}$, $R_l = \{(i_1, i_2) : i_1 = 1, \dots, n \, , \, i_2 = 1, \dots, (n-2)\}$, $n_{R_l} = n(n-2)$.
c) $J_l = \{(0,0), (-1,1), (0,1), (1,1)\}$, $R_l = \{(i_1, i_2) : i_1 = 2, \dots, (n-1) \, , \, i_2 = 1, \dots, (n-1)\}$, $n_{R_l} = (n-1)(n-2)$.

In a) and b) the design points t_{i+k} with $k \in J_l$ constitute a straight line, whereas the design points in c) are 'T-shaped' (Figure 1).

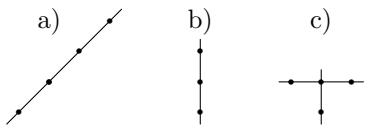


Fig. 1. Configurations as described in Example 1.

REMARK 2. We mention that for reasonable configurations generally it holds that $0 \in J_l$ for all $l = 1, \ldots, L$, and R_l will consist mainly in the set $\times_{k=1}^{d}\{1, \ldots, n_k\}$, except for $O(n^{d-1})$ points located on the edge of the configuration, so that $n_{R_l} = N + O(n^{d-1}) = n^d + O(n^{d-1})$. Furthermore, often $L = d$, although this is not always the best choice as we will see.

Hall et al. (1991) showed that for $d = 2$ the MSE within the class of difference-based estimators is minimized asymptotically by estimators where the local residuals are on a straight line with the same optimal weights as for $d = 1$ given in Theorem 1 (or a disconnected combination of it), i.e. for $s^{(l)} \in \mathbb{Z}^2$ the residuals are supported on

$$J_l = \{\kappa\, s^{(l)} : \kappa = 0, \ldots, r_l\}, \qquad l = 1, \ldots, L\,. \tag{12}$$

A similar result holds for $d = 3$, as we will show in Theorem 3.

EXAMPLE 2. (The case $d = 3$). *Let $n_1 = n_2 = n_3 = n$. The difference scheme according to $J_l = \{(0,0,0), (0,0,1), (0,0,-1), (0,1,0), (0,-1,0), (1,0,0), (-1,0,0)\}$, $R_l = \{(i_1, i_2, i_3) : i_1, i_2, i_3 = 2, \ldots, n-1\}$, and $n_{R_l} = (n-2)^3$ (Figure 2).*

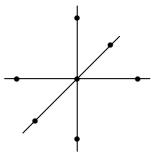


Fig. 2. Configuration as described in Example 2.

THEOREM 3. *Let $d = 1, 2, 3$. If $g \in Lip_\gamma[0,1]^d$, $\gamma > d/4$, then under the conditions (1) and (9) we have for any difference estimator defined in (11),*

$$MSE[\hat{\sigma}^2] = \frac{\sigma^4}{N}\left(\gamma_4 - 1 \,+\, 2\sum_{k\neq 0}\left(\sum_{l=1}^{L}\mu_l \sum_j d_j^{(l)}\, d_{j+k}^{(l)}\right)^2\right) \,+\, o(N^{-1})\,. \tag{13}$$

A variation of the last theorem together with a central limit theorem for random quadratic forms of de Jong (1987) gives also $\sqrt{N}$-consistency of $\hat{\sigma}^2$ as long as $d \leq 3$, that is $N^{1/2}(\hat{\sigma}^2 - \sigma^2)$ is asymptotically centered normal with finite variance as given on the right-hand side in (13).

However, as pointed out for $d = 2$ by Hall et al. (1991) and by Thompson et al. (1991), the bias contribution may become very large for configurations located on a straight line

for finite sample sizes, particularly when the signal fluctuates sharply. Hence, in Hall et al. (1991) it was recommended to use 'compact' configurations, whereas 'long linear' configurations should be avoided. This is highlighted in the next theorem, which provides a second order expansion of the bias of $\hat{\sigma}^2$ for arbitrary $d \in \mathbb{N}$.

THEOREM 4. *Let $d \in \mathbb{N}$ and assume model (1), such that $g \in C^{(2)}[0,1]^d$. Furthermore, assume that (8), (9) and (10) hold. Then for any difference-based estimator $\hat{\sigma}^2$ defined in (11) with residuals located on index sets J_l as in (12), we have that*

$$Bias^2\left[\hat{\sigma}^2\right] = \left(\sum_{l=1}^{L}\left\{n_l^{-2}\mu_l\, C(r_l)\,\| <\nabla g, s^{(l)}> \|_2^2 + o(n_l^{-2})\right\}\right)^2, \tag{14}$$

where $C(r_l) = \left\{\sum_{j=0}^{r_l} j d_j^{(l)}\right\}^2$, $\nabla g = (\frac{\partial g}{\partial t_1}, \ldots, \frac{\partial g}{\partial t_d})'$ denotes the vector of partial derivatives, and the norm $\|\cdot\|_2^2$ is with respect to the design density $\prod_{k=1}^d f_k$. If $n_1 = \ldots = n_d = n$, this simplifies to

$$Bias^2\left[\hat{\sigma}^2\right] = n^{-4}\left(\sum_{l=1}^{L}\mu_l C(r_l)\,\| <\nabla g, s^{(l)}> \|_2^2\right)^2 + o(n^{-4}).$$

Together with the definition of (12), the proof of the last theorem follows similar lines as the calculations in Dette et al. (1998), p.755-756. Note that by means of Remark 1 a similar result can be obtained for the case of random grid points.

EXAMPLE 3. *In the case $d = 2$ with equidistant grid points $(n \times n)$ the configuration (for $L = 1$) $J_1 = \{\kappa\,(1,0) : \kappa = -2, \ldots, 2\}$, as well as the configuration (for $L = 2$) $\tilde{J}_1 = \{\kappa_1\,(1,0) : \kappa_1 = -1, 0, 1\}$, $\tilde{J}_2 = \{\kappa_2\,(0,1) : \kappa_2 = -1, 0, 1\}$, will lead to the same asymptotic MSE, $n^{-2}\sigma^4(\gamma_4 - 1 + 1/4) + O(n^{-4})$, but the finite sample term of order $O(n^{-4})$ for the (long linear) configuration J_1 is 9 times larger (here $r = 4$) than that of the (compact) configuration $\tilde{J}_1$, $\tilde{J}_2$ (here $r = 2$), as one can easily deduce from Theorem 4.*

If d increases, a first order approximation of the bias becomes even worse, as illustrated in the next example.

EXAMPLE 4. *For the optimal difference scheme of Hall et al. (1990) (applied to each of the L directions, separately) we get $C(r_l) = (2r_l + 1)(r_l + 1)/12$. As an example, we consider the generalized von Neumann (1941) estimator (see also Rice (1984)), $r = 1$, $d_0^{(l)} = -d_1^{(l)} = 2^{-1/2}$, $L = d$. Here the index set $J_l = \{0, e_l\}$, where e_l, $l = 1, \ldots, d$, denotes the standard basis in $\mathbb{R}^d$. Assume that $n_1 = \ldots = n_d = n$, $\mu_l = d^{-1}$. Then Theorem 4 yields*

$$Bias^2\left[\hat{\sigma}^2\right] = n^{-4}(2d)^{-2}\left\{\sum_{l=1}^{d}\left\|\frac{\partial g}{\partial t_l}\right\|_2^2\right\}^2 + o(n^{-4}),$$

which is of no better order than $O(n^{-4})$, provided g is not constant. For any difference scheme, however, the variance contribution to the MSE is of the order of the inverse number of grid points, $O(n^{-d})$.

Hence, from this simple example it follows that even asymptotically for $d \geq 4$, in general, the bias will dominate the MSE, in contrast to the case $d \leq 3$. This implies in particular that the von Neumann (1941) estimator is not any more $\sqrt{N}$-consistent if $d \geq 4$. In order to correct for this bias the polynomial weighting scheme in (7) will become necessary.

