Bernoulli **8**(1), 2002, 1–37

Bootstrap of kernel smoothing in nonlinear time series

JÜRGEN FRANKE[1], JENS-PETER KREISS[2] and ENNO MAMMEN[3]

[1]*Department of Mathematics, University of Kaiserslautern, D-67653 Kaiserslautern, Germany. E-mail: franke@mathematik.uni-kl.de*
[2]*Institute for Mathematical Stochastics, Technische Universität Braunschweig, Pockelsstrasse 14, D-38106 Braunschweig, Germany. E-mail: j.kreiss@tu-bs.de*
[3]*Institute for Applied Mathematics, Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 294, D-69120 Heidelberg, Germany. E-mail: mammen@statlab.uni-heidelberg.de*

Kernel smoothing in nonparametric autoregressive schemes offers a powerful tool in modelling time series. We show that the bootstrap can be used for estimating the distribution of kernel smoothers. This can be done by mimicking the stochastic nature of the whole process in the bootstrap resample or by generating a simple regression model. Consistency of these bootstrap procedures will be shown.

Keywords: bandwidth selection; bootstrap; kernel estimates; local polynomial estimates; nonparametric heteroscedastic autoregression; nonparametric time series

1. Introduction

Nonlinear modelling of time series is a promising approach in applied time series analysis. Many parametric models can be found in Priestley (1988) and Tong (1990). In this paper we consider nonparametric models of nonlinear autoregression. Motivated by econometric applications, we allow for heteroscedastic errors:

$$X_t = m(X_{t-1}, \ldots, X_{t-p}) + \sigma(X_{t-1}, \ldots, X_{t-q})\varepsilon_t, \qquad t = 0, 1, 2, \ldots. \tag{1}$$

Here the (ε_t) are independent and identically distributed (i.i.d.) random variables with mean 0 and variance 1. Furthermore, m and σ are unknown smooth functions. Ergodicity and mixing properties of such processes have been discussed in Diebolt and Guegan (1990). For the sake of simplicity, we consider only the case $p = q = 1$. In this particular case, (1) can be interpreted as discrete version of the general diffusion process with arbitrary (nonlinear) trend m and volatility function σ,

$$\mathrm{d}S_t = m(S_t) + \sigma(S_t)\mathrm{d}W_t,$$

where W_t is a standard Wiener process. The class of processes (1) also contains as a special case the qualitative threshold ARCH (QTARCH) processes. These processes were proposed by Gouriéroux and Montfort (1992) as models for financial time series.

Estimation of m and σ can be done by kernel smoothing of Nadaraya–Watson type. For the estimation of σ we consider two estimates:

$$\hat{m}_h(x) = \frac{(\hat{p}_h(x))^{-1}}{T-1} \sum_{t=1}^{T-1} K_h(x - X_t) X_{t+1}, \tag{2}$$

$$\hat{\sigma}^2_{1,h'}(x) = \frac{(\hat{p}_{h'}(x))^{-1}}{T-1} \sum_{t=1}^{T-1} K_{h'}(x - X_t) X^2_{t+1} - \hat{m}^2_{h'}(x), \tag{3}$$

$$\hat{\sigma}^2_{2,h'}(x) = \frac{(\hat{p}_{h'}(x))^{-1}}{T-1} \sum_{t=1}^{T-1} K_{h'}(x - X_t) \hat{r}^2_{t+1}. \tag{4}$$

Here $K_h(\cdot)$ denotes $h^{-1}K(\cdot/h)$ for a kernel K. The residuals $X_{t+1} - \hat{m}_h(X_t)$ are denoted by $\hat{r}_{t+1}$. In the definition of $\hat{\sigma}^2_{2,h'}(x)$ the residuals $\hat{r}_{t+1}$ could be replaced by $X_{t+1} - \hat{m}_{h'}(X_t)$ without changing the asymptotic first-order properties of $\hat{\sigma}^2_{2,h'}(x)$. The estimate $\hat{p}_h$ is a kernel estimate of the univariate stationary density p of the time series $\{X_t\}$:

$$\hat{p}_h(x) = \frac{1}{T-1} \sum_{t=1}^{T-1} K_h(x - X_t). \tag{5}$$

An early reference for Nadaraya–Watson smoothing in time series analysis is Collomb (1984). Asymptotic normality of $\hat{m}_h$, $\hat{\sigma}_{1,h}$ and $\hat{p}_h$ was shown in Robinson (1983) and Masry (1996). Uniform consistency results were given in Ango Nze and Portier (1994) and Masry (1997). Asymptotic expansions for bias and variance were derived in Auestad and Tjøstheim (1990) and Masry and Tjøstheim (1994). Tests for parametric models based on the comparison of these estimates and parametric estimates were proposed in Hjellvik and Tjøstheim (1995); compare also Kreiss *et al*. (1998). The estimate $\hat{\sigma}_{2,h}$ was proposed by Fan and Yao (1998), who argued that it outperforms $\hat{\sigma}_{1,h}$. For further references on the now extensively discussed field of nonparametric time series analysis, see the review papers by Györfy *et al*. (1989), Tjøstheim (1994) and Härdle *et al*. (1997).

In this paper several bootstrap procedures will be considered which consistently approximate the laws of $\hat{m}_h$, $\hat{\sigma}^2_{1,h}$ and $\hat{\sigma}^2_{2,h}$. The first resampling scheme (the *autoregression bootstrap*) follows a proposal of Franke and Wendel (1992) and Kreutzberger (1993). This approach is similar to residual-based resampling of linear autoregressions as discussed by Kreiss and Franke (1992). It is based on generating a bootstrap process

$$X^*_t = \tilde{m}(X^*_{t-1}) + \tilde{\sigma}(X^*_{t-1})\varepsilon^*_t, \tag{6}$$

where $\tilde{m}$ and $\tilde{\sigma}$ are estimates of m and σ and where (given the original sample $X_0, \ldots, X_T$) $\varepsilon^*_1, \ldots, \varepsilon^*_T$ is a conditionally i.i.d. resample with conditional distribution $\tilde{P}_\varepsilon$. In Section 2.1 we will state conditions on the estimates $\tilde{m}$, $\tilde{\sigma}$ and $\tilde{P}_\varepsilon$ under which the autoregression bootstrap works. Choices of $\tilde{m}$, $\tilde{\sigma}$ and $\tilde{P}_\varepsilon$ that satisfy these conditions will be given in Section 2.5. For applications of the autoregression bootstrap to the construction of uniform confidence bands for $m(x)$, see Franke *et al*. (1999). Another bootstrap method (the 'Markovian local bootstrap') that mimics the Markovian stochastic structure has been proposed by Paparoditis and Politis (1999). In their approach a Markov process is generated that only takes values from the observed process.

In our second bootstrap approach (the *regression bootstrap*), a regression model is generated with (conditionally) fixed design $(X_0, \ldots, X_{T-1})$,

$$X_t^* = \tilde{m}(X_{t-1}) + \tilde{\sigma}(X_{t-1})\varepsilon_t^*, \tag{7}$$

where, again, a conditionally i.i.d. resample of error variables $\varepsilon_1^*, \ldots, \varepsilon_T^*$ is used. The conditional distribution of ε_j^* is again denoted by $\tilde{P}_\varepsilon$. On the right-hand side of (7) the original process X_t is used instead of a resampled process. Thus, in the bootstrap a nonparametric regression model is now simulated. This is the reason why we call this resampling method the *regression bootstrap*. A modification of the regression bootstrap is the local bootstrap; see Paparoditis and Politis (2000) and Ango Nze *et al.* (1999). The local bootstrap resamples a model where the conditional distribution of the innovation $\sigma(X_{t-1})\varepsilon_t$ (given the past) is assumed to depend smoothly on the value of X_{t-1}. In the resampling these conditional distributions are approximated by smoothing kernel estimates. Clearly, this approach uses a more complex resampling scheme than the regression bootstrap where only the conditional variance of the innovations depends on X_{t-1}. The mathematical analysis of these two approaches is very similar.

In the third bootstrap, a regression model is again generated with (conditionally) fixed design $(X_0, \ldots, X_{T-1})$,

$$X_t^* = \tilde{m}(X_{t-1}) + \eta_t^*. \tag{8}$$

Here $\eta_1^*, \ldots, \eta_T^*$ is an independent resample where η_t^* has (conditional) mean zero and variance $(X_t - \hat{m}_h(X_{t-1}))^2$. This procedure has been called the *wild bootstrap* by Mammen (1992) and Härdle and Mammen (1993). For applications of the wild bootstrap to the construction of uniform confidence bands, see Neumann and Kreiss (1998).

Another bootstrap procedure that has been used for dependent data is the blockwise bootstrap of Künsch (1989). By construction it is clear that this approach cannot consistently estimate the bias of a nonparametric smoother. We now explain this for a simplified modification of the blockwise bootstrap which we call the pair bootstrap. This bootstrap method works as follows. Generate i.i.d. random variables $N_1, \ldots, N_T$ with uniform distribution on $\{1, 2, \ldots, T-1\}$. A resample in the pair bootstrap consists of the pairs $(X_{N_1}, X_{N_1+1}), \ldots, (X_{N_T}, X_{N_T+1})$. For the resample the Nadaraya–Watson estimate of $m(x)$ is given by

$$\hat{m}_h^*(x) = \frac{\sum_t K_h(x - X_{N_t})X_{N_t+1}}{\sum_t K_h(x - X_{N_t}).}$$

The conditional expectation of $\hat{m}_h^*(x)$ (given the original sample) is equal to $\hat{m}_h(x)$. Furthermore, under ergodicity conditions it can easily be checked that the conditional distribution of $\sqrt{Th}(\hat{m}_h^*(x) - \hat{m}_h(x))$ converges to a normal limit with mean zero and variance that is equal to the asymptotic variance of $\hat{m}_h(x)$. Thus, the pair bootstrap correctly mimics the variance part of $\hat{m}_h(x)$, but it does not correctly catch the bias part. Clearly, it is possible to add an estimate of the bias to $\sqrt{Th}(\hat{m}_h^*(x) - \hat{m}_h(x))$ and thus to achieve a consistent bootstrap estimate.

The asymptotic treatment of the regression bootstrap and the wild bootstrap is quite simple because it only requires the mathematical analysis of a nonparametric regression model in the bootstrap world. The mathematics for the autoregression bootstrap will turn out to be more difficult. In this bootstrap procedure a complicated resampling structure has to be generated and classical approaches based on mixing conditions are not easily available because the stochastic structure of the bootstrap is random and not fixed. For attempts to carry over mixing methods to bootstrap processes, see Franke *et al*. (1999), Bickel and Bühlmann (1999) and Ango Nze *et al*. (1999).

Our results can be extended to the case of higher-order nonparametric regression models; see (1). This is straightforward and has been omitted for the sake of notational simplicity. Furthermore, multivariate nonparametric time series can be treated similarly. Another generalization is the bootstrap for local polynomial estimators of m and σ. It has been argued that these estimators outperform Nadaraya–Watson smoothers. For a discussion of local polynomials in nonparametric regression, see Stone (1977), Tsybakov (1986), Fan (1992; 1993) and Fan and Gijbels (1992; 1995). Härdle and Tsybakov (1997) applied the idea of local polynomial fitting to autoregression models. For a local polynomial estimate that generalizes $\hat{\sigma}^2_{2,h}$, see Fan and Yao (1998). The bootstrap results presented in this paper also hold for local polynomials. It is only for the sake of simplicity that we restrict our attention in the following to Nadaraya–Watson kernel estimates $\hat{m}_h$, $\hat{\sigma}_{1,h}$ and $\hat{\sigma}_{2,h}$; cf. (2)–(4). A short discussion of why our results can be extended to local polynomials is given in Section 2.6.

This paper is organized as follows. In Sections 2.1–2.3 we show that the autoregression, regression and wild bootstraps work under appropriate conditions. A discussion of the pilot estimates $\tilde{m}$, $\tilde{\sigma}^2$ and $\tilde{P}_\varepsilon$ is given in Section 2.5. In particular, we give a choice that works for the autoregression bootstrap. Applications of our results to bandwidth choice are discussed in Section 2.4. There a generalization of our results in Sections 2.1–2.3 will be given that implies consistency of a local bandwidth selector. Extensions of our results to local polynomials are discussed in Section 2.6. Simulation results will be given in Section 3. Section 4 contains the proofs of our results in Section 2.

2. Main results: consistency of the bootstrap

In this section we present our main results and give assumptions under which our three bootstrap procedures are consistent.

2.1. Autoregression bootstrap

We consider a stationary and geometrically ergodic process of the form

$$X_t = m(X_{t-1}) + \sigma(X_{t-1})\varepsilon_t. \tag{9}$$

The unique stationary distribution is denoted by π. Stationarity and geometric ergodicity follow, for example, from the following two simple sufficient conditions:

- The distribution of the i.i.d. innovations ε_t possesses a Lebesgue density p_ε, which satisfies $\inf_{x\in C} p_\varepsilon(x) > 0$ for all compact sets C.

- m, σ and σ^{-1} are bounded on compact sets and $\limsup_{|x|\to\infty} |x|^{-1}\mathrm{E}|m(x) + \sigma(x)\varepsilon_1| < 1$.

This is a direct consequence of Theorems 1 and 2 in Diebolt and Guegan (1990); compare also Meyn and Tweedie (1993) and Doukhan (1994, pp. 106–107). The assumptions ensure that the stationary distribution π possesses a strictly positive Lebesgue density, which we denote by p. From (9) we obtain

$$p(x) = \int_{\mathbb{R}} \frac{1}{\sigma(u)} p_\varepsilon\left(\frac{x - m(u)}{\sigma(u)}\right) p(u)\mathrm{d}u. \tag{10}$$

For a stationary solution of (9), geometric ergodicity implies that the process is strongly mixing (α-mixing) with geometrically decreasing mixing coefficients (cf. Doukhan 1994, Sections 2.4 and 1.3). Moreover, this property carries over to processes of the type $Y_t = f_t(X_t)$.

To keep our proofs simple, we use somewhat stronger assumptions:

(AB1) The time series $X_1, \ldots, X_T$ is a realization of the stationary version of (9) with i.i.d. innovations $\varepsilon_1, \ldots, \varepsilon_T$ with mean 0 and variance 1. The function m is Lipschitz continuous with constant L_m, and the function σ is Lipschitz continuous with constant L_σ.

(AB2) $L_m + L_\sigma \mathrm{E}|\varepsilon_1| < 1$.

(AB3) There exists a constant $\sigma_0 > 0$ such that $\sigma(x) \geqslant \sigma_0$, for all $x \in \mathbb{R}$.

(AB4) The distribution P_ε of the innovations ε_t has a density p_ε with the following properties: $\inf_{x\in C} p_\varepsilon(x) > 0$ for all compact sets C; p_ε is twice continuously differentiable; p_ε, p'_ε and p''_ε are bounded; and $\sup_{x\in\mathbb{R}} |x p'_\varepsilon(x)| < \infty$. Furthermore, $\mathrm{E}|\varepsilon_1|^\gamma < \infty$ for a constant $\gamma > 2$.

(AB5) m is twice continuously differentiable with bounded derivatives.

(AB6) K has compact support $[-1, 1]$, say. K is symmetric, has a bounded derivative on $(-1, 1)$, and satisfies $\int K(v)\mathrm{d}v = 1$. The bandwidth h satisfies $h \to 0$ and $Th^5 \to B^2 > 0$.

Assumption (AB6) assumes that the bandwidths are of order $O(T^{-1/5})$. This rate of convergence has been motivated by optimality considerations. Our results can be extended to other rates of convergence.

We now state our assumptions on the estimates $\tilde{m}$, $\tilde{\sigma}^2$ and $\tilde{P}_\varepsilon$. These estimates are used in the generation of bootstrap resamples. Choices of estimates that satisfy these conditions are discussed in Section 2.5.

(AB7) The bootstrap innovations ε_t^* have (conditional) mean 0 and variance 1, and $d_{\mathrm{K}}(\tilde{P}_\varepsilon, P_\varepsilon) = o_P(1)$, where d_{K} denotes the Kolmogorov distance which is defined for two probability measures Q_1, Q_2 as

$$d_{\mathrm{K}}(Q_1, Q_2) = \sup_{x\in\mathbb{R}} |Q_1(X \leqslant x) - Q_2(X \leqslant x)|.$$

The initial value X_0^* satisfies $\mathrm{E}^*|X_0^*| = O_P(1)$. Furthermore, $\mathrm{E}^*|\varepsilon_t^*|^\gamma = O_P(1)$ for a constant $\gamma > \frac{5}{2}$.

