Control and Cybernetics

vol. **42** (2012) No. 2

Performance of variance function estimators for autoregressive time series of order one: asymptotic normality and numerical study[1]

by

Piotr Borkowski, Jan Mielniczuk[2]

Institute of Computer Science,
Polish Academy of Sciences,
Jana Kazimierza 5, 01-248 Warsaw, Poland
email: piotrb@ipipan.waw.pl
email: miel@ipipan.waw.pl

Abstract: We study performance of several conditional variance estimators for an autoregressive time series which include local linear smoothers with various bandwidths, local likelihood and difference-based estimators. In the theoretical part, asymptotic normality of the local linear estimator of variance with no mixing assumptions imposed on the underlying process is proved. Moreover, numerical examples performed reveal that a two-stage local linear smoother with a bandwidth, proposed by Ruppert, Sheather and Wand, used to estimate the regression function and a simple rule of thumb bandwidth for variance estimation performs best for variances without much structure, whereas the bandwidth considered by Fan and Yao works very well for much more variable variances.

Keywords: autoregressive process, bandwidth, heteroscedasticity, integrated squared error, local linear and local maximum likelihood estimator, difference-based estimator, geometric moment contraction, variance function, volatility.

1. Introduction

We focus here on the following real valued time series $(X_t)_{t\in\mathbb{Z}}$ satisfying

$$X_{t+1} = m(X_t) + \sigma(X_t)\varepsilon_{t+1}, \qquad t \in \mathbb{Z}, \tag{1}$$

where $m(\cdot)$ and $\sigma(\cdot)$ are some real functions, $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an iid sequence such that $\mathbb{E}(\varepsilon_t) = 0$ and $\mathrm{Var}(\varepsilon_t) = 1$ and ε_{t+1} is independent of the history $\mathcal{F}_t$ of the

[1] Submitted: February 2009; Accepted: March 2012

[2] This author is also with Warsaw University of Technology, Faculty of Mathematics and Information Science

process up to time t, $\mathcal{F}_t = \sigma(\ldots, X_{-1}, X_0, \ldots, X_t)$. Such a time series is called autoregressive; in the case when m is not an affine function, it is described as a nonlinear autoregression. Note that $(X_t)_{t\in\mathbb{Z}}$ is a Markov process of order 1. A general autoregressive case involves in equation (1) $m(X_t, X_{t-1}, \ldots, X_{t-p})$ and $\sigma(X_t, X_{t-1}, \ldots, X_{t-q})$ for some $p, q \in \mathbb{N}$. The conditions under which it has a stationary limiting distribution have been extensively studied, for a discussion see e.g. Ango Nze (1992), Diaconis and Freedman (1999), conditions A1-A5 in Härdle and Tsybakov (1997), and Section 2 in Neumann and Kreiss (1998). We refer also to Lu and Jiang (2001) for multivariate extensions and the references there. Throughout the paper we assume that $(X_t)_{t\in\mathbb{Z}}$ is strictly stationary and its sample path $X_1, X_2, \ldots, X_n$ is given. Note that assumption of stationarity amounts to assuming that X_1 is generated according to the limiting stationary distribution. We assume that $(X_t)_{t\in\mathbb{Z}}$ is a causal process i.e. it can be represented as a transformation of a Bernoulli shift $\eta_t = (\ldots, \varepsilon_{t-1}, \varepsilon_t)$: $X_t = J(\eta_t)$ for a certain measurable J. We refer to Wu (2005) for a general discussion of causal processes. An intensively studied special case of (1) is the autoregressive process with ARCH(1) errors for which $\sigma^2(x) = c_0 + b_1x^2, \quad c_0 \geq 0, \quad b_1 \geq 0$. For $m(x) = 0$ condition $b_1 < 1$ implies strict stationarity.

It follows from equation (1) that $\mathbb{E}(X_{t+1}|X_t) = m(X_t)$ i.e. $m(\cdot)$ is the regression of X_{t+1} given X_t, and

$$\mathrm{Var}(X_{t+1}|X_t) = \mathbb{E}((X_{t+1} - \mathbb{E}(X_{t+1}|X_t))^2|X_t) = \mathbb{E}((X_{t+1} - m(X_t))^2|X_t) = \sigma^2(X_t),$$

provided marginal distribution of X_t has a finite second moment. Thus, $\sigma^2(\cdot)$ coincides with the conditional variance function of X_{t+1} given X_t.

In the paper we discuss estimation of the conditional variance function $\sigma^2(\cdot)$. This is often of independent interest from estimation of the regression, especially when one would like to assess heteroscedasticity of the considered dependence structure or evaluate volatility or risk. Frequently, the conditional variance is used to evaluate some related characteristics of the conditional distribution, as e.g. in Value at Risk (VaR) estimation (see, e.g., McNeil et al., 2005). Some preliminary estimates of variance are also needed to construct variance stabilizing transformation or weighted regression estimators. In image analysis algorithms for noise reduction, segmentation and clustering are based on variance estimation (see e.g. Sijbers et al. (2007)).

Let us also note that (1) can be considered as a discrete time approximation of a diffusion process when drift and diffusion functions are time invariant. Indeed, the Euler approximation with step Δ to a process $(X_t)_{t\in\mathbb{R}}$ such that $dX_t = \mu(X_t)dt + \sigma(X_t)dW_t$, with W_t being the Wiener process, is

$$X_{(i+1)\Delta} = X_{i\Delta} + \mu(X_{i\Delta})\Delta + \sqrt{\Delta}\sigma(X_{i\Delta})\varepsilon_{(i+1)\Delta}, \quad i \in \mathbb{Z}, \tag{2}$$

where $\varepsilon_{(i+1)\Delta} \sim N(0,1)$ and is independent of $X_{i\Delta}$, so that (2) corresponds to an obvious modification of (1). Thus, reliable estimation of $\sigma(\cdot)$ is important in this context (see Fan, 2005).

We also consider random design regression model such that the underlying regression function and the conditional variance function are the same as in (1) i.e. a bivariate sequence $(X_t, Y_t), t \in \mathbb{Z}$ consisting of *independent* and identically distributed random pairs such that

$$Y_t = m(X_t) + \sigma(X_t)\varepsilon_{t+1}, \qquad t \in \mathbb{Z}, \tag{3}$$

where (ε_t) satisfies conditions stated above. It is easy to see that indeed $E(Y_t|X_t) = m(X_t)$ and $\mathrm{Var}(Y_t|X_t) = \sigma^2(X_t)$. Models (1) and (3) will be denoted $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively.

In model $\mathcal{M}_1$, marginal density depends in a complex way on the regression and the conditional variance function. This makes data generation in $\mathcal{M}_2$ with the same marginal density as in $\mathcal{M}_1$ infeasible. Thus, in order to investigate how dependence of autoregression process influences performance of variance estimators we compared their performance in model $\mathcal{M}_1$ and in a modified model $\mathcal{M}_2$. Namely, we modified (3) in the following way. For a given sample path of autoregressive stationary process (1) $X_1 = x_1, \ldots, X_n = x_n$ we consider

$$Y_t^* = m(x_t) + \sigma(x_t)\varepsilon_t^*, \quad t = 1, 2, \ldots, n, \tag{4}$$

where $\varepsilon_1^*, \ldots, \varepsilon_n^*$ are independent copies of ε_1. Thus, given the values $X_1 = x_1, \ldots, X_n = x_n$, $Y_1^*, \ldots, Y_n^*$ are conditionally independent. Model (4) will be called $\mathcal{M}_3$. Fig. 1 shows the difference between data generation in models $\mathcal{M}_1$ and $\mathcal{M}_3$. The difference between models $\mathcal{M}_2$ and $\mathcal{M}_3$ is that in the latter model the predictors for different observations are not independent. However, they have the same distribution as in $\mathcal{M}_2$ and $\mathcal{M}_1$. Note also that, due to the construction, $(Y_1^*, \ldots, Y_{n-1}^*)$ and $(X_2, \ldots, X_n)$ are responses pertaining to the same regression, conditional variance and values of predictors.

$$\begin{tikzcd}
X_1 \arrow[r, "\varepsilon_2"] \arrow[d, "\varepsilon_1^*"] & X_2 \arrow[r, "\varepsilon_3"] \arrow[d, "\varepsilon_2^*"] & \ldots \arrow[r, "\varepsilon_n"] & X_n \arrow[d, "\varepsilon_n^*"] \\
Y_1^* & Y_2^* & \ldots & Y_n^*
\end{tikzcd}$$

Figure 1: Data generation in models $\mathcal{M}_2$ and $\mathcal{M}_3$

In all considered setups we do not assume any prior knowledge of the parameters of the models: $m(\cdot)$, $\sigma(\cdot)$ and the marginal distribution of X_t, and thus we adopt a nonparametric approach in order to estimate $\sigma^2(\cdot)$. We prove asymptotic normality of a local linear smoother of the variance without imposing hard-to-verify mixing assumptions on the underlying process (X_t). Our assumptions on dependence of (X_t) boil down to the conditions that a conditional density of X_{t+1} given X_t is a Lipschitz function and (X_t) is geometrically moment contracting (see Section 3 for definition). Easily verifiable conditions for the last property exist. Moreover, we investigate how conditional variance estimators designed for the random design setup $\mathcal{M}_2$ perform for the autoregressive model $\mathcal{M}_1$. Apart from scattered examples there is no in-depth study addressing this issue. In particular, no clear picture is available on comparison of their performance. In order to fill this gap we provide, in particular, some evidence that estimators with parameters chosen for the iid case indeed also work for autoregressive case. This phenomenon named *whitening by windowing principle* usually describes coincidence of asymptotic laws or the same asymptotic behavior of some theoretical measure of performance in both cases. Here we argue that it also holds for samples of moderate size when Empirical Integrated Squared Error (EISE) is considered as a goodness-of-fit measure. Moreover, we show that in considered examples a two-stage local linear smoother with a bandwidth, proposed by Ruppert, Sheather and Wand (1995), used to estimate the regression function and a simple rule of thumb bandwidth used to estimate the variance performs best for variances without much structure, whereas the bandwidth considered in Fan and Yao (1998) works very well for much more variable variances. The simulations reveal also that the local linear estimator of the variance works comparably to its benchmark version for which the regression function is known.

The paper is structured as follows. In Section 2 the considered estimators of variance function are introduced and motivated. In Section 3 we prove asymptotic normality of a two-stage local-linear estimator of $\sigma^2(\cdot)$ under a new set of conditions which do not include mixing of autoregressive process. In Section 4 we describe the performed simulation study and its results.