THEOREM 5. *Assume that $d \in \mathbb{N}$ and assume that $g \in C^{(m)}[0,1]^d$ for $m = [d/4] + 1$. (Here $[x]$ denotes the largest integer smaller than $x \in \mathbb{R}$.) Under the model (1) and assumption (9) for any polynomial difference-based estimator with weighting scheme (7) of order r, such that $r \geq m$, the asymptotic expansion (13) of the MSE holds. Moreover, these estimators are $\sqrt{N}$-consistent.*

We have seen that for increasing dimension d the control of the bias becomes the major task for difference-based estimators, particularly for fluctuating signals. This was highlighted in Theorem 4, where it is shown that the second-order term of the bias dominates the finite sample MSE in these cases. Nevertheless, the first order bias term serves still as a good approximation if the signal fluctuates only slowly, and it might be of interest whether asymptotically the MSE can be minimized as for $d = 1$ (Theorem 1). For $d \geq 2$ this task is more involved because the particular configuration of the difference schemes (see Definition 2) may play a role. In the next theorem we will determine the specific configuration which minimizes asymptotically the MSE, provided L is fixed. Due to our preceding discussion it becomes necessary to treat the cases $1 \leq d \leq 3$ and $d \geq 4$ separately. It turns out, however, that for both cases it is sufficient for the minimization of the MSE to consider sets J_l, $l = 1, \ldots, L$, of non-parallel straight lines as in (12). Note that the next theorem has to be applied with some caution and will yield only a valid approximation for slowly fluctuating signals, because it is based on a *first order* expansion of the MSE, which is *asymptotically* valid, as long as $d \leq 3$.

THEOREM 6. *Let $d = 2$ ($d = 3$) and $g \in Lip_\gamma[0,1]^d$ with $\gamma > 1/2$ ($\gamma > 3/4$). Assume the model (1) with (9) and (10). In the class of variance estimators as in Definition 2 with fixed $L \in \mathbb{N}$, $r_l = \#J_l - 1$, $l = 1, \ldots, L$, and $r = \sum_{l=1}^{L} r_l$, the asymptotically optimal MSE*

$$MSE\left[\hat{\sigma}^2_{opt,r}\right] = \frac{\sigma^4}{N}\left(\gamma_4 - 1 + \frac{1}{r}\right) + o(N^{-1}),$$

is achieved for a difference estimator $\hat{\sigma}^2_{opt,r}$ which has weights of the form $\mu_l = r_l/r$, $l = 1, \ldots, L$, and the J_l are non-parallel straight lines. The generalized difference schemes $(d_j^{(l)})_{j \in J_l}$ are exactly as in the one-dimensional case described in Hall et al. (1990), i.e. they fulfill the one-dimensional asymptotic optimality criterion $\sum_{j \in J_l(k)} d_j^{(l)} d_{j+k}^{(l)} = -1/(2\, r_l)$ as in (5) for all $0 \neq k$ with $J_l(k) = \{j \in J_l : l + k \in J_l\} \neq \emptyset$.

We mention that for $d \geq 4$ a similar result can be shown, where, however, the class of difference estimators has to be restricted to those which are unbiased estimators for polynomials of order $m = [d/4]+1$ (see Wagner (1999)). The last theorem can be illustrated with the help of the following example.

EXAMPLE 5. *Assume $d = 2$ and an equidistant design. In Figure 3 an example is given for a particular configuration in the case $L = 4$, $r_1 = r_2 = 4$ (horizontal and vertical lines)*

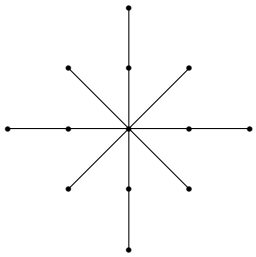


Fig. 3. Example for a difference scheme estimator of the variance with $d = 2$ and $L = 4$.

and $r_3 = r_4 = 2$ (diagonal lines). In order to achieve the asymptotically minimal MSE in the class of generalized difference estimators where $r = 12$,

$$MSE\left[\hat{\sigma}^2_{opt,12}\right] = \frac{\sigma^4}{N}\left(\gamma_4 - 1 + \frac{1}{12}\right) + o(N^{-1}),$$

one has to choose in a first step the weights according to Theorem 6 as $\mu_1 = \mu_2 = 1/3$ and $\mu_3 = \mu_4 = 1/6$. In a second step the optimal difference schemes of Hall et al. (1990) have to be used for $r = 4$ and $r = 2$, respectively. Observe that here $d \neq L$.

REMARK 3. As pointed out by a referee it is an interesting task to discuss the role of L in generalized difference estimators. This is in principle very difficult, because it cannot be separated from the question of optimal length r_l along each direction l. A comprehensive answer is beyond the scope of this paper, but a qualitative answer can be given by means of Theorem 4, provided the expansion (14) can be considered as sufficiently accurate, which is the case for not too strongly oscillating signals. Then, the bias will be minimized when the directions $s^{(l)}$ are chosen such that the expression on the right-hand side in (14) is minimized. This is a discrete minimization problem, and the mimimum would be achieved (if ∇g was known) by choosing $L = 1$ and $s^{(1)} \in \mathbb{Z}^d$ as the mimimizer of $\| < s^{(1)}, \nabla g > \|_2^2$. However, in general higher order terms in the expansion of the bias will involve mixed derivatives of g, and here $L = 1$ is not necessarily the best choice. Nevertheless, Theorem 4 justifies in general configurations, so that residuals are computed along directions where the gradient of g is small. If g fluctuates sharply in all directions, this can only be achieved by small numbers of local residuals r_l but large L. If g is such that the gradient is small along a specific direction, a large number of residuals along this direction for $L = 1$ will give a good result. This is in accordance with the bias trimming algorithms of Thompson et al. (1991) (for $d = 2$), because these algorithms identify in a first step those grid points where ∇g is expected to be large and eliminate these from further calculations.

Finally, note that with increasing variance σ^2 (noise level) the optimal estimators fare better, because the bias does not depend on σ^2 (Theorem 4), in general. However, the practical merits of this finding are limited since in this case the overall quality of the estimators will be bad.

4. A numerical comparison

4.1. Kernel estimators

In this section we compare the difference-based estimators with the (asymptotically efficient) d-dimensional generalization of Hall and Marron's (1990) kernel estimator and the corresponding local linear estimator. For the sake of brevity we consider product kernels $K(x) = \prod_{i=1}^{d} K_1(x_i)$ of order $r \in \mathbb{N}$ (here $x = (x_1, \ldots, x_d)' \in \mathbb{R}^d$) such that $K_1 : \mathbb{R} \to \mathbb{R}$ is symmetric with compact support. Furthermore,

$$\begin{aligned} \int_{\mathbb{R}^d} K(x)\,dx &= 1\,, \qquad \int_{\mathbb{R}^d} x_i^l\, x_j^k\, K(x)\,dx = 0, \quad i \neq j\,;\ 0 \leq l, k < r\,, \\ \int_{\mathbb{R}^d} x_i^r\, K(x)\,dx &= \mu_r(K) \neq 0, \quad i = 1, \ldots, d\,. \end{aligned}$$

Finally, the bandwidth matrix is assumed to be diagonal, $H = diag(h_1, \ldots, h_d)$.

THEOREM 7. *Consider the nonparametric regression model (1) with $g \in C^{(r+1)}[0,1]^d$ and design points $t_i \in [0,1]^d$, $i = 1, \ldots, N$, located on a d-dimensional grid, so that (9) holds. For the diagonal bandwidth matrix H it holds that $h_{min} := \min_{k=1,\ldots,d} h_k > 0$, $N h_{min}^d \to \infty$, $\lambda_{k,N} := h_k / h_{min} \to \lambda_k \in (0, \infty)$, $k = 1, \ldots, d$, as $N \to \infty$. Let K be a d-dimensional kernel of order r. Then, the MSE of the variance estimator*

$$\hat{\sigma}_K^2 = \frac{1}{v} \sum_{i=1}^{N} \left(Y_i - \sum_{j=1}^{N} w_{ij}\, Y_j \right)^2 \tag{15}$$

with $v = N - 2\sum_i w_{ii} + \sum_{i,j} w_{ij}^2$ and $w_{i,j} = K(H^{-1}(t_i - t_j)) / \sum_{\ell=1}^{N} K(H^{-1}(t_i - t_\ell))$ is given by

$$MSE\left[\hat{\sigma}_K^2\right] = \frac{1}{N}(\gamma_4 - 1)\,\sigma^4 + C_1\,(N^2 \prod_{k=1}^{d} h_k)^{-1} + C_2^2\, h_{min}^{4r} + o\Big((N^2 \prod_{k=1}^{d} h_k)^{-1} + h_{min}^{4r}\Big)\,.$$

Here

$$\begin{aligned} C_1 &= 2\sigma^4 \int_{\mathbb{R}^d} \left((K * K)(x) - 2\,K(x)\right)^2\,dx\,, \\ C_2 &= \kappa_r^2 \int_{\mathbb{R}^d} \Big((g\,f)^{(r)}(x) - g(x)\,f^{(r)}(x) \Big)^2\,(f(x))^{-1}\,dx\,, \qquad \kappa_r = \frac{(-1)^r}{r!}\,\mu_r(K), \end{aligned}$$

where for a function $f \in C^{(r)}[0,1]^d$ we use

$$f^{(r)}(x) = \sum_{k=1}^{d} \lambda_k^r\, \frac{\partial^r}{(\partial t_k)^r}\, f(t)\Big|_{t=x}\,.$$

The proof is omitted and follows in principle the pattern of the one-dimensional case as found in Hall and Marron (1990). Note that due to the special choice of the product kernel no mixed derivatives are involved in C_2.