Here we denote the conditional expectation (given the original sample) by E^*. The conditional probability is denoted by P^*.

(AB8) There exists a sequence $\gamma_T \to \infty$ such that $\sup_{|x|>\gamma_T} |\tilde{m}(x)| = O_P(1)$, $\sup_{|x|>\gamma_T} |\tilde{\sigma}(x)| = O_P(1)$ and $\sup_{|x|>\gamma_T} |\tilde{\sigma}^{-1}(x)| = O_P(1)$.

(AB9) With γ_T as in (AB8), $\sup_{|x|\leq\gamma_T} |\tilde{\sigma}(x) - \sigma(x)| = o_P(\gamma_T^{-1})$. For all $C > 0$ and for $j = 0$ and $j = 1$, $\sup_{|x|\leq\gamma_T} |\tilde{m}^{(j)}(x) - m^{(j)}(x)| = o_P(1)$ and $\sup_{|x|\leq C\gamma_T} |\tilde{p}_\varepsilon^{(j)}(x) - p_\varepsilon^{(j)}(x)| = o_P(1)$.

(AB10) $\sup_{|x-x_0|\leq h} |\tilde{m}^{(2)}(x) - m^{(2)}(x)| = o_P(1)$.

For the bootstrap of the variance estimates we need the following additional assumptions:

(AB11) $\mathrm{E}|\varepsilon_1|^\gamma < \infty$ for a constant $\gamma > 4$. The conditional fourth moment $\mathrm{E}[\varepsilon_1^4 | X_1 = x]$ is a continuous function in $x = x_0$. The function σ is twice continuously differentiable with bounded derivatives. The bandwidth h' satisfies $h' \to 0$ and $T(h')^5 \to (B')^2 > 0$.

(AB12) $\mathrm{E}^*|\varepsilon_t^*|^\gamma = O_P(1)$ for a constant $\gamma > 5$. Furthermore, $\mathrm{E}^*|\varepsilon_t^*|^4 = \mathrm{E}|\varepsilon_t|^4 + o_P(1)$. For $j = 0$ and $j = 1$, $\sup_{|x|\leq\gamma_T} |\tilde{\sigma}^{(j)}(x) - \sigma^{(j)}(x)| = o_P(\gamma_T^{-1})$. Furthermore, $\sup_{|x-x_0|\leq h} |\tilde{\sigma}^{(2)}(x) - \sigma^{(2)}(x)| = o_P(1)$.

Analogously to (2)–(5), the bootstrap sample $X_0^*, \ldots, X_T^*$ defines, for each point x, kernel estimates

$$\hat{m}_h^*(x) = \frac{(\hat{p}_h^*(x))^{-1}}{T-1} \sum_{t=1}^{T-1} K_h(x - X_t^*) X_{t+1}^*, \tag{11}$$

$$\hat{\sigma}_{1,h'}^{*2}(x) = \frac{(\hat{p}_{h'}^*(x))^{-1}}{T-1} \sum_{t=1}^{T-1} K_{h'}(x - X_t^*) X_{t+1}^{*2} - \hat{m}_{h'}^{*2}(x), \tag{12}$$

$$\hat{\sigma}_{2,h'}^{*2}(x) = \frac{(\hat{p}_{h'}^*(x))^{-1}}{T-1} \sum_{t=1}^{T-1} K_{h'}(x - X_t^*) \hat{r}_{t+1}^{*2}, \tag{13}$$

$$\hat{p}_h^*(x) = \frac{1}{T-1} \sum_{t=1}^{T-1} K_h(x - X_t^*), \tag{14}$$

with residuals $\hat{r}_{t+1}^* = X_{t+1}^* - \hat{m}_h^*(X_t^*)$ (or $= X_{t+1}^* - \hat{m}_{h'}^*(X_t^*)$).

The conditional distribution of $\sqrt{Th}\{\hat{m}_h^*(x) - \tilde{m}(x)\}$ given $X_1, \ldots, X_T$ is denoted by $\mathcal{L}_\mathrm{B}(x)$. This is the bootstrap estimate of $\mathcal{L}(x)$, the distribution of $\sqrt{Th}\{\hat{m}_h(x_0) - m(x_0)\}$. For $j = 1, 2$, the distribution of $\sqrt{Th}\{\hat{\sigma}_{j,h}^2(x_0) - \sigma^2(x_0)\}$ is denoted by $\mathcal{L}_j^\sigma(x)$. The bootstrap estimates of these distributions are given by the conditional distributions of $\sqrt{Th}\{\hat{\sigma}_{j,h}^{*2}(x_0) - \tilde{\sigma}^2(x_0)\}$. These estimates are denoted by $\mathcal{L}_{j,\mathrm{B}}^\sigma(x)$. Consistency of these estimates is stated in the following theorem.

Theorem 1. *Assume* (AB1)–(AB10) *for some* $x_0 \in \mathbb{R}$. *Then*

$$d_\mathrm{K}(\mathcal{L}_\mathrm{B}(x_0), \mathcal{L}(x_0)) \to 0 \textit{ in probability},$$

where, as above, d_K denotes the Kolmogorov distance. Under the additional assumption of (AB11)–(AB12),

$$d_K(\mathcal{L}^{\sigma}_{j,B}(x_0), \mathcal{L}^{\sigma}_{j}(x_0)) \to 0 \textit{ in probability}, \qquad \textit{for } j = 1, 2.$$

2.2. Regression bootstrap

We will discuss the regression and wild bootstraps in a larger class of models. We assume that one observes a stationary stochastic process (X_t, Y_t) and that one wishes to estimate the conditional mean $m(x) = \mathrm{E}[Y|X = x]$ and the conditional variance $\sigma^2(x) = \mathrm{E}[(Y - m(x))^2|X = x]$. This includes the set-up of the previous subsection where $Y_t = X_{t+1}$. The estimates $\hat{m}_h$, $\hat{\sigma}^2_{1,h}$ and $\hat{\sigma}^2_{2,h}$ are now defined as

$$\hat{m}_h(x) = \frac{(\hat{p}_h(x))^{-1}}{T-1} \sum_{t=1}^{T-1} K_h(x - X_t) Y_t, \tag{15}$$

$$\hat{\sigma}^2_{1,h}(x) = \frac{(\hat{p}_h(x))^{-1}}{T-1} \sum_{t=1}^{T-1} K_h(x - X_t) Y_t^2 - \hat{m}_h^2(x), \tag{16}$$

$$\hat{\sigma}^2_{2,h}(x) = \frac{(\hat{p}_h(x))^{-1}}{T-1} \sum_{t=1}^{T-1} K_h(x - X_t) \hat{r}^2_{t+1}. \tag{17}$$

As above, the residuals $Y_t - \hat{m}_h(X_t)$ are denoted by $\hat{r}_{t+1}$. We here use the value $t+1$ as the index of the residual to keep the notation consistent with the autoregression model of the previous subsection. For the same reason we denote $[Y_t - m(X_t)]/\sigma(X_t)$ by ε_{t+1}. The conditional estimate $\hat{p}_h$ is defined as in (5). In this and the following subsection we make no further assumptions on the stochastic structure of the process (X_t, Y_t), e.g. mixing conditions or Markov assumptions. We only assume that the estimates $\hat{m}_h(x)$ and $\hat{\sigma}^2_{j,h}(x)$, $j = 1, 2$, have a normal limit and that $\hat{p}_h(x)$ and related kernel density estimates are consistent; see assumptions (RB2), (RB3) and (RB8) below. To check conditions, consult the literature discussed in the introduction. For Markov models, see also Section 2.1.

In the regression bootstrap, one generates i.i.d. resamples $\varepsilon_1^*, \ldots, \varepsilon_T^*$ and puts

$$Y^*_{t-1} = \tilde{m}(X_{t-1}) + \tilde{\sigma}(X_{t-1})\varepsilon^*_t.$$

Here, again, $\tilde{m}$ and $\tilde{\sigma}$ are estimates. The original sample $X_0, \ldots, X_{T-1}$ acts in the resampling as a fixed design. We now define

$$\hat{m}_h^*(x) = \frac{(\hat{p}_h(x))^{-1}}{T-1} \sum_{t=1}^{T-1} K_h(x - X_t) Y_t^*,$$

$$\hat{\sigma}_{1,h}^{*2}(x) = \frac{(\hat{p}_h(x))^{-1}}{T-1} \sum_{t=1}^{T-1} K_h(x - X_t) Y_t^{*2} - \hat{m}_h^{*2}(x),$$

$$\hat{\sigma}_{2,h}^{*2}(x) = \frac{(\hat{p}_h(x))^{-1}}{T-1} \sum_{t=1}^{T-1} K_h(x - X_t) \hat{r}_{t+1}^{*2},$$

where the bootstrap residuals $\hat{r}_{t+1}^*$ are defined by $\hat{r}_{t+1}^* = Y_t^* - \hat{m}_h^*(X_t)$. The conditional distribution of $\sqrt{Th}\{\hat{m}_h^*(x) - \tilde{m}(x)\}$ is denoted by $\mathcal{L}_{\mathrm{RB}}(x)$ and the conditional distribution of $\sqrt{Th}\{\hat{\sigma}_{j,h}^{*2}(x) - \tilde{\sigma}^2(x)\}$ is denoted by $\mathcal{L}_{j,\mathrm{RB}}^{\sigma}(x)$ $[j = 1, 2]$. These are now the bootstrap estimates for $\mathcal{L}(x)$ and $\mathcal{L}_j^{\sigma}(x)$.

For the regression bootstrap we make the following assumptions:

(RB1) For the kernel K and the bandwidth h, condition (AB6) holds.

(RB2) The density p of X_t has one derivative at x_0, and the regression function m has two continuous derivatives at x_0. For all $c > 0$,

$$T^{-4/5} \#\{1 \leqslant t \leqslant T-1 : x_0 - cT^{-1/5} \leqslant X_t \leqslant x_0 + cT^{-1/5}\} \to 2cp(x_0)$$

in probability, where # denotes the number of elements in a set. Furthermore,

$$T^{-1} h^{-2} \sum_{t=1}^{T-1} (X_t - x_0) K_h(X_t - x_0) \to p'(x_0) \int u^2 K(u) \mathrm{d}u \text{ in probability.}$$

Conditions (RB1) and (RB2) imply $\hat{p}_h(x_0) = p(x_0) + o_P(1)$.

(RB3) $\sup_{|x-x_0| \leqslant h} |\tilde{m}^{(j)}(x) - m^{(j)}(x)| = o_P(1)$, for $0 \leqslant j \leqslant 2$, and $\sup_{|x-x_0| \leqslant h} |\tilde{\sigma}(x) - \sigma(x_0)| = o_P(1)$.

(RB4) $\sqrt{Th}[\hat{m}_h(x_0) - m(x_0)]$ has an asymptotic normal distribution with mean $b(x_0) = B \cdot \int v^2 K(v) \mathrm{d}v \cdot [p'(x_0) m'(x_0)/p(x_0) + \frac{1}{2} m''(x_0)]$ and variance $\tau^2(x_0) = \sigma^2(x_0) \int K^2(v) \mathrm{d}v / p(x_0)$.

(RB5) The bootstrap innovations ε_t^* have (conditional) mean 0 and variance 1, and $\mathrm{E}^* |\varepsilon_t^*|^{\gamma} = O_P(1)$ for a constant $\gamma > 2$.

For the bootstrap of the variance estimates we need the following additional assumptions:

(RB6) The conditional fourth moment $\mathrm{E}[\varepsilon_1^4 | X_1 = x]$ is a continuous function in $x = x_0$. The function σ is twice continuously differentiable at $x = x_0$. The bandwidth h' satisfies $h' \to 0$ and $T(h')^5 \to (B')^2 > 0$.

(RB7) $\mathrm{E}^* |\varepsilon_t^*|^{\gamma} = O_P(1)$ for a constant $\gamma > 5$. Furthermore, $\mathrm{E}^* |\varepsilon_t^*|^4 = \mathrm{E}\big[|\varepsilon_t|^4 | X_{t-1} = x\big] + o_P(1)$. For $0 \leqslant j \leqslant 2$, $\sup_{|x-x_0| \leqslant h} \big|\tilde{\sigma}^{(j)}(x) - \sigma^{(j)}(x)\big| = o_P(1)$.

(RB8) For $j = 1$ and $j = 2$, the estimate $\sqrt{Th'}[\hat{\sigma}_{j,h'}(x_0) - \sigma(x_0)]$ has an asymptotic normal distribution with mean $b_{\sigma,j}(x_0)$ and variance $\tau_{\sigma}^2(x_0)$, where

$$b_{\sigma,1}(x) = B' \int v^2 K(v) \mathrm{d}v \left\{ \frac{\mathrm{d}\sigma^2}{\mathrm{d}x}(x) \frac{p'(x)}{p(x)} + m'(x)^2 + \frac{1}{2} \frac{\mathrm{d}^2\sigma^2}{(\mathrm{d}x)^2}(x) \right\},$$

$$b_{\sigma,2}(x) = B' \int v^2 K(v) \mathrm{d}v \left\{ \frac{\mathrm{d}\sigma^2}{\mathrm{d}x}(x) \frac{p'(x)}{p(x)} + \frac{1}{2} \frac{\mathrm{d}^2\sigma^2}{(\mathrm{d}x)^2}(x) \right\},$$

$$\tau_\sigma^2(x) = \int K^2(v) \mathrm{d}v \sigma^4(x) \mathrm{E}[\varepsilon_1^4 - 1 | X_0 = x] \frac{1}{p(x)}.$$

Consistency of the regression bootstrap is stated in the following theorem.

Theorem 2. *Assume* (RB1)–(RB5) *for an* $x_0 \in \mathbb{R}$. *Then*

$$d_\mathrm{K}(\mathcal{L}_\mathrm{RB}(x), \mathcal{L}(x)) \to 0 \ (\textit{in probability}).$$

Under the additional assumption of (RB6)–(RB8) *we have, for* $j = 1, 2$,

$$d_\mathrm{K}(\mathcal{L}_{j,\mathrm{RB}}^\sigma(x), \mathcal{L}_j^\sigma(x)) \to 0 \ (\textit{in probability}).$$

2.3. Wild bootstrap

The wild bootstrap starts by generating an i.i.d. sample $\bar{\eta}_1, \ldots, \bar{\eta}_T$ with mean 0 and variance 1. (For higher-order performance the distribution of $\bar{\eta}_t$ is often chosen such that additionally $\mathrm{E}\bar{\eta}_t^3 = 1$; for a discussion of this point and for choices of the distribution of $\bar{\eta}_t$, compare Mammen (1992).) Now put $\eta_t^* = (X_t - \hat{m}_h(X_{t-1}))\bar{\eta}_t$. The wild bootstrap resample is defined as

$$X_t^* = \tilde{m}(X_{t-1}) + \eta_t^*.$$

As in the previous subsection, this resample can be used for calculating $\hat{m}_h^*(x)$. The conditional distribution of $\sqrt{Th}\{\hat{m}_h^*(x) - \tilde{m}(x)\}$ is denoted by $\mathcal{L}_\mathrm{WB}(x)$. In particular, the wild bootstrap is appropriate in cases of irregular variance functions $\sigma(x)$. Such models may arise when $\sigma(x)$ only acts as nuisance parameter and the main interest lies in estimating m.

For the wild bootstrap we make the following assumption:

(WB1) Assumptions (RB1), (RB2) and (RB4) apply, and $\sup_{|x-x_0|\leqslant h} |\tilde{m}^{(j)}(x) - m^{(j)}(x)| = o_P(1)$, for $0 \leqslant j \leqslant 2$, and $\mathrm{E}^*|\bar{\eta}_t|^\gamma = O_P(1)$ for a constant $\gamma > 2$.

Theorem 3. *Assume* (WB1) *for an* $x_0 \in \mathbb{R}$. *Then*

$$d_\mathrm{K}(\mathcal{L}_\mathrm{WB}(x), \mathcal{L}(x)) \to 0 \ \textit{in probability}.$$

Condition (RB3) typically only makes sense if σ is continuous at x_0. Otherwise an estimate $\tilde{\sigma}(x)$ with $x \neq x_0$ in an h-neighbourhood of x_0 could not be expected to converge to $\sigma(x_0)$. Continuity of σ is not required for the wild bootstrap. Thus fewer smoothness assumptions on σ are made for the wild bootstrap than for the regression bootstrap.