2. Estimators of the conditional variance function

We will briefly describe the most frequently used estimators of $\sigma^2(\cdot)$ in the random design regression model $\mathcal{M}_2$. Assume that the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, pertaining to a stationary solution of (3), is available.

The main idea underlying the first discussed approach to estimate $\sigma^2(\cdot)$ is to view it as the regression function of $(Y_t - m(X_t))^2$ given X_t and use one of many available regression tools for its estimation. As $m(\cdot)$ is unknown, it also has to be approximated, and thus the estimators are obtained by two-step procedure in which appropriate estimator $\hat{m}$ is constructed first and residuals $e_i = Y_i - \hat{m}(X_i)$ are obtained. Then, the regression function $(Y_t - m(X_t))^2$ given X_t is estimated using squared residuals $e_i^2 = (Y_i - \hat{m}(X_i))^2$. The differences between the methods consist in the way the two regressions involved are estimated. Here we discuss two main classes of estimators based on this methodology: local linear and log-linear approach. These are counterparts of two basic procedures of regression estimation, namely weighted least squares and maximum likelihood, applied to the squared residuals. Moreover, we present modified Rice estimator based on apparently different methodology, which, as it turns out, is also based on squared residuals from a specific regression fit. Finally, let us mention an idea of modeling the mean and the variance functions simultaneously by penalized likelihood method with penalization related to roughness of m and σ (see Yau and Kohn, 2003, and Yuan and Wahba, 2004).

2.1. Local linear estimator

The method applies local linear smoother on both levels: when estimating the regression and the conditional variance function. Namely, $\hat{m}(x) = \hat{\beta}_0(x)$, where $\hat{\beta}_0(x)$ is defined as

$$(\hat{\beta}_0(x), \hat{\beta}_1(x)) = \text{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1(X_i - x))^2 K_{h_2}(X_i - x), \tag{5}$$

where $K_h(s) = h^{-1}K(s/h)$, K is a probability density function chosen by the experimenter and h is a smoothing parameter (a bandwidth). Recognizing that (5) is a weighted least squares problem with the design $n \times 2$ matrix X consisting of the column of 1s and the column consisting of $X_i - x, \quad i = 1, 2, \dots, n$, where the $n \times n$ diagonal matrix of weights $W = \text{diag}(K_{h_2}(X_1 - x), \dots, K_{h_2}(X_n - x))$, the solution can be written as

$$(\hat{\beta}_0(x), \hat{\beta}_1(x)) = (X'WX)^{-1}X'WY$$

with $Y = (Y_1, Y_2, \dots, Y_n)'$ and X' denoting transposition of X. An explicit form of $\hat{\beta}_0(x)$ is given by

$$\hat{m}(x) = \frac{s_2(x,h_2)t_0(x,h_2) - s_1(x,h_2)t_1(x,h_2)}{s_0(x,h_2)s_2(x,h_2) - s_1(x,h_2)^2}, \tag{6}$$

where $s_j(x,h) = \sum_{k=1}^{n}(X_k - x)^j K_h(X_k - x)$ and $t_j(x,h) = \sum_{k=1}^{n}(X_k - x)^j K_h$ $(X_k - x)Y_i$. Many proposals of bandwidth choice for estimation of $m(\cdot)$ in the homoscedastic case exist. For a choice of h_2 used in regression estimation we consider here first the method proposed by Ruppert, Sheather and Wand (1995) (whenceforth called RSW bandwidth) and implemented as dpill procedure in R package *kernsmooth*.

The estimator of $\sigma^2(\cdot)$ is defined as the local linear smoother applied to the transformed data (X_i, e_i^2), where

$$e_i^2 = (Y_i - \hat{m}(X_i))^2. \tag{7}$$

The linear smoother at the second stage uses a kernel $\widetilde{K}$, possibly different from the kernel K and a bandwidth h_1. As the choice of kernel does not significantly change the properties of the resulting estimator, we choose $\widetilde{K} = K$ in our investigations. A choice of h_1 is based on asymptotic considerations which imply that under appropriate conditions (see Yu and Jones, 2004)

$$\mathbb{E}(\hat{\sigma}^2(x)|X_1, \dots X_n) - \sigma^2(x) = \frac{1}{2}(\sigma^2)''(x) \int s^2 K(s)\, ds\, h_1^2 + o_P(h_1^2) \tag{8}$$

and

$$\mathrm{Var}(\hat{\sigma}^2(x)|X_1, \dots, X_n) = \sigma^4(x) \frac{\mathbb{E}(\varepsilon^2 - 1)^2}{n h_1 f(x)} \int K^2(s)\, ds + o_P((nh_1)^{-1}), \tag{9}$$

where $f(\cdot)$ is a density of distribution of X_t. Note that the main terms of asymptotic bias and variance of $\hat{\sigma}^2(\cdot)$ are the same as if $m(\cdot)$ were known. This can be easily seen by considering $(Y - m(X))^2$ as the response in the regression model

$$(Y - m(X))^2 = \sigma^2(X) + \sigma^2(X)(\varepsilon^2 - 1) \tag{10}$$

and using standard results on regression function estimation for the local linear smoothers (see Fan and Gijbels, 1996). Note that $\mathbb{E}(\sigma^2(X)(\varepsilon^2 - 1)|X = x) = 0$. Thus $\hat{\sigma}^2$ behaves asymptotically as a benchmark estimator being a local linear smoother based on $(Y_i - m(X_i))^2 = \sigma^2(X_i)\varepsilon_i^2$ and whence it is adaptive to an unknown regression function m. In this context it is important to remark that Wang et al. (2008) proved that minimax rates of convergence of variance estimator for fixed design model are the same for known and unknown m provided it has at least 1/4 derivatives; see also Gendre (2008). Note also that as $\hat{\sigma}^2(\cdot)$ is a local smoother, its asymptotic properties for the model (10) and fixed x are the same as for its homoscedastic analogue with $\sigma^2 \equiv \sigma^2(x)\mathbb{E}(\varepsilon^2 - 1)$.

Assume that f is positive on a compact interval $[a, b]$. Then the bandwidth $h_{1,opt}$ minimizing the main term of asymptotic expansion of a weighted MISE $\mathbb{E}\int_a^b (\hat{\sigma}^2(x) - \sigma^2(x))^2 f(x)\, dx$ is of the form

$$h_{1,opt} = \left(\frac{\int_a^b [(\sigma^2)''(x)]^2 f(x)\, dx\, m_2^2(K)}{\int_a^b \sigma^4(x)\, dx\, R(K) C_\varepsilon} \right)^{-1/5} n^{-1/5}, \tag{11}$$

where $C_\varepsilon = \mathbb{E}(\varepsilon^2 - 1)^2$, $R(K) = \int K^2(s)\, ds$ and $m_2(K) = \int s^2 K(s)\, ds$. The following plug-in estimators of $h_{1,opt}$ are considered:

(i) $h_{1,opt}^{(2)}$ is plug-in estimate obtained by fitting a third order polynomial $\sum_{i=0}^3 c_i x^i$ to $\{e_j^2\}_{j=1}^n$ in order to obtain a preliminary estimator of $\sigma^2(\cdot)$. We call the estimated coefficients $\hat{c}_i$ and use $\int_a^b (\sum_{j=0}^3 \hat{c}_j x^j)^2\, dx$ as an estimator of $\int_a^b \sigma^4(s)\, ds$ and $n^{-1}\sum_{i=1}^n (2\hat{c}_2 + 6\hat{c}_3 X_i)^2 I\{X_i \in [a, b]\}$ as an estimator of

$\int_a^b [(\sigma^2)''(s)]^2 f(s)\,ds$. In the simulation examples we assumed that ε is $N(0,1)$-distributed and thus $C_\varepsilon = 2$. The resulting estimator, which uses $h_{1,opt}^{(2)}$ at the second stage, will be denoted $\hat{\sigma}_{LL2}^2$.
(ii) $h_{1,opt}^{(1)}$ is defined analogously, but a preliminary fit $\exp\{\sum_{i=0}^3 c_i x^i\}$ to $\{e_j^2\}$ is applied in order to obtain an estimator of $\sigma^2(\cdot)$. The resulting estimator will be denoted $\hat{\sigma}_{LL1}^2$. This bandwidth is also used by Yu and Jones (2004) for a local log-linear fit described below. Note that as weighted MISE

$$MISE = \mathbb{E}\int_a^b (\hat{\sigma}^2(x) - \sigma^2(x))^2 f(x)\,dx$$

is considered as a theoretical measure of performance, in simulations we used a mean of empirical integrated squared errors (EISEs), where EISE equals to $m^{-1}\sum_{i=1}^n (\hat{\sigma}^2(X_i) - \sigma^2(X_i))^2 I_{\{X_i \in [a,b]\}}$, with $m = \#\{X_i \in [a,b]\}$.

In order to gauge the effect of bandwidth choice on performance of locally linear variance estimators we also consider two additional bandwidth proposals. The first one is preasymptotic substitution method used in the context of variance estimation by Fan and Yao (1998) (whenceforth abbreviated to FY) and introduced in Fan and Gijbels (1995). The method is used twice: first to estimate the regression function and then the variance, using the squared residuals from the regression fit. Authors' open source program has been used for the calculations (source code can be found at `http://www.orfe.princeton.edu/~jqfan/papers/pub/constband.c`). The resulting estimator will be called $\hat{\sigma}_{LL3}^2$. Moreover, $\hat{\sigma}_{LL4}^2$ will denote a *k-nn* local linear smoother using bandwidths based on $k(n)$ nearest neighbours. More specifically, both bandwidths are defined as the distance from x to its $k(n)$th nearest neighbour, where $k(n)$ is a predetermined sequence of integers.