REMARK 4. As a byproduct of the last theorem we obtain for $4r > d$ from $C_2^2\, h^{4r} = C_1 N^{-2} h^{-d}$ the optimal diagonal bandwidth matrix of the form $H = h\, I_d$, which asymptotically minimizes the MSE, as

$$h_0 = \left(C_1 (C_2\, N)^{-2}\right)^{\frac{1}{4r+d}} = O\left(N^{-\frac{2}{4r+d}}\right) . \tag{16}$$

In this case, with h_0 in (16), one can show that $\sqrt{N}\left(\hat{\sigma}_K^2 - \sigma^2\right)$ is asymptotically centered normal with variance $(\gamma_4 - 1)\,\sigma^4$, and hence $\sqrt{N}$-efficient.

4.2. A simulation study

In the following, a simulation study for the cases $d = 2, 3, 4$ will be presented (see Thompson et al. (1991), or Dette et al. (1998), for an extensive numerical study when $d = 1$). To this end we assumed normally distributed errors; similar results where found for skewed errors which are not displayed. All functions under consideration were defined in $[0, 1]^d$, and equidistant designs were used. For the kernel estimator we chose product kernels generated by the Epanechnikov kernel, $K_1(x) = 3/4\ (1 - x^2)\, \mathbf{1}\{|x| \le 1\}$, and the bandwidth matrix $H = h\, I_d$. In addition to the kernel estimator (15), a local linear estimator with the same kernel was investigated (Wand and Jones (1995)). The bandwidths required for these estimators were obtained by cross-validation. Finally, for each setting the 'oracle' estimator was calculated, i.e. the estimator of the variance when the true regression function g is known (the 'ideal' estimator in Thompson et al. (1991)). This serves as a benchmark for the best possible estimator. In each simulation scenario 500 (or 1000) runs were performed, where the random generator g05ddc from C++ NAG was used.

4.2.1. The case $d = 2$

From our previous discussion it can be expected that the oscillation and the smoothness of the signal g will affect the effective choice of the weighting scheme significantly. Therefore, we have considered the following regression functions (RF; Figure 4):

$$\begin{aligned}
g_1(x,y) &= y\, \sin(8\pi\, x) && \text{(heavily oscillating in } x\text{-direction)}\\
g_2(x,y) &= y\, \sin(2\pi\, x) && \text{(oscillating in } x\text{-direction)}\\
g_3(x,y) &= \sin(2\pi(x+y)) && \text{(oscillating in both directions)}\\
g_4(x,y) &= \sin(5\pi(x+y)) && \text{(heavily oscillating in both directions)}\\
g_5(x,y) &= x\, y && \text{(polynomial function)}\\
g_6(x,y) &= \exp(-(x+y)/2) && \text{(monotone function)}\\
g_7(x,y) &= \max\{g_{7h}(x), g_{7h}(y)\} - g_{7h}(x) g_{7h}(y) && \text{(chess-board-like function)}\\
g_{7h}(x) &= \begin{cases} 1, & 1/3 < x \le 2/3 \\ 0, & \text{else} \end{cases} && (17)\\
g_8(x,y) &= \max\{g_{8h}(x), g_{8h}(y)\} && \text{(spiky function)}\\
g_{8h}(x) &= \begin{cases} 1/8, & x = 0 \\ 5/8, & x = 0.5 \\ 2/8, & x = 1 \\ 0, & \text{else} \end{cases} && (18)
\end{aligned}$$

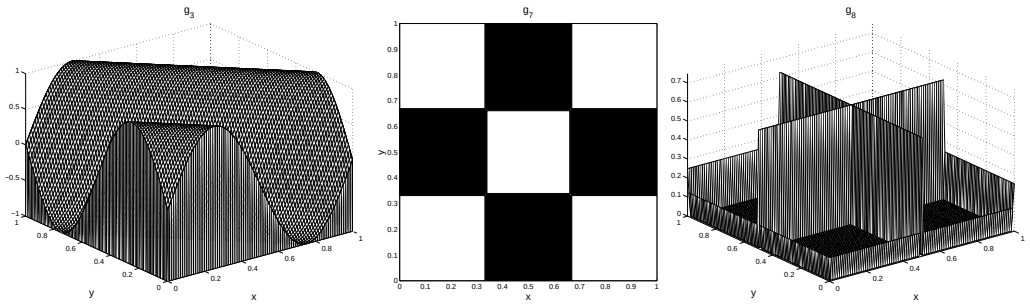


Fig. 4. Functions $g_3(x,y)$, $g_7(x,y)$ (the areas where $g_7(x,y)=1$ are marked black), and $g_8(x,y)$.

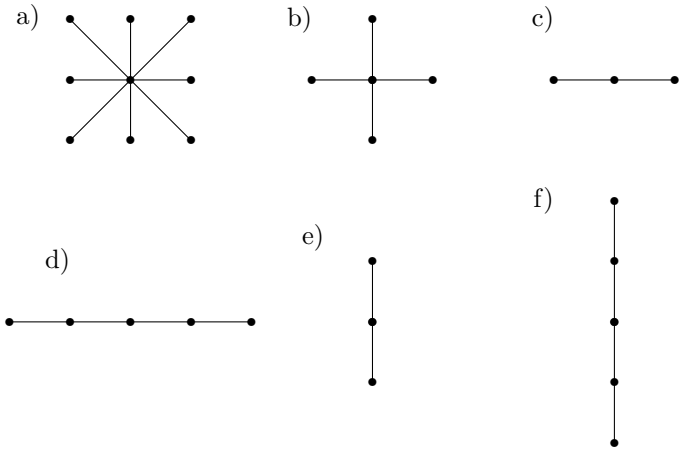


Fig. 5. Residual configurations for various difference estimators.

The following difference estimators were considered:

(a) $\hat{\sigma}^2_{S,p}$ with residuals as in Fig. 5a), where on each line the polynomial difference scheme (PolDif) $(d_0,d_1,d_2)=(1,-2,1)/\sqrt{6}$ was chosen.
(b) $\hat{\sigma}^2_{K,p}$ with residuals as in Fig. 5b) and PolDif as in (a).
(c) $\hat{\sigma}^2_{x2,p}$ with residuals as in Fig. 5c) and PolDif as in (a).
(d) $\hat{\sigma}^2_{x4,p}$ with residuals as in Fig. 5d) and PolDif $(d_0,d_1,d_2,d_3,d_4)=(1,-4,6,-4,1)/\sqrt{70}$.
(e) $\hat{\sigma}^2_{y2,p}$ with residuals as in Fig. 5e) and PolDif as in (a).
(f) $\hat{\sigma}^2_{y4,p}$ with residuals as in Fig. 5f) and PolDif as in (d).

Furthermore, the same estimators with optimal difference schemes (OptDif) of order 2, $(d_0,d_1,d_2)=(0.809,-0.5,-0.309)$ and of order 4, $(d_0,d_1,d_2,d_3,d_4)=(0.2708,-0.0142,0.6909,-0.4858,-0.4617)$, respectively, were considered (Hall et al. (1990)). The resulting estimators are denoted as $\hat{\sigma}^2_{S,o}$, $\hat{\sigma}^2_{K,o},\ldots,\hat{\sigma}^2_{y4,o}$. In the following tables we use the notation 4.22_2 for $4.22*10^{-2}$, and so on.