Furthermore, the autoregression bootstrap requires even more smoothness assumptions than the regression bootstrap.

2.4. Bootstrap bandwidth choice

In this subsection we consider the important problem of choosing the smoothing parameter adaptively from the data. Various alternatives, from crossvalidatory to plug-in procedures, are discussed in the literature, e.g. for time series by Härdle and Vieu (1992) or in a recent comparative study for density estimates by Jones *et al*. (1996). One particular procedure, discussed by Härdle and Bowman (1988) for regression estimates and Franke and Härdle (1992) for spectral density estimates, is based on minimizing a bootstrap approximation of the error criterion, e.g. the pointwise mean square error or the mean integrated square error (MISE), as a function of bandwidth h. In the following, we investigate whether the proposed bootstrap methods can be used to select data-dependent bandwidths for kernel smoothers of m and σ in model (1).

For a fixed value of x_0 and for $h = h_\eta(T) = \eta T^{-1/5}$, with $\eta \in [a, b] \subset (0, \infty)$, we consider the process

$$\eta \to Z_T(\eta),$$

where, for $\eta \in [a, b]$,

$$\begin{aligned} Z_T(\eta) &= \sqrt{Th_\eta(T)}(\hat{m}_{h_\eta(T)}(x_0) - m(x_0)) \\ &= \frac{\left(T^{-2/5}/\sqrt{\eta}\right)\sum_t K\left(T^{1/5}(x_0 - X_t)/\eta)(X_{t+1} - m(x_0)\right)}{(T^{-4/5}/\eta)\sum_t K(T^{1/5}(x_0 - X_t)/\eta)}. \end{aligned}$$

The bootstrap can be used for the estimation of the distribution of the process Z_T. We will discuss this for the autoregression bootstrap. The autoregression bootstrap process is defined for $\eta \in [a, b]$ as

$$Z_T^*(\eta) := \frac{\left(T^{-2/5}/\sqrt{\eta}\right)\sum_t K\left(T^{1/5}(x_0 - X_t^*)/\eta)(X_{t+1}^* - \tilde{m}(x_0)\right)}{(T^{-4/5}/\eta)\sum_t K(T^{1/5}(x_0 - X_t^*)/\eta)}.$$

The following result states consistency of the autoregression bootstrap.

Theorem 4. *Under* (AB1)–(AB10)*, $Z_T(\cdot)$ converges weakly in $C[a, b]$ to a Gaussian process with mean function*

$$\mu(\eta) = \eta^{5/2} \int v^2 K(v) \mathrm{d}v \left(\frac{p'(x_0) m'(x_0)}{p(x_0)} + \frac{1}{2} m''(x_0) \right)$$

and covariance function

$$R(\eta_1, \eta_2) = \frac{(p(x_0))^{-1}}{\sqrt{\eta_1 \eta_2}} \int K\left(\frac{v}{\eta_1}\right) K\left(\frac{v}{\eta_2}\right) \mathrm{d}v \sigma^2(x_0).$$

Furthermore, $Z_T^*(\cdot)$ *converges weakly in* $C[a, b]$ *to the same Gaussian process (in probability).*

Under the additional assumption of (AB11)–(AB12) *an analogous result on the consistency of bootstrap for the process* $\eta \to \sqrt{Th_\eta(T)}(\hat{\sigma}_{h_\eta(T)}(x_0) - \sigma(x_0))$ *applies.*

This result has two immediate consequences. It implies that a kernel smoother $\hat{m}_{\hat{h}}(x_0)$ with data-adaptive bandwidth has the same asymptotic limit as a smoother $\hat{m}_h(x_0)$ with deterministic bandwidth h of order $T^{-1/5}$ as long as $(\hat{h} - h)/h$ converges to zero in probability.

Another implication is that the bootstrap accurately approximates the distribution of the pointwise squared error $Th(\hat{m}_h(x) - m(x))^2$ for $h = \eta T^{-1/5}$ uniformly in $\eta \in [a, b]$. This does not immediately imply that the bootstrap approximates the pointwise mean squared error. Such a result holds after a slight modification of the processes $Z_T(\eta)$ and $Z_T^*(\eta)$. This minor technical complication is caused by the fact that the denominator of the Nadaraya–Watson estimate may become quite close to 0, though with small probability; therefore, this problem does not occur for the bootstrap bandwidth choice for e.g. Priestley–Chao estimates in a regression setting with equidistant design as discussed by Härdle and Bowman (1988). To overcome this complication, one possibility is to truncate $Th(\hat{m}_h(x) - m(x))^2$ by a slowly growing constant. Another way out would be to modify the definition of $\hat{m}_h(x)$ such that $Th(\hat{m}_h(x) - m(x))^2$ has a bounded second moment. This could be done by adding a positive sequence to the denominator of $\hat{m}_h(x)$ that converges to zero at a proper rate.

2.5. Choice of the estimates $\tilde{m}$, $\tilde{\sigma}^2$ and $\tilde{P}_\varepsilon$

In this subsection we discuss the choice of the pilot estimates $\tilde{m}$, $\tilde{\sigma}^2$ and $\tilde{P}_\varepsilon$ for the stochastic model of Section 2.1 where we discussed the autoregression bootstrap. For this set-up the pilot estimates $\tilde{m}$ and $\tilde{\sigma}^2$ can be chosen as truncated Nadaraya–Watson smoothers and the distribution $\tilde{P}_\varepsilon$ as the empirical distribution of residuals. The same choices of $\tilde{m}$ (and $\tilde{\sigma}^2$ in the case of the regression bootstrap) work for the regression and wild bootstraps under the assumptions of Section 2.1 or under appropriate mixing conditions. We refer to Masry (1997) for uniform convergence results under mixing conditions. For constants C_m, c_σ and C_σ, put $\overline{C}_m = C_m |\overline{X}|$, $\bar{c}_\sigma = c_\sigma \bar{r}^2$ and $\overline{C}_\sigma = C_\sigma \bar{r}^2$, where $\overline{X}$ is the sample mean and $\bar{r}^2$ is the sample mean of the squared residuals $\hat{r}_t^2$. We put

$$\tilde{m}(x) = \begin{cases} \hat{m}_g(x) & \text{if } |\hat{m}_g(x)| \leqslant \overline{C}_m, \\ \overline{C}_m & \text{if } \hat{m}_g(x) > \overline{C}_m, \\ -\overline{C}_m & \text{if } \hat{m}_g(x) < -\overline{C}_m, \end{cases}$$

$$\tilde{\sigma}(x) = \begin{cases} \hat{\sigma}_{j,g'}(x) & \text{if } \bar{c}_\sigma \leqslant \hat{\sigma}_{j,g'}(x) \leqslant \overline{C}_\sigma, \\ \overline{C}_\sigma & \text{if } \hat{\sigma}_{j,g'}(x) > \overline{C}_\sigma, \\ \bar{c}_\sigma & \text{if } \hat{\sigma}_{j,g'}(x) < \bar{c}_\sigma, \end{cases}$$

where $j = 1$ or 2 and where g and g' are bandwidths that typically differ from h and h'; see below. We now define $\tilde{P}_\varepsilon$. For a sequence $\gamma_T \to \infty$, we put $I_T = \{1 \leqslant t \leqslant T : |\hat{r}_t| \leqslant \gamma_T\}$. Then $\tilde{P}_\varepsilon$ is defined as the empirical distribution of $\hat{r}_t - \bar{r}_{I_T}$ for $t \in I_T$. Here $\bar{r}_{I_T}$ is the average of $\hat{r}_t$ for $t \in I_T$. We need the following conditions:

(P1) Assumptions (AB1)–(AB6) hold.
(P2) The kernel K has three bounded derivatives, and for the bandwidths g and g' we have that $g,\ g' \to 0$, $ng^5 \to \infty$ and $n^{-1}(g')^{-5/3}$ is bounded.
(P3) $\mathrm{E}|\varepsilon_1|^5 < \infty$.

For the bootstrap of $\hat{\sigma}_{j,h}$ we need the following assumptions:

(P4) Condition (AB11) holds.
(P5) $\mathrm{E}|\varepsilon_1|^{10} < \infty$.

Theorem 5: *Assume* (P1)–(P3) *and that C_m and C_σ are large enough and c_σ is small enough. Then there exists a sequence $\gamma_T \to \infty$ such that* (AB7)–(AB10) *hold for $\tilde{m}$, $\tilde{\sigma}^2$ and $\tilde{P}_\varepsilon$. Suppose additionally that* (P4)–(P5) *hold; then there exists a sequence $\gamma_T \to \infty$ such that* (AB7)–(AB10) *and* (AB12) *hold for $\tilde{m}$, $\tilde{\sigma}^2$ and $\tilde{P}_\varepsilon$.*

The first statement implies that with the above choices of $\tilde{m}$, $\tilde{\sigma}^2$ and $\tilde{P}_\varepsilon$ the autoregression bootstrap works for $\hat{m}_h(x_0)$; the second statement shows consistency of the autoregression bootstrap for $\hat{\sigma}_{j,h}(x_0)$ with $j = 1$ or 2.

2.6. Bootstrap for local polynomials

We now briefly outline how our results can be extended to local polynomials. In Theorems 1–3 consistency of the bootstrap approaches is shown for Nadaraya–Watson smoothing. In the proofs of these theorems the distributions of the Nadaraya–Watson estimates are compared for the bootstrap resamples and for the original samples. In proving consistency of the bootstrap it is shown that the conditional distributions based on the bootstrap resample and the unconditional distributions based on the original data have the same asymptotic normal limit. For $\hat{m}_h$ this will be done by showing for its nominator and denominator

- that $\hat{p}_h(x)$ and its counterpart in the bootstrap world have the same (deterministic) limit (namely, the stationary density $p(x)$ of X_t); and
- that $\sqrt{Th}(T-1)^{-1}\sum_{t=1}^{T-1}[K_h(x-X_t)X_{t+1} - \mathrm{E}K_h(x-X_t)X_{t+1}]$ and its counterpart in the bootstrap world have the same normal limit.

The rth-order local polynomial estimator $\hat{m}_h^{\mathrm{locpol}}$ of m is defined as $\hat{a}_0$, where $(\hat{a}_0, \ldots, \hat{a}_{r-1})^{\mathrm{T}}$ minimizes

$$\sum_{t=1}^{T-1} K_h(x-X_t)\left(X_{t+1} - \sum_{j=0}^{r-1} a_j\left(\frac{x-X_t}{h}\right)^j\right)^2.$$

This can be rewritten as

$$\hat{a} = \hat{S}(x)^{-1}\hat{\tau}(x),$$

where the matrix $\hat{S}(x)$ has elements $\hat{S}(x)_{j,k} = (T-1)^{-1}\sum_{t=1}^{T-1}[(x-X_t)/h]^{j+k}\ K_h(x-X_t)$, $j, k = 0, \ldots, r-1$, and where the vector $\hat{\tau}(x)$ has elements $\hat{\tau}(x)_j = (T-1)^{-1}\sum_{t=1}^{T-1} X_{t+1}[(x-X_t)/h]^j K_h(x-X_t)$, $j = 0, \ldots, r-1$. Consistency of the bootstrap can be shown by an extension of the approach for Nadaraya–Watson smoothing. One now shows that $\hat{S}(x)$ and its counterpart in the bootstrap world have the same (deterministic) limit and that $\sqrt{Th}[\hat{\tau}(x) - \mathrm{E}\hat{\tau}(x)]$ and its counterpart in the bootstrap world have the same normal limit. Theorem 4 can be generalized to local polynomials by simple changes of arguments. For $r = 1$, the local polynomial estimator $\hat{m}_h^{\mathrm{locpol}}$ coincides with the Nadaraya–Watson estimator. For $r = 2$, the local polynomial estimator is called the local linear estimator $\hat{m}_h^{\mathrm{loclin}}$.

3. Simulations

In this section we discuss the behaviour of the bootstrap for finite sample size. This will be done by simulations for the proposed wild and autoregression bootstraps. We have considered simulated realizations of the process ($t = 1, \ldots, T$)

$$X_t = 4\sin(X_{t-1}) + \sqrt{0.5 + 0.25X_{t-1}^2}\,\varepsilon_t, \tag{18}$$

where the random variables (ε_t, $t = 1, \ldots, T$) are assumed to be i.i.d. with standard normal law. The standard deviation of the one-dimensional stationary distribution of the time series (X_t) is approximately 3.3, while the standard deviation of the noise process $(\sqrt{0.5 + 0.25X_{t-1}^2}\,\varepsilon_t)$ is about 1.8.

We have compared the distribution of a kernel estimate $\hat{m}_h(x)$ (cf. (2)) of $m(x) = 4\sin(x)$ at $x = \pi/2$ with its bootstrap counterparts. Based on a Monte Carlo simulation, we display for three different sample sizes, $T = 50$, 100 and 200, the density of the distribution of

$$\frac{\sqrt{Th}(\hat{m}_h(\pi/2) - m(\pi/2))}{\hat{V}_h(\pi/2)}, \tag{19}$$

where $\hat{V}_h^2(x)$ denotes the following estimator of the variance of $\hat{m}_h(x)$:

$$\hat{V}_h^2(x) = \frac{(h/(T-1))\sum_{t=1}^{T-1} K_h^2(x - X_t)(X_{t+1} - \hat{m}_h(X_t))^2}{\hat{p}_h^2(x)}. \tag{20}$$

This standardization ensures that the asymptotic distribution is a normal one with unit variance but with usually non-vanishing mean.

The wild bootstrap approximation (cf. Section 2.3) of (19) is given through the distribution of

$$\frac{\sqrt{Th}(\hat{m}_h^*(\pi/2) - \tilde{m}(\pi/2))}{\hat{V}_h(\pi/2)}, \tag{21}$$

where we make use of the same standardization as above. $\tilde{m}$ has to be chosen so as to satisfy (WB1). We propose to use a so-called oversmoothed kernel estimator, i.e. $\tilde{m} = \hat{m}_g$, with a pilot bandwidth g such that $g \gg h$. As a rule of thumb the choice $g \approx 1.5h$ or $g \approx 2h$ is often suggested. The quality of the wild bootstrap approximation is not judged on a single underlying realization $X_1, \ldots, X_T$. We have simulated 100 underlying time series (each of length T) and have computed for each data set the wild bootstrap distribution, cf. (21), which is of course a conditional distribution given the underlying data. Each bootstrap distribution has been obtained via simulation based on 1000 wild bootstrap replications. Since a plot of all 100 bootstrap densities looks rather confusing, we decided to display only three representative cases. For each of the 100 bootstrap densities g_i^* we computed the MISE to the true underlying density f, i.e. $\text{MISE}_i = \int (g_i(x) - f(x))^2 \, dx$. We display the wild bootstrap density which corresponds to the median of the MISE values (as an average case), the one which corresponds to the upper quartile of the MISE values (as a non-favourable case), and the one which corresponds to the lower quartile of MISE values (as a favourable case).

Figure 1(b) displays these three wild bootstrap densities together with the underlying density, which we intend to approximate, for a sample size of $T = 50$ and $h = 0.90$ (pilot bandwidth $g = 1.20$), while Figures 2(b) and 3(b) give the same results for $T = 100$, $h = 0.80$, $g = 1.10$ and $T = 200$, $h = 0.65$, $g = 0.90$, respectively. It can be seen that the quality of the wild bootstrap approximation is not very good for sample sizes 50 and 100. Nevertheless, it can be seen that the quality of the approximation increases with increasing sample size and seems to be reasonable for $T = 200$.

By way of comparison, we have investigated the quality of the normal approximation. Because of the standardization, we used normal densities with unit variance. Estimation of the bias has been done by exactly the same method, namely oversmoothing, as we used for the wild bootstrap procedure. As above, we have computed an estimator of the asymptotic bias based on 100 independent data sets, and plots of the three normal approximations which correspond to the lower and upper quartile as well as the median of the MISE distances of the normal approximations to the underlying true density are displayed. The results for the three different sample sizes are given in Figures 1(a), 2(a) and 3(a). It can be seen that the wild bootstrap does not outperform the normal approximation for sample sizes 50 and 100. Only for the largest sample size of 200 do we obtain a slightly better behaviour of the wild bootstrap approximation compared with the asymptotic normal approximation.