There are several papers studying similar proposals to those discussed above. One possibility is to study kernel estimators instead of linear smoothers, this, however, results in significantly poorer performance near the boundaries. 'Direct' estimator of $\sigma^2(\cdot)$ is constructed by plugging local-linear estimators of the conditional second moment and the regression into equality $\sigma^2(x) = \mathbb{E}(Y^2|X = x) - (\mathbb{E}(Y|X = x))^2$; such approach is adopted e.g. in Härdle and Tsybakov (1997). This turns out to lead to a more biased estimator than the presented above (see Fan and Yao, 1998, p. 649). Chen et al. (2009) estimated $\sigma(\cdot)$ by means of the linear smoother using $\log(r_i + n^{-1})$ instead of r_i^2. For the Euler approximation (2) Stanton (1997) used a kernel estimator relying on approximate equality $\mathbb{E}((X_{(i+1)\Delta} - X_{i\Delta})^2/\Delta|X_{i\Delta}) \approx \sigma^2(X_{i\Delta})$. However, this approximation holds only for small Δ and in general one has to base the estimation on squared residuals $(X_{(i+1)\Delta} - X_{i\Delta} - \hat{\mu}(X_{i\Delta}))^2/\Delta$. Let us also mention that Ćwik et al. (2000) studied construction of confidence bands for integrated conditional variance using kernel and linear smoothers with fixed and *k-nn* bandwidths.

2.2. Local log-linear estimator

The estimator is based on the idea of localizing the (conditional) loglikelihood in the case when errors are normally distributed. Observe that we have then

$$-2\log\mathcal{L}(Y_1,Y_2,\ldots,Y_n|X_1,X_2,\ldots,X_n)=\sum_{i=1}^{n}\left(\frac{(Y_i-m(X_i))^2}{\sigma^2(X_i)}+\log\sigma^2(X_i)\right)+n\log 2\pi. \quad (12)$$

Omitting the constant and assuming that $\sigma^2(s)$ behaves approximately as $\exp(v_0+v_1(s-x))$ in the neighborhood of x, we obtain the following localized version of (12) introduced in Yu and Jones (2004):

$$\sum_{i=1}^{n}\Big((Y_i-m(X_i))^2\exp(-v_0-v_1(X_i-x))+v_0+v_1(X_i-x)\Big)K_{h_1}(X_i-x). \quad (13)$$

When the squared errors $(Y_i-m(X_i))^2$ are replaced by squared residuals e_i^2 defined in (7), criterion (13) is known as pseudolikelihood. The above expression is minimized with respect to v_0 and v_1. Call the resulting values $\hat{v}_0(x)$ and $\hat{v}_1(x)$. Then, $\hat{\sigma}^2(x):=\exp\hat{v}_0(x)$. In the following m is chosen as the local linear smoother with RSW bandwidth. The main result in Yu and Jones (2004) states that the formula for asymptotic variance of log-linear smoother remains as in (9) whereas the asymptotic bias has the form

$$\mathbb{E}(\hat{\sigma}^2(x)|X_1,\ldots X_n)-\sigma^2(x)=\frac{1}{2}(\log\sigma^2)''(x)\sigma^2(x)m_2(K)\,h_1^2+o_P(h_1^2).$$

This results in $h_{1,opt}$ as in (11) but with $\int_a^b[(\sigma^2)''(x)]^2f(x)\,dx$ replaced by

$$\int_a^b[(\log\sigma^2)''(x)]^2\sigma^2(x)f(x)\,dx.$$

Bandwidth $h_{1,opt}$ is estimated by plug-in method analogously to (ii) in Section 2.1 i.e. $\exp\{\sum_{i=0}^3 c_ix^i\}$ is fitted to $\{e_j^2\}$ in order to obtain a preliminary estimator of $\sigma^2(\cdot)$. We call the final variance estimator log-linear or local maximum likelihood estimator $\hat{\sigma}^2_{LML}$. As a technical aside consider the problem of minimization of (13) and note that taking derivatives with respect to v_0 and v_1 yields

$$\exp(-v_0)=A_1/A_4, \qquad \exp(-v_0)=A_2/A_3,$$

where

$$\begin{aligned}
A_1 &= \sum_{i=1}^{n}K_{h_1}(x-X_i), A_2=\sum_{i=1}^{n}K_{h_1}(x-X_i)(x-X_i),\\
A_3 &= \sum_{i=1}^{n}K_{h_1}(x-X_i)(x-X_i)(Y_i-\hat{m}(X_i))^2\exp(-v_1(X_i-x)),\\
A_4 &= \sum_{i=1}^{n}K_{h_1}(x-X_i)(Y_i-\hat{m}(X_i))^2\exp(-v_1(X_i-x)).
\end{aligned}$$

Instead of minimizing (13) with respect to v_0 and v_1 one finds a zero of $A_1/A_4 - A_2/A_3$ and then, denoting the resulting values of A_i by $\hat{A}_i$, one gets $\hat{\sigma}^2_{LML} = \hat{A}_4/\hat{A}_1$.

2.3. Modified Rice estimators

Assume that f is supported on the finite interval $[a, b]$ and let $Y_{[1]}, Y_{[2]}, \ldots, Y_{[n]}$ be concomitants of ordered explanatory variables $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$. The original Rice (1984) estimator has been designed to estimate the variance $\sigma^2 \equiv \sigma^2(\cdot)$ in the homoscedastic case and has the following form

$$\hat{\sigma}^2_R = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{[i+1]} - Y_{[i]})^2, \tag{14}$$

which is an average of squares of differenced concomitants. For a more general definition involving weighted differences see, e.g., Dette et al. (1998). The modified Rice estimator is based on the following approximate equality: for $X_i = x \approx x' = X_j$ and $i \neq j$ we have

$$\mathbb{E}((Y_i - Y_j)^2 | X_i = x, X_j = x') \approx \mathbb{E}((\varepsilon_i \sigma(x) - \varepsilon_j \sigma(x'))^2) \approx 2\sigma^2(x). \tag{15}$$

Thus, by regressing $2^{-1}(Y_{[j+1]} - Y_{[j]})^2$ on $X_{j:n}$ one gets heuristically justified estimator of $\sigma^2(\cdot)$. We define a local version of (14), called henceforth modified Rice estimator $\hat{\sigma}^2_{MR1}(\cdot)$, as a local linear smoother based on a sample $(X_{j:n}, 2^{-1}(Y_{[j+1]} - Y_{[j]})^2)$, $j = 1, 2, \ldots, n-1$. The obvious advantage of such estimator is that the regression estimators do not enter its definition and thus estimation of m is entirely avoided. A kernel estimator of $\sigma^2(\cdot)$ based on the same idea is defined in Levine (2006), see also Cai and Wang (2008) and Cai et al. (2009). Wang et al. (2008) proved asymptotic minimaxity of a wavelet estimator of variance function based on $(Y_{[j+1]} - Y_{[j]})^2$.

In order to put (14) and $\hat{\sigma}^2_{MR1}$ into broader perspective note, however, that $Y_{[i]} - Y_{[i+1]}$ is a regression residual at $X_{i:n}$ when $\hat{m}(X_{i:n})$ is defined as $Y_{[i+1]}$, i.e. value of the response at the adjacent point is used as regression estimate. We also consider the following refinement of this procedure due to Gasser et al. (1986) in which residual at $X_{i:n}$ is defined as the difference between $Y_{[i]}$ and the value at $X_{i:n}$ of the line joining the points $(X_{i-1:n}, Y_{[i-1]})$ and $(X_{i+1:n}, Y_{[i+1]})$, namely

$$\begin{aligned} D_i^0 &= Y_{[i]} - \frac{X_{i+1:n} - X_{i:n}}{X_{i+1:n} - X_{i-1:n}} Y_{[i-1]} - \frac{X_{i:n} - X_{i-1:n}}{X_{i+1:n} - X_{i-1:n}} Y_{[i+1]} \\ &=: Y_{[i]} + a_i Y_{[i-1]} + b_i Y_{[i+1]}. \end{aligned} \tag{16}$$

The normalized version of D_i^0 is defined as $D_i = (1 + a_i^2 + b_i^2)^{-1/2} D_i^0$. Note that $-a_i Y_{[i-1]} - b_i Y_{[i+1]}$ is a predictor of regression value $m(X_{i:n})$. We define $\hat{\sigma}^2_{MR2}(\cdot)$ as a local linear smoother based on $(X_{i:n}, D_i^2)$.

Observe that in the case of autoregression $2^{-1}(Y_{[j+1]} - Y_{[j]})^2$ equals $2^{-1}(X_{d(j+1)+1} - X_{d(j)+1})^2$, where $d(j)$ is j^{th} antirank i.e. the index of an observation equal to j^{th} order statistic $X_{j:n}$. Then, approximate equality (15) still holds as $\sigma(X_{j:n})\varepsilon_{d(j)+1}$ and $\sigma(X_{j+1:n})\varepsilon_{d(j+1)+1}$ are conditionally independent given $X_{j:n}$ and $X_{j+1:n}$.

As the form of the asymptotic bias and the variance of $\hat{\sigma}^2_{MR1}(\cdot)$ and $\hat{\sigma}^2_{MR2}(\cdot)$ are not known (this is even true for (14) in the case of random design regression), we used dpill procedure applied to $(X_{i:n}, 2^{-1}(Y_{[j+1]} - Y_{[j]})^2)$ and $(X_{i:n}, D_i^2)$ to calculate bandwidths for the respective estimators considered in numerical experiments. One expects such bandwidths to undersmooth as, e.g., the variance of $\hat{\sigma}^2_{MR1}$ is likely to be larger than (9) due to non negligible contribution of *positive* covariances of $(1/2)(Y_{[j+1]} - Y_{[j]})^2$ and $(1/2)(Y_{[j]} - Y_{[j-1]})^2$. In order to account for this we compared the variances of local linear smoother and Rice estimator for the fixed uniform design case. A routine calculation shows that the last variance is $(nh_1)^{-1}\sigma^4(x)\mathbb{E}(\varepsilon^4)\int K^2(s)ds/f(x) + o_P((nh_1)^{-1})$, which means that the ratio of asymptotic variances is $3/2$ for the normal errors. In view of this, we used RSW bandwidth multiplied by $(3/2)^{1/5}$ as a smoothing parameter for modified Rice estimators. Crossvalidation has been also tried as a method of bandwidth choice, but it consistently yielded very small bandwidths in this case, possibly due to dependence of observations.