In Table 1, results for a heavily oscillating signal in the x-direction (g_1) are displayed, where $\sigma^2=0.25,0.5$. Observe that here the y direction is linear. From this table it can

Table 1. The case $d = 2$. Bias^2 and variance for regression function g_1

			PolDif				OptDif				Oracle
			$r=2$		$r=4$		$r=2$		$r=4$		
σ^2	(n_1,n_2)	var.est.	$\hat\sigma^2_{x2,p}$	$\hat\sigma^2_{y2,p}$	$\hat\sigma^2_{x4,p}$	$\hat\sigma^2_{y4,p}$	$\hat\sigma^2_{x2,o}$	$\hat\sigma^2_{y2,o}$	$\hat\sigma^2_{x4,o}$	$\hat\sigma^2_{y4,o}$	
0.25	(10,10)	Bias^2	2.49_1	1.43_5	6.60_1	3.69_5	4.51_2	1.21_4	7.38_2	5.97_4	9.13_6
		Variance	1.77_2	2.78_3	4.44_2	4.95_3	4.75_3	1.81_3	7.23_3	2.32_3	1.12_3
	(10,25)	Bias^2	2.34_1	5.44_6	6.21_1	1.09_2	4.13_2	4.74_6	6.72_2	1.97_5	3.51_7
		Variance	7.44_3	1.09_3	1.86_2	3.13_3	1.85_3	6.39_4	2.73_3	6.80_4	4.84_4
	(25,10)	Bias^2	1.02_3	1.30_8	1.60_5	5.19_3	2.92_2	6.18_5	5.97_2	4.56_4	1.29_7
		Variance	1.08_3	1.15_3	1.52_3	1.00_3	1.44_3	7.53_4	2.29_3	9.61_4	5.02_4
	(30,30)	Bias^2	2.29_4	2.90_7	1.57_6	3.75_7	1.54_2	1.66_6	4.91_2	9.43_6	5.77_7
		Variance	2.96_4	2.74_4	4.03_4	3.99_4	3.13_4	1.81_4	5.55_4	1.77_4	1.40_4
	(100,100)	Bias^2	1.60_9	2.18_8	1.00_9	5.23_8	1.68_4	$<1.0_{12}$	1.36_3	1.89_8	1.46_8
		Variance	2.48_5	2.25_5	3.41_5	3.07_5	1.60_5	1.53_5	1.51_5	1.42_5	1.22_5
0.5	(10,10)	Bias^2	2.44_1	3.90_6	6.49_1	1.03_5	4.47_2	1.04_4	7.30_2	5.46_4	1.18_5
		Variance	4.24_2	1.15_2	1.01_1	1.85_2	1.34_2	7.50_3	1.84_2	9.07_3	5.25_3
	(10,25)	Bias^2	2.35_1	8.53_6	6.22_1	4.21_2	4.30_2	4.80_6	7.01_2	1.52_5	4.24_6
		Variance	1.60_2	4.51_3	3.91_2	1.32_2	5.23_3	2.71_3	7.22_3	2.59_3	1.90_3
	(25,10)	Bias^2	1.05_3	2.22_5	1.87_5	2.00_2	2.94_2	1.21_4	5.89_2	6.18_4	1.52_5
		Variance	4.35_3	4.43_3	5.94_3	3.56_3	4.52_3	3.04_3	6.22_3	3.77_3	2.27_3
	(30,30)	Bias^2	1.50_4	6.24_7	4.04_6	1.71_6	1.49_2	1.79_7	4.80_2	9.73_8	2.55_6
		Variance	1.15_3	1.11_3	1.65_3	1.64_3	1.06_3	7.66_4	1.60_3	7.20_4	5.86_4
	(100,100)	Bias^2	3.50_9	2.03_7	1.31_8	2.87_7	1.76_4	3.24_8	1.37_3	6.79_8	2.50_9
		Variance	1.01_4	9.36_5	1.40_4	1.31_4	5.96_5	6.04_5	5.51_5	5.45_5	4.61_5

be concluded that for larger sample sizes the use of asymptotically optimal weights along the x-axis direction leads to a much larger bias compared to the polynomial estimators, as was expected. Of course, if the optimal weighting scheme is chosen in the y-direction, the resulting estimator outperforms the polynomial weighting scheme, particularly for sample sizes where $n_1, n_2 \geq 30$. Furthermore, increasing the order of the optimal difference-based estimator in the x-direction results in a decrease of the MSE when the sample size grows ($n_1 = n_2 = 100$). For smaller sample sizes ($n_1, n_2 \leq 30$) this is not observed. Note that this is in accordance with our theoretical finding in Theorem 4, where the bias increases as r increases. For large sample sizes, in most cases the variance dominates the squared bias in this setting. However, for the case of optimal difference schemes along the x-direction the variance is dominated by the squared bias. For the choice of weights along the x-direction, the polynomial difference schemes perform better than the corresponding optimal difference schemes. Further, from Table 1 it becomes apparent that for the polynomial weighting scheme $r = 2$ is in most cases superior to $r = 4$, which is due to the smaller variance. This is in accordance with the recommendation by Hall et al. (1991) to use short and compact configurations. Finally, we mention that these findings are independent of the noise level; similar results hold for $\sigma^2 = 1$ (not displayed).

In Table 2 the different regression functions are compared. Here only short and compact configurations of the residuals as in Figure 5a) and b) were considered. In most cases, the variance dominates the squared bias, except for strongly oscillating signals (g_4). The variances are mostly of comparable magnitude in each row of the table. The polynomial weighting scheme yields a smaller bias than the optimal weighting scheme in case of oscillating and non-smooth signals, in particular for larger sample sizes. It becomes apparent

Table 2. The case $d = 2$. Bias^2 and variance for $\sigma^2 = 0.5$

(n_1, n_2)	RF	var.est.	$\hat{\sigma}^2_{S,p}$	$\hat{\sigma}^2_{K,p}$	$\hat{\sigma}^2_{S,o}$	$\hat{\sigma}^2_{K,o}$	$\hat{\sigma}^2_{Kern}$	$\hat{\sigma}^2_{locLin}$	Oracle
$(10, 10)$	g_2	Bias^2	7.49_5	1.74_5	3.60_3	2.08_3	1.72_4	1.07_4	4.76_6
		Variance	8.52_3	1.02_2	7.37_3	7.60_3	6.64_3	6.77_3	4.99_3
	g_3	Bias^2	3.94_3	2.09_4	9.21_2	6.96_2	1.97_3	1.31_3	2.39_5
		Variance	8.48_3	9.48_3	1.18_2	1.14_2	7.24_3	8.29_3	4.63_3
	g_4	Bias^2	2.87_1	2.08_1	2.80_1	6.23_1	2.48_1	2.63_1	4.89_8
		Variance	2.56_2	2.47_2	2.24_2	4.26_2	1.62_2	1.66_2	4.77_3
	g_5	Bias^2	6.10_7	6.13_6	6.93_5	5.16_5	5.43_5	3.27_5	2.73_8
		Variance	9.53_3	1.15_2	7.36_3	8.32_3	5.52_3	5.44_3	5.00_3
	g_6	Bias^2	1.04_6	1.76_7	4.79_6	2.59_8	6.45_5	2.02_5	7.93_6
		Variance	1.01_2	1.23_2	7.87_3	8.63_3	5.78_3	5.49_3	4.85_3
	g_7	Bias^2	1.67_2	8.20_3	4.36_2	3.74_2	7.70_3	8.48_3	1.29_5
		Variance	1.34_2	1.59_2	1.31_2	1.40_2	8.69_3	9.22_3	5.38_3
	g_8	Bias^2	4.20_3	2.28_3	3.70_3	1.80_3	1.71_3	1.82_3	8.90_7
		Variance	1.19_2	1.23_2	8.41_3	8.52_3	6.63_3	7.68_3	4.99_3
$(30, 30)$	g_2	Bias^2	3.59_7	1.00_8	4.67_5	2.10_5	1.84_5	2.73_5	1.12_8
		Variance	6.89_4	9.01_4	5.85_4	6.23_4	5.20_4	5.87_4	5.10_4
	g_3	Bias^2	1.35_3	6.22_4	5.63_3	3.00_3	2.11_3	2.16_3	7.10_7
		Variance	8.41_4	8.86_4	8.15_4	8.16_4	8.21_4	7.82_4	6.00_4
	g_4	Bias^2	7.98_4	6.70_5	4.71_2	2.90_2	1.12_3	9.62_4	1.67_6
		Variance	8.41_4	9.75_4	9.24_4	9.06_4	1.06_3	9.93_4	6.00_4
	g_5	Bias^2	1.82_6	1.76_6	4.86_7	1.01_6	6.37_8	2.80_9	6.03_7
		Variance	7.67_4	9.43_4	6.56_4	6.90_4	5.82_4	5.59_4	5.87_4
	g_6	Bias^2	5.48_8	1.37_8	1.0_{10}	2.22_8	4.88_8	3.53_7	1.17_7
		Variance	7.80_4	9.47_4	7.01_4	7.36_4	6.28_4	5.41_4	5.70_4
	g_7	Bias^2	5.68_6	6.56_6	1.97_3	9.61_4	9.24_5	7.19_5	1.92_6
		Variance	7.50_4	8.91_4	6.66_4	7.14_4	6.20_4	6.65_4	5.51_4
	g_8	Bias^2	7.98_4	6.70_5	4.71_2	2.90_2	1.12_3	9.62_4	1.67_6
		Variance	8.41_4	9.75_4	9.24_4	9.06_4	1.06_3	9.93_4	6.00_4

