RESEARCH ARTICLE

Nonparametric estimation for dependent data[†]

Jan Johannes[a*] and Suhasini Subba Rao[b]

[a]*Institute of Applied Mathematics, University Heidelberg, 69120 Heidelberg, Germany*;
[b]*Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.*

(*26 February 2008*)

In this paper we consider nonparametric estimation for dependent data, where the observations do not necessarily come from a linear process. We study density estimation and also discuss associated problems in nonparametric regression using the 2-mixing dependence measure. We compare the results under the 2-mixing with those derived under the assumption that the process is linear.

Keywords: Density estimation; nonparametric regression; 2-mixing; nonlinear processes.

AMS Subject Classification: Primary: 62G05; 62M10; Secondary: 62G07; 62G08.

1. Introduction

Nonparametric estimation for dependent observations has a long history in statistics. Rosenblatt [42] first studied density estimation for dependent data. Since then several authors have considered nonparametric estimation under various assumptions (notable early articles include Robinson [39] and Hart [29]). For example, Hall and Hart [25], Giraitis et al. [22], Mielniczuk [34] and Estevas and Vieu [18] consider density estimation for linear processes which have long memory, whereas Cheng and Robinson [9] consider density estimation for random variables which are nonlinear functions of a linear process. A notable result, is that they show if the observations were from a linear process and have short memory, then the usual rate of convergence, known for independent observations, also holds for dependent observations. On the other hand, for long memory processes, the rate of convergence is different. Interestingly, despite long memory influencing the rate of convergence, there is no influence of long memory on the bandwidth choice, which is same regardless of short or long memory. In other words, if the observations come from a linear process, a larger bandwidth does not improve the rate of convergence

[†]This work was partially supported by the DFG (DA 187/12-3).
[*]Corresponding author. Email: johannes@statlab.uni-heidelberg.de
ISSN: 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/1048525YYxxxxxxxx
http://www.informaworld.com

of the density estimator. Similar results can also be derived for nonparametric regression problems (c.f. Hall and Hart [26], Cheng and Robinson [10], Hall et al. [28], Claesken and Hall [11] and Csörgö and Mielniczuk [13, 14, 15] and Bryk and Mielniczuk [7]). However, usually it is assumed that the observations come from a linear process or are functions of a linear or Gaussian process (often refered to as generalised Gaussian processes). In the case of linearity, the joint density of the observations can be characterized (in some sense) in terms of the autocovariances. It is this representation that allows for the mean squared error of the nonparametric estimator to be derived in terms of the autocovariance function. However this result does not necessarily hold when the process is nonlinear.

The assumption of linearity can be relaxed by using the notion of 2-mixing (see Bosq [3] and Bradley [5]). Unlike the autocovariance function, 2-mixing can be considered as a measure of dependence between two random variables (see Definition 2.3, below) and the *2-mixing size* quantifies this dependence: a large mixing size indicates little dependence, whereas a small mixing size indicates large dependence. Since the strong mixing size gives a lower bound for the 2-mixing size it can be established for several types of processes, for example, linear processes, see Athreya and Pantula [1], Chanda [8], Gorodetskii [24] and Appendix A.3, and also nonlinear processes, see, Pham [35], Masry and Tjøstheim [33], Cline and Pu [12], Bousamma [4] and Basrak et al. [2] (where many of the these results show geometric ergodicity, which implies 2-mixing of the process). Assuming that the 2-mixing size (or strong mixing size) is *sufficiently* large, Robinson [39] (see assumption A3.1 and A3.2) and Bosq [3] (see, also, Vieu [46], Viano et al. [45], Mielniczuk [34], Fan and Yao [19] etc.) obtain the rate of convergence for nonparametric kernel estimators. However, despite, there existing a huge body of literature on rates of convergence for nonparametric kernel estimators based on the assumption of linearity of the process, and some on rates of convergence for processes which are 2-mixing with a sufficiently large 2-mixing size, as far as we are aware, very little exists on rates of convergence of nonparametric kernel estimators for nonlinear processes whose 2-mixing size *can be small*. This is particularly pertinent, as nonlinear processes with small mixing sizes can arise in several applications, for example the ARCH(∞) process is a nonlinear process which is used to model finance time series and can have a small mixing size (see Fryzlewicz and Subba Rao [20] for the details). In this paper we address this issue, and obtain rates of covergence for nonparametric kernel estimators for dependent data and formulate the results in terms of the 2-mixing size. We study both density estimation and also problems in nonparametric regression.

In Section 3 we consider kernel density estimation, in particular we obtain the sampling properties of the Rosenblatt-Parzen kernel estimator and obtain a bound for the mean squared error under the assumption that the time series are stationary and 2-mixing. We show that, like the long memory process, the 2-mixing size can influence the rate of convergence. But unlike the long memory process, a much larger 2-mixing size may be required to obtain the usual rate of convergence. Moreover, the bandwidths which minimise the obtained bounds are influenced by

the 2-mixing size - the smaller the 2-mixing size the larger the bandwidth. We demonstrate that several problems could arise if one were to falsely suppose that observations were from a linear process, when they do not. For example, if the usual optimal bandwidth for linear processes were used on nonlinear processes, the mean squared error may no longer converge to zero. Thus our results give a warning to practitioners who apply well known results for the linear process, without checking whether the process is linear or not.

In Section 4 we consider nonparametric regression for dependent data. We discuss this with reference to two models. First we suppose the response and explanatory variables (X_t, Z_t) satisfy (i) $X_t = \varphi(Z_t) + h(Z_t)\eta_t$, where $\{\eta_t\}$ and $\{Z_t\}$ are independent of each other (Hart [30] considers a particular example of this model, where $h(\cdot)$ is a constant), and secondly we assume the conditional expectation satisfies (ii) $(X_t|Z_t) = \varphi(Z_t)$. We observe that the latter model includes the former model as a special case. We estimate $\varphi(\cdot)$ using the classical kernel estimator and derive rates of convergence similar to those obtained for the density estimator. But in the case of model (i) the rate of convergence depends on two factors, the 2-mixing size of $\{Z_t\}$ and the rate of decay of the autocovariance function of $\{\eta_t\}$, whereas for model (ii) the rate of convergence is determined by the mixing size of the multivariate random process $\{(X_t, Z_t)\}$.

All the proofs can be found in the appendix. Also some 2-mixing inequalities for linear processes used here are included in the appendix.

2. Notation

In this section we introduce some definitions that will be used in the paper. Note we will assume all the necessary densities exist. We start by defining the kernel.

Definition 2.1 A kernel K is of order r (see Scott [43]), if K is a univariate, even function such that

$$\int du\ K(u) = 1, \quad \int du\ u^i K(u) = 0$$

for all $i = 1, \ldots, r-1$ and there exists a constant S_K such that

$$\int du\ |u|^r K(u) = S_K.$$

Let $K_b(z) := b^{-d}K(z/b)$, where $b > 0$ is a bandwidth. Below we define the smoothness class (c.f. Robinson [38]) which we use to bound the bias of the estimators.

Definition 2.2 For $s, \triangle > 0$, the space $\mathfrak{G}^s_{\triangle}$ is the class of functions $g : \quad \rightarrow$ satisfying: g is everywhere $(m-1)$-times differentiable for $m - 1 < s \leqslant m$; where

for some $\rho > 0$ and for all x, the inequality

$$\sup_{y:|y-x|<\rho} \frac{|g(y) - g(x) - Q(y-x)|}{|y-x|^s} \leqslant \triangle,$$

holds true with $Q = 0$ when $m = 1$ and for $m > 1$, Q is an $(m-1)$th-degree homogeneous polynomial in $y - x$, whose coefficients are the derivatives of g of orders 1 to $m-1$ evaluated at x; and $\triangle$ is a finite constant.

The dependence of the process $\{Y_t\}$ is quantified in terms of 2-mixing size, which we define below.

Definition 2.3

(i) A stationary process $\{Y_t\}$ is said to be 2-mixing with size $\mathfrak{v}$ if for all $t \neq 0$

$$\sup_{A\in\sigma(Y_t),B\in\sigma(Y_0)} |P(A \cap B) - P(A)P(B)| \leq C|t|^{-\mathfrak{v}}.$$

for some $C < \infty$ independent of t.

(ii) The covariance of a stationary process $\{Y_t\}$ has size $\mathfrak{u}$ if for all $t \neq \tau$, $|\mathrm{cov}(Y_t, Y_0)| \leq C|t|^{-\mathfrak{u}}$ for some $C < \infty$ independent of t.