Additionally, we have investigated the autoregression bootstrap for the same underlying situation as above. The implementation of the autoregression bootstrap is much more in-

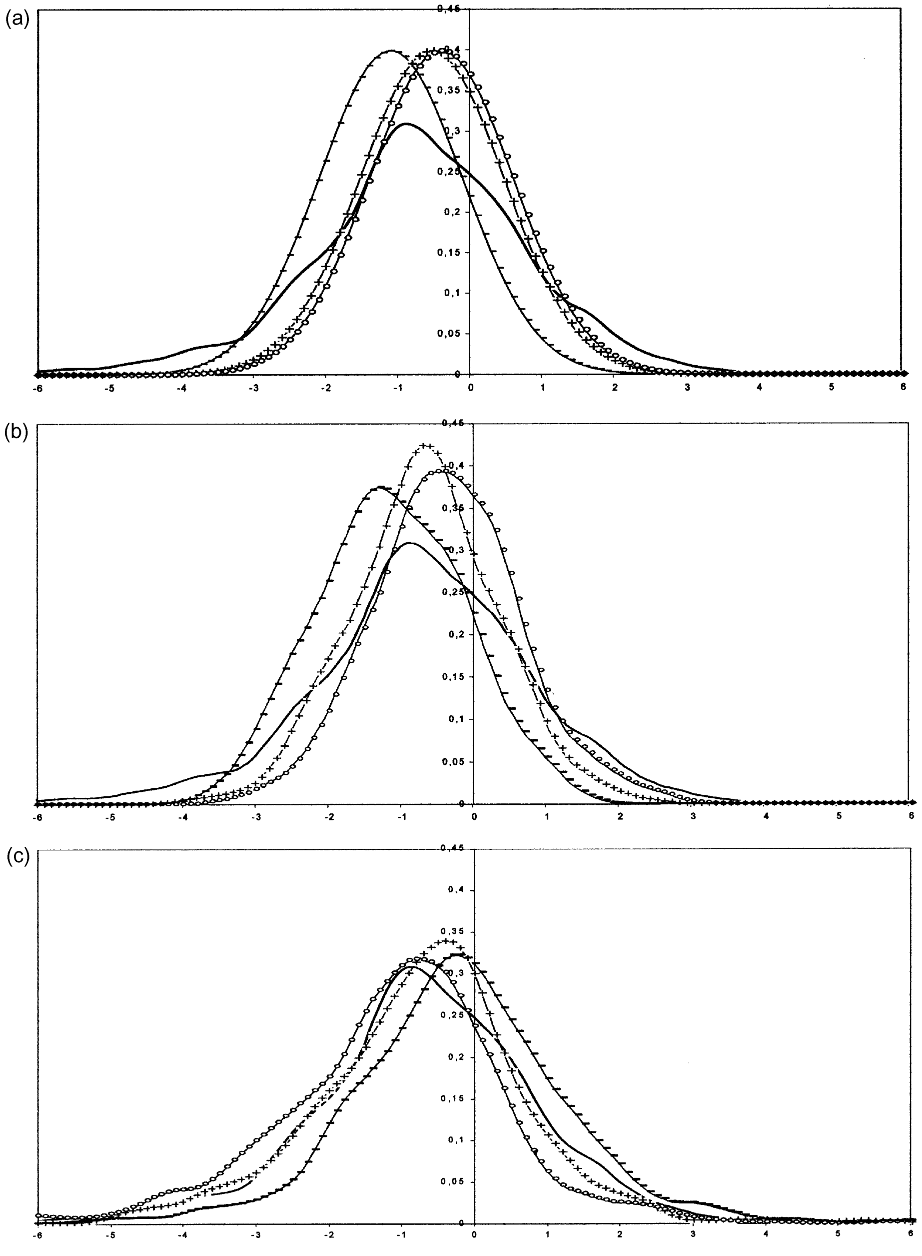

Figure 1. Simulated density of standardized kernel estimator (solid line) with bandwidth $h = 0.90$ and sample size $T = 50$, together with densities of three representative (a) standardized normal, (b) wild bootstrap and (c) autoregression bootstrap approximations with MISE to the density of the kernel estimator, corresponding to lower quartile (+++), median (○○○) and upper quartile (- - -). A pilot bandwidth of $g = 1.20$ has been used.

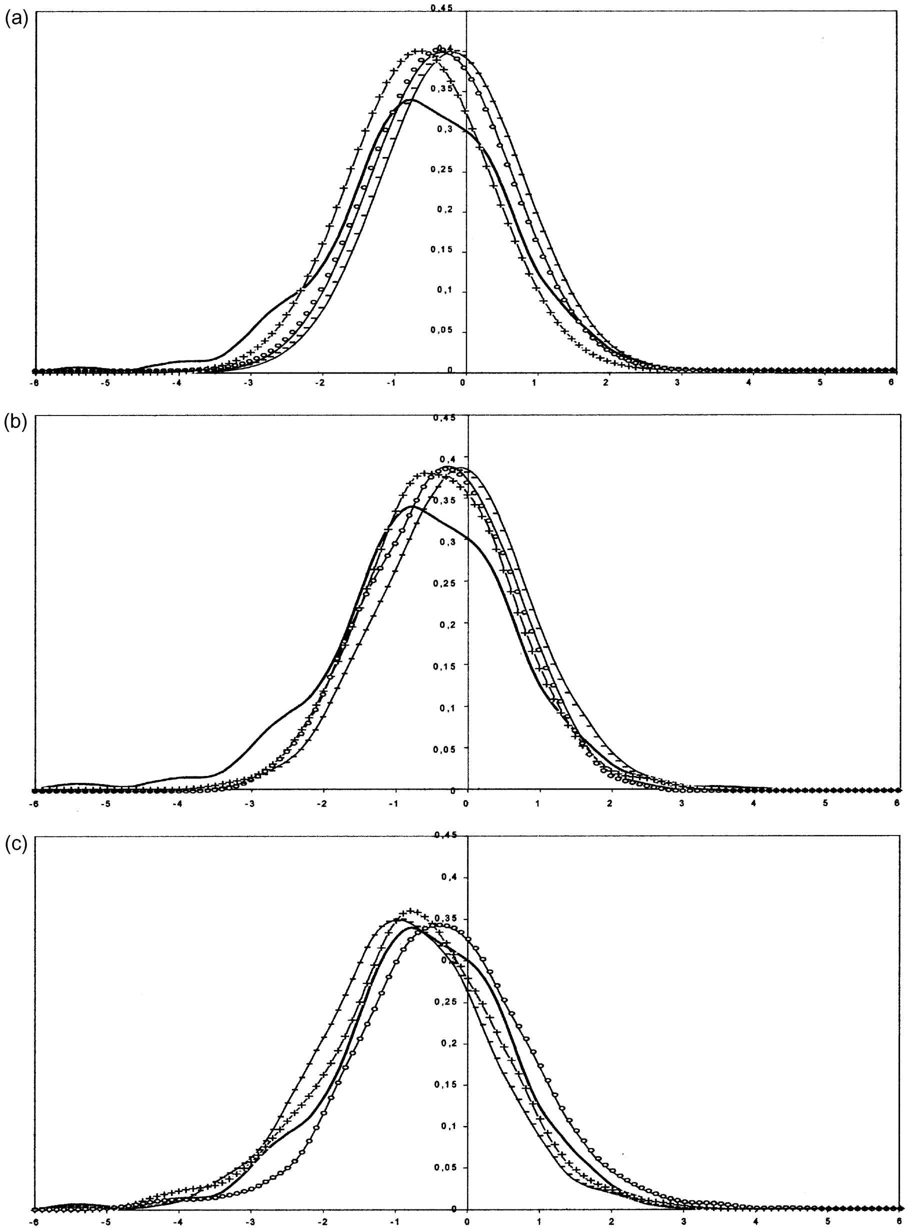

Figure 2. Simulated density of standardized kernel estimator (solid line) with bandwidth $h = 0.80$ and sample size $T = 100$, together with densities of three representative (a) standardized normal, (b) wild bootstrap and (c) autoregression bootstrap approximations with MISE to the density of the kernel estimator, corresponding to lower quartile (+++), median (∘∘∘) and upper quartile (- - -). A pilot bandwidth of $g = 1.10$ has been used.

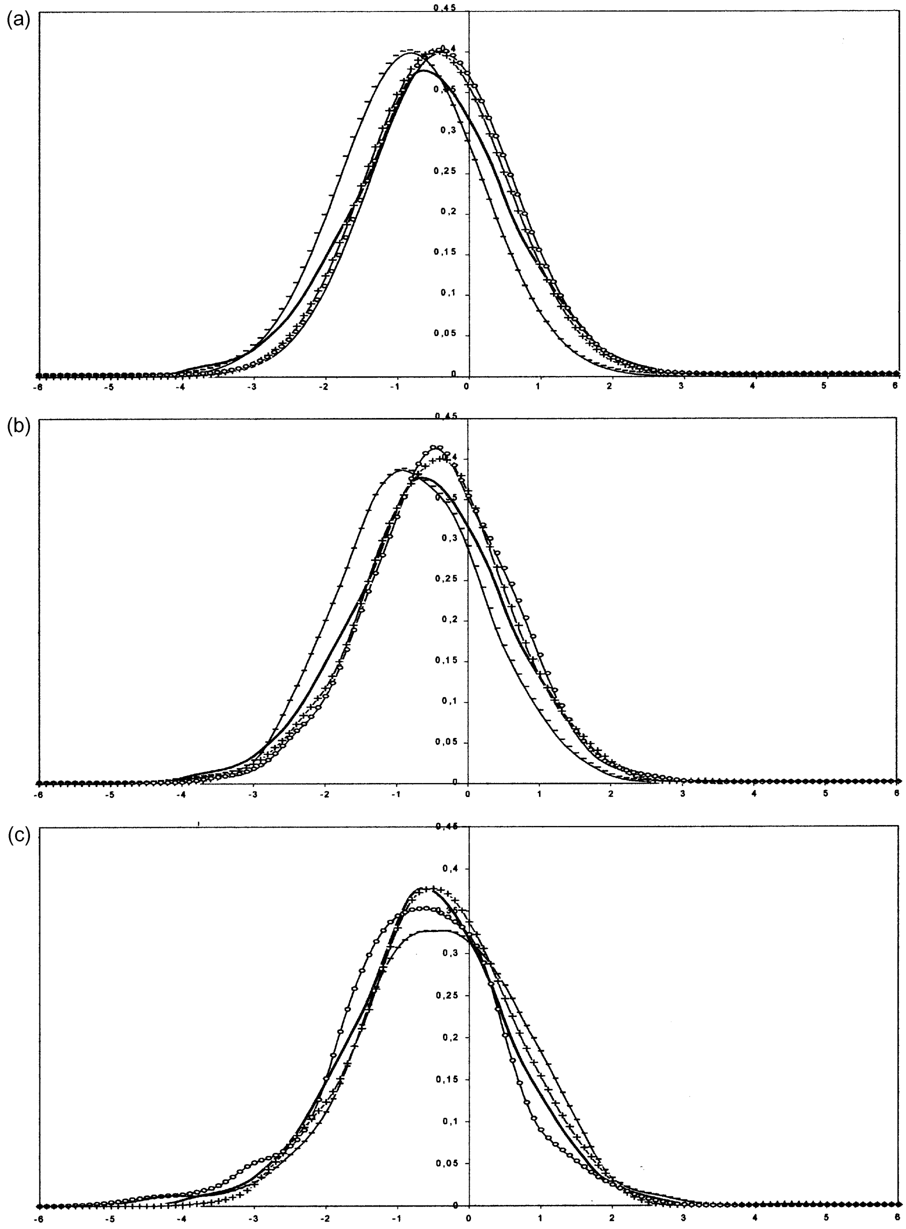

Figure 3. Simulated density of standardized kernel estimator (solid line) with bandwidth $h = 0.65$ and sample size $T = 200$, together with densities of three representative (a) standardized normal, (b) wild bootstrap and (c) autoregression bootstrap approximations with MISE to the density of the kernel estimator, corresponding to lower quartile (+++), median (○○○) and upper quartile (- - -). A pilot bandwidth of $g = 0.90$ has been used.

volved since we have to estimate the conditional variance, too. The autoregression bootstrap mimics the complete dependence structure of the underlying process, since we also make use of recursion (18) in the autoregression bootstrap world (cf. Section 2.1). As initial estimates $\tilde{m}$ and $\tilde{\sigma}$ we have used (as above) oversmoothed kernel estimates $\hat{m}_g$ and $\hat{\sigma}_g$ defined in (2) and (4). Along exactly the same lines as above, we have simulated autoregression bootstrap approximations of the density of the distribution of (19) based on 100 independent underlying time series $X_1, \ldots, X_T$. Based on MISE values, we have selected, as above, three representative density estimates, given in Figure 1(c) ($T = 50$), Figure 2(c) ($T = 100$) and Figure 3(c) ($T = 200$). It can clearly be seen that the autoregression bootstrap outperforms the normal approximation as well as the wild bootstrap approximation for all sample sizes. The autoregression bootstrap works quite well, even for a sample size as low as 50.

This clearly demonstrates that the more sophisticated autoregression resampling scheme, which takes full account of the dependence structure, results in much better approximations of the underlying situation. In contrast, the wild bootstrap (and similarly the regression bootstrap) does not take the dependence structure into account at all. This is correct from an asymptotic point of view, as we have shown in Sections 2.2 and 2.3, but moderate to large sample sizes are necessary in order to yield reasonable approximations which outperform the normal approximation. However, it should be mentioned that the proposed wild bootstrap even works in situations where the underlying nonparametric autoregression (18) does not hold. But this assumption is necessary for the correctness of the autoregression bootstrap.

Finally, it has to be mentioned that we have used the kernel

$$K(u) = \tfrac{15}{16}(1 - u^2)^2 1_{[-1,1]}(u)$$

and the smoothing parameter 0.70 in order to create from 1000 Monte Carlo replications the smooth density plots in Figures 1–3. The value 0.70 is the smoothing parameter which minimizes the MISE in kernel density estimation where the underlying distribution is a normal one with unit variance and unknown mean. This corresponds exactly to the asymptotic situation we have to hand.

The MISE values of all plotted densities in Figures 1–3 are gathered together in Table 1.

4. Proofs

In the following proofs we assume, for the sake of notational simplicity, that $h = h'$.

4.1. Proof of Theorem 1

First we give a proof of the first statement of Theorem 1. For this purpose we have to treat the estimate $\hat{m}_h(x_0)$. This estimate can be split into a variance and a bias term,

Table 1. MISE values of densities plotted in Figures 1–3

	Normal approximation	Wild bootstrap	Autoregression bootstrap
$T = 50,\ h = 0.90$			
Lower quartile	0.0185	0.0165	0.0036
Median	0.0228	0.0214	0.0078
Upper quartile	0.0345	0.0310	0.0121
$T = 50,\ h = 0.90$			
Lower quartile	0.0067	0.0059	0.0024
Median	0.0103	0.0089	0.0045
Upper quartile	0.0193	0.0175	0.0101
$T = 50,\ h = 0.90$			
Lower quartile	0.0026	0.0022	0.0016
Median	0.0050	0.0042	0.0030
Upper quartile	0.0108	0.0096	0.0055

$$\sqrt{Th}(\hat{m}_h(x_0) - m(x_0)) = \frac{\sqrt{Th}\hat{r}_{\mathrm{V},h}(x_0)}{\hat{p}_h(x_0)} + \frac{\sqrt{Th}\hat{r}_{\mathrm{B},h}(x_0)}{\hat{p}_h(x_0)},$$

where

$$\hat{r}_{\mathrm{V},h}(x_0) = \frac{1}{T-1}\sum_t K_h(x_0 - X_t)\sigma(X_t)\varepsilon_{t+1},$$

$$\hat{r}_{\mathrm{B},h}(x_0) = \frac{1}{T-1}\sum_t K_h(x_0 - X_t)(m(X_t) - m(x_0)).$$

Similarly, we decompose the bootstrap estimate $\hat{m}_h^*(x_0)$,

$$\sqrt{Th}(\hat{m}_h^*(x_0) - \tilde{m}(x_0)) = \frac{\sqrt{Th}\hat{r}^*_{\mathrm{V},h}(x_0)}{\hat{p}^*_h(x_0)} + \frac{\sqrt{Th}\hat{r}^*_{\mathrm{B},h}(x_0)}{\hat{p}^*_h(x_0)},$$

where

$$\hat{r}^*_{\mathrm{V},h}(x_0) = \frac{1}{T-1}\sum_t K_h(x_0 - X_t^*)\tilde{\sigma}(X_t^*)\varepsilon^*_{t+1},$$

$$\hat{r}^*_{\mathrm{B},h}(x_0) = \frac{1}{T-1}\sum_t K_h(x_0 - X_t^*)(\tilde{m}(X_t^*) - \tilde{m}(x_0)).$$

In the following Lemmas 4.4–4.6 we compare the random variables $\hat{r}_{\mathrm{V},h}(x_0)$, $\hat{r}_{\mathrm{B},h}(x_0)$ and $\hat{p}_h(x_0)$ with $\hat{r}^*_{\mathrm{V},h}(x_0)$, $\hat{r}^*_{\mathrm{B},h}(x_0)$ and $\hat{p}^*_h(x_0)$. The first statement of Theorem 1 immediately follows by application of these lemmas.