3. Asymptotic normality of local linear smoother of conditional variance

In this section we study asymptotic normality of the two-step local linear variance estimator, defined in Section 2.1 under assumptions tailored for the autoregressive process (1). We assume that (X_n) has Bernoulli shift representation $X_n = J(\ldots, \varepsilon_{n-1}, \varepsilon_n)$ for some measurable J. Let $||X||_p = (\mathbb{E}|X|^p)^{1/p}$. (X_n) is called geometric moment contracting (GMC) if for some $q > 1$ and $0 < r < 1$ $||X_n - X_n^*||_q = \mathcal{O}(r^n)$, where $X_n^* = J(\ldots, \varepsilon_{-1}, \varepsilon_0^*, \ldots, \varepsilon_{n-1}, \varepsilon_n)$ and ε_0^* is an independent copy of ε_0. A simple sufficient condition for GMC is $||L_\varepsilon||_q < 1$ where $L_\varepsilon := \sup_{x\neq x'} |R(x,\varepsilon) - R(x',\varepsilon)|/|x - x'|$ and $R(x,\varepsilon) = m(x) + \sigma(x)\varepsilon$, see Wu and Shao (2004). GMC condition is needed to ensure that supremum $\sup_x ||f_1(x|X_n) - f_1(x|X_n^*)||_q$ be summable over n, where $f_1(x|y)$ be a conditional density of X_{t+1} given X_t.

We will prove the result on asymptotic normality of the local linear estimator of $\sigma^2(\cdot)$ for stationary autoregressive time series. The version of this result was proved in Fan and Yao (1998) for a stationary bivariate process (X_i, Y_i) such that $Y_i = m(X_i) + \sigma(X_i)\varepsilon_{i+1}$ provided it is absolutely regular with coefficients $\beta(j)$ satisfying $\sum_{j=1}^{\infty} j^2\beta^{\delta/(1+\delta)}(j) < \infty$, where $\delta \in [0,1)$ is such that $E|Y|^{4(1+\delta)} < \infty$. Here we prove this result directly for the autoregressive process without using any mixing conditions and a slightly different set of conditions on bandwidths. The proof exploits the martingale structure of the main term in the decomposition of $\hat{\sigma}^2(\cdot)$. Let $\xrightarrow{\mathcal{D}}$ denote convergence in distribution and $x \in \mathbb{R}$ be an arbitrary fixed point. We assume throughout that $h_1, h_2 \to 0$ and $nh_1, nh_2 \to \infty$.

THEOREM 1 *Assume that $X_n = J(\dots, \varepsilon_{n-1}, \varepsilon_n)$ satisfies (1) and moreover:*
(i) m and σ are twice continuously differentiable in a neighborhood of x, σ and m/σ are Lipschitz continuous on $\mathbb{R}$;
(ii) a density function $f_\varepsilon(x)$ of ε_t is Lipschitz continuous and bounded;
(iii) $(X_t)_{t\in\mathbb{Z}}$ is geometric moment contracting;
(iv) $\inf_{x\in\mathbb{R}} \sigma(x) > 0$ and $f(x) > 0$;
(v) K is symmetric, bounded, compactly supported with support in $[-1, 1]$;
(vi) $\mathbb{E}|\varepsilon_1|^p < \infty$ for some $p \geq 4$;
(vii) h_1 and h_2 satisfy the following conditions:

(a) $h_1/h_2 = \mathcal{O}(1)$ (b) $n^{1-2/p}h_2/\log n \to \infty$ (c) $n{h_2}^5 \to 0$. Then

$$(nh_1)^{1/2}(\hat{\sigma}^2(x) - \sigma^2(x)) \xrightarrow{D} N(0, v(x)), \tag{17}$$

where

$$v(x) = \frac{\sigma^4(x)\mathbb{E}((\varepsilon_1^2 - 1)^2)\int K^2(s)ds}{f(x)}. \tag{18}$$

Remark. (a) The proof of Theorem 1 indicates that condition (a) of (vii) can be replaced by a weaker condition $h_1^{1/2}\log n/n^{1/2}h_2 \to 0$. Note that this condition, without a $\log n$ term is actually imposed in Fan and Yao (1998), where it is needed to deal with remainder terms of order $((nh_2)^{-1})$ in their (A2.5). Moreover, for $p = 4$ condition (b) reduces to $nh_2^2/\log^2 n \to \infty$ whereas Fan and Yao (1998) assume that liminf $nh_i^4 > 0$, $i = 1, 2$. Condition (c) is the usual condition for negligibility of asymptotic bias. It can be seen by analyzing the main term I_1 in the proof that when $nh_i^5 \to C_{h_i}$, $i = 1, 2$ then the result still holds with a mean of asymptotic distribution equal to $C_{h_1}^{1/2}(\sigma^2(x))'' \int K(s)s^2\,ds/2$.
(b) As it was mentioned, summability of $\sup_x ||f_1(x|X_n) - f_1(x|X_n^*)||_q$, which is a consequence of GMC property is needed in the proof. It also follows, however, from local property of $\hat{\sigma}^2(\cdot)$ that when a weaker sufficient condition is used, namely $\sup_{y\in U_x} ||f_1(y|X_n) - f_1(y|X_n^*)||_q$ is summable for some $q > 1$, where U_x denotes some open neighborhood of x, then assumptions of the Lipschitz continuity of $\sigma(\cdot)$, $m/\sigma(\cdot)$ and f_ε can be dropped (see (21) and (25)). Finally, note that the Lipschitz continuity of m/σ follows e.g. from the Lipschitz continuity of both σ and m when σ is bounded away from 0 and m is bounded.
(c) As (X_t) is weakly dependent, we conjecture that the bias and the variance of the local linear smoother $\hat{\sigma}^2(x)$ in the autoregressive model $\mathcal{M}_1$ have the same rates as in the random design model $\mathcal{M}_2$. This is also supported by empirical evidence in Section 4 showing that the effect of dependence on MISE of $\hat{\sigma}^2(x)$ is negligible. Then it follows from (8) and (9) that in model $\mathcal{M}_2$

$MSE(\hat{\sigma}^2(x)) \sim C_1/(nh_1)+C_2h_1^4$, the MSE-optimal bandwidth is of order $n^{-1/5}$ and the resulting MSE of order $n^{-4/5}$. It is known (see Wang et al., 2008, Remarks 2 and 3, p. 650) that for sufficiently smooth m the minimax rate of MSE over a ball of twice differentiable standard deviations is attained by the linear smoother and coincides with the rate for a fixed standard deviation. Given that, it is also conjectured that the minimax rate in the autoregressive case under the stated assumptions is $n^{-4/5}$ and is attained by the linear smoother. The problem of influence of bandwidth h_1 on the performance of $\hat{\sigma}^2(x)$ is also addressed by means of simulations in Section 4.

Proof. C will denote a generic constant, a value of which may vary. We outline how the main terms in the decomposition (A2.3) in Fan and Yao (1998) can be directly dealt with in the case of autoregressive process without resorting to assumptions on mixing. The decomposition is as follows:

$$\begin{aligned}&\hat{\sigma}^2(x)-\sigma^2(x)=\\&I_1+I_2-I_3+I_4+\mathcal{O}_P(h_1)(|I_1+I_2-I_3+I_4|+|I_1'+I_2'-I_3'+I_4'|),\end{aligned}\tag{19}$$

where

$$\begin{aligned}I_1&=\frac{1}{nh_1f(x)}\sum_{i=1}^n K\Big(\frac{X_i-x}{h_1}\Big)\{\sigma^2(X_i)-\sigma^2(x)-(\sigma^2(x))'(X_i-x)\},\\I_2&=\frac{1}{nh_1f(x)}\sum_{i=1}^n K\Big(\frac{X_i-x}{h_1}\Big)\sigma^2(X_i)(\varepsilon_{i+1}^2-1),\\I_3&=\frac{2}{nh_1f(x)}\sum_{i=1}^n K\Big(\frac{X_i-x}{h_1}\Big)\sigma(X_i)\varepsilon_{i+1}\{\hat{m}(X_i)-m(X_i)\},\\I_4&=\frac{1}{nh_1f(x)}\sum_{i=1}^n K\Big(\frac{X_i-x}{h_1}\Big)\{\hat{m}(X_i)-m(X_i)\}^2,\end{aligned}\tag{20}$$

and I_j' is defined in the same way as I_j with additional multiplicative factor $h_1^{-1}(X_i-x)$ in the ith summand.

Note that it is enough to prove that $(nh_1)^{1/2}I_2 \xrightarrow{D} N(0,v(x))$, $(nh_1)^{1/2}I_i \xrightarrow{P} 0$ for $i=1,3,4$ and $(nh_1)^{1/2}I_i' \xrightarrow{P} 0$ for $1\le i\le 4$.

In order to deal with the term I_1 observe that as K is bounded and compactly supported, Markov's inequality implies that $(nh_1)^{-1}\sum K((X_i-x)/h_1)=\mathcal{O}_P(1)$. This, together with two term Taylor decomposition and $nh_1^5\to 0$ yields that $I_1=o_P((nh_1)^{-1/2})$.

Consider now term I_2. Let $f_1(\cdot|y)$ denote a conditional density of X_{k+1} given $X_k=y$. Obviously in view of (1), $f_1(x|y)=\sigma(y)^{-1}f_\varepsilon((x-m(y))/\sigma(y))$ and $f_1(x|y)$ bounded in x as $f_\varepsilon(\cdot)$ is bounded and $\inf_{s\in\mathbb{R}}\sigma(s)>0$. Moreover, since both $m(\cdot)$ and $\sigma(\cdot)$ satisfy the Lipschitz condition, it is easily seen that

$$|f_1(x|y)-f_1(x|y^*)|\le C(1+|x-m(y)|)|y-y^*|.\tag{21}$$

In order to show that $(nh_1)^{1/2} I_2 \xrightarrow{D} N(0, v(x))$, observe that since I_2 is a sum of martingale differences with respect to $\mathcal{F}_i = \sigma(\ldots, \varepsilon_{i-1}, \varepsilon_i)$, in view of martingale CLT (see, e.g., Chow and Teicher, 1998), it is enough to check that

$$\begin{aligned} &\frac{1}{nh_1} \sum_{i=1}^{n} \mathbb{E}\Big(K^2\Big(\frac{X_i - x}{h_1}\Big)\sigma^4(X_i)(\varepsilon_{i+1}^2 - 1)^2 | \mathcal{F}_i\Big) = \\ &\frac{\lambda^2}{nh_1} \sum_{i=1}^{n} \sigma^4(X_i) K^2\Big(\frac{X_i - x}{h_1}\Big) \xrightarrow{P} f(x)^2 v(x), \end{aligned} \tag{22}$$

where $\lambda^2 := E(\varepsilon_{i+1}^2 - 1)^2$ and to verify conditional Lindeberg's condition. LHS of (22) equals to

$$\begin{aligned} &\frac{\lambda^2}{nh_1} \sum_{i=1}^{n} \Big[K^2\Big(\frac{X_i - x}{h_1}\Big)\sigma^4(X_i) - E(K^2\Big(\frac{X_i - x}{h_1}\Big)\sigma^4(X_i)|\mathcal{F}_{i-1})\Big] + \\ &+ \frac{\lambda^2}{nh_1} \sum_{i=1}^{n} [E(K^2\Big(\frac{X_i - x}{h_1}\Big)\sigma^4(X_i)|\mathcal{F}_{i-1}) - E(K^2\Big(\frac{X_i - x}{h_1}\Big)\sigma^4(X_i))] + \\ &+ \frac{\lambda^2}{h_1} E(K^2\Big(\frac{X_1 - x}{h_1}\Big)\sigma^4(X_1)) =: \lambda^2 (\sum_{i=1}^{n} M_i + \sum_{i=1}^{n} N_i + D_n). \end{aligned} \tag{23}$$