We note that the notion of the strong mixing is defined in a similar way (see Bradley [5] for properties of strong mixing and Rio [37] for applications in central limit theorems). However the crucial difference between 2-mixing and strong mixing are the sigma-algebras over which the supremum is taken. In the definition of strong-mixing the sigma algebra is over the entire left and right tails of $\{X_t\}$, whereas the sigma-algebras in the definition 2-mixing are more restrictive. We observe that the covariance is a measure of linear dependence, whereas 2-mixing is a generalization of this, and can be considered as a measure of dependence. 2-mixing is quite a general notion, which is satisfied by several processes. For example, under certain conditions on the innovations, most linear models are 2-mixing (see Appendix A.3, and Athreya and Pantula [1], Cline and Pu [12] and Chanda [8], where strong mixing is shown). Further, under additional conditions on the innovations and the parameters, ARCH/GARCH processes are also strongly mixing (c.f. Masry and Tjøstheim [33], Bousamma [4] and Basrak et al. [2]) which implies that they also 2-mixing. Most of the results and bounds in this paper are derived using 2-mixing. In general, the larger the mixing size the faster the rate of convergence. For example, in the case of iid observations (the 2-mixing size can be treated as ∞) using just a few observations, information over the entire domain of the density function can be obtained. On the other hand, a sample which has a small mixing size (so tends to be clustered about certain points) will require a much larger number of observations to give the same information.

For brevity, we use the standard notation $\wedge$ to denote minimum and $\vee$ to denote maximum.

3. Kernel density estimation

Suppose we observe the stationary time series $\{Z_1, \ldots, Z_T\}$, and let f denote the density of Z_t. The most popular estimator of f, is the Rosenblatt-Parzen kernel estimator

$$\hat{f}(u) = \frac{1}{T}\sum_{t=1}^{T} K_b(Z_t - u), \tag{1}$$

where $K_b(z)$ is defined below Definition 2.1. In this section we investigate the sampling properties of the kernel density estimator defined above. The dependence of the process $\{Z_t\}$ is quantified in terms of its 2-mixing size (see Definition 2.3).

We first derive a bound for the mean squared error (MSE) $|\hat{f}(z) - f(z)|^2$ using only minimal assumptions on the distribution of $\{Z_t\}$.

Proposition 3.1 Suppose the univariate stationary process $\{Z_t\}$ is 2-mixing with size $\mathfrak{v}$ and the marginal density f of Z_t and its second derivative f'' are both uniformly bounded. Let $\hat{f}$ be defined as in (1), where K is a rectangular kernel, i.e., $K(x) = 1$ if $x \in [-1/2, 1/2]$ and zero otherwise. Then we have

$$|\hat{f}(z) - f(z)|^2 = O(b^4 + T^{-[\mathfrak{v}\wedge 1]} b^{\frac{-[(\mathfrak{v}\vee 1)+1]}{\mathfrak{v}\vee 1}}) = \begin{cases} O(b^4 + T^{-1}b^{-\frac{\mathfrak{v}+1}{\mathfrak{v}}}) & \mathfrak{v} > 1 \\ O(b^4 + T^{-\mathfrak{v}}b^{-2}) & \mathfrak{v} \leq 1 \end{cases}$$

Proof. To prove the result we will bound the risk using the standard variance bias decomposition. First the bias: as we are using a rectangular kernel and f'' is uniformly bounded, it is clear that $\hat{f}(z) = f(z) + O(b^2)$. To obtain a bound for the variance we require a bound for the covariances inside the variance expansion $T^2 \cdot \mathrm{var}(\hat{f}(z)) = \sum_{t,\tau} \mathrm{cov}[K_b(Z_t - z), K_b(Z_\tau - z)]$. Since $\{Z_t\}$ is 2-mixing with size $\mathfrak{v}$ by using the covariance inequality in Bradley [6] (see also Rio [36]) we have

$$\begin{aligned} &|\mathrm{cov}[K_b(Z_t - z), K_b(Z_\tau - z)]| \\ &\leq 4 \cdot \int_0^\infty \int_0^\infty \min\left(C|t-\tau|^{-\mathfrak{v}}, P(|K_b(Z_t - z)| > x), P(|K_b(Z_\tau - z)| > y)\right) dxdy. \end{aligned} \tag{2}$$

Studying $P(|K_b(Z_t - z)| > x)$ and recalling that $K(\cdot)$ is a rectangular kernel we can show that

$$P(|K_b(Z_t - z)| > x) = \begin{cases} 0, & \text{if } x > 1/b; \\ P(Z_t \in [z - b/2, z + b/2]), & \text{otherwise.} \end{cases}$$

By using the mean value theorem we have $P(X_t \in [z - b/2, z + b/2]) = bf(\tilde{z})$, for

some $\tilde{z} \in [z - b/2, z + b/2]$. Substituting this into (2) leads to

$$|\mathrm{cov}[K_b(Z_t - z), K_b(Z_\tau - z)]| \leq 4 \cdot \int_0^{1/b} \int_0^{1/b} \min\left(C|t-\tau|^{-\mathfrak{v}}, b \cdot f(\tilde{z})\right) dxdy$$
$$= 4 \cdot b^{-2} \min\left(C|t-\tau|^{-\mathfrak{v}}, b \cdot f(\tilde{z})\right). \quad (3)$$

Altogether this yields the bound

$$T^2 \cdot \mathrm{var}(\hat{f}(z)) \leq 4 \sum_{t,\tau} b^{-2} \min\left(C|t-\tau|^{-\mathfrak{v}}, b \cdot f(\tilde{z})\right).$$

Examining the minimum inside the summand above, we partition the sum into two parts which we bound separately (for the details see the proof of Theorem 3.3, in the Appendix). Finally recalling that $| \quad \hat{f}(z) - f(z)|^2 = O(b^4)$ leads to the desired result. □

We observe, in the proof above, that besides the 2-mixing condition we do not have any assumptions on the joint distribution of (Z_t, Z_τ). The cost of using such weak assumptions is that the usual bound $O(b^4 + (bT)^{-1})$ for the MSE, obtained for independent observations, is not achieved. Even for large $\mathfrak{v}$ the 2-mixing size has an influence on the bound. However, introducing some assumptions on the joint densities of $\{Z_t\}$ allows us to tighten the bound derived in (3) and, hence for a sufficiently large mixing size $\mathfrak{v}$ to recover the usual bound $O(b^4 + (bT)^{-1})$ for the MSE (we note that the rest of the proofs in this section and the subsequent sections require more subtle arguments, and these can be found in the appendix).

Assumption 3.2 Densities and kernels

(i) The marginal density f is uniformly bounded.
(ii) For each $t, \tau \in$ let $f^{(t)}$ denote the joint density of (Z_t, Z_0). Define[1] $F^t := f^t - f \otimes f$. Then $\|F^{(t)}\|_{p_F}$ is uniformly bounded in t for some $p_F > 2$ and we define $q_F = 1 - 2/p_F$.
(iii) The kernel K is uniformly bounded and has a finite first and second moment, i.e., $\|K\|_1 < \infty$ and $\|K\|_2 < \infty$.

We use these assumptions to derive an uniform bound for the MSE of the density estimator.

Theorem 3.3 Let us suppose the stationary time series $\{Z_t\}$ is 2-mixing with size $\mathfrak{v}$ and Assumption 3.2 is fulfilled for some $q_F \in (0, 1)$. In addition assume that $f \in \mathfrak{G}^s_\triangle$ for some $\triangle, s > 0$ (see Definition 2.2). Let $\hat{f}$ be defined as in (1), where K is a kernel of order s. Then we have uniformly for all $z \in$

$$|\hat{f}(z) - f(z)|^2 = O\Big(b^{2s} + b^{-1} \cdot T^{-1} + b^{-2-q_F(1-[\mathfrak{v}\vee 1])} \cdot T^{-[\mathfrak{v}\wedge 1]}\Big), \quad T \to \infty.$$

[1] We use the notation $f \otimes g(x, y) = f(x)g(y)$ and $\|f\|_p = (\int |f(x)|^p dx)^{1/p}$.

For ease of presentation we have only stated the result for univariate $\{Z_t\}$, however it is straightforward to extend this result for multivariate $\{Z_t\}$.

Remark 3.4 We note that in the bound given in Theorem 3.3 the second term dominates the third term when $\mathfrak{v} > 1 + 1/q_F$. Conversely, when $\mathfrak{v} < 1 + 1/q_F$ the third term dominates the second term. Moreover, the third term can be partitioned into two further cases, when $1 < \mathfrak{v} \leq 1 + 1/q_F$ and when $\mathfrak{v} \leq 1$. This means that Theorem 3.3 can be written as

(i) if $\mathfrak{v} > 1 + 1/q_F$, then $|\hat{f}(z) - f(z)|^2 = O\Big(b^{2s} + \frac{1}{bT}\Big)$;

(ii) if $1 < \mathfrak{v} \leq 1 + 1/q_F$, then $|\hat{f}(z) - f(z)|^2 = O\Big(b^{2s} + \frac{1}{b^{(2+q_F(1-\mathfrak{v}))}T}\Big)$;

(iii) if $\mathfrak{v} \leq 1$ then $|\hat{f}(z) - f(z)|^2 = O\Big(b^{2s} + \frac{1}{b^2 T^{\mathfrak{v}}}\Big)$;

as $T \to \infty$.