that for the difference-based estimators the star-shaped configuration tends to be better than the cross-shaped configuration in case of smooth functions which are not too heavily oscillating. This is due to a reduction of variance because of the larger number of residuals taken into account by the first-named estimators. In terms of the MSE, the polynomial weighting scheme tends to outperform the optimal weighting scheme for oscillating functions (g_3, g_4) and for the chess-board type function g_7. In general, the kernel estimator and the local linear estimator are comparable or even more efficient than the difference-based estimators, except for heavily oscillating signals, where the polynomial difference estimator should be used (g_8). However, it has to be taken into account that the computing time for the kernel and the local linear estimator is much higher than for the difference-based estimators, which is due to the cross validation of the bandwidths. For instance, for $n_1, n_2 = 30$ the computation of a single estimator takes 15ms on a Pentium 4 with 512MB RAM and 1.8GHz for a difference estimator, as compared to about 15s (factor 1000) for the kernel estimator.

4.2.2. The case $d = 3$

In the following we briefly consider the case $d = 3$, with the regression functions

$$\begin{aligned} g_1(x,y,z) &= \sin(2\pi(x+y+z)) \\ g_2(x,y,z) &= \sin(5\pi(x+y+z)) \\ g_3(x,y,z) &= \max\{g_{7h}(x), g_{7h}(y), g_{7h}(z)\} - g_{7h}(x)g_{7h}(y)g_{7h}(z), \qquad g_{7h} \text{ from (17)} \\ g_4(x,y,z) &= \max\{g_{8h}(x), g_{8h}(y), g_{8h}(z)\}, \qquad g_{8h} \text{ from (18)}, \end{aligned}$$

and 'star-shaped' difference-based estimators of order $r = 2$,

$$\hat{\sigma}^2_{W,p} := \frac{1}{26} \sum_{r,s,t=-1}^{1} \frac{1}{n_R} \sum_{j \in R} \left(d_0 Y_{j-(r,s,t)'} + d_1 Y_j + d_2 Y_{j+(r,s,t)'} \right)^2$$

with polynomial difference scheme. The same estimator using an optimal difference scheme is called $\hat{\sigma}^2_{W,o}$. The cardinality of $R := \times_{k=1}^{3} \{2, \ldots, n_k - 1\}$ is $n_R := \#R$. Furthermore, we investigated 'cross-shaped' estimators

$$\hat{\sigma}^2_{K,p} := \frac{1}{3} \sum_{l=1}^{3} \frac{1}{n_{R_l}} \sum_{j \in R_l} (d_0 Y_{j-\mathbf{1}_l} + d_1 Y_j + d_2 Y_{j+\mathbf{1}_l})^2$$

with polynomial difference scheme and its counterpart $\hat{\sigma}^2_{K,o}$ with optimal difference scheme. The cardinality of $R_l := \left\{ j = (j_1, j_2, j_3)' \in \times_{k=1}^{3} \{1, \ldots, n_k\} \,:\, j_l \in \{2, \ldots, n_l - 1\} \right\}$ is $n_{R_l} := \#R_l$ for $l = 1, 2, 3$.

In Tables 3 and 4 selected results for $d = 3$ are displayed. It can be seen that the polynomial weighting scheme mostly outperforms the optimal weighting scheme in case of oscillating signals, and that the cross-shaped estimators $\hat{\sigma}^2_{K,*}$ tend to outperform the corresponding star-shaped estimators $\hat{\sigma}^2_{W,*}$ $(* = p, o)$. In most cases, the squared bias dominates the variance of the estimators. Interestingly, this fails to hold for the chess-board type function g_3 with polynomial weighting scheme as the sample size increases. Again, the kernel estimator and the local linear estimator yield comparable results for the sample sizes under consideration and perform better than the difference-based estimators in case of not heavily oscillating signals, whereas the estimator $\hat{\sigma}^2_{K,p}$ yields the best results in case of oscillating signals, especially for larger sample sizes. Note that, for example, for $n_1, n_2, n_3 = 10$ the kernel estimator takes about 1300 times the computing time of the difference estimators. For larger sample sizes (50,50,50) and (100,100,100) the optimal weighting difference estimator is always outperformed by the bias reducing polynomial weighting estimator. Here it becomes nicely apparent that, the more d increases, the more the bias becomes the dominating term for the MSE.

Table 3. The case $d = 3$. Bias^2 and variance for $\sigma^2 = 0.25$

(n_1, n_2, n_3)	RF	var.est.	$\hat{\sigma}^2_{W,p}$	$\hat{\sigma}^2_{W,o}$	$\hat{\sigma}^2_{K,p}$	$\hat{\sigma}^2_{K,o}$	$\hat{\sigma}^2_{Kern}$	$\hat{\sigma}^2_{locLin}$	Oracle
$(5,5,5)$	g_1	Bias^2	2.58_1	2.68_1	1.18_1	5.69_1	2.48_1	2.41_1	3.50_6
		Variance	1.48_2	9.81_3	4.20_3	1.05_2	5.73_3	5.82_3	1.01_3
	g_2	Bias^2	2.99_1	2.55_1	9.64_1	4.57_1	2.56_1	2.57_1	7.25_8
		Variance	1.94_2	1.08_2	1.99_2	1.04_2	5.21_3	5.29_3	1.01_3
	g_3	Bias^2	5.27_2	5.24_2	5.08_2	5.14_2	1.72_2	2.32_2	2.42_6
		Variance	9.73_3	5.21_3	4.09_3	3.25_3	2.79_3	2.89_3	9.73_4
	g_4	Bias^2	6.68_3	3.07_3	3.32_3	2.10_3	1.46_3	2.03_3	3.51_6
		Variance	6.08_3	2.80_3	2.25_3	1.61_3	1.84_3	1.83_3	1.02_3
$(10,10,10)$	g_1	Bias^2	1.40_2	1.14_1	3.51_4	6.99_2	1.73_3	1.53_3	1.76_8
		Variance	2.95_4	5.30_4	1.66_4	3.09_4	1.56_4	1.95_4	1.15_4
	g_2	Bias^2	2.57_1	2.72_1	2.11_1	6.03_1	2.49_1	2.50_1	1.00_9
		Variance	1.04_3	8.90_4	6.11_4	1.31_3	7.06_4	6.37_4	1.21_4
	g_3	Bias^2	8.76_3	2.38_2	2.41_3	1.21_2	1.73_3	1.79_3	2.04_7
		Variance	3.52_4	3.50_4	1.98_4	2.15_4	2.09_4	2.20_4	1.22_4
	g_4	Bias^2	4.99_3	4.93_3	1.06_3	1.10_3	9.52_4	1.17_3	1.06_7
		Variance	4.28_4	2.98_4	2.24_4	1.69_4	2.88_4	3.41_4	1.25_4

Table 4. The case $d = 3$. Bias^2 and Variance for $\sigma^2 = 0.25$

(n_1, n_2, n_3)	RF	var.est.	$\hat{\sigma}^2_{W,p}$	$\hat{\sigma}^2_{W,o}$	$\hat{\sigma}^2_{K,p}$	$\hat{\sigma}^2_{K,o}$	Oracle
$(50,50,50)$	g_1	Bias^2	2.57_4	1.19_3	6.18_5	3.12_4	4.0_{10}
		Variance	1.31_6	1.29_6	1.40_6	1.18_6	1.03_6
	g_2	Bias^2	2.57_4	2.61_4	5.98_5	6.23_5	2.0_{10}
		Variance	1.45_6	1.25_6	1.45_6	1.16_6	1.02_6
	g_3	Bias^2	5.39_8	4.34_4	2.20_9	1.05_4	$<1.0_{12}$
		Variance	1.19_6	1.11_6	1.36_6	1.13_6	1.02_6
	g_4	Bias^2	6.60_5	1.34_2	7.46_7	3.89_3	$<1.0_{12}$
		Variance	1.26_6	1.42_6	1.41_6	1.22_6	1.04_6
$(100,100,100)$	g_1	Bias^2	6.12_5	2.97_4	1.44_5	7.28_5	1.0_{10}
		Variance	1.53_7	1.49_7	1.73_7	1.47_7	1.31_7
	g_2	Bias^2	6.87_5	7.37_5	1.55_5	1.57_5	2.0_{10}
		Variance	1.48_7	1.36_7	1.62_7	1.33_7	1.19_7
	g_3	Bias^2	2.0_{10}	2.69_5	1.0_{10}	6.25_6	$<1.0_{12}$
		Variance	1.38_7	1.30_7	1.62_7	1.34_7	1.24_7
	g_4	Bias^2	2.85_7	9.94_4	3.80_9	2.44_4	$<1.0_{12}$
		Variance	1.37_7	1.30_7	1.65_7	1.33_7	1.19_7