For the second statement of Theorem 1 it remains to show $d_K(\mathcal{L}^{\sigma}_{j,B}(x_0), \mathcal{L}^{\sigma}_{j}(x_0)) = o_P(1)$ for $j = 1, 2$. For $j = 1$, this follows with similar arguments to those for $d_K(\mathcal{L}_B(x_0), \mathcal{L}(x_0)) = o_P(1)$. For $j = 2$, note that

$$\hat\sigma^2_{2,h}(x_0) = \hat\sigma^2_{2,1,h}(x_0) - 2\hat\sigma^2_{2,2,h}(x_0) + \hat\sigma^2_{2,3,h}(x_0),$$

where

$$\hat\sigma^2_{2,1,h}(x_0) = \frac{(\hat p_h(x_0))^{-1}}{T-1}\sum_{t=1}^{T-1} K_h(x_0 - X_t)\sigma^2(X_t)\varepsilon^2_{t+1}, \tag{22}$$

$$\hat\sigma^2_{2,2,h}(x_0) = \frac{(\hat p_h(x_0))^{-1}}{T-1}\sum_{t=1}^{T-1} K_h(x_0 - X_t)\sigma(X_t)\varepsilon_{t+1}[\hat m_h(X_t) - m(X_t)], \tag{23}$$

$$\hat\sigma^2_{2,3,h}(x_0) = \frac{(\hat p_h(x_0))^{-1}}{T-1}\sum_{t=1}^{T-1} K_h(x_0 - X_t)[\hat m_h(X_t) - m(X_t)]^2. \tag{24}$$

Similarly, one can write

$$\hat\sigma^{*2}_{2,h}(x_0) = \hat\sigma^{*2}_{2,1,h}(x_0) - 2\hat\sigma^{*2}_{2,2,h}(x_0) + \hat\sigma^{*2}_{2,3,h}(x_0),$$

where

$$\hat\sigma^{*2}_{2,1,h}(x_0) = \frac{(\hat p^*_h(x_0))^{-1}}{T-1}\sum_{t=1}^{T-1} K_h(x_0 - X_t)\tilde\sigma^2(X^*_t)(\varepsilon^*_{t+1})^2, \tag{25}$$

$$\hat\sigma^{*2}_{2,2,h}(x_0) = \frac{(\hat p^*_h(x_0))^{-1}}{T-1}\sum_{t=1}^{T-1} K_h(x_0 - X^*_t)\tilde\sigma(X^*_t)\varepsilon^*_{t+1}[\hat m^*_h(X^*_t) - \tilde m(X^*_t)], \tag{26}$$

$$\hat\sigma^{*2}_{2,3,h}(x_0) = \frac{(\hat p^*_h(x_0))^{-1}}{T-1}\sum_{t=1}^{T-1} K_h(x_0 - X_t)[\hat m^*_h(X^*_t) - \tilde m(X^*_t)]^2. \tag{27}$$

With similar arguments to those for the first statement of Theorem 1, it can be easily verified that $d_K(\mathcal{L}(\sqrt{hT}[\hat\sigma^2_{2,1,h}(x_0) - \sigma^2(x_0)]), \mathcal{L}^*(\sqrt{hT}[\hat\sigma^{*2}_{2,1,h}(x_0) - \tilde\sigma^2(x_0)])) = o_P(1)$. The claim

$$d_K(\mathcal{L}(\sqrt{hT}[\hat\sigma^2_{2,h}(x_0) - \sigma^2(x_0)]), \mathcal{L}^*(\sqrt{hT}[\hat\sigma^{*2}_{2,h}(x_0) - \tilde\sigma^2(x_0)])) = o_P(1)$$

now immediately follows from

$$\sqrt{Th}\hat\sigma^2_{2,j,h}(x_0) = o_P(1),$$

$$\sqrt{Th}\hat\sigma^{*2}_{2,j,h}(x_0) = o_P(1),$$

for $j = 2, 3$. This statement will be shown in Lemma 4.7. This shows the second statement of Theorem 1.

In the proofs of Lemmas 4.4–4.7, we will make use of the fact that X_t and X^*_t have approximately the same distribution. For the proof of this property, we now use strong approximation methods. For this purpose we construct samples of error variables $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_T$ that have conditional distribution P_ε (given the sample $X_1, \ldots, X_T$). We will use these error variables to construct a process $\{\tilde{X}_t\}$ with conditional distribution equal to the unconditional distribution of $\{X_t\}$. We will show that $\mathrm{E}^*|X_t^* - \tilde{X}_t| = o_P(1)$; see Lemma 4.2. This implies the claim that X_t and X_t^* have approximately the same distribution.

We now choose samples of error variables $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_T$ with the following properties:

(i) $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_T$ are conditionally i.i.d. (given the original data $X_0, \ldots, X_T$);
(ii) $\tilde{\varepsilon}_t$ has a conditional distribution (given the original data $X_0, \ldots, X_T$) which is identical to the unconditional distribution of $\{\varepsilon_t\}$, i.e. $\mathcal{L}^*(\tilde{\varepsilon}_t) = \mathcal{L}(\varepsilon_t)$;
(iii) $\mathrm{E}^*(\tilde{\varepsilon}_t - \varepsilon_t^*)^2 = d_2^2(\tilde{\varepsilon}_t, \varepsilon_t^*) = d_2^2(\varepsilon_t, \varepsilon_t^*)$, where d_2 denotes the Mallows distance.

For two random variables X and Y, the Mallows distance is defined as

$$d_2^2(X, Y) = d_2^2(\mathcal{L}(X), \mathcal{L}(Y)) = \inf\{\mathrm{E}(U - V)^2 | \mathcal{L}(U) = \mathcal{L}(X), \mathcal{L}(V) = \mathcal{L}(Y)\}.$$

Existence of random variables $\tilde{\varepsilon}_t$ with (i)–(iii) follows from the fact that the infimum in the definition of d_2 is attained; see Bickel and Freedman (1981).

From (AB7) it follows that $d_2^2(\varepsilon_t, \varepsilon_t^*) = d_2^2(P_\varepsilon, \tilde{P}_\varepsilon) = o_P(1)$. This holds because $d_{\mathrm{K}}(P_\varepsilon, \tilde{P}_\varepsilon) = o_P(1)$ and $\mathrm{E}^*(\varepsilon^*)^2 = \mathrm{E}\varepsilon^2$; see Bickel and Freedman (1981). This leads to

$$\mathrm{E}^*(\tilde{\varepsilon}_t - \varepsilon_t^*)^2 = o_P(1). \tag{28}$$

We now define the process $\tilde{X}_t$. We define $\tilde{X}_0$ such that $\mathcal{L}^*(\tilde{X}_0) = \mathcal{L}(X_0)$ and we put

$$\tilde{X}_t = m(\tilde{X}_{t-1}) + \sigma(\tilde{X}_{t-1})\tilde{\varepsilon}_t.$$

For the comparison of the distributions of $\tilde{X}_t$ and X_t^*, we show that under our assumptions $|X_t^*| \leqslant \gamma_T$ holds with probability tending to one.

Lemma 4.1. *Under assumptions* (AB1)–(AB10),

$$\max_{0 \leqslant t \leqslant T} P^*(|X_t^*| \geqslant \gamma_T) \to 0 \textit{ in probability}.$$

Proof. We start by showing that

$$\max_{0 \leqslant t \leqslant T} \mathrm{E}^*|X_t^*| = O_P(1). \tag{29}$$

With $D_m = \sup_{|x|>\gamma_T} |\tilde{m}(x)| = O_P(1)$ and $D_\sigma = \sup_{|x|>\gamma_T} |\tilde{\sigma}(x)| = O_P(1)$, we have

$$\begin{aligned} |X_{t+1}^*| &\leqslant [D_m + D_\sigma|\varepsilon_{t+1}^*|]1_{\{|X_t^*|>\gamma_T\}} + |\tilde{m}(X_t^*) + \tilde{\sigma}(X_t^*)\varepsilon_{t+1}^*|1_{\{|X_t^*|\leqslant\gamma_T\}} \\ &\leqslant D_m + D_\sigma|\varepsilon_{t+1}^*| + |m(X_t^*) + \sigma(X_t^*)\varepsilon_{t+1}^*| + o_P(1), \end{aligned}$$

where in the second equality boundedness of $\mathrm{E}^*|\varepsilon_s^*|$ has been used. This follows from (AB7). For $L = L_m + L_\sigma \mathrm{E}^*|\varepsilon_s^*|$ we have, because of (28), $L = L_m + L_\sigma \mathrm{E}|\varepsilon_s| + o_P(1) < \lambda + o_P(1)$ with a constant $0 < \lambda < 1$; see (AB2). Put $D = D_m + D_\sigma \mathrm{E}^*|\varepsilon_s^*| + |m(0)| + \sigma(0)\mathrm{E}^*|\varepsilon_s^*|$. Then $D = O_P(1)$ and, by Lipschitz continuity of m and σ, we have, iterating with respect to t,

$$\mathrm{E}^*|X_{t+1}^*| \leqslant L \cdot \mathrm{E}^*|X_t^*| + D \leqslant \ldots \leqslant L^{t+1}\mathrm{E}^*|X_0^*| + \sum_{k=0}^{t} L^k \cdot D$$

$$\leqslant \mathrm{E}|X_0| + \frac{D}{1-L} = O_P(1).$$

This shows (29). We now prove

$$\max_{0\leqslant t\leqslant T} P^*\{X_t^* > \gamma_T\} = o_P(1). \tag{30}$$

With similar arguments one shows $\max_{0\leqslant t\leqslant T} P^*\{X_t^* < -\gamma_T\} = o_P(1)$. This implies the statement of the lemma. For the proof of (30) we use the notation

$$q_T(x) = \frac{\gamma_T - m(x)}{\sigma(x)}, \qquad \tilde{q}_T(x) = \frac{\gamma_T - \tilde{m}(x)}{\tilde{\sigma}(x)}.$$

Observe that $\tilde{q}_T(x) \geqslant D_\sigma^{-1}\gamma_T - D'_\sigma D_m$ for $|x| > \gamma_T$ (where $D'_\sigma = \sup_{|x|>\gamma_T} |\tilde{\sigma}^{-1}(x)| = O_P(1)$) and that, by assumption,

$$\tilde{q}_T(x) = q_T(x) + o_P(1) \text{ uniformly in } |x| \leqslant \gamma_T.$$

Therefore,

$$\begin{aligned} P^*\{X_t^* > \gamma_T\} &= P^*\{\varepsilon_t^* > \tilde{q}_T(X_{t-1}^*)\} \\ &\leqslant P^*\{\varepsilon_t^* > D_\sigma^{-1}\gamma_T - D'_\sigma D_m, |X_{t-1}^*| > \gamma_T\} \\ &\quad + P^*\{\varepsilon_t^* > q_T(X_{t-1}^*) + o_P(1), |X_{t-1}^*| \leqslant \gamma_T\}. \end{aligned}$$

The first summand is bounded by $P^*\{\varepsilon_t^* > D_\sigma^{-1}\gamma_T - D'_\sigma D_m\}$, which converges to 0 for $\gamma_T \to \infty$, using (AB7). Denoting by P_{t-1}^* the (conditional) probability of X_{t-1}^*, the second summand is

$$\begin{aligned} \int_{-\gamma_T}^{\gamma_T} P^*\{\varepsilon_t^* > q_T(x) + o_P(1)\} P_{t-1}^*(\mathrm{d}x) &\leqslant \int_{-\gamma_T}^{\gamma_T} \frac{\mathrm{E}^*|\varepsilon_t^*|}{q_T(x) + o_P(1)} P_{t-1}^*(\mathrm{d}x) \\ &= \int_{-\gamma_T}^{\gamma_T} \frac{\mathrm{E}^*|\varepsilon_t^*|}{q_T(x)} p_{t-1}^*(x)\mathrm{d}x\{1 + o_P(1)\} \\ &= \int_{-\gamma_T}^{\gamma_T} \frac{\sigma(x)}{\gamma_T - m(x)} P_{t-1}^*(\mathrm{d}x)\{1 + o_P(1)\}\mathrm{E}^*|\varepsilon_t^*|. \end{aligned}$$

For the inequality, we have used Markov's inequality and the fact that $q_T(x)$ is positive and bounded away from 0 uniformly in $|x| \leqslant \gamma_T$. Now, by Lipschitz continuity of m and σ, $\gamma_T - m(x) \geqslant \gamma_T - L_m \cdot |x| - |m(0)| \geqslant (1 - L_m)\gamma_T - |m(0)|$ for $|x| \leqslant \gamma_T$, $\sigma(x) \leqslant L_\sigma \cdot |x| + \sigma(0)$. Therefore, for a suitable constant C^*, the last integral is bounded by

$$\frac{C^*}{(1-L_m)\gamma_T - |m(0)|} \int_{-\gamma_T}^{\gamma_T} \{L_\sigma|x| + \sigma(0)\} P_{t-1}^*(\mathrm{d}x) \leqslant C^* \frac{L_\sigma \mathrm{E}^*|X_{t-1}^*| + \sigma(0)}{(1-L_m)\gamma_T - |m(0)|} \to 0$$

for $\gamma_T \to \infty$. Because of (29), this shows (30). □

Lemma 4.2. *Under assumptions* (AB1)–(AB10)*, for a constant* $0 < \lambda < 1$ *and for random variables* $S_1 = o_P(1)$*,* $S_2 = O_P(1)$ *and* $L < \lambda + o_P(1)$ *that do not depend on t, we have, for* $1 \leqslant t \leqslant T$*, that*

$$\sup_{1\leqslant t\leqslant T} \mathrm{E}^*|X_t^* - \tilde{X}_t| = S_1 + L^{t-1}S_2.$$

Proof. We have

$$\begin{aligned}
&\mathrm{E}^*|X_t^* - \tilde{X}_t| \\
&\quad = \mathrm{E}^*|\tilde{m}(X_{t-1}^*) - m(\tilde{X}_{t-1}) + (\tilde{\sigma}(X_{t-1}^*) - \sigma(\tilde{X}_{t-1}))\varepsilon_t^* + \sigma(\tilde{X}_{t-1})(\varepsilon_t^* - \tilde{\varepsilon}_t)| \\
&\quad \leqslant \mathrm{E}^*|\tilde{m}(X_{t-1}^*) - m(\tilde{X}_{t-1})| + \mathrm{E}^*|(\tilde{\sigma}(X_{t-1}^*) - \sigma(\tilde{X}_{t-1}))\varepsilon_t^*| + \mathrm{E}^*|\sigma(\tilde{X}_{t-1})(\varepsilon_t^* - \tilde{\varepsilon}_t)|.
\end{aligned}$$

For the treatment of the first summand we use

$$\begin{aligned}
\mathrm{E}^*|\tilde{m}(X_{t-1}^*) - m(\tilde{X}_{t-1})| \leqslant{}& \mathrm{E}^*|\tilde{m}(X_{t-1}^*)|1\{|X_{t-1}^*| > \gamma_T\} + \mathrm{E}^*|\tilde{m}(X_{t-1}^*) \\
&- m(X_{t-1}^*)|1\{|X_{t-1}^*| \leqslant \gamma_T\} + \mathrm{E}^*|m(X_{t-1}^*) \\
&- m(\tilde{X}_{t-1})|1\{|X_{t-1}^*| \leqslant \gamma_T\} + \mathrm{E}^*|m(\tilde{X}_{t-1})|1\{|X_{t-1}^*| > \gamma_T\}.
\end{aligned}$$