Studying the three bounds, we see that the bound increases linearly with $\mathfrak{v}$ for $0 \leq \mathfrak{v} \leq 1$, after this point there is a change in behavior and the increase is more gradual. The bound plateaux when $\mathfrak{v} > 1+1/q_F$, after this point we have the usual nonparametric bound obtained for iid observations. There is also a continuity in the three bounds. More precisely, when $\mathfrak{v}$ is at the boundary of 1 and $1+q_F^{-1}$, there is a continuous transition between the bounds. □

We now consider the rate of convergence by using a bandwidth b^* which balances the three terms in the bound of Theorem 3.3.

Corollary 3.5 Under the assumptions of Theorem 3.3 if $b^* \approx T^{-\gamma/(2s+1)}$ with

$$\gamma := \begin{cases} 1, & \mathfrak{v} > 1 + 1/q_F; \\ [\mathfrak{v} \wedge 1] \cdot \frac{2s+1}{2s+(2+q_F(1-[\mathfrak{v}\vee 1]))}, & 1 + 1/q_F \geq \mathfrak{v}. \end{cases} \tag{4}$$

Then uniformely for all $z \in$ we have $|\hat{f}(z) - f(z)|^2 = O\Big(T^{-\frac{2s}{2s+1}\cdot\gamma}\Big)$ as $T \to \infty$.

In other words, if $b^* \approx T^{-\gamma/(2s+1)}$, then we have

$$|\hat{f}(z) - f(z)|^2 := \begin{cases} O\Big(T^{-\frac{2s}{2s+1}}\Big), & \mathfrak{v} > 1 + 1/q_F; \\ O\Big(T^{-\frac{2s}{2s+1}\cdot(\frac{2s+1}{2s+(2+q_F(1-\mathfrak{v}))})}\Big), & 1 + 1/q_F \geq \mathfrak{v} > 1; \\ O\Big(T^{(\mathfrak{v}\cdot\frac{2s+1}{2s+2})}\Big), & 1 \geq \mathfrak{v}. \end{cases} . \tag{5}$$

We note that if $\sup_z |f(z)| < \infty$ and $\sup_t \sup_z |f^{(0,t)}(z)| < \infty$ (both the density and the joint densities are uniformly bounded), then uniformly in all t, $\|F^{(t)}\|_\infty < \infty$. This means $q_F = 1$, and the bound can be divided into the three cases where $\mathfrak{v} \leq 1$, $1 \leq \mathfrak{v} \leq 2$ and $\mathfrak{v} \geq 2$. On the other hand when $\|F^{(t)}\|_{p_F} < \infty$ for only a finite p_F, then $q_F < 1$ and $\mathfrak{v} > 1 + q_F^{-1} > 2$ to be sure of the usual nonparametric bound.

Referring to Corollary 3.5, we observe that when $\mathfrak{v} < 1 + q_F^{-1}$, then the used bandwidth b^* is much larger than usual bandwidths encountered in nonparametric regression ($b \approx T^{\frac{-1}{2s+1}}$). We discuss this further in Section 3.2.

3.1. *A comparison of the MSEs for linear processes*

In this section we compare the MSE in Theorem 3.3 with the results obtained under the stronger condition that the observations $\{Z_t\}$ come from a linear process (which is a much stronger assumption than 2-mixing, see Appendix A.3). We will use the results in Appendix A.3 and show that if the process were linear, and not just mixing, that then the rate of convergence is better than the rate obtained in Corollary 3.5. However, in Section 3.2 we demonstrate that by misspecifying the process to be linear, can lead to several problems with the density estimator, including bounds which do not converge to zero.

Let us suppose $\{Z_t\}$ has a linear process representation and satisfies

$$Z_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}, \tag{6}$$

where the innovations $\{\varepsilon_t\}$ are iid. Under the assumptions in Lemma A.8 (see the Appendix), it can be shown that $\mathrm{cov}(K_b(Z_0), K_b(Z_t)) = O(\mathrm{cov}(Z_0, Z_t))$. Using this as the basis, Hall and Hart [25], Giraitis et al. [22], Mielniczuk [34] and Estevas and Vieu [18]) have shown that the MSE is

$$|\hat{f}(z) - f(z)|^2 = O\left(b^{2s} + \frac{1}{bT} + \frac{1}{T} R_T\right), \tag{7}$$

where $R_T = \sum_{t=1}^{T} |\mathrm{cov}(Z_0, Z_t)|$. It is clear that both $\mathrm{cov}(Z_0, Z_t)$ and R_T depend on the rate of decay of the parameters $\{a_j\}$. We observe if $|a_j| \leq Cj^{-\theta}$, then

$$\begin{aligned} &|\mathrm{cov}(Z_0, Z_t)| = O(T^{-2\theta+1}) \quad \text{and} \quad R_T = O(T^{-(2\theta-1)+1}) \text{ if } 1/2 < \theta \leq 1 \\ &|\mathrm{cov}(Z_0, Z_t)| = O(T^{-\theta}) \qquad \text{and} \quad R_T = O(T^{-\theta+1}) \qquad \text{if } \theta > 1. \end{aligned}$$

Substituting these rates into (7) we see that the bound of the MSE depends on θ. We recall that a process $\{Z_t\}$ is called a short memory process if $\sum_t |\mathrm{cov}(Z_0, Z_t)| < \infty$, otherwise it is called a long memory process. Now studying (7) we see that R_T does not depend on the bandwidth b. In other words long memory has *no influence* on the choice of the optimal bandwidth. To summarize, the rate of convergence for observations coming from a linear process is

$$|\hat{f}(z) - f(z)|^2 \leq \begin{cases} O(T^{-(2\theta-1)}), & \text{if } 2\theta - 1 \leq \frac{2s}{2s+1}; \\ O(T^{\frac{-2s}{2s+1}}), & \text{if } 2\theta - 1 > \frac{2s}{2s+1}. \end{cases} \tag{8}$$

Hall and Hart [25] have shown that the rates above are optimal. On the other hand let us recall Corollary 3.5, above, in particular the case $\mathfrak{v} \leq 1$, where we have shown

$$|\hat{f}(z) - f(z)|^2 = O(T^{-\mathfrak{v} + \frac{2\mathfrak{v}}{2s+2}}). \tag{9}$$

It is difficult to directly compare (8) and (9), since (8) is in terms of its long memory parameter whereas (9) is in terms of its mixing size $\mathfrak{v}$. However in the special case that $\{Z_t\}$ is Gaussian (and thus linear), there is a one-to-one correspondence, for

example, if $2\theta - 1 \leq 1$ then the covariance size and mixing size are the same, and $\mathfrak{v} = (2\theta - 1) = \mathfrak{u}$. We illustrate the case when the mixing and the covariance sizes are the same in Figure 1 (for both large and small s). In the non-Gaussian case,

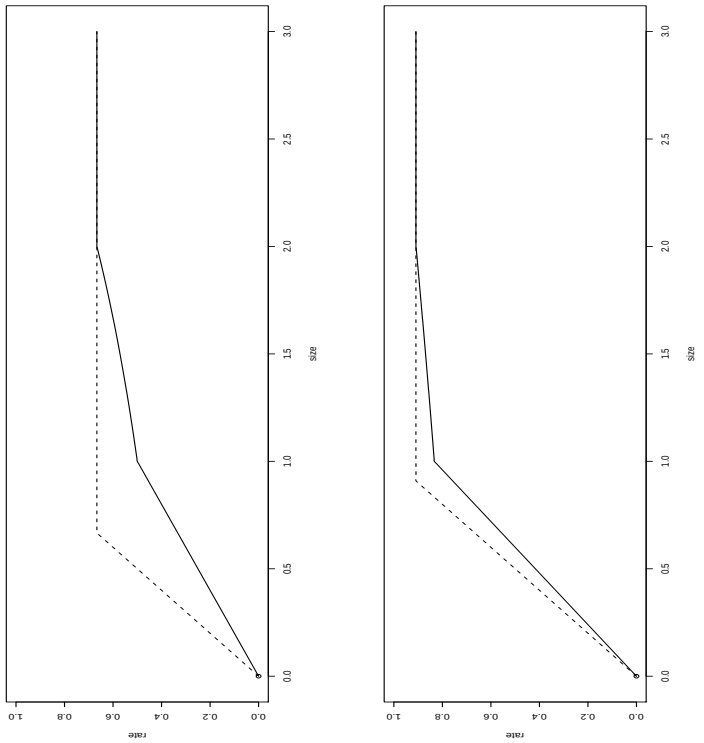

Figure 1. The top and bottom plot corresponds to $s = 1$ and $s = 5$, respectively. The x-axis is the covariance and mixing size (assuming both are the same) and the y-axis is the index δ in the MSE $|\hat{f}(x) - f(x)|^2 = O(T^{-\delta})$. The solid line is the MSE using 2-mixing and dotted line is the MSE when $\{Z_t\}_t$ is a linear process. We have assumed that $q_F = 1$ (in other words $\|F^{(t)}\|_\infty < \infty$).

where the 2-mixing and covariance size do not necessarily coincide ($\mathfrak{v} \neq \mathfrak{u}$), if all the moments of $\{Z_t\}$ exist, then for $1/2 < \theta \leq 1$ we have that $(2\theta-1)/2 \leq \mathfrak{v} \leq (2\theta-1)$ (see (A32) in the appendix). In this case, it is not clear whether the rate (9) is worse than (8). However substituting the lower bound $\mathfrak{v} \geq (2\theta - 1)/2$ into Corollary 3.5 yields a rate which is less than (8). In summary better rates of convergence can often be obtained if the observations come from a linear process. On the other hand, 2-mixing is a *weaker condition, that is satisfied by a far wider class of processes.* We consider below the MSE for processes which are not linear, and show that misspecifying the model, and assuming linearity, when the process is nonlinear could severely affect the MSE.