Table 5. *The case $d = 4$. Bias^2 and variance for $(n_1, n_2, n_3, n_4) = (5, 5, 5, 5)$*

σ^2	RF	var.est.	$\hat{\sigma}^2_{W,p}$	$\hat{\sigma}^2_{W,o}$	$\hat{\sigma}^2_{K,p}$	$\hat{\sigma}^2_{K,o}$	$\hat{\sigma}^2_{Kern}$	$\hat{\sigma}^2_{locLin}$	Oracle
0.25	g_1	Bias^2	5.04_2	1.62_1	8.46_4	1.05_1	3.57_3	3.08_3	7.64_7
		Variance	2.22_3	1.77_3	2.86_4	6.23_4	3.14_4	2.89_4	2.02_4
	g_2	Bias^2	2.46_1	2.53_1	1.13_1	5.63_1	2.51_1	2.47_1	1.02_6
		Variance	5.20_3	2.88_3	8.43_4	2.22_3	1.06_3	1.03_3	2.17_4
	g_3	Bias^2	2.85_2	2.85_2	3.06_2	3.05_2	8.63_3	9.98_3	4.05_7
		Variance	2.28_3	1.09_3	5.94_4	4.77_4	3.83_4	4.41_4	1.94_4
	g_4	Bias^2	3.55_3	1.72_3	1.53_3	1.04_3	1.01_3	1.21_3	1.41_7
		Variance	1.46_3	6.29_4	3.40_4	2.75_4	3.40_4	3.82_4	1.93_4
0.5	g_1	Bias^2	5.06_2	1.60_1	9.63_4	1.05_1	4.62_3	4.44_3	5.18_6
		Variance	6.89_3	5.21_3	1.13_3	1.94_3	1.44_3	1.41_3	7.95_4
	g_2	Bias^2	2.42_1	2.50_1	1.11_1	5.55_1	2.48_1	2.50_1	4.89_7
		Variance	1.09_2	5.91_3	1.88_3	4.82_3	2.53_3	2.35_3	8.83_4
	g_3	Bias^2	2.85_2	2.83_2	3.04_2	3.06_2	1.03_2	1.17_2	1.77_6
		Variance	6.67_3	3.56_3	1.87_3	1.63_3	1.45_3	1.52_3	8.17_4
	g_4	Bias^2	3.63_3	1.60_3	1.32_3	8.92_4	1.52_3	1.51_3	4.11_6
		Variance	4.50_3	2.07_3	1.27_3	1.00_3	1.16_3	1.05_3	8.13_4

4.2.3. The case $d = 4$

For the case $d = 4$ we considered the regression functions

$$\begin{aligned}
g_1(x,y,z,s) &= \sin(\pi(x+y+z+s)) \\
g_2(x,y,z,s) &= \sin(2\pi(x+y+z+s)) \\
g_3(x,y,z,s) &= \max_{t\in\{x,y,z,s\}} g_{7h}(t) - g_{7h}(x)g_{7h}(y)g_{7h}(z)g_{7h}(s), \qquad g_{7h} \text{ from (17)} \\
g_4(x,y,z,s) &= \max\{g_{8h}(x), g_{8h}(y), g_{8h}(z), g_{8h}(s)\}, \qquad g_{8h} \text{ from (18)},
\end{aligned}$$

and difference-based estimators analogous to the case $d = 3$. Tables 5 and 6 show selected results from our simulation study. We mention that the kernel and local polynomial estimators are computationally feasible only for small sample sizes such as $n_i = 5, i = 1, ..., 4$, due to the cross validation procedure. In this case, the calculation of the kernel estimator takes 13s, for $n_i = 6$ it takes 54s, and for $n_i = 7$ more than 3min.

Again, as for $d = 3$, the estimator $\hat{\sigma}^2_{K,p}$ performs best almost always among the difference-based estimators. From Table 6 we find that in accordance with Theorem 3.6, as the sample size increases the use of the polynomial weighting scheme corrects for the bias, whereas the generalized von Neumann (1941) estimator completely fails for $g_1 - g_3$. Observe that g_4 is close to a constant function.

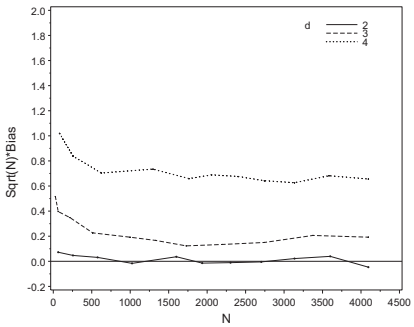
4.2.4. Comparison between dimensions

To illustrate the inconsistency of the optimal weighting scheme estimators (including the von Neumann (1941) estimator) for increasing dimension (see Example 4), we considered the functions $\tilde{g}_d(x) := 2^{1/2}\pi^{-1}\sin(\pi\sum_{l=1}^d x_l)$, for $d = 2, 3, 4$. Note that $\|\partial g/\partial t_l\|_2^2 = 1$, $l = 1, \ldots, d$, in order to render the functions comparable.

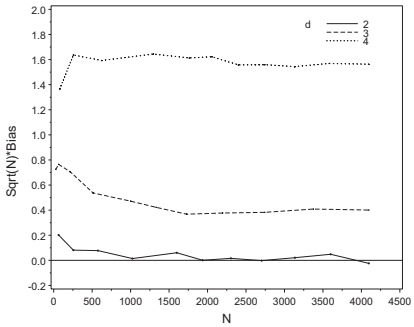
Figure 6 shows that the bias (normed by $\sqrt{N}$, which stems from the C.L.T., see the comment below Theorem 3) gets larger for increasing values of d except for the polynomial weighting scheme estimators. In contrast, the variance plays only a minor role (Figure 7).

Table 6. *The case* $d = 4$. *Bias*^{2} *and variance for* $\sigma^2 = 0.25$

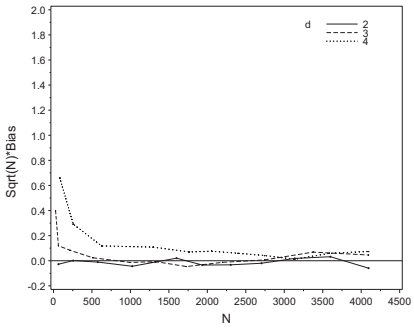
(n_1, n_2, n_3, n_4)	RF	var.est.	$\hat{\sigma}^2_{W,p}$	$\hat{\sigma}^2_{W,o}$	$\hat{\sigma}^2_{K,p}$	$\hat{\sigma}^2_{K,o}$	Oracle
$(10, 10, 10, 10)$	g_1	Bias^2	3.64_4	2.65_2	1.21_6	5.39_3	5.40_9
		Variance	2.72_5	3.34_5	1.64_5	1.56_5	1.15_5
	g_2	Bias^2	2.93_2	1.36_1	3.38_4	7.05_2	1.50_8
		Variance	3.74_5	6.07_5	1.64_5	2.92_5	1.21_5
	g_3	Bias^2	4.13_3	1.06_2	9.48_4	4.80_3	4.70_9
		Variance	4.10_5	3.65_5	1.76_5	1.68_5	1.16_5
	g_4	Bias^2	5.67_3	5.20_3	6.95_4	6.98_4	1.26_8
		Variance	4.89_5	3.34_5	1.94_5	1.48_5	1.19_5
$(50, 50, 50, 50)$	g_1	Bias^2	3.0_{10}	4.71_5	1.0_{10}	6.48_6	1.0_{10}
		Variance	2.30_8	2.19_8	2.55_8	2.15_8	1.98_8
	g_2	Bias^2	1.72_7	7.21_4	5.0_{10}	1.04_4	$< 1.0_{12}$
		Variance	2.07_8	1.99_8	2.33_8	2.00_8	1.86_8
	g_3	Bias^2	1.56_4	7.04_4	2.33_5	1.18_4	$< 1.0_{12}$
		Variance	2.44_8	2.43_8	2.59_8	2.18_8	1.96_8
	g_4	Bias^2	4.11_4	4.15_4	5.69_5	5.89_5	$< 1.0_{12}$
		Variance	2.76_8	2.50_8	2.65_8	2.15_8	1.98_8