Now, the first term on the right-hand side converges to 0 (uniformly for $0 \leqslant t \leqslant T$, in probability). This follows from Lemma 4.1. The same holds for the second term by assumption (AB9). For the third term we have, from Lipschitz continuity of m,

$$\mathrm{E}^*|m(X_{t-1}^*) - m(\tilde{X}_{t-1})|1\{|X_{t-1}^*| \leqslant \gamma_T\} \leqslant L_m \mathrm{E}^*|X_{t-1}^* - \tilde{X}_{t-1}|.$$

For the last term we obtain, for constants $C > 0$,

$$\begin{aligned}
&\mathrm{E}^*|m(\tilde{X}_{t-1})|1\{|X_{t-1}^*| > \gamma_T\} \leqslant \mathrm{E}^*|m(\tilde{X}_{t-1})|1\{|X_{t-1}^*| > \gamma_T\} \\
&\quad \leqslant \mathrm{E}^*|m(\tilde{X}_{t-1}|1\{|X_{t-1}^*| > \gamma_T, |\tilde{X}_{t-1}| > C\} + \mathrm{E}^*|m(\tilde{X}_{t-1})|1\{|X_{t-1}^*| > \gamma_T, |\tilde{X}_{t-1}| \leqslant C\} \\
&\quad \leqslant \mathrm{E}^*(L_m|m(\tilde{X}_{t-1})| + |m(0)|)1\{|\tilde{X}_{t-1}| > C\} + \mathrm{E}^*(L_m C + |m(0)|)1\{|X_{t-1}^*| > \gamma_T\}.
\end{aligned}$$

The second term on the right-hand side of this last inequality is of order $o_P(1)$, uniformly for $0 \leqslant t \leqslant T$, because of Lemma 4.1. The first term, $\mathrm{E}^*(L_m|m(\tilde{X}_{t-1})| + |m(0)|)1\{|\tilde{X}_{t-1}| > C\} = \mathrm{E}(L_m|m(X_{t-1})| + |m(0)|)1\{|X_{t-1}| > C\}$, turns out not to depend on t and can be made arbitrarily small by choice of C. This leads to

$$\max_{0\leqslant t\leqslant T} \mathrm{E}^*|m(\tilde{X}_{t-1})|1\{|X_{t-1}^*| > \gamma_T\} = o_P(1).$$

So we arrive at

$$\mathrm{E}^*|\tilde{m}(X^*_{t-1}) - m(\tilde{X}_{t-1})| \leqslant L_m \mathrm{E}^*|X^*_{t-1} - \tilde{X}_{t-1}| + R^m_T,$$

where $R^m_T = o_P(1)$ is a random variable that does not depend on t. Exactly along the same lines, we obtain

$$\mathrm{E}^*|\tilde{\sigma}(X^*_{t-1}) - \sigma(\tilde{X}_{t-1})|\,|\varepsilon^*_t| \leqslant \mathrm{E}^*|\varepsilon^*_1|[L_\sigma \mathrm{E}^*|X^*_{t-1} - \tilde{X}_{t-1}|] + R^\sigma_T,$$

where $R^\sigma_T = o_P(1)$ is a random variable that does not depend on t. Finally, by assumption (28),

$$\mathrm{E}^*\sigma(\tilde{X}_{t-1})|\varepsilon^*_t - \tilde{\varepsilon}_t| = \mathrm{E}^*\sigma(\tilde{X}_1)\mathrm{E}^*|\varepsilon^*_1 - \tilde{\varepsilon}_1| = o_P(1).$$

Thus, we have shown that, with $L = L_m + \mathrm{E}^*|\varepsilon^*_1|L_\sigma$ and a random variable $R = o_P(1)$,

$$\begin{aligned}\mathrm{E}^*|X^*_t - \tilde{X}_t| &\leqslant L\mathrm{E}^*|X^*_{t-1} - \tilde{X}_{t-1}| + R \\ &\leqslant \sum_{\nu=0}^{t-2} L^\nu R + L^{t-1}\{\mathrm{E}^*[\tilde{X}_0] + \mathrm{E}^*[X^*_0]\}.\end{aligned}$$

The lemma follows from $L < 1 + o_P(1)$ (see Lemma 4.1) and $\mathrm{E}^*[\tilde{X}_0] = \mathrm{E}[X_0]$. □

From Lemma 4.2 we obtain the following corollary.

Lemma 4.3. *Under the assumptions of Theorem* 1*, we have*

$$\mathrm{E}^*\left\{\frac{1}{T}\sum_{t=1}^{T}|\tilde{X}_t - X^*_t|\right\} = o_P(1).$$

Lemma 4.4. *Under assumptions* (AB1)–(AB10)*,*

$$d_{\mathrm{K}}[\mathcal{L}(\sqrt{Th}\hat{r}_{\mathrm{V},h}(x_0)),\ N(0, \tau^2(x_0))] = o(1), \tag{31}$$

$$d_{\mathrm{K}}[\mathcal{L}^*(\sqrt{Th}\hat{r}^*_{\mathrm{V},h}(x_0)),\ N(0, \tau^2(x_0))] = o_P(1), \tag{32}$$

where $\tau^2(x_0) = \sigma^2(x_0)p(x_0)\int K^2(v)\mathrm{d}v$.

Proof. We only give a proof of (32). Claim (31) can be proved along the same lines; see also Härdle and Tsybakov (1997). The results of Bosq (1996), Masry (1996) or Masry and Fan (1997) cannot be easily applied because they require conditions on the conditional density of (X_0, X_l) given (X_1, X_{l+1}) that are not easy to check in our case. For (32) it suffices to verify the assumptions of a version of the central limit theorem for martingale difference arrays (Brown 1971), namely

$$\frac{h}{T}\sum_t \mathrm{E}^*[K_h^2(x_0 - X^*_t)\tilde{\sigma}^2(X^*_t)(\varepsilon^*_{t+1})^2|\mathcal{F}^*_t] \to \sigma^2(x_0)p(x_0)\int K^2(v)\mathrm{d}v \tag{33}$$

in probability, and, for all $\delta > 0$, again in probability,

$$\frac{h}{T}\sum_t \mathrm{E}^*\left[K_h^2(x_0 - X_t^*)\tilde{\sigma}^2(X_t^*)(\varepsilon_{t+1}^*)^2 \mathbf{1}\left\{\frac{h}{T}K_h^2(x_0 - X_t^*)\tilde{\sigma}^2(X_t^*)(\varepsilon_{t+1}^*)^2 > \delta\right\}\middle|\mathcal{F}_t^*\right] \to 0. \tag{34}$$

Here $\mathcal{F}_t^* = \sigma(X_1^*, \ldots, X_t^*)$. Note that K and σ are bounded in a neighbourhood of x_0 and that $\sup_{|x-x_0|\leqslant h} |\tilde{\sigma}^2(x) - \sigma^2(x)| = o_P(1)$; see (AB9). So assertion (34) can be concluded from

$$\frac{1}{Th}\sum_t \mathrm{E}^*((\varepsilon_{t+1}^*)^2 \mathbf{1}\{(\varepsilon_{t+1}^*)^2 > c\delta hT\}) \leqslant T^{-1}(c\delta hT)^{-(\gamma-2)/2}\mathrm{E}^*(\varepsilon_{t+1}^*)^\gamma = o_P(1),$$

for all $c > 0$ with a constant $\gamma > \frac{5}{2}$; see (AB7). To see (33), we write

$$\frac{h}{T}\sum_t \mathrm{E}^*[K_h^2(x_0 - X_t^*)\tilde{\sigma}^2(X_t^*)(\varepsilon_{t+1}^*)^2|\mathcal{F}_t^*]$$

$$= \frac{\mathrm{E}^*(\varepsilon_1^*)^2}{Th}\sum_t \left(K^2\left(\frac{x_0 - X_t^*}{h}\right)\tilde{\sigma}^2(X_t^*) - \mathrm{E}^*\left[K^2\left(\frac{x_0 - X_t^*}{h}\right)\tilde{\sigma}^2(X_t^*)\middle|\mathcal{F}_{t-1}^*\right]\right)$$

$$+ \frac{\mathrm{E}^*(\varepsilon_1^*)^2}{Th}\sum_t \int K^2\left(\frac{x_0 - \tilde{m}(X_{t-1}^*) - \tilde{\sigma}(X_{t-1}^*)u}{h}\right)\tilde{\sigma}^2(\tilde{m}(X_{t-1}^*) + \tilde{\sigma}(X_{t-1}^*)u)\tilde{P}_\varepsilon(\mathrm{d}u).$$

Now the second conditional moment of the first summand (given $X_0, \ldots, X_T$) is of order $O_P(1/(T^2h^4))$. This can be easily seen by the fact that K is bounded and that $\tilde{\sigma}^2$ is stochastically bounded in an h-neighbourhood of x_0; see above. So the first summand is of order $O_P(1/(Th^2)) = o_P(1)$. It suffices to consider the second summand. Because $\mathrm{E}^*(\varepsilon_1^*)^2 = 1$, it is equal to

$$\frac{1}{T}\sum_t \int_{[-1,1]} K^2(v)\tilde{\sigma}^2(x_0 + hv)\tilde{p}_\varepsilon\left(\frac{x_0 - \tilde{m}(X_{t-1}^*)}{\tilde{\sigma}(X_{t-1}^*)} + \frac{hv}{\tilde{\sigma}(X_{t-1}^*)}\right)\frac{1}{\tilde{\sigma}(X_{t-1}^*)}\mathrm{d}v.$$

The argument of $\tilde{p}_\varepsilon$ is bounded in absolute value by

$$\frac{|x| + \sup_{|x|\leqslant\gamma_T}|\tilde{m}(x) - m(x)| + \sup_{|x|>\gamma_T}|\tilde{m}(x)|\sup_{|x|\leqslant\gamma_T}|m(x)| + h}{\min[\inf_{|x|\leqslant\gamma_T}\sigma(x), \inf_{|x|>\gamma_T}\tilde{\sigma}(x)] - \sup_{|x|\leqslant\gamma_T}|\tilde{\sigma}(x) - \sigma(x)|}.$$

This is of order $O_P(\gamma_T)$ by assumptions (AB1), (AB8) and (AB9). Therefore, by assumption (AB9), we can replace $\tilde{p}_\varepsilon$ by p_ε. Using uniform convergence of $\tilde{\sigma}$ to σ on compact sets (see assumption (AB9)), we obtain that the last expression is equal to

$$\frac{1}{T}\sum_t \int K^2(v)\sigma^2(x_0 + hv)p_\varepsilon\left(\frac{x_0 - \tilde{m}(X_{t-1}^*)}{\tilde{\sigma}(X_{t-1}^*)} + \frac{hv}{\tilde{\sigma}(X_{t-1}^*)}\right)\frac{1}{\tilde{\sigma}(X_{t-1}^*)}\,\mathrm{d}v + o_p(1).$$

We now use the fact that σ and p_ε have a bounded derivative and that $\tilde{\sigma}$ is stochastically bounded from below. This shows that the last expression is equal to

$$\int K^2(v)\mathrm{d}v \cdot \sigma^2(x_0) \cdot \frac{1}{T}\sum_t p_\varepsilon\left(\frac{x_0 - \tilde{m}(X^*_{t-1})}{\tilde{\sigma}(X^*_{t-1})}\right)\frac{1}{\tilde{\sigma}(X^*_{t-1})} + o_p(1).$$

Finally, we have to verify that this term converges in probability to $\tau^2(x_0)$. For this claim it suffices to show that

$$\frac{1}{T}\sum_t p_\varepsilon\left(\frac{x_0 - \tilde{m}(X^*_{t-1})}{\tilde{\sigma}(X^*_{t-1})}\right)\frac{1}{\tilde{\sigma}(X^*_{t-1})} = \frac{1}{T}\sum_t p_\varepsilon\left(\frac{x_0 - m(\tilde{X}_{t-1})}{\sigma(\tilde{X}_{t-1})}\right)\frac{1}{\sigma(\tilde{X}_{t-1})} + o_P(1), \qquad (35)$$

$$\frac{1}{T}\sum_t p_\varepsilon\left(\frac{x_0 - m(\tilde{X}_{t-1})}{\sigma(\tilde{X}_{t-1})}\right)\frac{1}{\sigma(\tilde{X}_{t-1})} = p(x_0) + o_P(1). \qquad (36)$$

Recall that $(\tilde{X}_t)$ is a process with conditional distribution equal to the unconditional distribution of (X_t). The process $(\tilde{X}_t)$ was constructed in the discussion before Lemma 4.1. The expectation of the left-hand side of (36) is equal to $p(x_0)$. Thus, (36) follows from the ergodicity of the process (X_t). Claim (35) means that the bootstrap process has in some sense an ergodic behaviour. Such a result will be needed at several points later on. We present the arguments in some detail here. The proof of (35) can be split into the following steps:

$$\frac{1}{T}\sum_t p_\varepsilon\left(\frac{x_0 - \tilde{m}(X^*_{t-1})}{\tilde{\sigma}(X^*_{t-1})}\right)\frac{1}{\tilde{\sigma}(X^*_{t-1})}\mathbf{1}\{|X^*_{t-1}| > \gamma_T\} = o_P(1); \qquad (37)$$

$$\frac{1}{T}\sum_t \left\{p_\varepsilon\left(\frac{x_0 - \tilde{m}(X^*_{t-1})}{\tilde{\sigma}(X^*_{t-1})}\right)\frac{1}{\tilde{\sigma}(X^*_{t-1})} - p_\varepsilon\left(\frac{x_0 - m(X^*_{t-1})}{\sigma(X^*_{t-1})}\right)\frac{1}{\sigma(X^*_{t-1})}\right\}\mathbf{1}\{|X^*_{t-1}| \leqslant \gamma_T\} = o_P(1); \qquad (38)$$

$$\frac{1}{T}\sum_t p_\varepsilon\left(\frac{x_0 - m(X^*_{t-1})}{\sigma(X^*_{t-1})}\right)\frac{1}{\sigma(X^*_{t-1})}\mathbf{1}\{|X^*_{t-1}| > \gamma_T\} = o_P(1); \qquad (39)$$

$$\frac{1}{T}\sum_t \left|p_\varepsilon\left(\frac{x_0 - m(X^*_{t-1})}{\sigma(X^*_{t-1})}\right)\frac{1}{\sigma(X^*_{t-1})} - p_\varepsilon\left(\frac{x_0 - m(\tilde{X}_{t-1})}{\sigma(\tilde{X}_{t-1})}\right)\frac{1}{\sigma(\tilde{X}_{t-1})}\right| = O\left(\frac{1}{T}\sum_t |X^*_t - \tilde{X}_t|\right). \qquad (40)$$

With (37)–(40), claim (35) follows from Lemma 4.3. This completes the proof of (32).

To see (37), observe that p_ε is bounded and that $\sup_{|x|>\gamma_T} \tilde{\sigma}^{-1}(x) = O_P(1)$. Thus the left-hand side of (37) is bounded by $(1/T)\sum_t \mathbf{1}\{|X^*_{t-1}| > \gamma_T\} O_P(1)$, which is $o_P(1)$ by Lemma 4.1. Similarly, (39) follows from Lemma 4.1 and the boundedness of p_ε and σ^{-1}. Claim (38) follows from the boundedness of p_ε and p'_ε and from the fact that $\tilde{m}$ and $\tilde{\sigma}$ converge uniformly on $[-\gamma_T, \gamma_T]$. For the proof of (40), note that by the boundedness of p'_ε and σ^{-1}

and by assumptions (AB1), (AB4) and (AB5), the function $\sigma(\cdot)^{-1} p_\varepsilon([x - m(\cdot)]/\sigma(\cdot))$ is Lipschitz continuous. □

The next lemma discusses kernel estimates of the stationary density.

Lemma 4.5. *Under assumptions* (AB1)–(AB10)*,*

$$\hat{p}_h(x_0) \to p(x_0) \text{ in probability,} \tag{41}$$

$$\hat{p}_h^*(x_0) \to p(x_0) \text{ in probability.} \tag{42}$$

Proof. Claim (41) follows from Theorem 1 in Masry (1996). It can also be obtained along the lines of the following proof of (42).