Obviously, $D_n \to f(x)\sigma^4(x) \int K^2(u)du$ and thus we check that the two first sums tend to 0 in probability. Consider the first sum. As f_1 and K are bounded and $\sigma(\cdot)$ is bounded in a neighborhood of x, it is easily seen that $|M_i| \le C/nh_1$ and $\sum_{i=1}^{n} \mathbb{E}(M_i^2 | \mathcal{F}_{i-1}) \le C/nh_1$. Then it follows from Freedman's (1975) inequality that for $\lambda_n = nh_1 a$, with sufficiently small positive a we have

$$P(\sum_{i=1}^{n} M_i \ge \varepsilon) \le \mathbb{E} \exp(\lambda_n \sum_{i=1}^{n} M_i) / \exp(\lambda_n \varepsilon) \le \exp(nh_1(Ca^2 - a\varepsilon)).$$

For $a < \varepsilon/C$ the bound tends to 0. For the second term observe that

$$\begin{aligned} &\mathbb{E}|\sum_{i=1}^{n} N_i| \le \frac{1}{nh_1} \int K^2\Big(\frac{u - x}{h_1}\Big)\sigma^4(u)\mathbb{E}|H_n(u)|du \le \\ &\frac{1}{nh_1} \int K^2\Big(\frac{u - x}{h_1}\Big)\sigma^4(u)||H_n(u)||_q du \le \frac{C}{n} \sup_{u \in [x - h_1, x + h_1]} ||H_n(u)||_q, \end{aligned} \tag{24}$$

where $H_n(u) = \sum_{i=1}^{n} f_1(u|X_{i-1}) - f(u)$. Lemma 3(i) in Wu et al. (2010) in conjunction with Jensen's inequality implies that the last quantity is bounded by $C\Theta_q^{1/\min(2,q)}(n)/n$, where

$$\begin{aligned}\Theta_q(n) &= \sum_{j\in\mathbb{Z}}(\sum_{i=1-j}^{n-j}|\theta_q(i)|)^{\min(2,q)},\\ \theta_q(i) &= 2\sup_{u\in[x-h_1,x+h_1]}||f_1(u|X_i)-f_1(u|X_i^*)||_q,\end{aligned} \tag{25}$$

$X_i^* = J(\ldots,\varepsilon_{-1},\varepsilon_0^*,\varepsilon_1,\ldots,\varepsilon_i)$ with ε_0^* being an independent copy of ε_0.

In view of (21) $\theta_q(i) = \mathcal{O}(||X_i - X_i^*||_q)$ and it follows from assumption (iii) that $||X_i - X_i^*||_q = \mathcal{O}(r^i)$ for some $0 < r < 1$. Then it is easily seen that $\Theta_q(n) = \mathcal{O}(n)$ and thus $\mathbb{E}|\sum_{i=1}^n N_i| \to 0$. Lindeberg's condition routinely follows from boundedness of K and local boundedness of $\sigma(\cdot)$. Consider now I_3. Using (A2.2) in Fan and Yao (1998) we obtain the decomposition $I_3 = I_{31} + I_{32} + I_{33}$, where $I_{31} = a_n^{-1}\sum_{i,j=1}^n \phi_{ij}$,

$$\phi_{ij} = K\Big(\frac{X_i - X_j}{h_2}\Big)\sigma(X_i)\sigma(X_j)\varepsilon_{i+1}\varepsilon_{j+1}\Big\{\frac{1}{f(X_i)}K\Big(\frac{X_i - x}{h_1}\Big)+\frac{1}{f(X_j)}K\Big(\frac{X_j - x}{h_1}\Big)\Big\},$$

$a_n = n^2h_1h_2f(x)$ and terms I_{32} and I_{33} are defined in (A2.6) of Fan and Yao (1998). Moreover, its is proved in Fan and Yao (1998) that $I_{32} = o_P(h_2^2)$ and $I_{33} = o_P(h_1^2 + h_2^2)$.

We consider $\mathbb{E}(\sum_{i,j=1}^n \phi_{ij})^2 = \sum_{i,j,k,l=1}^n \mathbb{E}\phi_{ij}\phi_{kl}$. Observe that if i, j, k, l are all different then $\mathbb{E}(\phi_{ij}\phi_{kl}) = 0$. In order to deal with the remaining sum it is enough to prove that the following sums $\sum_{i,j,k=1}^n \mathbb{E}\phi_{ij}\phi_{kj}$, $\sum_{i,j,k=1}^n \mathbb{E}\phi_{jj}\phi_{ik}$, $\sum_{\substack{i,j=1\\ i\neq j}}^n \mathbb{E}\phi_{ii}\phi_{ij}$, $\sum_{i=1}^n \mathbb{E}\phi_{ii}^2$ are $o(\frac{a_n^2}{nh_1})$. We outline the proof for the first sum, the remaining ones are dealt with in a similar fashion.

It equals

$$\sum_{i,j,k=1}^n K\Big(\frac{X_k - X_j}{h_2}\Big)K\Big(\frac{X_i - X_j}{h_2}\Big)\sigma(X_i)\sigma(X_k)\sigma^2(X_j)\varepsilon_{i+1}\varepsilon_{k+1}\varepsilon_{j+1}^2W_{i,j}W_{k,j},$$

where $W_{i,j} = f(X_i)^{-1}K((X_i - x)/h_1) + f(X_j)^{-1}K((X_j - x)/h_1)$. Breaking down the product $W_{i,j}W_{k,j}$ into summands, we decompose the original sum into four sums. One of them equals

$$\begin{aligned}&\sum_{j=1}^n\Big\{\sum_{i,k=1}^n K\Big(\frac{X_k - X_j}{h_2}\Big)K\Big(\frac{X_i - X_j}{h_2}\Big)\sigma(X_i)\varepsilon_{i+1}\sigma(X_k)\varepsilon_{k+1}\Big\}\cdot\\ &\cdot\sigma^2(X_j)\varepsilon_{j+1}^2K^2\Big(\frac{X_j - x}{h_1}\Big)/f^2(X_j) =\\ &\sum_{j=1}^n\Big\{\sum_{i=1}^n K\Big(\frac{X_i - X_j}{h_2}\Big)\sigma(X_i)\varepsilon_{i+1}\Big\}^2\sigma^2(X_j)\varepsilon_{j+1}^2K^2\Big(\frac{X_j - x}{h_1}\Big)/f^2(X_j) \leq\\ &\leq \sup_{|x-y|\leq h_1} W^2(y)\sum_{i=1}^n \sigma^2(X_j)\varepsilon_{j+1}^2K^2\Big(\frac{X_j - x}{h_1}\Big)/f^2(X_j),\end{aligned}$$

where $W(y) = \sum_{i=1}^n K((X_i - y)/h_2)\sigma(X_i)\varepsilon_{i+1}$.

As $\sum_{i=1}^n \sigma^2(X_j)\varepsilon_{j+1}^2 K^2\Big(\frac{X_j-x}{h_1}\Big)/f^2(X_j) = \mathcal{O}_P((nh_1))$, what can be shown by using Markov's inequality, it is enough to prove that $\sup_{|x-y|\le h_1} W^2(y) = o_P((nh_2))^2$. This follows from Wu et al. (2010), who proved in their Proposition 2 that $\sup_{x\in[-L,L]}|D_n(y)| = \mathcal{O}_{a.s}(\frac{n^{1/p}\log n}{h_2} + (\frac{n\log n}{h_2})^{1/2})$ for any $L > 0$, where $D_n(y) = h_2^{-1}W(y)$. It is easily seen that the condition (vii)(b) implies the required order of $\sup_{|x-y|\le h_1} W^2(y)$. The remaining terms in decomposition of $\mathbb{E}\phi_{ij}\phi_{kj}$ are dealt with similarly by a slight adaptation of a result of Wu et al. (2010) to sums
$\sum_{i=1}^n K((X_i - y)/h_2)K((X_i - y)/h_1)\sigma(X_i)\varepsilon_{i+1}/f(X_i)$.

Terms I_{32} and I_{33} are dealt with as in Fan and Yao (1998), yielding $I_{32}+I_{33} = o_P(h_1^2 + h_2^2)$. Finally, we bound I_4 using

$$|I_4| \le \sup_{|y-x|\le h_1} |\hat{m}(y) - m(y)|^2 \frac{1}{nh_1 f(x)} \sum_{i=1}^n K\Big(\frac{X_i - x}{h_1}\Big)$$

and observing that

$$\sup_{|y-x|\le L} |\hat{m}(y) - m(y)| = \mathcal{O}_P\Big(\frac{\log n}{n^{1-1/p}h_2} + \Big(\frac{\log n}{nh_2}\Big)^{1/2} + h_2^2\Big)$$

which follows from representation of $\hat{m}(x) - m(x)$ in (A2.2) of Fan and Yao (1998) and application of Proposition 2 in Wu et al. (2010) again. This, in view of (vii)(b), implies $(nh_1)^{1/2}I_4 \xrightarrow{P} 0$. Terms I_i' are dealt with in an analogous manner.