von Neumann estimator $(r = 1)$



$\hat{\sigma}^2_{K,o}$ $(r = 2)$

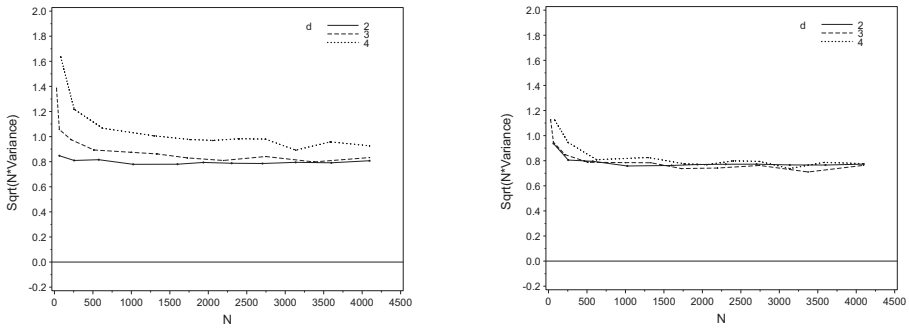


$\hat{\sigma}^2_{K,p}$ $(r = 2)$

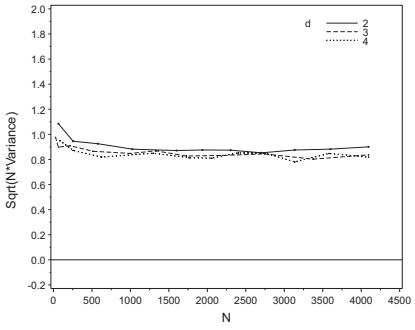
Fig. 6. $\sqrt{N}Bias$ in dependence on N.

5. Conclusions

In summary, we found that polynomial difference estimators with 'compact' configurations and not too large length of local residuals ($r \leq 4$ was sufficient in all settings we have



von Neumann estimator ($r = 1$) $\hat{\sigma}^2_{K,o}$ ($r = 2$)



$\hat{\sigma}^2_{K,p}$ ($r = 2$)

Fig. 7. $\sqrt{N\,Variance}$ in depencence on N.

considered) perform well for small and moderate sample sizes, provided a fluctuating signal cannot be excluded a priori. The polynomial difference estimators are simple to calculate and $\sqrt{N}$-consistent in arbitrary dimensions of the regressor space. If the signal is known to be slowly oscillating the use of a kernel type estimator becomes feasible due to its superior efficiency. However, here a computationally more feasible bandwidth selection than cross validation is required (see Herrmann et al. (1995)). We did not pursue this topic in this paper. The local linear estimator was not found to be significantly different to the kernel estimator. This might be due to the fact that, when the variance is constant, boundary effects do not play a big role in contrast to the case when σ^2 is a function (see Ruppert et al. (1997)). Hence, the more simple kernel estimator can be used instead, without significant loss of performance. Note, however, that the computing time for these estimators increases drastically with higher dimensions as compared to the difference estimators. This yields a considerable practical burden, for example in image analysis where computation in real time is often required. For $d > 2$, the use of optimal difference schemes cannot be recommended at all, including the generalization of von Neumann's (1941) estimator. Note that these estimators even fail to be $\sqrt{N}$-consistent if the dimension of the regressor space is larger than three.

Acknowledgements: The authors acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG Mu1230/8-1). We would like to thank L. Boysen and H. Dette for helpful comments. Parts of this paper can be found in the PhD thesis of T. Wagner at the

Fakultät für Mathematik, Ruhr-Universität Bochum. The authors gratefully acknowledge the comments of two referees and an associate editor which have led to a much improved presentation of this paper.

6. Appendix: Proofs

Proof of Theorem 3. As in the proof of Theorem 2.1 in Hall et al. (1991),

$$\begin{aligned}
E[\tilde\sigma_l^2\,\tilde\sigma_j^2] &= \frac{1}{n_{R_l} n_{R_j}} \sum_{i_1\in R_l} \sum_{i_2\in R_j} \sum_{j_{11},j_{12}\in J_l} \sum_{j_{21},j_{22}\in J_j} d_{j_{11}}^{(l)} d_{j_{12}}^{(l)}\, d_{j_{21}}^{(j)}\, d_{j_{22}}^{(j)}\, E[\varepsilon_{i_1+j_{11}}\varepsilon_{i_1+j_{12}}\varepsilon_{i_2+j_{21}}\varepsilon_{i_2+j_{22}}] \\
&= \sigma^4 + N^{-1}\Big(E[\varepsilon^4] - 3\sigma^4 + 2\,\sigma^4 \sum_{j_{11}\in J_l} \sum_{j_{21}\in J_j} \tilde{\sum_k} d_{j_{11}}^{(l)}\, d_{j_{11}+k}^{(l)}\, d_{j_{21}}^{(j)}\, d_{j_{21}+k}^{(j)}\Big) \;+\; o(N^{-1}) \\
&= \sigma^4 + \frac{\sigma^4}{N}\Big(\gamma_4 - 1 + 2\tilde{\sum_{k\neq 0}}\Big(\sum_{j_{11}\in J_l} d_{j_{11}}^{(l)}\, d_{j_{11}+k}^{(l)}\Big)\Big(\sum_{j_{21}\in J_j} d_{j_{21}}^{(j)}\, d_{j_{21}+k}^{(j)}\Big)\Big) \;+\; o(N^{-1})\,,
\end{aligned}$$

where $\tilde{\sum}_k$ and $\tilde{\sum}_{k\neq 0}$ denote summation over $\{k \in \mathbb{Z}^d : j_{11}+k \in J_l \wedge j_{21}+k \in J_j\}$ and $\{k \in \mathbb{Z}^d\backslash\{0\} : j_{11}+k \in J_l \wedge j_{21}+k \in J_j\}$, respectively. Because the bias is of order $O(n^{-2\gamma}) = o(n^{-d/2}) = o(N^{-1/2})$ in the same way as in the one-dimensional case, the variance contribution of order $O(N^{-1})$ dominates the MSE. The variance equals

$$\begin{aligned}
Var[\hat\sigma^2] &= Var\Big[\sum_{l=1}^{L} \mu_l\,\hat\sigma_l^2\Big] \\
&= \sum_{l,k=1}^{L} \mu_l\,\mu_k\,\big(Cov[\tilde\sigma_l^2,\tilde\sigma_k^2] \;+\; O(n^{-2\gamma})\,Cov[\xi_l,\xi_k] \;+\; O(n^{-\gamma})\,Cov[\xi_l,\tilde\sigma_k^2]\big)\,,
\end{aligned}$$

with

$$\begin{aligned}
Cov[\xi_l,\xi_k] &= \frac{1}{n_{R_l}\,n_{R_k}} \sum_{i\in R_l} \sum_{j\in R_k} \sum_{\nu\in J_l} \sum_{\eta\in J_k} d_\nu^{(l)}\, d_\eta^{(k)}\, E[\varepsilon_{i+\nu}\,\varepsilon_{j+\eta}] \\
&= \frac{1}{n_{R_l}\,n_{R_k}} \sum_{i\in R_l} \sum_{\substack{i+\nu=j+\eta \\ \nu\in J_l,\,\eta\in J_k,\,j\in R_k}} d_\nu^{(l)}\, d_\eta^{(k)}\, E[\varepsilon_{i+\nu}\,\varepsilon_{j+\eta}] = O(N^{-1})\,,
\end{aligned}$$

because $\#R_l = n_{R_l} = N + O(n^{d-1})$ for $l = 1,\ldots,L$. In the same way, $Cov[\xi_l,\tilde\sigma_k^2] = O(N^{-1})$. Because of $E[\tilde\sigma_l^2] = \sigma^2$ the result follows from the condition $\sum \mu_i = 1$. □

Proof of Theorem 5. For the sake of brevity we give only a sketch of the proof. A similar calculation as in the proof of Theorem 3 shows that the variance of any difference estimator $\hat\sigma^2$ is given asymptotically by the right-hand side of (13). Therefore, it remains to show that the bias contribution to the MSE is of order $o(N^{-1})$. To this end note that the generalized difference scheme $(d_j^{(l)})_{j\in J_l}$, $l = 1,\ldots,L$, satisfies (7), such that $\sum r_l = r \geq m$. Then a Taylor expansion shows that the bias is of order $O(n^{-2m})$, by a similar computation as in the one-dimensional case (Dette et al. (1998)). From $m = [d/4]+1$, the result follows. □