For a proof of (42), observe that

$$\mathrm{E}^* \left(\frac{1}{T} \sum_t \{ K_h(x - X_t^*) - \mathrm{E}^*[K_h(x - X_t^*)|\mathcal{F}_{t-1}^*]\} \right)^2 = O_P\left(\frac{1}{Th}\right) = o_P(1)$$

and

$$\mathrm{E}^*[K_h(x - X_t^*)|\mathcal{F}_{t-1}^*] = \int_{[-1,1]} K(v)\tilde{p}_\varepsilon\left(\frac{x - \tilde{m}(X_{t-1}^*) - vh}{\tilde{\sigma}(X_{t-1}^*)}\right)\frac{1}{\tilde{\sigma}(X_{t-1}^*)}\mathrm{d}v.$$

The argument of $\tilde{p}_\varepsilon$ is bounded in absolute value by $O_P(\gamma_T)$; see the proof of Lemma 4.4. By assumption (AB9), $\tilde{p}_\varepsilon$ converges uniformly on $[-C\gamma_T, C\gamma_T]$ towards p_ε. Thus, it suffices to consider

$$\frac{1}{T}\sum_t \int K(v) p_\varepsilon\left(\frac{x_0 - \tilde{m}(X_{t-1}^*) - vh}{\tilde{\sigma}(X_{t-1}^*)}\right)\frac{1}{\tilde{\sigma}(X_{t-1}^*)}\,\mathrm{d}v$$
$$= O_P(h) + \int K(v)\mathrm{d}v \frac{1}{T}\sum_t p_\varepsilon\left(\frac{x_0 - \tilde{m}(X_{t-1}^*)}{\tilde{\sigma}(X_{t-1}^*)}\right)\frac{1}{\tilde{\sigma}(X_{t-1}^*)}.$$

Lemma 4.5 now follows by application of (35). □

Finally, it remains to treat the bias terms.

Lemma 4.6. *Under assumptions* (AB1)–(AB10)*,*

$$\sqrt{Th}\,\hat{r}_{\mathrm{B},h}(x_0) \to b(x_0) \text{ in probability,} \tag{43}$$

$$\sqrt{Th}\,\hat{r}_{\mathrm{B},h}^*(x_0) \to b(x_0) \text{ in probability,} \tag{44}$$

where $b(x_0) = B \cdot \int v^2 K(v)\mathrm{d}v \cdot [p'(x_0)m'(x_0) + \frac{1}{2}p(x_0)m''(x_0)]$.

Proof. We only give the proof of (44). Claim (43) follows as in Masry (1996) or by a modification of the following proof of (44).

A Taylor expansion gives

$$\sqrt{Th}\hat{r}^*_{\mathrm{B},h}(x_0)$$
$$= \sqrt{\frac{h}{T}}\sum_t K_h(x_0 - X_t^*)(X_t^* - x_0)\tilde{m}'(x_0) + \frac{1}{2}\sqrt{\frac{h}{T}}\sum_t K_h(x_0 - X_t^*)(X_t^* - x_0)^2\tilde{m}''(\hat{X}_t), \tag{45}$$

where $\hat{X}_t$ denotes a suitable value between x_0 and X_t^*. We will show that

$$\sqrt{\frac{h}{T}}\sum_t \mathrm{E}^*[K_h(x - X_t^*)(X_t^* - x)|\mathcal{F}^*_{t-1}] \to Bp'(x)\int v^2 K(v)\mathrm{d}v, \tag{46}$$

$$\sqrt{\frac{h}{T}}\sum_t \mathrm{E}^*[K_h(x - X_t^*)(X_t^* - x)^2|\mathcal{F}^*_{t-1}] \to Bp(x)\int v^2 K(v)\mathrm{d}v. \tag{47}$$

Claim (44) then follows from convergence of $\tilde{m}'$ and $\tilde{m}''$ (see (AB9)–(AB10)), and from the fact that the conditional variances of the both terms on the right-hand side of (45) are of order $o_P(1)$.

For the proof of (46), note that the left-hand side of (46) is equal to

$$\sqrt{\frac{h^3}{T}}\sum_t \int vK(v)\tilde{p}_\varepsilon\left(\frac{x - \tilde{m}(X^*_{t-1}) + hv}{\tilde{\sigma}(X^*_{t-1})}\right)\frac{1}{\tilde{\sigma}(X^*_{t-1})}\,\mathrm{d}v.$$

A Taylor expansion for $\tilde{p}_\varepsilon$ yields, for a suitable value $\hat{Z}_t^*$ between $(x - \tilde{m}(X^*_{t-1}))/\tilde{\sigma}(X^*_{t-1})$ and $(x - \tilde{m}(X^*_{t-1}) + hv)/\tilde{\sigma}(X^*_{t-1})$, that this expression is equal to

$$\sqrt{\frac{h^5}{T}}\sum_t \int v^2 K(v)\tilde{p}'_\varepsilon(\hat{Z}_t^*)\frac{1}{\tilde{\sigma}(X^*_{t-1})}\,\mathrm{d}v.$$

In the proof of Lemma 4.5 we have already seen that $|\hat{Z}_t^*| = O_P(\gamma_T)$ for a suitable constant $C > 0$. Since $\tilde{p}'_\varepsilon(x)$ converges uniformly to $p'_\varepsilon(x)$ for $|x| \leqslant C\gamma_T$, for all $C > 0$ (see (AB9)), and since p''_ε is bounded, the left-hand side of (46) is asymptotically equal to

$$\sqrt{Th^5}\int v^2 K(v)\mathrm{d}v\,\frac{1}{T}\sum_t p'_\varepsilon\left(\frac{x - \tilde{m}(X^*_{t-1})}{\tilde{\sigma}(X^*_{t-1})}\right)\frac{1}{\tilde{\sigma}(X^*_{t-1})}.$$

Claim (46) now follows by similar arguments to those in the proof of (35). Claim (47) can be obtained along the same lines. □

For the proof of the second statement of Theorem 1, we use the following lemma. Recall that

$$\hat{\sigma}^2_{2,h}(x_0) = \hat{\sigma}^2_{2,1,h}(x_0) - 2\hat{\sigma}^2_{2,2,h}(x_0) + \hat{\sigma}^2_{2,3,h}(x_0),$$

$$\hat{\sigma}^{*2}_{2,h}(x_0) = \hat{\sigma}^{*2}_{2,1,h}(x_0) - 2\hat{\sigma}^{*2}_{2,2,h}(x_0) + \hat{\sigma}^{*2}_{2,3,h}(x_0),$$

where $\hat{\sigma}^2_{2,j,h}(x_0)$ and $\hat{\sigma}^{*2}_{2,j,h}$ are defined in (22)–(27), $j = 1, 2, 3$.

Lemma 4.7. *Under assumptions* (AB1)–(AB12), *for* $j = 1$ *and* $j = 2$,

$$\sqrt{Th}\hat{\sigma}^2_{2,j,h}(x_0) = o_P(1), \tag{48}$$

$$\sqrt{Th}\hat{\sigma}^{*2}_{2,j,h}(x_0) = o_P(1). \tag{49}$$

Proof. For a proof of (48) under mixing conditions, see Fan and Yao (1998). Their mixing conditions are difficult to check for the bootstrap process where the distribution of the simulated process is random and not fixed. For $j = 2$, we now give a proof of (49) that makes no use of mixing conditions. For $j = 3$, claim (49) can be shown by similar arguments.

For the proof of (49) for $j = 2$, we now write

$$\hat{\sigma}^{*2}_{2,2,h}(x_0) = \frac{1}{T(T-1)} \sum_{s,t=1}^{T-1} K_h(x_0 - X^*_t) K_h(X^*_s - X^*_t) \tilde{\sigma}(X^*_t) \varepsilon^*_{t+1} \frac{\tilde{m}(X^*_s) - \tilde{m}(X^*_t)}{\hat{p}^*_h(x_0)\hat{p}^*_h(X^*_t)}$$

$$+ \frac{1}{T(T-1)} \sum_{s,t=1}^{T-1} K_h(x_0 - X^*_t) K_h(X^*_s - X^*_t) \tilde{\sigma}(X^*_t) \tilde{\sigma}(X^*_s) \varepsilon^*_{t+1} \varepsilon^*_{s+1} [\hat{p}^*_h(x_0)\hat{p}^*_h(X^*_t)]^{-1}. \tag{50}$$

We now argue that the first term on the right-hand side of (50) differs only by terms of order $o_P(1/\sqrt{Th})$ from

$$\frac{1}{T(T-1)} \sum_{s,t=1}^{T-1} K_h(x_0 - X^*_t) K_h(X^*_s - X^*_t) \tilde{\sigma}(X^*_t) \varepsilon^*_{t+1} \frac{\tilde{m}(X^*_s) - \tilde{m}(X^*_t)}{p(x_0)p(X_t)}. \tag{51}$$

This follows from

$$\sup_{|x - x_0| \leq h} |\hat{p}^*_h(x) - p(x)| = O_P\left(h^2 + \sqrt{\log T / Th}\right). \tag{52}$$

Claim (52) can be shown with similar arguments to those for (55) below. The expression in (51) can be decomposed into the following three terms:

$$\frac{1}{T(T-1)}\sum_{s<t} K_h(x_0 - X_t^*)K_h(X_s^* - X_t^*)\tilde{\sigma}(X_t^*)\varepsilon_{t+1}^* \frac{\tilde{m}(X_s^*) - \tilde{m}(X_t^*)}{p(x_0)p(X_t^*)}$$

$$+ \frac{1}{T(T-1)}\sum_{s>t} K_h(x_0 - X_t^*)[p(x_0)p(X_t^*)]^{-1}\tilde{\sigma}(X_t^*)\varepsilon_{t+1}^*$$

$$\{K_h(X_s^* - X_t^*)[\tilde{m}(X_s^*) - \tilde{m}(X_t^*)] - \mathrm{E}^*[K_h(X_s^* - X_t^*)[\tilde{m}(X_s^*) - \tilde{m}(X_t^*)]|X_{s-1}^*]\}$$

$$+ \frac{1}{T(T-1)}\sum_{s>t} K_h(x_0 - X_t^*)[p(x_0)p(X_t^*)]^{-1}\tilde{\sigma}(X_t^*)\varepsilon_{t+1}^*$$

$$\mathrm{E}^*[K_h(X_s^* - X_t^*)[\tilde{m}(X_s^*) - \tilde{m}(X_t^*)]|X_{s-1}^*],$$

where $\mathrm{E}^*[\ldots|X_{s-1}^*]$ denotes the conditional expectation, given $X_0, \ldots, X_T$ and X_{s-1}^*. By calculation of the second conditional moments of these terms, one can show that these terms are of order $o_P(1/\sqrt{Th})$. This implies that the first term on the right-hand side of (50) is of order $o_P(1/\sqrt{Th})$. It remains to show that $A = o_P(1/\sqrt{Th})$, where A is the second term on the right-hand side of (50).

For the proof of this claim, we will use the fact that, for $0 < \rho < \frac{4}{5}$,

$$\sup_{|x-x_0|\leqslant h}\left|\frac{1}{T}\sum_{s=0}^{T-1} K_h(x - X_s^*)\tilde{\sigma}(X_s^*)\varepsilon_{s+1}^*\right| = O_P\left(\sqrt{\frac{\log T}{Th}}\right), \tag{53}$$

$$\sup_{|x-x_0|\leqslant h}|\hat{p}_h^*(x) - p(x_0)| = O_P\left(h + \sqrt{\frac{\log T}{Th}}\right), \tag{54}$$

$$\sup_{|x-x_0|\leqslant h, |x-x'|\leqslant T^{-\rho}h}\left|\frac{1}{T}\sum_{s=0}^{T-1}[K_h(x - X_s^*) - K_h(x' - X_s^*)]\tilde{\sigma}(X_s^*)\varepsilon_{s+1}^*\right| = O_P\left(T^{-\rho}\sqrt{\frac{\log T}{Th}}\right). \tag{55}$$

A proof of claim (55) will be given below. Claims (53) and (54) follow by similar arguments.

For $-T^{\rho} < j < T^{\rho} + 1$, with $0 < \rho < \frac{2}{5}$, we now define

$$A_j = \frac{(p(x_0))^{-2}}{T(T-1)}\sum_{s,t=1}^{T-1} 1(x_{j-1} < X_t^* \leqslant x_j)K_h(x_0 - X_t^*)K_h(X_s^* - x_j)\tilde{\sigma}(X_t^*)\tilde{\sigma}(X_s^*)\varepsilon_{t+1}^*\varepsilon_{s+1}^*,$$

with $x_j = x_0 + jT^{-\rho}h$. From (53)–(55) we obtain

$$A = \sum_{-T^{\rho}<j<T^{\rho}+1} A_j + o_P\left(\frac{1}{\sqrt{Th}}\right). \tag{56}$$

We now apply

$$\mathrm{E}^*|A_j| \leqslant p(x_0)^{-2}\left\{\mathrm{E}^*\left[\frac{1}{T}\sum_{s=1}^{T-1} K_h(x_j - X_s^*)\tilde{\sigma}(X_s^*)\varepsilon_{s+1}^*\right]^2\right\}^{1/2}$$

$$\times\left\{\mathrm{E}^*\left[\frac{1}{(T-1)}\sum_{t=1}^{T-1} 1(x_{j-1} < X_t^* \leqslant x_j)K_h(x_0 - X_t^*)\tilde{\sigma}(X_t^*)\varepsilon_{t+1}^*\right]^2\right\}^{1/2}$$

$$= O_P\left(\frac{T^{-\rho/2}}{Th}\right).$$

This implies $\mathrm{E}^*|A| = O_P(T^{\rho/2}/Th)$ and, in particular, our claim $A = o_P((Th)^{-1/2})$, because $\rho < \frac{4}{5}$.

It remains to prove (55). Put $\eta_s' = \varepsilon_s^* 1(|\varepsilon_s^*| \leqslant T^{1/5})$ and $\eta_s = \eta_s{}' - \mathrm{E}^*\eta_s{}'$. Then we have, because of (AB12), that

$$\begin{aligned} P^*(\eta_s' \neq \varepsilon_s^* \text{ for some } 0 \leqslant s \leqslant T) &\leqslant P^*(|\varepsilon_s^*| > T^{1/5} \text{ for some } 1 \leqslant s \leqslant T) \\ &\leqslant (T+1)P^*(|\varepsilon_1^*| > T^{1/5}) \\ &\leqslant (T+1)T^{-8/5}\mathrm{E}^*|\varepsilon_1^*|^8 \\ &= o_P(1). \end{aligned}$$

Furthermore, $|\mathrm{E}^*\eta_s'| = |\mathrm{E}^*\varepsilon_s^* - \eta_s{}'| = |\mathrm{E}^*\varepsilon_s^* 1(|\varepsilon_s^*| > T^{1/5})| \leqslant T^{-4/5}\mathrm{E}^*|\varepsilon_s^*|^5 1(|\varepsilon_s^*| > T^{1/5}) = O_P(T^{-4/5})$. So for (55) it suffices to show that

$$\sup_{|x-x_0|\leqslant h, |x-x'|\leqslant T^{-\rho}h} |B(x, x')| = O_P\left(T^{-\rho}\sqrt{\frac{\log T}{hT}}\right), \tag{58}$$

where

$$B(x, x') = \frac{1}{T}\sum_{s=0}^{T-1}[K_h(x - X_s^*) - K_h(x' - X_s^*)]\tilde{\sigma}(X_s^*)\eta_{s+1}.$$

We will show that, for all $C > 0$, there exists a sequence $C' > 0$ with

$$\sup_{|x-x_0|\leqslant h, |x-x'|\leqslant T^{-\rho}h} P^*\left(B(x, x') > C'T^{-\rho}\sqrt{\frac{\log T}{hT}}\right) < T^{-C}, \tag{59}$$

$$\sup_{|x-x_0|\leqslant h, |x-x'|\leqslant T^{-\rho}h} P^*\left(B(x, x') < -C'T^{-\rho}\sqrt{\frac{\log T}{hT}}\right) < T^{-C}. \tag{60}$$

Inequalities (59) and (60) imply that (58) holds with a supremum that runs only over a grid of $o(T^C)$ points (x, x'). By an appropriate choice of grid, this shows (58) with a supremum that runs over all points of the set. We give now a proof of (59). Claim (60) can be shown by similar arguments.