3.2. *The MSE for nonlinear processes*

As far as we are aware, theory is required to bridge the gap for processes which are nonlinear but have a small mixing size. One of the main aims of Theorem 3.3 is to fill in the gap in the theory, and to derive a bound for the MSE when the observations come from nonlinear processes with small 2-mixing size.

The joint densities of processes which are nonlinear do not necessarily satisfy the density decomposition in Lemma A.8. Without this result it cannot be shown that $\mathrm{cov}(K_b(Z_0), K_b(Z_t)) = O(\mathrm{cov}(Z_0, Z_t))$, and the rates in (8) do not necessarily hold. Instead, to prove the results, under Assumption 3.2, we use classical mixing inequalities to tighten the bound given in (3) (see the proof of Proposition 3.1). More precisely, to prove Theorem 3.3 we show that

$$|\mathrm{cov}(K_b(Z_t - z), K_b(Z_\tau - z))| \leq C \cdot b^{-2} \min\left(|t-\tau|^{-\mathfrak{v}}, b^{(1+q_F)}\right),$$

where C is a finite constant (see Lemma A.1, for the proof).

Looking at some of the implications of Theorem 3.3, we demonstrate below that several problems could arise if one were to falsely suppose that the observations come from a linear process, when they do not.

(i) In the case of linear processes, the optimal bandwidth has the same order as the bandwidth for iid random variables (regardless of long memory). The same is not necessarily true when all that is known is that the process is 2-mixing. Moreover, if the mixing size satisfies $\mathfrak{v} \leq 1$ and the bandwidth is such that $b^2 T^{\mathfrak{v}} < \infty$, then we see from Theorem 3.3 that the bound does not converge to zero. An important example, is when the 'usual' bandwidth for linear or iid data is used (that is $b \approx T^{-\frac{1}{2s+1}}$). In this case, substituting $b \approx T^{-\frac{1}{2s+1}}$ into Theorem 3.3 leads to the result

$$|\hat{f}(z) - f(z)|^2 = \begin{cases} O(T^{\frac{-2s}{2s+1}}) & \mathfrak{v} > 1 + 1/q_F \\ O(T^{\frac{1+q_F(1-\mathfrak{v})-2s}{2s+1}}) & 1 < \mathfrak{v} \leq 1 + 1/q_F \\ O(T^{\frac{2-\mathfrak{v}(2s+1)}{2s+1}}) & 0 \leq \mathfrak{v} \leq 1 \end{cases}$$

Studying the rates above we see when $\mathfrak{v} < 1 + 1/q_F$, the rates are lower than the rates given using the bandwidth which balances bias and variance (compare the above with the rates in Corollary 3.5). Moreover, in the case that $\mathfrak{v} \leq \frac{2}{2s+1}$, the bound cannot be used to show consistency of the estimator - since the bound does not even converge to zero.

In short, to estimate the density at any given point, the number of observations (approximately bT) needs to be much larger than in the iid case.

(ii) Rather surprisingly even when $\sum_{t=1}^{\infty} |\mathrm{cov}(Z_0, Z_t)| < \infty$, the 'usual $O(T^{-\frac{2s}{2s+1}})$' rate, may not hold, unlike for linear processes. However, the usual rate does hold when $\mathfrak{v} \geq 1 + 1/q_F > 2$. Therefore, even when the mixing and covariance size are the same, a far larger mixing size may be require to obtain the 'usual $O(T^{-\frac{2s}{2s+1}})$' rate of convergence.

Our results give a cautionary warning to practitioners who apply the optimal bandwidths for linear processes to nonlinear process. In the subsequent sections, where we consider nonparametric regression problems, the assumptions and proofs will be more involved, however the underlying message is the same. That is, more than just the second order autocovariance function may have influence on the rate of convergence, and the rate of convergence can be severely compromised if the

usual bandwidths were used.

Example 3.6 It is almost impossible to estimate the 2-mixing size from the observations, in contrast to long memory (c.f. Geweke and Porter-Hudak [21], Künsch [31] and Robinson [41]). However to conclude this section we give an example of a nonlinear process whose 2-mixing size is less than $1+\delta$, for some $\delta > 0$. Let us consider the ARCH(∞) process (see Robinson [40]), where $\{Z_t\}$ satisfies

$$Z_t = \sigma_t \varepsilon_t \qquad \sigma_t^2 = a_0 + \sum_{j=1}^{\infty} a_j Z_{t-j}^2,$$

with $\mathbb{E}(\varepsilon_t) = 0$ (estimation of ARCH(∞) parameters is considered in Subba Rao [44]). Giraitis et al. [23] have shown that if for large t, $a_t \approx t^{-(1+\delta)}$ (for some $\delta < 0$) and $[\mathbb{E}(\varepsilon_t^4)]^{1/2} \sum_{j=1}^{\infty} a_j < 1$, then $|\mathrm{cov}(Z_0^2, Z_t^2)| \approx t^{-(1+\delta)}$. That is, the absolute sum of the covariances is finite, but 'only just' if δ is small. Furthermore, if we assume that $|\varepsilon_t| < 1$, then it is straightforward to show that Z_t is a bounded random variable. This means that using the mixing inequality for bounded random variables (see Hall and Heyde [27]) we can show

$$|\mathrm{cov}(Z_0^2, Z_t^2)| \leq C \sup_{A \in \sigma(Z_0), B \in \sigma(Z_t)} |P(A \cap B) - P(A)P(B)|,$$

for some $C < \infty$. Altogether, this gives a lower bound for the 2-mixing size of the ARCH(∞) process with $a_t \approx t^{-(1+\delta)}$, and $\mathfrak{v} \leq (1+\delta)$.
In other words the 2-mixing size for some ARCH(∞) process is small, and far from the geometric rate often assumed in nonparametric estimation. An upper bound for the 2-mixing size can be found in Fryzlewicz and Subba Rao [20]. □

4. Nonparametric regression

In this section we consider nonparametric regression, with random design, where the observations are dependent. It is worth mentioning that there has been extensive research done on nonparametric regression with fixed design and dependent errors (c.f. Hall and Hart [26], Csörgö and Mielniczuk [13], and the references therein). In this case typically, one observes Y_t, where $Y_t = \varphi(\frac{t}{T}) + \varepsilon_t$ and $\{\varepsilon_t\}_t$ are stationary random variables with varying degrees of dependence. It has been shown that the rate of convergence depends on the covariance of $\{\varepsilon_t\}_t$, in particular their absolute sum, $\sum_{t=1}^{\infty} |\mathrm{cov}(\varepsilon_0, \varepsilon_t)|$.

In the random design model, one observes the stationary two-dimensional vector time series $\{(X_t, Z_t)\}_t$, where

$$X_t = \varphi(Z_t) + \varepsilon_t \tag{10}$$

with $\mathbb{E}(X_t | Z_t = z) = \varphi(z)$ and $\varepsilon_t = X_t - \mathbb{E}(X_t|Z_t)$. The randomness in this model is determined by two factors: the design $\{Z_t\}$ and the errors $\{\varepsilon_t = X_t - \mathbb{E}(X_t|Z_t)\}$. Therefore, unlike the fixed design model, the rate of convergence of any estimator

of φ must depend on the sampling properties of the design density estimator. Thus, it is clear that similar results to those in Section 3 should also apply to an estimator of φ.