Proof of Theorem 6. The condition of non-parallel straight lines implies that for any k, such that $0 \neq k \in \mathbb{Z}^d$, at most one of the sets $J_l(k) = \{j \in J_l : j + k \in J_l\}$ is non-empty for one $l \in \{1, \ldots, L\}$. Therefore, the part of the MSE which depends on the generalized difference scheme equals

$$\begin{aligned}
\sum_{k\neq 0}\Big(\sum_{l=1}^{L}\mu_l \sum_{j} d_j^{(l)}\, d_{j+k}^{(l)}\Big)^2 &= \sum_{k\neq 0}\sum_{l_1=1}^{L}\sum_{l_2=1}^{L}\mu_{l_1}\,\mu_{l_2}\Big(\sum_{j_1\in J_{l_1}(k)} d_{j_1}^{(l_1)}\, d_{j_1+k}^{(l_1)}\Big)\Big(\sum_{j_2\in J_{l_2}(k)} d_{j_2}^{(l_2)}\, d_{j_2+k}^{(l_2)}\Big) \\
&= \sum_{k\neq 0}\sum_{l=1}^{L}\mu_l^2\Big(\sum_{j\in J_l(k)} d_j^{(l)}\, d_{j+k}^{(l)}\Big)^2 \\
&\quad + \sum_{k\neq 0}\sum_{l_1\neq l_2}\mu_{l_1}\mu_{l_2}\Big(\sum_{j_1\in J_{l_1}(k)} d_{j_1}^{(l_1)} d_{j_1+k}^{(l_1)}\Big)\Big(\sum_{j_2\in J_{l_2}(k)} d_{j_2}^{(l_2)} d_{j_2+k}^{(l_2)}\Big) \quad (19) \\
&= \sum_{l=1}^{L}\mu_l^2 \sum_{k\neq 0}\Big(\sum_{j\in J_l(k)} d_j^{(l)}\, d_{j+k}^{(l)}\Big)^2 ,
\end{aligned}$$

because the sum in (19) is zero for $l_1 \neq l_2$ and $k \neq 0$. Further, $\#\{k \neq 0 : J_l(k) \neq \emptyset\} = 2\, r_l$, and hence the MSE can be expanded as

$$MSE\left[\hat{\sigma}^2\right] \;=\; \frac{\sigma^4}{N}\left(\gamma_4 - 1 \,+\, 2\sum_{l=1}^{L}\Big(\frac{r_l}{r}\Big)^2 \sum_{k\neq 0}\Big(\sum_{j\in J_l(k)} d_j^{(l)}\, d_{j+k}^{(l)}\Big)^2\right) \,+\, o(N^{-1}).$$

Now a similar argument as in the case $d = 1$ (Hall et al. (1990)) shows the asymptotic optimality of $\hat{\sigma}^2_{opt,r}$, and we obtain

$$\begin{aligned}
MSE\left[\hat{\sigma}^2_{opt,r}\right] &= \frac{\sigma^4}{N}\left(\gamma_4 - 1 \,+\, 2\sum_{l=1}^{L}\Big(\frac{r_l}{r}\Big)^2 (2r_l)\Big(-\frac{1}{2r_l}\Big)^2\right) \,+\, o(N^{-1}) \\
&= \sigma^4\, N^{-1}\left(\gamma_4 - 1 \,+\, 1/r\right) \,+\, o(N^{-1}) ,
\end{aligned}$$

where we have used that

$$\sum_{k\neq 0}\Big(\sum_{j\in J_l(k)} d_j^{(l)} d_{j+k}^{(l)}\Big)^2 = -\frac{1}{2\, r_l} .$$

It remains to show that $\hat{\sigma}^2_{opt,r}$ minimizes asymptotically the MSE in the class of variance estimators of Definition 2. We indicate the proof for $d = 2$, the case of general d can be treated analogously. Let

$$Y = (Y_{11}, ..., Y_{1n_2}, ..., Y_{n_1 1}, ..., Y_{n_1 n_2})^T ,$$

then

$$\hat{\sigma}^2 = \sum_{l=1}^{2}\frac{\mu_l}{tr(D_l)} Y^T D_l Y =: Y^T U Y ,$$

where $D_l = \tilde{D}_l^T \tilde{D}_l$, and

$$\tilde{D}_l = \begin{pmatrix} d_0^{(l)} & \dots & d_r^{(l)} & 0 & \dots & 0 \\ & \ddots & & \ddots & & \\ & & \ddots & & \ddots & \\ 0 & \dots & 0 & d_0^{(l)} & \dots & d_r^{(l)} \end{pmatrix} \in \mathbb{R}^{(n_1 n_2 - r_l) \times n_1 n_2} \ .$$

We have $tr(D_l) = n_{R_l} = N + O(n^{d-1})$, where $n = \min\{n_1, n_2\}$. Furthermore,

$$tr(U) = 1 + o(1), \qquad tr((diag\, U)^2) = N^{-1} + o(N^{-1}) \ .$$

Hence, we find that

$$\begin{aligned} MSE(\hat{\sigma}^2) &= \sigma^4 \Big((\gamma_4 - 3) N^{-1} + 2 tr(U^2) \Big) + O(N^{-1}) \\ &= \sigma^4 \Big((\gamma_4 - 1) N^{-1} + 2 \sum_{i \neq j} u_{i,j}^2 \Big) + O(N^{-1}) \ , \end{aligned}$$

where $u_{i,j}$ denote the elements of U. Finally, $\sum_{i \neq j} u_{i,j}^2$ is minimized if $u_{i,j} = -1/(2\,r)$, $i \neq j$, up to the order of $O(n^{d-1})$ terms. Now, the non-diagonal elements of D_l are $-1/(2\,r_l)$, $l = 1, ..., L$, for $\hat{\sigma}^2_{opt,r}$ up to terms of order $O(n^{d-1})$, hence the minimum is attained for $\mu_l = r_l / r$ as required. □

References

Bissantz, N. and A. Munk (2002). A graphical selection method for parametric models in noisy inhomogeneous regression. *Mon. Not. R. Astron. Soc. 336*, 131–138.

Carroll, R. and D. Ruppert (1988). *Transforming and Weighting in Regression.* London: Chapman and Hall.

de Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probab. Th. Rel. Fields 75*, 261–277.

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

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 variance and residual pattern in nonlinear regression. *Biometrika 73*, 625–633.

Hall, P. and R. J. Carroll (1989). Variance function estimation in regression: the effect of estimating the mean. *J. R. Statist. Soc. 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., J. Kay, and D. Titterington (1991). On estimation of noise variance in two-dimensional signal processing. *Adv. Appl. Prob. 23*, 476–495.

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

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

Herrmann, E., M. Wand, J. Engel, and T. Gasser (1995). A bandwidth selector for bivariate kernel regression. *J. R. Statist. Soc. B 57*, 171–180.

Kay, J. (1988). *On the choice of the regularisation parameter in image restoration.*, Volume 301 of *Lecture Notes in Computer Science.* Springer.

Lee, J. S. (1981). Refined filtering of image noise using local statistics. *Comp. Graphics Image Processing 15*, 380–389.

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

Munk, A. (2002). Testing the lack of fit in nonlinear regression models with random toeplitz-forms. *Scand. J. Statist. 29*, 501–535.

Neumann, M. (1994). Fully data-driven nonparametric variance estimation. *Statistics 25*, 189–212.

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

Ripley, B. (1981). *Spatial Statistics.* New York: Wiley.

Ruppert, D., M. Wand, U. Holst, and O. Hössjer (1997). Local polynomial variance-function estimation. *Technometrics 39*, 262–272.

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

Spokoiny, V. (2002). Variance estimation for high-dimensional regression models. *J. Mult. Anal. 82*, 111–133.

Stoer, J. (1979). *Einführung in die Numerische Mathematik I.* Berlin: Springer.

Thompson, A., J. Kay, and D. Titterington (1991). Noise estimation in signal restoration using regularization. *Biometrika 78*, 475–488.

von Neumann, J. (1941). Distribution of the ratio of the mean squared successive difference to the variance. *Ann. Math. Statist. 12*, 367–395.

Wagner, T. (1999). *Untersuchungen über die Varianz in nichtparametrischer Regression.* Ph. D. thesis, Fakultät für Mathematik, Ruhr-Universtät Bochum, Germany.

Wand, M. and M. Jones (1995). *Kernel Smoothing.* London: Chapman and Hall.