4. Simulation study

4.1. The setting

In the performed numerical experiments the following three regression functions $m_i\,, i = 1, 2, 3$, and eight conditional standard deviations $\sigma_j\,, j = 1, \dots, 8$, have been considered:

(m_1) $m(x) = 0.8x$;
(m_2) $m(x) = (0.8 - 1.1\exp(-30x^2))x$;
(m_3) $m(x) = 0.8xI_{\{0<x\}} - 0.3xI_{\{x\le 0\}}$;

(σ_1) $\sigma(x) = 0.5$;
(σ_2) $\sigma(x) = 0.5((x + 1)I_{[-1,0]} + (-x + 1)I_{(0,1]})$;
(σ_3) $\sigma(x) = ((1 + (1 - x)^2)/8)^{1/2}$;
(σ_4) $\sigma(x) = 0.25I_{(-\infty,0.5]} + 0.5((x+1)I_{(-0.5,0]} + (-x+1)I_{(0,0.5)}) + 0.25I_{[0.5,\infty)}$;
(σ_5) $\sigma(x) = 0.4I_{(-\infty,-0.5)} + 0.8I_{[-0.5,0.5)} + 0.6I_{[0.5,\infty)}$;

(σ_6) $\sigma(x) = 0.75\exp(-x^2/8)$;
(σ_7) $\sigma(x) = 0.525 + 0.225[0.3\cos(\pi x) + 0.4\sin(3\pi(x - 0.175))]$;
(σ_8) $\sigma(x) = \frac{1}{2}\phi(0,1) + \sum_{l=0}^{4} \frac{1}{10}\phi(l/2 - 1, (\frac{1}{10})^2)$, where $\phi(0,1)$ is the standard Gaussian density.

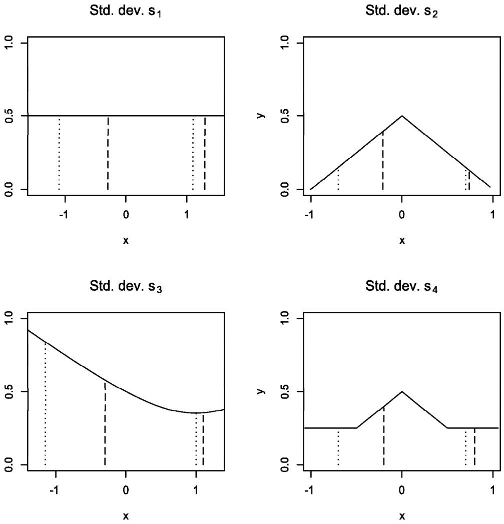


Figure 2: Standard deviations $\sigma_1 - \sigma_4$

Each combination (m, σ) has been investigated yielding $3 \times 8 = 24$ models $\mathcal{M}_1$ in total. The plots of conditional standard deviation functions are given in Figs. 2 and 3.

Standard deviations σ_7 and σ_8 are much more volatile that the remaining ones, which makes their estimation significantly more difficult. This also holds for a discontinuous standard deviation σ_5.

In all cases the error random variable ε has the standard normal distribution. For standard deviation functions σ_1, σ_2 and σ_4 the maximal value of variance of errors $\sup_x \operatorname{var}(\varepsilon\sigma(x))$ equals to 0.25 and is attained at 0. For the last standard deviation functions σ_5 through σ_8 the maximal variance is 0.64, 0.56, 0.42 and 0.36, respectively. For σ_3 maximal variance is attained at a random left endpoint of an interval on which conditional variance is estimated. The motivation to consider larger variances in the last four examples has been to make estimation of $\sigma^2(\cdot)$ feasible on a larger part of its support.

Namely, in the autoregression problem the support of the stationary distribution depends in an involved way on the regression and the variance function as well as on the distribution of errors. However, it has been suggested by our numerical experiments that increasing the variance of errors results in lengthen-

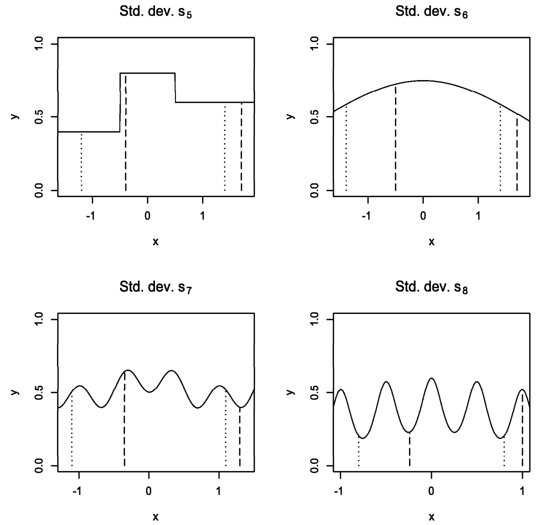


Figure 3: Standard deviations $\sigma_5 - \sigma_8$

ing of its support. In order to support this empirical observation we investigated the influence of augmenting $\sigma(\cdot)$ on the length of the central part of support of the marginal distribution. In particular, we have considered model $\mathcal{M}_1$ with standard deviation σ_2 multiplied by C, where $C = 0.2, 0.4, \ldots, 2$. Fig. 4 shows the 10-90 inter percentile range $IPR = q_{0.9} - q_{0.1}$ of the central part of support of stationary distribution for regression functions m_1 and m_3 (the results for m_2 are very similar to those for m_1). The monotone dependence of IPR on C is evident.

For sample sizes even as large as 500, observations are very scattered in both tails what significantly worsens the performance of variance estimators in these regions. In order not to have performance measures unduly influenced by this, we considered estimation on the central part of the support. It is defined analogously to the definition above as the interval $[a, b] = [X_{[0.1n]:n}, X_{[0.9n]:n}]$ with endpoints being 10th to 90th empirical percentile of a sample at hand. This is analogous to the approach of Fan and Yao (1998), who in their example 2 for the model $\mathcal{M}_2$ rejected approximately 10% of the largest and the smallest observations.

In Figs. 2 and 3 the average values of the 10th and 90th percentile based on 1000 simulations are also depicted. Dotted lines indicate the left and the right endpoints of the central part of supports of the stationary distribution for regression m_1. The supports pertaining to m_2 were omitted as they almost coincide with those for m_1. Dashed lines correspond to the support of the stationary distribution pertaining to m_3.

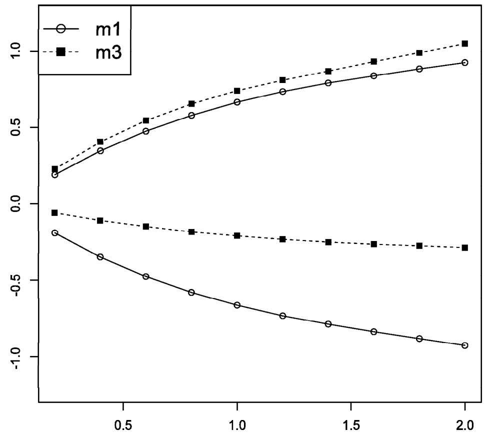


Figure 4: Endpoints of the central part of the support of the stationary distribution against C for $(m_1, C\sigma_2)$ and $(m_3, C\sigma_2)$ where $C = 0.2(0.2)2$ (see text)

4.2. Considered estimators

In the simulation study we considered local linear smoothers $\hat{\sigma}^2_{LLi}$, $i = 1, \ldots, 4$, the local maximum likelihood estimator $\hat{\sigma}^2_{LML}$, and the modified Rice estimators $\hat{\sigma}^2_{MRi}$, $i = 1, 2$. For the k-nn linear smoother k was always equal to $0.3m$, m being the number of points falling into $[a, b]$, but the bandwidth was modified near the boundaries of the interval $[a, b]$. Namely, the k-nn rule was applied for $x \in [q_{0.2}, q_{0.8}]$ where $q_{0.2}$ and $q_{0.8}$ are, respectively, 20th and 80th empirical percentile of a sample. For the remaining x in $[a, b]$ neighborhoods of size equal to the size of the neighborhood of the closest point in $[q_{0.2}, q_{0.8}]$, were considered. The performance of estimator was assessed using Empirical Integrated Squared Error EISE

$$EISE = \frac{1}{m} \sum_{X_i \in [a,b]} (\hat{\sigma}^2(X_i) - \sigma^2(X_i))^2, \tag{26}$$

where $m = \#\{i : X_i \in [a, b]\}$. For each particular combination of the regression and the variance $k = 1000$ replications were performed. Sample size n was equal 500 throughout. Burn-in period to ensure stationarity of autoregressive series was $T = 500$.

In the simulation study we focused on several aspects of performance. Besides the main aim of comparing the performance of the considered estimators for the models under study we have also investigated the effect of dependence of predictors occurring when the regression model $\mathcal{M}_3$ is replaced by the autoregression model $\mathcal{M}_1$, the impact of estimating regression function $m(\cdot)$ on the variance estimation and the importance of the choice of smoothing parameter.

These aspects will be consecutively discussed.

4.3. Effect of dependence

In order to assess the effect of dependence we compared the behavior of introduced estimators of variance based on $(X_1, Y_1^*), \dots, (X_{n-1}, Y_{n-1}^*)$, generated from model $\mathcal{M}_3$, and $(X_1, X_2), \dots, (X_{n-1}, X_n)$ from model $\mathcal{M}_1$ using bandwidths designed for the random-design regression model. The main observation which follows is that the performance of variance estimators in both situations is very similar, i.e. the effect of dependence is negligible. Tables 1 and 2 present means and standard deviations for the case (m_2, σ_2) and (m_2, σ_7), respectively. All values are multiplied by 10^4. Results of comparison $\mathcal{M}_1$ with $\mathcal{M}_3$ for all other cases are similar. In Fig. 5 boxplots for relative difference of mean EISE (MEISE) in both models $\mathcal{M}_1$ and $\mathcal{M}_3$ for all considered estimators are displayed. Namely, each boxplot is based on 24 values (3 regressions × 8 variances) of

$$I = 100\% \times (MEISE_{\mathcal{M}_1} - MEISE_{\mathcal{M}_3})/MEISE_{\mathcal{M}_3}.$$

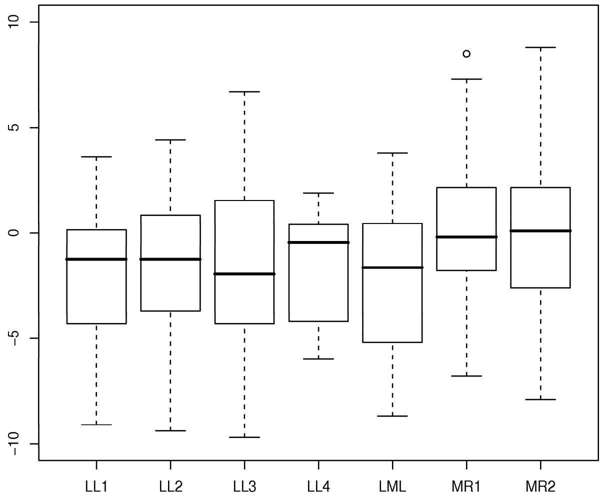


Figure 5: Boxplots for index I (see text)

It turns out that the absolute values of I do not exceed 10% and are less than 5% in most cases. All values of $|I|$ exceeding 5% corresponding to regressions m_1 and m_2 are negative, indicating that in these cases dependence between responses actually *increases* the accuracy of the estimator. The three smallest values of I (around -9.5%) correspond to estimation of the constant variance in the case of regression m_3 by local linear smoothers LLi ($i = 1, 2, 3$). Standard deviations of EISE are in most cases slightly larger for, $\mathcal{M}_3$ model.