We now define the classical Nadaraya-Watson estimator of $\varphi(\cdot)$ and study its sampling properties, under various assumptions on $\{(X_t, Z_t)\}$. Let $p(x, z)$ be the joint density of (X_t, Z_t). The estimator is

$$\hat{\varphi}(z) = \frac{\hat{g}(z)}{\hat{f}(z)}, \tag{11}$$

where $\hat{g}(z) := \frac{1}{T}\sum_{t=1}^{T} X_t K_b(Z_t - z)$ and $\hat{f}(z) := \frac{1}{T}\sum_{t=1}^{T} K_b(Z_t - z)$ are estimators of $g(z) = \int xp(x, z)dx$ and $f(z)$, which is the density of Z_t.

We first consider the sampling properties for a particular class of models which satisfy (10). Suppose the vector time series $\{(X_t, Z_t)\}$ satisfies the representation

$$X_t = \varphi(Z_t) + h(Z_t)\eta_t \tag{12}$$

for some $h : \quad \to \quad ^{+}$, where the time series $\{Z_t\}$ and $\{\eta_t\}$ are independent of each other. This class of models is similar to the fixed design model $X_t = \varphi(\frac{t}{T}) + \eta_t$, but in (12) the design is random and the conditional variance $\mathrm{var}(X_t|Z_t) = h(Z_t)^2\mathrm{var}(\eta_t)$, depends on the design. This model arises in various applications and we consider one application in Remark 4.5. We will show in the theorem below that the rate of convergence depends both on the mixing size of the design $\{Z_t\}$, but also on the size of the covariances of the process $\{\eta_t\}$ (which we denote by $\mathfrak{u}$, see Definition 2.3).

We require the following assumptions.

Assumption 4.1 Densities, moments and kernels

(i) For some $p > 2$ the functions $h^2 \cdot f$ and $|\varphi|^p \cdot f$ are uniformly bounded and we define $q := 1 - 2/p$.

(ii) Let $f^{(t)}$ and $F^{(t)}$ be defined as in Assumption 3.2 (ii),

$$g^{(t)}(z_1, z_2) := \quad [X_t X_0 | Z_t = z_1, Z_0 = z_2] \cdot f^{(t)}(z_1, z_2).$$

and $G^{(t)} := g^{(t)} - g \otimes g$. Then $\|F^{(t)}\|_{p_F}$ and $\|G^{(t)}\|_{p_G}$ are uniformly bounded in t for some $p_F, p_G > 2$. We define $q_F := 1 - 2/p_F$, $q_G := 1 - 2/p_G$ and $q_{FG} := q_F \wedge q_G$.

(iii) The kernel K has finite first and p-th moment.

Studying Assumption 4.1(i), we see that it allows for various types of growth of the regression function φ and the conditional variance h. The type of growth depends on the rate the density f decays to zero. For example, if f were the Gaussian density, then exponential growth of φ and h is possible. However, as we shall demonstrate in the theorem below, the larger the p, such that $\sup_x h(x)^p \cdot f(x) < \infty$ and $\sup_x |\varphi(x)|^p \cdot f(x) < \infty$, then the faster the rate of convergence of $|\hat{\varphi}(z) - \varphi(z)|^2$.

Theorem 4.2 Suppose the stationary time series $\{(X_t, Z_t)\}$ satisfies (12), $\{Z_t\}$

is 2-mixing with size $\mathfrak{v}$ and the autocovariance of the time series $\{\eta_t\}$ has size $\mathfrak{u}$. Let Assumption 4.1 be fulfilled for some $q, q_{FG} \in (0,1)$. In addition assume that $\varphi \cdot f$, $f \in \mathfrak{G}^s_{\triangle}$ for some $\triangle, s > 0$ and that f is bounded away from zero. Let the estimator $\hat{\varphi}(z)$ be defined as in (11), where K is a kernel of order s. Then we have for all $z \in$

$$|\hat{\varphi}(z)-\varphi(z)|^2 = O_P\Big(b^{2s}+b^{-1}.T^{-(\mathfrak{u}\wedge 1)}+b^{-1-q-q_{FG}(1-[(q\mathfrak{v})\vee 1])}.T^{-[(q\mathfrak{v})\wedge 1]}\Big), \quad T\to\infty.$$

Remark 4.3 We observe that the bound obtained in Theorem 4.2 are similar to the bound derived for the density estimator in Theorem 3.3, where

$$|\hat{f}(z) - f(z)|^2 = O\Big(b^{2s} + b^{-1}\cdot T^{-1} + b^{-2+q_F(1-[\mathfrak{v}\vee 1])}\cdot T^{-[\mathfrak{v}\wedge 1]}\Big). \tag{13}$$

The difference is the inclusion of the covariance size $\mathfrak{u}$ of the errors and the q which 'balances' the tails of $1/\varphi$ and f (see Assumption 4.1(i)). However, we observe that we can partition the bound in Theorem 4.2 into three cases, which are similar to the three cases considered in Remark 3.4. Most notably, we observe if $\mathfrak{u} > 1$ and $\mathfrak{v} > 1/q_{FG} + 1/q$ then we obtain the usual bound $O_P(b^{2s} + b^{-1}\cdot T^{-1})$ for the squared error. It is interesting to note that in the case $h^p \cdot f$ and $|\varphi|^p \cdot f$ are uniformly bounded for all p, then $q = 1$ (eg. h and φ are bounded functions and f is exponential density). In this case the bounds given in (13) and Theorem 4.2 are quite similar. The main difference is the appearance of q_{FG} rather than q_F and, the term $b^{-1}T^{-(\mathfrak{u}\wedge 1)}$ which replaces $b^{-1}T^{-1}$. □

Corollary 4.4 Under the assumptions of Theorem 4.2 if $b^* \approx T^{-\gamma/(2s+1)}$ with

$$\gamma := \begin{cases} \min(\mathfrak{u}, 1), & q\mathfrak{v} > 1 + 1/q_{FG}; \\ \min\Big(\mathfrak{u}, [(q\mathfrak{v})\wedge 1]\cdot \frac{2s+1}{2s+1+q+q_{FG}(1-[(q\mathfrak{v})\vee 1])}\Big), & 1+1/q_{FG} \geq q\mathfrak{v}. \end{cases}, \tag{14}$$

then we have $|\hat{\varphi}(z) - \varphi(z)|^2 = O_P\Big(T^{-\frac{2s}{2s+1}\cdot\gamma}\Big)$ for all $z \in$.

Let us now compare Theorem 4.2 with the bound obtained for the deterministic design $X_t = \varphi(\frac{t}{T}) + \varepsilon_t$, where $\mathfrak{u}$ is the covariance size of the errors. In the case of the fixed design, the bound for the deviation of the kernel estimator is $O(b^{2s} + T^{-(\mathfrak{u}\wedge 1)}b^{-1})$ (c.f. Hall and Hart [26]). We see that the bound in Theorem 4.2 include this term, but also the additional term $O(b^{-1-q-q_{FG}(1-[(q\mathfrak{v})\vee 1])}\cdot T^{-[(q\mathfrak{v})\wedge 1]})$, which is the influence of the design, in particular, $\mathfrak{v}$. If the mixing size of the design were sufficiently large, then the fixed design and random design estimators have the same rate of convergence, $O(T^{-\frac{2s}{2s+1}})$.

Example 4.5 Examples of processes which satisfy (12) are stochastic volatility models (c.f Linton and Mammen [32]), where one observes $\{Y_t\}$, which satisfies the representation

$$Y_t = \sigma(Z_t)\eta_t.$$

Here $\{\eta_t\}$ are iid random variables, $(\eta_t^2) = 1$ and $\{Z_t\}$ are explanatory variables

which can include past values of Y_t. Usually in finance the object is to estimate the conditional volatility σ^2. By noting that Y_t^2 can be written as

$$Y_t^2 = \sigma(Z_t)^2 + (\eta_t^2 - 1)\sigma(Z_t)^2,$$

we see that Y_t^2 satisfies (12) with $X_t = Y_t^2$, $\varepsilon_t = (\eta_t^2 - 1)$ and $h(\cdot) = \sigma(\cdot)^2$. Therefore we can estimate the volatility $\sigma(\cdot)^2$ using (11), where $\hat{\sigma}(\cdot)^2$, is the kernel estimator of $\sigma(\cdot)^2$. Furthermore, Theorem 4.2 can be applied to obtain the rate of convergence. More precisely, let $\mathfrak{v}$ be the mixing size of $\{Z_t\}$, and noting that $\mathrm{cov}\{(\eta_t^2 - 1), (\eta_s^2 - 1)\} = 0$, when $t \neq s$, which implies $\mathfrak{u} = \infty$, we obtain

$$|\hat{\sigma}(z)^2 - \sigma(z)^2|^2 = O_P\Big(b^{2s} + b^{-1-q-q_{FG}(1-[(q\mathfrak{v})\vee 1])} \cdot T^{-[(q\mathfrak{v})\wedge 1]}\Big). \qquad \square$$

From Corollary 4.4 we see that there are two factors which affect the rate of convergence: the mixing size $\mathfrak{v}$ of the random design $\{Z_t\}$ and the size $\mathfrak{u}$ of the covariance function of $\{\eta_t\}$. There are however several models of interest, which do not satisfy condition (12). In this case Theorem 4.2 cannot be applied and it is of interest to investigate what happens in the general case.