We will show (59) by splitting $B(x, x')$ into two terms $B(x, x') = B_1(x, x') + B_2(x, x')$, with

$$B_1(x, x') = \sum_{s=0}^{(T-1)/2} w_{2s}(x, x')\eta_{2s+1}, \tag{61}$$

$$w_s(x, x') = T^{-1}[K_h(x - X_s^*) - K_h(x' - X_s^*)]\tilde{\sigma}(X_s^*), \tag{62}$$

where, without loss of generality, we have assumed that T is odd. We will show that, for all $C > 0$, there exists a sequence $C' > 0$ with

$$\sup_{|x-x_0|\leqslant h, |x-x'|\leqslant T^{-\rho}h} P^*\left(B_1(x, x') > C'T^{-\rho}\sqrt{\frac{\log T}{hT}}\right) < T^{-C}. \tag{63}$$

With a similar expression for $B_2(x, x')$, this implies our claim (59). With $w_s'(x, x') = \sqrt{hT \log T}\, T^\rho w_s(x, x')$, we have, for $t > 0$, that

$$P^*\left(B_1(x, x') > C'T^{-\rho}\sqrt{\frac{\log T}{hT}}\right) = P^*\left(\sum_{s=0}^{(T-1)/2} w_{2s}(x, x')\eta_{2s+1} > C'T^{-\rho}\sqrt{\frac{\log T}{hT}}\right)$$

$$\leqslant \mathrm{E}^* \exp\left(t\sum_{s=0}^{(T-1)/2} w_{2s}'(x, x')\eta_{2s+1}\right)\exp(-tC' \log T).$$

For two constants A_1 and A_2, we now use the inequalities

$$\mathrm{E}^*[\exp(tw_{2s}'(x, x')\eta_{2s+1})|\mathcal{F}_{2s}] \leqslant 1 + t^2 A_1 w_{2s}'(x, x')^2, \tag{64}$$

$$\sup_{|x-x_0|\leqslant h, |x-x'|\leqslant T^{-\rho}h} \mathrm{E}^*[1 + t^2 A_1 w_{2s}'(x, x')^2|\mathcal{F}_{2s-1}] \leqslant 1 + t^2 A_2 T^{-1} \log T \tag{65}$$

$$\leqslant \exp(t^2 A_2 T^{-1} \log T),$$

which hold almost surely. For the proof of (64) one uses the fact that $w_{T-1}'(x, x')\eta_T$ is almost surely bounded. By iterative application of (64) and (65) we obtain that

$$\sup_{|x-x_0|\leqslant h, |x-x'|\leqslant T^{-\rho}h} P^*\left(B_1(x, x') > C'T^{-\rho}\sqrt{\frac{\log T}{hT}}\right)$$

$$\leqslant \exp\left(\frac{T-1}{2} t^2 A_2 T^{-1} \log T - tC' \log T\right). \tag{66}$$

By appropriate choice of t and C', the right-hand side of (66) is smaller than n^{-C}. This shows (63). □

4.2. Proof of Theorems 2 and 3

Consistency of the bootstrap for $\mathcal{L}_{\mathrm{B}}(x_0)$ and $\mathcal{L}^{\sigma}_{1,\mathrm{B}}(x_0)$ (for Theorem 2) follows by standard smoothing arguments. As in the proof of Theorem 1, one decomposes the estimates $\hat{m}_h$ and $\hat{\sigma}^2_{1,h}$ and their counterparts in the bootstrap world into a variance and a bias component. Smoothness of the pilot estimate $\tilde{m}$ (or $\tilde{\sigma}$) entails correct approximations of the bias terms by bootstrap. Asymptotic normality of the variance term follows by a standard application of the central limit theorem. For an analysis of the regression bootstrap of $\hat{\sigma}^2_{2,h}(x_0)$ one decomposes $\hat{\sigma}^2_{2,h}(x_0)$ and $\hat{\sigma}^{*2}_{2,h}(x_0)$ as in equations (22)–(27). Now one shows that the unconditional distribution of $\hat{\sigma}^2_{2,1,h}$ and the conditional distribution of $\hat{\sigma}^{*2}_{2,1,h}$ are asymptotically equivalent. This can be seen with the same arguments as for $\hat{m}_h$. Then Theorem 2 follows by showing that $\sqrt{Th}\hat{\sigma}^2_{2,j,h}(x_0)$ and $\sqrt{Th}\hat{\sigma}^{*2}_{2,j,h}(x_0)$ are of order $o_P(1)$ for $j = 2, 3$. This can be shown as in the proof of Lemma 4.7.

4.3. Proof of Theorem 4

First one shows that the finite-dimensional distributions of $Z_T(\eta)$ converge weakly to the finite-dimensional distributions of a Gaussian process with mean function μ and covariance function R. This can be done along the lines of the proof of Theorem 1. It remains to prove the tightness of the process $Z_T(\cdot)$. This follows from the tightness of the following three processes:

$$Z_{T,1}(\eta) = \frac{T^{-2/5}}{\sqrt{\eta}} \sum_t K\left(\frac{x - X_t}{\eta} T^{1/5}\right)\sigma(X_t)\varepsilon_{t+1}, \tag{67}$$

$$Z_{T,2}(\eta) = \frac{T^{-4/5}}{\eta} \sum_t K\left(\frac{x - X_t}{\eta} T^{1/5}\right), \tag{68}$$

$$Z_{T,3}(\eta) = \frac{T^{-2/5}}{\sqrt{\eta}} \sum_t K\left(\frac{x - X_t}{\eta} T^{1/5}\right)(m(X_t) - m(x)). \tag{69}$$

To prove the tightness of these processes, we make use of a tightness criterion of Billingsley (1968, Theorem 12.3).

For the first process $Z_{T,1}$, we have, for $\eta_1, \eta_2 \in [a, b]$, that

$$\mathrm{E}(Z_{T,1}(\eta_1) - Z_{T,1}(\eta_2))^2 = \int \left(\frac{1}{\sqrt{\eta_1}} K\left(\frac{v}{\eta_1}\right) - \frac{1}{\sqrt{\eta_2}} K\left(\frac{v}{\eta_2}\right)\right)^2 \sigma^2(x - T^{-1/5}v)p(x - T^{-1/5}v)\mathrm{d}v.$$

The differentiability of K, boundedness of σ^2 and p, and the fact that $\eta_1, \eta_2 \geqslant a > 0$ imply that this expression is bounded by some constant times $(\eta_1 - \eta_2)^2$. This implies tightness of $Z_{T,1}$.

For the tightness of $Z_{T,2}$, one first verifies the tightness criterion of Billingsley (1968) for the process

$$\frac{T^{-4/5}}{\eta}\sum_t\left\{K\left(\frac{x-X_t}{\eta}T^{1/5}\right)-\mathrm{E}\left[K\left(\frac{x-X_t}{\eta}T^{1/5}\right)\middle|\mathcal{F}_{t-1}\right]\right\}.$$

This can be done as for $Z_{T,1}$. It remains to consider

$$\frac{T^{-4/5}}{\eta}\sum_t\mathrm{E}\left[K\left(\frac{x-X_t}{\eta}T^{1/5}\right)\middle|\mathcal{F}_{t-1}\right]=\frac{1}{T}\sum_t\int K(v)p_\varepsilon\left(\frac{x-m(X_{t-1})-\eta vT^{-1/5}}{\sigma(X_{t-1})}\right)\frac{1}{\sigma}X_{t-1}\,\mathrm{d}v.$$

It can be shown that the difference between this last expression and

$$\frac{1}{T}\sum_t\int K(v)\mathrm{d}v\;p_\varepsilon\left(\frac{x-m(X_{t-1})}{\sigma(X_{t-1})}\right)\frac{1}{\sigma}X_{t-1}$$

converges uniformly in η to zero. The last term does not depend on η and converges in probability to

$$\mathrm{E}p_\varepsilon\left(\frac{x-m(X_1)}{\sigma(X_1)}\right)\frac{1}{\sigma(X_1)}=p(x);$$

cf. (10).

The tightness of $Z_{T,3}(\cdot)$ can be shown by using a Taylor expansion for $m(X_t)-m(x)$ and similar arguments to those for $Z_{T,2}(\cdot)$.

4.4. Proof of Theorem 5

We will show that (P1)–(P3) imply, for all $c>0$, that

$$\sup_{|x|\leqslant c}|\tilde{m}'(x)-m'(x)|=o_P(1),\tag{70}$$

$$\sup_{|x-x_0|\leqslant h}|\tilde{m}^{(2)}(x)-m^{(2)}(x)|=o_P(1).\tag{71}$$

Note that by a trivial argument this would imply that (70) holds with c replaced by a sequence γ_T that converges slowly enough to ∞. The other claims of (AB9) and (AB10) can be shown by similar arguments. Furthermore, (AB8) trivially holds and (AB7) immediately follows from $\sup_{|x|\leqslant\gamma_T}|\hat{m}_h(x)-m(x)|=o_P(1)$. This can be shown in the same way as (70). So for the first statement of Theorem 5 it remains to show (70) and (71). The second statement follows by similar arguments.

We now come to the proof of (70). We will show that

$$\sup_{|x|\leqslant c}\left|\frac{1}{T-1}\sum_{t=1}^{T-1}K_g'(X_t-x)\sigma(X_t)\varepsilon_{t+1}\right|=o_P(1).\tag{72}$$

It is easy to see that this expression and bias considerations imply (70). For a proof of (72), put $\eta_t'=\varepsilon_t\mathbf{1}(|\varepsilon_t|\leqslant\sqrt{Tg^3})$. Then, under our assumptions, $\eta_t'=\varepsilon_t$ for all $2\leqslant t\leqslant T$ with probability tending to one; compare (57). This implies that, with probability tending to one $(1/(T-1))\sum_{t=1}^{T-1}K_g'(X_t-x)\sigma(X_t)\varepsilon_{t+1}=\sum_{t=1}^{T-1}W_t(x)$, where $W_t(x)=(1/(T-1))$

$K'_g(X_t - x)\sigma(X_t)\eta_{t+1}$. Now, by construction, $W_t(x)$ is absolutely bounded for $1 \leqslant t \leqslant T-1$ and for all x. This enables us to show (72) by the same methods as in the proof of (59).

It remains to show (71). For $|z| \leqslant 1$ consider the process $R(z) = \tilde{m}^{(2)}(x_0 + zh) - m^{(2)}(x_0 + zh)$. It is easy to check that $R(z) = o_P(1)$ for all $|z| \leqslant 1$. Claim (71) follows by showing the tightness of the process R. This can be done as in the proof of Theorem 4.

Acknowledgement

Financial support was received by the Deutsche Forschungsgemeinschaft, SFB 373 ('Quantifikation und Simulation ökonomischer Prozesse'), Humboldt-Universität zu Berlin, and from Project MA1026/6-1, and is greatly appreciated. The authors also wish to thank the referees. Their detailed reports led to a considerable improvement of the paper.

References

Ango Nze, P. and Portier, D. (1994) Estimation of the density and of the regression functions of an absolutely regular stationary process. *Publ. Inst. Statist. Univ. Paris*, **38**, 59–88.

Ango Nze, P., Bühlmann, P. and Doukhan, P. (1999) Nonparametric regression estimation under weak dependence beyond mixing and association. Preprint.

Auestad, B. and Tjøstheim, D. (1990) Identification of nonlinear time series: First order characterization and order determination. *Biometrika*, **77**, 669–687.

Bickel, P.J. and Bühlmann, P. (1999) A new mixing notion and functional central limit theorems for a sieve bootstrap in time series. *Bernoulli*, **5**, 413–446.

Bickel, P.J. and Freedman, D.A. (1981) Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**, 1196–1217.

Billingsley, P. (1968) *Convergence of Probability Measures*. New York: Wiley.

Bosq, D. (1996) *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Lecture Notes in Statist. 110. New York: Springer-Verlag.

Brown, B.M. (1971) Martingale central limit theorems. *Ann. Math. Statist.*, **42**, 54–66.

Collomb, G. (1984) Propriétés de convergence presque complète du predicteur à noyau. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, **66**, 441–460.

Diebolt, J. and Guegan, D. (1990) Probabilistic properties of the general nonlinear Markovian process of order one and applications to time series modelling. Technical report 125, LSTA, Université Paris XIII.

Doukhan, P. (1994) *Mixing: Properties and Examples*, Lecture Notes in Statist. 85. Berlin: Springer-Verlag.

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

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

Fan, J. and Gijbels, I. (1992) Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, **20**, 2008–2036.

Fan, J. and Gijbels, I. (1995) *Local Polynomial Modelling and its Applications – Theory and Methodologies*. London: Chapman & Hall.

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

Franke, J. and Härdle, W. (1992) On bootstrapping kernel spectrum estimates. *Ann. Statist.*, **20**, 121–145.

Franke, J. and Wendel, M. (1992) A bootstrap approach for nonlinear autoregressions – some preliminary results. In K.-H. Jöckel, G. Rothe and W. Sendler (eds), *Bootstrapping and Related Techniques*, Lecture Notes in Econom. and Math. Systems 376. Berlin: Springer-Verlag.

Franke, J., Kreiss, J.-P., Mammen, E. and Neumann, M.H. (1999) Properties of the nonparametric autoregressive bootstrap. Preprint.

Gouriéroux, C. and Montfort, A. (1992) Qualitative threshold ARCH models. *J. Econometrics*, **52**, 159–199.

Györfy, L. Härdle, W., Sarda, P. and Vieu, P. (1989) *Nonparametric Curve Estimation from Time Series*, Lecture Notes in Statist. 60. Berlin: Springer-Verlag.

Härdle, W. and Bowman, A. (1988) Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.*, **83**, 102–110.

Härdle, W. and Mammen, E. (1993) Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, **21**, 1926–1947.

Härdle, W. and Vieu, P. (1992) Kernel regression smoothing of time series. *J. Time Ser. Anal.*, **13**, 209–232.

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

Härdle, W., Lütkepohl, H. and Chen, R. (1997) A review of nonparametric time series analysis. *Internat. Statist. Rev.*, **65**, 49–72.

Hjellvik, V. and Tjøstheim, D. (1995) Nonparametric tests of linearity for time series. *Biometrika*, **82**, 351–368.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1996) A brief survey of bandwidth selection for density estimation. *J. American Statist. Assoc.*, **91**, 401–407.

Kreiss, J.-P. and Franke, J. (1992) Bootstrapping stationary autoregressive moving-average models. *J. Time Ser. Anal.*, **13**, 297–317.

Kreiss, J.-P., Neumann, M.H. and Yao, Q. (1998) Bootstrap tests for simple structure in nonparametric time series regression. Technical Report, Technische Universität Braunschweig.

Kreutzberger, E. (1993) Bootstrap für nichtlineare AR(1)-Prozesse. Thesis, University of Kaiserslautern, Germany.

Künsch, H.R. (1989) The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, **17**, 1217–1241.

Mammen, E. (1992) *When Does Bootstrap Work? Asymptotic Results and Simulations*, Lecture Notes in Statist. 77. Berlin: Springer-Verlag.

Masry, E. (1996) Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Process. Appl*., **65**, 81–101. Correction (1997): **67**, 281.

Masry, E. (1997) Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *J. Time Ser. Anal.*, **17**, 571–599.

Masry, E. and Fan, J. (1997) Local polynomial estimation of regression functions for mixing processes. *Scand. J. Statist*., **24**, 165–179.

Masry, E. and Tjøstheim, D. (1994) Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality. *Econometric Theory*, **11**, 258–289.

Meyn, S.P. and Tweedie, R.L. (1993) *Markov Chains and Stochastic Stability*. Berlin: Springer-Verlag.

Neumann, M.H. and Kreiss, J.-P. (1998) Regression-type inference in nonparametric autorregression.

Ann. Statist., **26**, 1570–1613.

Paparoditis, E. and Politis, D.N. (1999) A Markovian local resampling scheme for nonparametric estimators in time series analysis. Preprint.

Paparoditis, E. and Politis, D.N. (2000) The local bootstrap for kernel estimators under general dependence conditions. *Ann. Inst. Statist. Math.*, **52**, 139–159.

Priestley, M.B. (1988) *Nonlinear and Nonstationary Time Series Analysis*. London: Academic Press.

Robinson, P.M. (1983) Nonparametric estimators for time series. *J. Time Ser. Anal.*, **4**, 185–207.

Stone, C.J. (1977) Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–620.

Tjøstheim, D. (1994) Non-linear time series: A selective review. *Scand. J. Statist.*, **21**, 97–130.

Tong, H. (1990) *Non-linear Time Series. A Dynamic System Approach.* Oxford: Oxford University Press.

Tsybakov, A.B. (1986) Robust reconstruction of functions by the local approximation method. *Problems Inform. Transmission*, **22**, 133–146.

Received February 1997 and revised July 2001