Table 1: Means and standard deviations of EISE ×10^4 for $(m_2(x), \sigma_2(x))$

	LL1	LL2	LL3	LL4	LML	MR1	MR2
$\mathcal{M}1$	4.32(3.25)	4.35(3.22)	5.25(4)	6.56(3.21)	4.44(3.28)	6.49(5.33)	8.13(6.92)
$\mathcal{M}3$	4.54(3.25)	4.54(3.22)	5.53(4.2)	6.9(3.62)	4.69(3.33)	6.96(5.12)	8.83(7.16)

Table 2: Means and standard deviations of EISE ×10^4 for $(m_2(x), \sigma_7(x))$

	LL1	LL2	LL3	LL4	LML	MR1	MR2
$\mathcal{M}1$	52(12.65)	53.29(12.02)	43.59(23.16)	54.98(10.08)	51.11(12.71)	57.75(21.74)	67.49(29.75)
$\mathcal{M}3$	51.7(12.83)	52.49(12.27)	43.18(23.35)	54.84(10.47)	50.68(13.08)	56.4(21.5)	66.13(27.94)

4.4. Comparison of variance estimators in the autoregressive case

The results on performance of considered estimators are given in Tables 3-5. It is seen that for variances, which do not have much structure ($\sigma_1 - \sigma_6$) among all estimators excluding LL4, the local linear smoother LL2 performs the best both in terms of the mean of EISE and its standard deviation. *k-nn* estimator LL4 clearly oversmoothes and this works in its favour for very regular cases σ_1, σ_3 and σ_6. Estimators LL1, LL2 and LML perform similarly for standard deviations $\sigma_1 - \sigma_6$. For such cases LL3 performs the worst among local linear smoothers. For σ_1 (uniform) and σ_6 (gaussian) its mean EISE is around 100% larger than for LL2.

The ranking of performance changes dramatically when significantly more variable standard deviations are considered as in cases of σ_7 and σ_8. In these cases LL3 performs the best, also local maximum estimator performs slightly better than LL2. The shape of regression function does not have much influence on behavior of variance estimators. An exception is much better performance of LL2 in the case of σ_3 and σ_5 for regression m_3, which is mainly caused by change of the interval on which variances are estimated. In the case of m_3, less variable part of the standard deviation is estimated. We also experimented with a modified LL2 estimator for which the order of the polynomial fitted to $(e_i^2)_{j=1}^n$ had been chosen by the Mallows criterion from possible orders 3, 4 and 5. This resulted in a moderate increase of a relative MEISE (less than 10%) in the cases corresponding to $\sigma_1 - \sigma_6$ with a significant decrease of it for σ_7 and σ_8. For σ_8 the decrease of MEISE was around 32%, 30% and 63% for regressions m_1, m_2 and m_3, respectively.

Modified Rice estimators perform consistently the worst. Surprisingly, calculation of residuals using a larger number of points, as in case of MR2, worsened the performance. This is possibly due to the fact that for considered examples the distribution of X is pronouncedly not uniform and as a result approximate equality in (15) is violated.

Table 3: Means and standard deviations of Empirical Integrated Squared Error for $m_1(x)$

	LL1	LL2	LL3	LL4	LML	MR1	MR2
σ_1	13.03(10.33)	11.53(9.04)	24.41(16.5)	10.88(7.93)	12.32(8.98)	22.9(18.84)	30.79(26.39)
σ_2	4.27(3.22)	4.29(3.22)	4.91(3.63)	7.15(3.48)	4.37(3.23)	6.41(4.72)	7.91(6.39)
σ_3	34.38(37.02)	31.91(35.59)	62.49(76.51)	25.9(31.26)	31.33(35.04)	66.39(121.19)	86.97(195.43)
σ_4	5.13(3.36)	4.96(3.3)	6.26(4.1)	7.28(3.25)	4.95(3.41)	7.53(4.78)	8.97(5.8)
σ_5	120.42(36.97)	117.94(35.79)	117.47(45.69)	133.61(33.87)	119.63(39.34)	139.3(47.56)	157.47(60.99)
σ_6	40.89(29.95)	39.97(28.61)	96.26(60.93)	34.72(24.61)	42.42(29.46)	69.59(50.85)	89.55(65.24)
σ_7	52.07(12.67)	53.44(11.95)	43.27(21.59)	55.42(10.25)	51.29(12.68)	57.67(21.04)	67(29.81)
σ_8	81.06(19.01)	89.16(12.2)	25.01(12.69)	96.07(6.73)	79.16(15.74)	78.78(24.44)	82.07(25.55)

Table 4: Means and standard deviations of Empirical Integrated Squared Error for $m_2(x)$

	LL1	LL2	LL3	LL4	LML	MR1	MR2
σ_1	12.96(10.18)	11.49(9)	24.84(17.75)	10.8(8.07)	12.3(9.06)	23.65(17.59)	31.05(25.26)
σ_2	4.32(3.25)	4.35(3.22)	5.25(4)	6.56(3.21)	4.44(3.28)	6.49(5.33)	8.13(6.92)
σ_3	35.91(44.25)	33.48(44.39)	67.1(92.37)	26.92(34.78)	33.21(42.68)	70.35(110.04)	89.25(135.23)
σ_4	5.1(3.37)	4.97(3.32)	6.49(4.54)	6.76(3)	4.96(3.39)	7.12(4.9)	8.81(6.67)
σ_5	119.52(36.41)	117.39(35.62)	117.82(47.91)	132.08(33.27)	119(38.75)	139.68(51.97)	159.38(70.53)
σ_6	40.6(29.38)	39.79(28.33)	95.14(57.07)	34.2(24.14)	42.17(28.76)	73.23(57.48)	97.49(75.4)
σ_7	52(12.65)	53.29(12.02)	43.59(23.16)	54.98(10.08)	51.11(12.71)	57.75(21.74)	67.49(29.75)
σ_8	80.88(19.31)	89.4(12.21)	25.48(14.47)	95.2(6.82)	78.93(15.73)	68.29(26.12)	71.94(26.77)

Table 5: Means and standard deviations of Empirical Integrated Squared Error for $m_3(x)$

	LL1	LL2	LL3	LL4	LML	MR1	MR2
σ_1	12.66(8.61)	11.2(7.97)	24.67(16.26)	10.74(7.34)	12.15(8.09)	24.62(18.51)	32.62(26.39)
σ_2	5.18(4.16)	5.24(4.09)	9.02(7.43)	4.73(3.83)	5.44(4.19)	8.41(7.21)	10.81(9.35)
σ_3	10.35(9.84)	9.18(9.21)	17.8(18.12)	8.14(8.1)	9.46(8.97)	18.5(22.45)	23.53(26.1)
σ_4	5.5(3.96)	5.38(3.96)	9.71(8.39)	5.24(3.82)	5.68(4.1)	8.3(6.71)	10.64(9.07)
σ_5	82.53(40.1)	80.41(38.79)	126.92(83.83)	74.87(39.17)	81.88(39.16)	121.73(83.59)	152.69(114.51)
σ_6	44.61(31.83)	41.52(30.22)	98.26(61.78)	38.04(28.32)	45.09(30.82)	86.07(67.35)	113.21(91.72)
σ_7	41.38(16.04)	46.21(14.61)	43.86(27.88)	43.4(12.99)	42.44(14.78)	49.25(28.51)	59.59(36.52)
σ_8	48.76(22.46)	63.93(18.72)	23.64(14.59)	70.98(8.79)	52.55(20.52)	47.8(19.89)	52.18(22.17)

4.5. Comparison with benchmark estimators

In order to assess the influence of estimating regression function m for the autoregressive model $\mathcal{M}_1$, we also compared performance of considered local linear smoothers with corresponding benchmark (ideal) estimators, for which m is assumed to be known. In the case of the local linear smoothers this amounts to estimation of variance based on squared errors and not squared residuals. Tables 6 and 7 show the means of EISE for LL2 and LL3 estimators. Overall, the performance of adaptive and benchmark estimators is very similar. Surprisingly, in some cases, notably for standard deviation σ_3, the ideal estimator *performed worse* than the adaptive one.

Table 6: Means for EISE for LL2 and corresponding benchmark

		σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
m_1	LL2	11.53	4.29	31.91	4.96	117.94	39.97	53.44	89.16
m_1	benchmark	11.88	4	34.49	4.72	116.13	38.66	53.29	88.85
m_2	LL2	11.49	4.35	33.48	4.97	117.39	39.79	53.29	89.4
m_2	benchmark	11.72	4.02	36.51	4.69	115.38	38.42	52.92	88.97
m_3	LL2	11.2	5.24	9.18	5.38	80.41	41.52	46.21	63.93
m_3	benchmark	11.43	5.03	9.54	5.26	80.62	41.07	46.39	62.92

Table 7: Means of EISE for LL3 and corresponding benchmark

		σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
m_1	LL3	24.41	4.91	62.49	6.26	117.47	96.26	43.27	25.01
m_1	benchmark	24.81	4.84	68.06	6.36	118.67	97.59	44.51	25.16
m_2	LL3	24.84	5.25	67.1	6.49	117.82	95.14	43.59	25.48
m_2	benchmark	25.25	5.1	73.51	6.48	118.38	96.23	43.51	25.7
m_3	LL3	24.67	9.02	17.8	9.71	126.92	98.26	43.86	23.64
m_3	benchmark	24.77	9.68	18.93	9.74	132.12	99.67	44.64	23.91

4.6. Influence of conditional error distribution

We also considered the uniform and Laplace distribution of errors ε (see Tables 8 and 9). The support of the uniform distribution is taken as $[-\sqrt{12}/2, \sqrt{12}/2]$ to ensure that the variance equals 1, and λ for the Laplace distribution is $1/\sqrt{2}$. The value of C_ε in (11) has been fixed at 2 as for normal errors. In the case of the uniform conditional errors performance of both local linear smoothers LL2 and LL3 has much improved, the change being more significant in the case of LL3. For the Laplace distribution the situation is reversed, both estimators perform worse than in the normal case, with the change being more significant again in the case of LL3. Note that the direction of the change of performance is consistent with the change in values of $C_\varepsilon = \mathbb{E}(\varepsilon^2 - 1)^2$ appearing in the asymptotic variance (18). Value of C_ε equals 2, 0.8 and 5, respectively, for the normal, uniform and Laplace distributions. LML estimator performs similarly to LL2.