Examples of models which do not necessarily satisfy (12) include the Cheng-Robinson model, where $\{X_t\}$ satisfies the representation $X_t = F(U_t) + G(U_t, Y_t)$ with $(G(U_t, Y_t)|U_t) = 0$ and $\{Y_t\}$ is a long memory process, which is independent of the weakly dependent design random variables $\{U_t\}$ (c.f. Cheng and Robinson [10], Csörgö and Mielniczuk (1999, 2001)). However, the results are derived under the assumption that $\{Y_t\}$ comes from a linear process and $G(\cdot)$ has a particular form.

An alternative approach is developed in Bosq [3], who considers nonparametric prediction for time series, where one observes the stationary time series $\{(X_t, Z_t)\}$ and the parameter of interest is $\varphi(z) = (X_t|Z_t = z)$. The sampling results in Bosq [3] are based on the assumption that the mixing size of $\{(X_t, Z_t)\}$ is sufficiently large, (thus excluding Cheng-Robinson type models) yielding an estimate which has the same rate as the kernel estimator for iid random variables.

We now consider the sampling properties of $\hat{\varphi}$, when the observations $\{(X_t, Z_t)\}$ satisfy the general model defined in (10), and dependence is quantified through its 2-mixing size, which can be arbitrary.

We will use the following assumptions.

Assumption 4.6 Densities, moments and kernels

(i) Let $|X_t|^p < \infty$ for some $p > 2$ and define $g^{(p)}(z) := [|X_t|^p|Z_t = z] \cdot f(z)$. Then the functions $g^{(p)}$ and f are uniformly bounded and we define $q := 1 - 2/p$.

(ii) Let $f^{(t)}$ and $F^{(t)}$ be defined as in Assumption 3.2 (ii) and let $g^{(t)}$ and $G^{(t)}$ be defined as in Assumption 3.2 (ii). Then $\|F^{(t)}\|_{p_F}$ and $\|G^{(t)}\|_{p_G}$ are uniformly bounded in t for some $p_F, p_G > 2$, where we define $q_F := 1 - 2/p_F$, $q_G := 1 - 2/p_G$ and $q_{FG} := q_F \wedge q_G$.

(iii) The kernel K has finite first and p-th moment.

We note that assumptions above are similar to Assumption 4.1. The difference lies in Assumption 4.1(i) and Assumption 4.6(i). Assumption 4.6(i) is in terms of moments whereas Assumption 4.1(i) is in terms of functions.

In the following theorem we derive an error bound for the estimator $\hat{\varphi}$.

Theorem 4.7 Suppose the stationary time series $\{(X_t, Z_t)\}$ satisfies (10), and is 2-mixing of size $\mathfrak{v}$. Furthermore, Assumption 4.6 is fulfilled for some $q_{FG}, q \in (0,1)$. In addition assume that $\varphi \cdot f$, $f \in \mathfrak{G}^s_{\triangle}$ for some $\triangle, s > 0$ and that f is bounded away from zero. Let the estimator $\hat{\varphi}(z)$ be defined as in (11), where K is a kernel of order s. Then we have for all $z \in$

$$|\hat{\varphi}(z)-\varphi(z)|^2 = O_P\Big(b^{2s}+b^{-1}\cdot T^{-1}+b^{-1-q-q_{FG}(1-[(q\mathfrak{v})\vee 1])}\cdot T^{-[(q\mathfrak{v})\wedge 1]}\Big), \quad T\to\infty.$$

We now obtain the rates of convergence by balancing the three terms in the bound of the last assertion.

Corollary 4.8 Under the assumptions of Theorem 4.7 if $b^* \approx T^{-\gamma/(2s+1)}$ with

$$\gamma := \begin{cases} 1, & q\mathfrak{v} > 1 + q/q_{FG}; \\ [(q\mathfrak{v})\wedge 1]\cdot \frac{2s+1}{2s+1+q+q_{FG}(1-[(q\mathfrak{v})\vee 1])}, & 1+q/q_{FG} \geq q\mathfrak{v}., \end{cases} \tag{15}$$

then we have $|\hat{\varphi}(z) - \varphi(z)|^2 = O_P\Big(T^{-\frac{2s}{2s+1}\cdot\gamma}\Big)$ for all $z \in$.

5. Discussion

In this paper we have considered nonparametric estimation for dependent data. Focusing on the case that the observations are nonlinear and highly dependent. We have obtained bounds for the kernel density estimator and also rates of convergence of two types of nonparametric regression models, both using the 2-mixing dependence measure. We show that when the assumption of linearity is relaxed, the rate of convergence does not necessarily depend on the autocovariance function of the observations.

As we are working under relatively weak conditions, we do not claim that the bounds obtained are minimax. However, the bounds can be considered as the worst case scenario for the nonparametric estimator. In future work, it would be of interest to investigate if the bounds in the paper are indeed close to minimax for certain nonlinear time series. It would also be of interest to develop bandwidth selection methods when the 2-mixing size of the observations is unknown.

Acknowledgments

The authors are grateful to Rainer Dahlhaus, Rafal Kulik and two anonymous referees for making several useful suggestions. This work has been supported by the Deutsche Forschungsgemeinschaft under DA 187/15-1.

Appendix A.

A.1. *Proofs: Nonparametric density estimation*

We now prove the results in Section 3.

Lemma A.1 Suppose the time series $\{Z_t\}$ is 2-mixing with size $\mathfrak{v}$ and Assumption 3.2 is fulfilled for some $q_F \in (0,1)$. If $1 \leq t, \tau \leq T$, then [1]

$$|\mathrm{cov}\{K_b(Z_t - z), K_b(Z_\tau - z)\}| \lesssim \min\Big(b^{-(1-q_F)}; b^{-2}|t-\tau|^{-\mathfrak{v}}\Big). \quad \text{(A1)}$$

Proof. Writing the covariance as an integral, and using the notation in Assumption 3.2 (ii) we have

$$\mathrm{cov}\{K_b(Z_t - z), K_b(Z_\tau - z)\} = \int K_b(u-z)K_b(v-z)F^{(t-\tau)}(u,v)dudv.$$

Now by using Hölder's inequality with $p_F^{-1} + \bar{p}_F^{-1} = 1$, and recalling that $q_F = 1 - 2/p_F$, it is clear that

$$|\mathrm{cov}\{K_b(Z_t - z), K_b(Z_\tau - z)\}| \leq \frac{1}{b^2} \cdot b^{2/\bar{p}_F}\|K\|_{\bar{p}_F}^2 \cdot \|F^{(t-\tau)}\|_{p_F} \lesssim b^{-(1-q_F)}.$$

Using Assumption 3.2 we have that $\|F^{(t-\tau)}\|_{p_F}$ is uniformly bounded and by using Lyaponov's inequality $\|K\|_{\bar{p}_F} < \infty$ for all $1 < \bar{p}_F < 2$. This gives us the first bound in (A1). On the other hand, under Assumption 3.2 (i) the kernel K is uniformly bounded and therefore, using the 2-mixing property of $\{Z_t\}$ together with Hall and Heyde [27], Theorem A.5, we obtain

$$|\mathrm{cov}\{K_b(Z_t - z), K_b(Z_\tau - z)\}| \lesssim b^{-2} \cdot |t-\tau|^{-\mathfrak{v}},$$

which gives the second bound in (A1). □

Proof of Theorem 3.3. We mention that parts of the following proof are motivated by techniques used in Bosq [3]. Consider the standard decomposition

$$|\hat{f}(z) - f(z)|^2 = \mathrm{var}(\hat{f}(z)) + |\quad \hat{f}(z) - f(z)|^2. \quad \text{(A2)}$$

Under the stated assumptions we will derive the following two bounds, which give together the result of the theorem. The bias is bounded by

$$|\quad \hat{f}(z) - f(z)|^2 \lesssim b^{2s}, \quad \text{(A3)}$$

while for the variance we have

$$\mathrm{var}(\hat{f}(z)) \lesssim T^{-1} \cdot b^{-1} + T^{-[\mathfrak{v}\wedge 1]} \cdot b^{-(2+q_F - q_F[\mathfrak{v}\vee 1])}. \quad \text{(A4)}$$

[1] We write $A \lesssim B$ when there exists a positive constant c independent of A and B such that $A \leqslant cB$.

Proof of (A3). We can write

$$\hat{f}(z) = \frac{1}{T}\sum_{t=1}^{T} \Big(K_b(Z_t - z)\Big) = \int du\, f(u)K_b(u-z).$$

Since $f \in \mathfrak{G}^s_{\triangle}$ and K is a kernel of order s with $\int du|u|^s K(u) \leq S_K$, using a Taylor expansion up to the order s leads to $\hat{f}(z) = f(z) + b^s R$ with reminder $|R| \leq \triangle S_K < \infty$, which proves (A3).

In order to proof (A4), we consider the expansion

$$\begin{aligned}\mathrm{var}(\hat{f}(z)) &= \frac{1}{T^2}\sum_{t=1}^{T} \mathrm{var}\left\{K_b(Z_t - z)\right\} + \frac{2}{T^2}\sum_{t>\tau} \mathrm{cov}\left\{K_b(Z_t - z), K_b(Z_\tau - z)\right\} \\ &=: A_1 + A_2. \end{aligned} \tag{A5}$$

We will show that $|A_1| \lesssim T^{-1}\cdot b^{-1}$ and

$$|A_2| \lesssim \begin{cases} T^{-\mathfrak{v}}\cdot b^{-2}, & \mathfrak{v} \leq 1; \\ T^{-1}\cdot\{b^{-1} + b^{-(2+q_F - q_F\mathfrak{v})}\}, & 1 < \mathfrak{v}. \end{cases} \tag{A6}$$

Furthermore, if $0 \leq \mathfrak{v} \leq 1/q_F + 1$ then $|A_1|$ is dominated by $|A_2|$. Whereas for $\mathfrak{v} > 1/q_F + 1$ the terms $|A_1|$ and $|A_2|$ are of the same order $O(T^{-1}b^{-1})$. Therefore, the bounds derived for $|A_2|$ will lead to (A4).

First let us consider A_1. Due to stationarity, we have the bound

$$T\cdot A_1 \leq \quad [K_b^2(Z_1 - z)] = \int du\, f(u)K_b^2(u-z).$$

Since under the stated assumptions $\|K\|_2 < \infty$ and the density f is uniformly bounded this leads to $A_1 \lesssim T^{-1}\cdot b^{-1}$.

The term $T\cdot|A_2|$ is bounded by the sum $4\sum_{t=2}^{T}|\mathrm{cov}\left\{K_b(Z_t - z), K_b(Z_1 - z)\right\}|$. If $\mathfrak{v} \leq 1$ then we estimate the sum using the second bound in Lemma A.1, i.e., $T\cdot|A_2| \lesssim T^{-\mathfrak{v}+1}b^{-2}$, which is the first bound in (A6). On the other hand if $\mathfrak{v} > 1$ we partition the sum into two parts which we estimate separately using the bounds in Lemma A.1, thus giving us

$$T\cdot|A_2| \lesssim \Big\{\sum_{t=2}^{h} b^{-(1-q_F)} + \sum_{t=h+1}^{T} b^{-2}t^{-\mathfrak{v}}\Big\} \lesssim \Big\{h\cdot b^{-(1-q_F)} + h^{-\mathfrak{v}+1}\cdot b^{-2}\Big\}.$$

Thereby using $h \approx b^{-q_F}$ we obtain $T\cdot|A_2| \lesssim b^{-1} + b^{-1(2+q_F - q_F\mathfrak{v})}$, i.e., the second bound in (A6). Thus we have proved (A4). □

Proof of Corollary 3.5 Under the assumption on the bandwidth the result is obtained by balancing the terms in the bound given in Theorem 4.2. □

A.2. *Proofs: Nonparametric regression*

We now prove the results in Section 4.

Lemma A.2 Suppose the stationary time series $\{X_t, Z_t\}$ satisfies (12), and $\{Z_t\}$ is 2-mixing with size $\mathfrak{v}$ and the autocovariances of the time series $\{\eta_t\}$ have size $\mathfrak{u}$ (see Definition 2.3). Suppose Assumption 4.1 is fulfilled for some $q, q_G \in (0,1)$. If $1 \leq t, \tau \leq T$, then

$$|\mathrm{cov}\{X_t K_b(Z_t - z), X_\tau K_b(Z_\tau - z)\}| \lesssim b^{-(1-q_G)}, \tag{A7}$$

$$|\mathrm{cov}\{\varphi(Z_t) K_b(Z_t - z), \varphi(Z_\tau) K_b(Z_\tau - z)\}| \lesssim b^{-(1+q)}|t-\tau|^{-q\mathfrak{v}}, \tag{A8}$$

$$|\mathrm{cov}\{h(Z_t) K_b(Z_t - z)\eta_t, h(Z_\tau) K_b(Z_\tau - z)\eta_\tau\}| \lesssim b^{-1}|t-\tau|^{-\mathfrak{u}}. \tag{A9}$$

Proof. Using the notation in Assumption 4.1 together with Hölder's inequality, and recalling that $q_G = 1 - 2/p_G$ with $p_G^{-1} + \bar{p}_G^{-1} = 1$, we have

$$|\mathrm{cov}\{X_t K_b(Z_t - z), X_\tau K_b(Z_\tau - z)\}| \lesssim b^{-(1-q_G)},$$

where we use that under Assumption 4.1, K has finite $1 < \bar{p}_G < p$ moment (by Lyaponovs inequality) and $\|G_{t,\tau}\|_{p_G}$ is uniformly bounded. This gives us (A7).

We now prove (A8). Under Assumption 4.1 the function $|\varphi|^p \cdot f$ is uniformly bounded and $\|K\|_p$ is finite for some $p = 2/(1-q) > 2$, therefore we have $[\ |\varphi(Z_1)K_b(Z_1 - z)|^p]^{2/p} \lesssim b^{-(q+1)}$. Using the 2-mixing property of $\{Z_t\}$ together with Hall and Heyde [27], Theorem A.6, we obtain (A8).

We now prove (A9). The series $\{Z_t\}$ and $\{\eta_t\}$ are independent, therefore expanding the term $A := \mathrm{cov}\{h(Z_t)K_b(Z_t - z)\eta_t, h(Z_\tau)K_b(Z_\tau - z)\eta_\tau\}$ gives

$$A = \mathrm{cov}(\eta_t, \eta_\tau) \cdot \quad [h(Z_t)K_b(Z_t - z)h(Z_\tau)K_b(Z_\tau - z)].$$

Since the covariance of the time series $\{\eta_t\}$ has size $\mathfrak{u}$, applying the Cauchy-Schwarz inequality gives

$$|A| \lesssim |t-\tau|^{-\mathfrak{u}} \cdot \quad |h(Z_1)K_b(Z_1 - z)|^2.$$

Under Assumption 4.1 the function $|h|^2 \cdot f$ is uniformly bounded and $\|K\|_2 < \infty$, therefore $|h(Z_1)K_b(Z_1 - z)|^2 \lesssim b^{-1}$, and hence we obtain (A9). □

Lemma A.3 Suppose the stationary time series $\{Z_t\}$ is 2-mixing with size $\mathfrak{v}$ and Assumption 4.1 is fulfilled for some $q, q_F \in (0,1)$. If $1 \leq t, \tau \leq T$, then

$$|\mathrm{cov}\{K_b(Z_t - z), K_b(Z_\tau - z)\}| \lesssim \min\Big(b^{-(1-q_F)}; b^{-(1+q)}|t-\tau|^{-q\mathfrak{v}}\Big). \tag{A10}$$

Proof. The proof is very similar to the proof of Lemma A.2 and we omit the details. □

Lemma A.4 Suppose the assumptions in Theorem 4.2 are satisfied. Let $\widehat{g}$ be defined as in (11). Then we have

$$|\widehat{g}(z)-g(z)| \lesssim b^{2s}+b^{-1}T^{-1}+b^{-(1+q+q_g(1-[(q\mathfrak{v})\vee 1]))}T^{-[(q\mathfrak{v})\wedge 1]}+b^{-1}T^{-(\mathfrak{u}\wedge 1)}. \quad \text{(A11)}$$

PROOF. Consider the standard variance bias decomposition

$$|\hat{g}(z)-g(z)|^2 = \mathrm{var}(\hat{g}(z)) + |\ \ \hat{g}(z)-g(z)|^2. \quad \text{(A12)}$$

Under the stated assumptions we will derive the following two bounds, which altogether give the estimate in (A11). The bias is bounded by

$$|\ \ \hat{g}(z)-g(z)|^2 \lesssim b^{2s}, \quad \text{(A13)}$$

while for the variance we have

$$\mathrm{var}(\hat{g}(z)) \lesssim b^{-1}T^{-1} + b^{-(1+q+q_G(1-[(q\mathfrak{v})\vee 1]))}T^{-[(q\mathfrak{v})\wedge 1]} + b^{-1}T^{-(\mathfrak{u}\wedge 1)}. \quad \text{(A14)}$$

We first prove (A13). We can write

$$\hat{g}(z) = \frac{1}{T}\sum_{t=1}^{T} \ \Big(\ \ [X_t|Z_t]K_b(Z_t-z)\Big) = \int du\ g(u)K_b(u-z).$$

Since $g \in \mathfrak{G}^s_{\triangle}$ and K is a kernel of order s with $\int du|u|^s K(u) \le S_K$, using a Taylor expansion up to the order s leads to $\hat{g}(z) = g(z) + b^s R$ with reminder $|R| \le \triangle S_K < \infty$, which proves (A13).