Table 8: Means of EISE for the uniform distribution

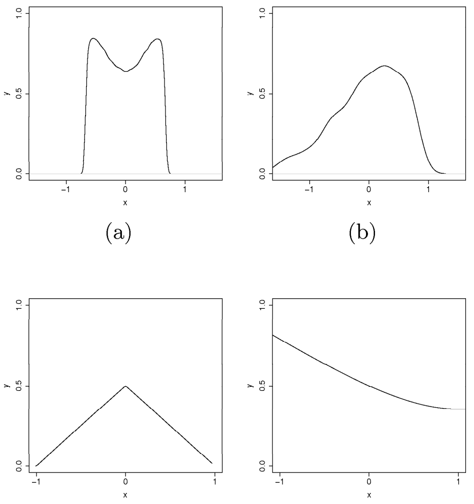
		σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
m_1	LL2	4.62	2.83	15.98	3.58	96.65	19.35	47.21	89.3
m_1	LL3	8.36	2.44	28.78	2.91	71.71	37.47	20.75	12.93
m_3	LL2	4.57	2.74	4.21	2.9	50.81	19.03	40.64	52.65
m_3	LL3	8.4	3.43	7.12	3.59	54.07	32.42	19.58	11.3

Table 9: Means of EISE for Laplace distribution

		σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
m_1	LL2	29.85	8.82	78.78	9.48	171.62	100.76	73.29	91.54
m_1	LL3	77.76	15.12	162.58	16.97	248.37	270.37	110.41	51.73
m_3	LL2	29.46	11.46	21.36	11.91	155.68	108.01	67.01	66.53
m_3	LL3	77.34	27.71	54.56	29.81	348.79	303.43	115.11	52.77

4.7. Choice of smoothing parameter

In order to check whether the performance of the local linear smoother LL2 can be still improved by an appropriate choice of bandwidth h_1 we have considered its EISE as a function of h_1 and minimized it over an equidistant grid of 50 points. The maximal value of h_1 considered equals half of the average length of the support of X_t. Estimator of m has been calculated as in definition of LL2.



(a)

(b)

(c)

(d)

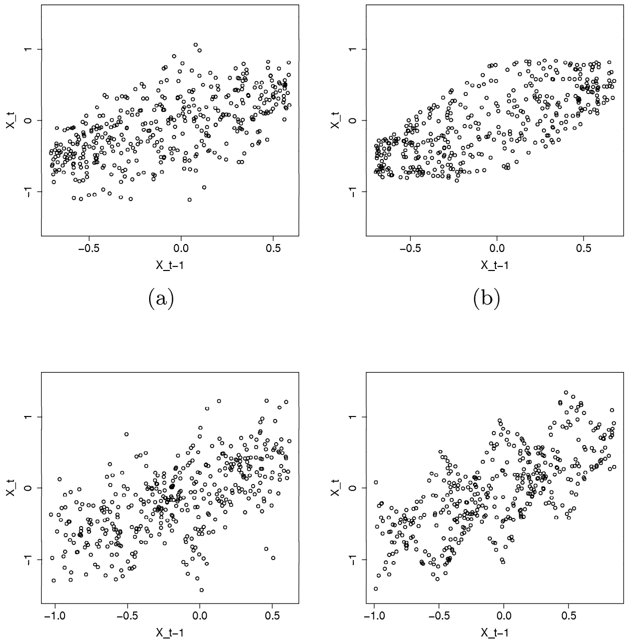
Figure 6: Marginal densities for (m_1, σ_2) and (m_1, σ_3) with corresponding standard deviations. (a) Marginal density for standard deviation σ_2, (b) Marginal density for standard deviations σ_3, (c) Standard deviation σ_2, (d) Standard deviation σ_3.

In Table 10 mean values of $I = 100\%(EISE - EISE_{min})/EISE_{min}$ are given. It is seen that the room of improvement is still significant both in the case of constant σ, when I is around 70%, and for highly variable standard deviation σ_8, when it is around 300%. Note that in Table 10 both estimators

Table 10: Relative change of EISE for LL2 when optimal h is used

	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
m_1	80.7	20.8	67.8	20.4	21.9	31.0	57.5	308.4
m_2	77.6	19.8	72.8	20.6	22.0	31.8	57.5	306.9
m_3	73.9	28.1	59.9	17.2	16.5	51.6	50.6	255.6

employed the same value of bandwidth for all x. In this context it is interesting to note the following interplay between the marginal density and the variance of conditional errors. Namely, Fig. 6 illustrates a *sweeping property* of standard deviations which asserts that the observations are swept from the regions where standard deviation is large, resulting in small values of marginal density in these regions. This indicates that, at least for standard deviations having large range of values, using local bandwidths can have beneficial effects.



(a)

(b)

(c)

(d)

Figure 7: Scatterplots of (X_{t-1}, X_t). (a) Standard deviation σ_2 with normally distributed ε_t, (b) Standard deviation σ_2 with uniformly distributed ε_t, (c) Standard deviation σ_8 with normally distributed ε_t, (d) Standard deviation σ_8 with uniformly distributed ε_t.

4.8. General conclusions from the simulation study

We can draw two conclusions from our simulation study, concerning which estimators should be applied in practice. One recommendation is to use LML estimator which enjoys uniformly good performance regardless of complexity of estimated standard deviations. The second, alternative proposal, is to evaluate the complexity of the standard deviation and, depending on the result, to apply either LL2 estimator for little varying variances LL3 estimator in the case of highly structured ones. The preliminary assessment of the variability of the variance function can be based on the scatterplot (X_{t-1}, X_t). Fig. 7 shows such plots for standard deviations σ_2 and σ_8 in the case of regression m_1, clearly indicating, especially for the uniform error distribution, much more structure in the second case.

Acknowledgements We are indebted to Dr. K. Yu for providing the main subroutine for calculation of the local maximum likelihood estimator.

5. References

Ango Nze, P. (1992) Critéres d'ergodicité de quelques modéles à représentation markovienne. *Comptes Rendus des Séances de l'Academie des Sciences Paris.* 315, 1301–1304 sér 1.

Cai, T.T., Levine, M., Wang, L. (2009) Variance function estimation in multivariate nonparametric regression with fixed design. *Journal of Multivariate Analysis*, **100** (1), 126-136.

Cai, T.T., Wang, L. (2008) Adaptive variance function estimation in heteroscedastic nonparametric regression. *Annals of Statistics*, **36**, 2025–2054.

Chen, L.-H., Cheng, M.-Y., Peng, L. (2009) Conditional variance estimation in heteroscedastic regression models. *Journal of Statistical Planning and Inference*, **139**, 236 – 245.

Chow, Y.S., Teicher, H. (1988) *Probability Theory. Indepedence, Interchangeability, Martingales.* Springer, New York.

Ćwik, J., Koronacki, J., Mielniczuk, J. (2000) Testing for difference between conditional variance functions of nonlinear time series. *Control and Cybernetics*, **29**, 33–50.

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

Diaconis, P., Freedman, D. (1999) Iterated random functions. *SIAM Review*, **41**, 45–76.

Fan, J. (2005) A selective overview of nonparametric methods in financial econometrics. *Statistical Science*, **20**, 317–337.

Fan, J., Gijbels, I. (1995) Data driven bandwidth selection in local polynomial fitting: variable bandwidth in spatial adaptation. *Journal of the*

Royal Statistical Society: Series B, **57**, 371–394.

Fan, J., Gijbels, I. (1996) *Local Polynomial Modelling and its Applications.* Chapman & Hall, London.

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

Freedman, D. A. (1975) On tail probabilities for martingales. *The Annals of Probability*, **3**, 100–118.

Gasser, U., Sroka, L., Jennen-Steinmetz, C. (1986) Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625–663.

Gendre, X. (2008) Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression. *Electronic Journal of Statistics*, **2**, 1345–1372.

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

Levine, M. (2006) Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: a possible approach. *Computational Statistics & Data Analysis*, **50**, 3404–3431.

Lu, Z., Jiang, Z. (2001) L1 geometric ergodicity of a multivariate nonlinear AR model with an ARCH term. *Statistics & Probability Letters*, **51**, 121-130.

McNeil, A.J., Frey, R. and Embrechts, P. (2005) *Quantitative Risk Management: Concepts, Techniques and Tools.* Princeton University Press, Princeton.

Neumann, M., Kreiss, J.P. (1998) Regression-type inference in nonparametric autoregression. *Annals of Statistics*, **26**, 1570–1613.

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

Ruppert, D., Sheather, S.J., Wand, M.P. (1995) An effective bandwidth selector for local least squares regression. *Journal of The American Statistical Association*, **90**, 1257–1270.

Sijbers, J., Poot, D., den Dekker, A. J. and Pintjenst, W. (2007) Automatic estimation of the noise variance from the histogram of a magnetic resonance image. *Physics in Medicine and Biology*, **52**, 1335.

Stanton, R. (1997) A nonparametric model of term structure dynamics and the market price of interest rate risk. *Journal of Finance*, **52**, 1973–2002.

Wang, L., Brown, L. D., Cai, T. and Levine, M. (2008) Effect of mean on variance function estimation in nonparametric regression. *Annals of Statistics*, **36**, 646–664.

Wu, W. B. (2005) Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences USA*, **102**, 14150–14154.

Wu, W. B., Huang, Y., Huang, Y. (2010) Kernel estimation for time series: an asymptotic theory. *Stochastic Processes and their Applications*, **120**, 2412 – 2431.

Wu, W. B., Shao, X. (2004) Limit theorems for iterated random functions. *Journal of Applied Probability*, **41**, 425–436.

Yau, P., Kohn, R. (2003) Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing*, **13**, 191–208.

Yu, K., Jones, M. C. (2004) Likelihood-based local linear estimation of the conditional variance function. *Journal of The American Statistical Association*, **99**, 139–144.

Yuan, M., Wahba, G. (2004) Doubly penalized likelihood estimator in heteroscedastic regression. *Statistics & Probability Letters*, **69**, 11–20.