In order to proof (A14), we consider the expansion

$$\begin{aligned}\mathrm{var}(\hat{g}(z)) &= \frac{1}{T^2}\sum_{t=1}^{T}\mathrm{var}\left\{X_tK_b(Z_t-z)\right\} + \frac{2}{T^2}\sum_{t>\tau}\mathrm{cov}\left\{X_tK_b(Z_t-z), X_\tau K_b(Z_\tau-z)\right\}\\ &=: A_1 + A_2. \end{aligned} \quad \text{(A15)}$$

We will show that $|A_1| \lesssim T^{-1}\cdot b^{-1} + T^{-(\mathfrak{u}\wedge 1)}\cdot b^{-1}$ and

$$|A_2| \lesssim \begin{cases} T^{-q\mathfrak{v}}\cdot b^{-(1+q)} + T^{-(\mathfrak{u}\wedge 1)}\cdot b^{-1}, & q\mathfrak{v} \le 1;\\ T^{-1}\cdot b^{-1} + T^{-1}\cdot b^{-(1+q+q_G(1-q\mathfrak{v}))} + T^{-(\mathfrak{u}\wedge 1)}\cdot b^{-1}, & 1 < q\mathfrak{v}. \end{cases} \quad \text{(A16)}$$

Furthermore, if $0 \le q\mathfrak{v} \le q/q_G + 1$ then we show that $|A_1|$ is dominated by $|A_2|$. Whereas for $q\mathfrak{v} > q/q_G + 1$ the terms $|A_1|$ and $|A_2|$ are of the same order $O(T^{-1}\cdot b^{-1} + T^{-(\mathfrak{u}\wedge 1)}\cdot b^{-1})$. Therefore, the bounds derived for $|A_2|$ will lead to the estimates in (A14).

First let us consider A_1. Due to stationarity of the process, we have the bound

$$T\cdot A_1 \le \ \ |X_1K_b(Z_1-z)|^2 \lesssim \ \ |\varphi(Z_1)K_b(Z_1-z)|^2 + \ \ |h(Z_1)K_b(Z_1-z)|^2.$$

Under the stated assumptions the functions $|\varphi|^p \cdot f$ with $p > 2$ and $|h|^2 \cdot f$ are uniformly bounded and the kernel $\|K\|_2 < \infty$, therefore $A_1 \lesssim T^{-1} \cdot b^{-1}$.

Let us now consider the term A_2, which is bound by

$$T \cdot |A_2| \leq 4 \sum_{t=2}^{T} |\mathrm{cov}\left\{X_t K_b(Z_t - z), X_1 K_b(Z_1 - z)\right\}|, \tag{A17}$$

where using representation (12) the t-th summand in (A17) can be estimated by

$$\begin{aligned} &|\mathrm{cov}\left\{\varphi(Z_t)K_b(Z_t - z), \varphi(Z_1)K_b(Z_1 - z)\right\}| \\ &\qquad + |\mathrm{cov}\left\{h(Z_t)K_b(Z_t - z)\eta_t, h(Z_1)K_b(Z_1 - z)\eta_1\right\}|. \end{aligned} \tag{A18}$$

If $q\mathfrak{v} \leq 1$ and $\mathfrak{u} \leq 1$ then we bound the sum (A17) using (A18), i.e.,

$$\begin{aligned} T \cdot |A_2| \lesssim &\sum_{t=2}^{T} |\mathrm{cov}\left\{\varphi(Z_t)K_b(Z_t - z), \varphi(Z_1)K_b(Z_1 - z)\right\}| \\ &+ \sum_{t=2}^{T} |\mathrm{cov}\left\{h(Z_t)K_b(Z_t - z)\eta_t, h(Z_1)K_b(Z_1 - z)\eta_1\right\}|. \end{aligned} \tag{A19}$$

We use the bounds (A8) and (A9) in Lemma A.2 to estimate each of the sums in (A19) separately, which gives

$$T \cdot |A_2| \lesssim T^{-q\mathfrak{v}+1} \cdot b^{-(1+q)} + T^{-\mathfrak{u}+1} \cdot b^{-1}. \tag{A20}$$

On the other hand if $q\mathfrak{v} > 1$ or if $\mathfrak{u} > 1$ we partition the sum (A17) into two parts, where we estimate the first part using the bound (A7) in Lemma A.2 and the second using (A18), thus giving us

$$\begin{aligned} T \cdot |A_2| \lesssim &\, h \cdot b^{-(1-q_G)} + \sum_{t=h+1}^{T} |\mathrm{cov}\left\{\varphi(Z_t)K_b(Z_t - z), \varphi(Z_1)K_b(Z_1 - z)\right\}| \\ &+ \sum_{t=h+1}^{T} |\mathrm{cov}\left\{h(Z_t)K_b(Z_t - z)\eta_t, h(Z_1)K_b(Z_1 - z)\eta_1\right\}|. \end{aligned} \tag{A21}$$

We use the bounds (A8) and (A9) in Lemma A.2 to estimate each of the sums in (A21) separately, which gives

$$T \cdot |A_2| \lesssim h \cdot b^{-(1-q_G)} + \begin{cases} T^{-q\mathfrak{v}+1} \cdot b^{-(1+q)} + h^{-\mathfrak{u}+1} \cdot b^{-1}, & q\mathfrak{v} \leq 1 \text{ and } \mathfrak{u} > 1; \\ h^{-q\mathfrak{v}+1} \cdot b^{-(1+q)} + T^{-\mathfrak{u}+1} \cdot b^{-1}, & q\mathfrak{v} > 1 \text{ and } \mathfrak{u} \leq 1; \\ h^{-q\mathfrak{v}+1} \cdot b^{-(1+q)} + h^{-\mathfrak{u}+1} \cdot b^{-1}, & q\mathfrak{v} > 1 \text{ and } \mathfrak{u} > 1. \end{cases} \tag{A22}$$

Thereby using $h \approx b^{-q_G}$ we obtain

$$T \cdot |A_2| \lesssim b^{-d} + \begin{cases} T^{-q\mathfrak{v}+1} \cdot b^{-d(1+q)} + b^{-(1+q_G(1-u))}, & q\mathfrak{v} \leq 1 \text{ and } \mathfrak{u} > 1; \\ b^{-(1+q+q_G(1-q\mathfrak{v}))} + T^{-\mathfrak{u}+1} \cdot b^{-1}, & q\mathfrak{v} > 1 \text{ and } \mathfrak{u} \leq 1; \\ b^{-(1+q+q_G(1-q\mathfrak{v}))} + b^{-(1+q_G(1-u))}, & q\mathfrak{v} > 1 \text{ and } \mathfrak{u} > 1. \end{cases} \tag{A23}$$

and hence, combining (A20) and (A23) gives the bound (A16) for the term A_2. □

We now state a slight variation of Theorem 3.3, where f can satisfy slightly weaker conditions. We use this result to prove Theorem 4.2.

Lemma A.5 Suppose the stationary time series $\{Z_t\}$ is 2-mixing with size $\mathfrak{v}$ and Assumption 4.1 is fulfilled for some $q, q_F \in (0,1)$. In addition assume, that the function f belongs to $\mathfrak{G}^s_{\triangle}$ for $s, \triangle > 0$. Let $\widehat{f}$ be defined as in (1), where the kernel is of order s. Then we have

$$|\widehat{f}(z) - f(z)| \lesssim b^{2s} + b^{-1}T^{-1} + b^{-(1+q+q_F(1-[(q\mathfrak{v})\vee 1]))}T^{-[(q\mathfrak{v})\wedge 1]}. \tag{A24}$$

Proof. Under the stated assumptions using Lemma A.3 the proof is very similar to the proof of Lemma A.4 and we omit the details. □

Proof of Theorem 4.2. Consider the decomposition

$$\begin{aligned} \hat{\varphi}(z) - \varphi(z) &= \frac{\hat{g}(z)}{\hat{f}(z)} - \frac{\hat{f}(z)}{\hat{f}(z)}\varphi(z) \\ &= \frac{\hat{g}(z) - \hat{f}(z)\varphi(z)}{f(z)} + \frac{f(z) - \hat{f}(z)}{\hat{f}(z)} \cdot \frac{\hat{g}(z) - \hat{f}(z)\varphi(z)}{f(z)}. \end{aligned}$$

We first note that Lemma A.5 gives $|f(z) - \hat{f}(z)|^2 = o(1)$, which implies that $|\hat{f}(z)^{-1}|$ is bounded in probability. Therefore the second term in the above expansion is of order $o_P(\{\hat{g}(z) - \hat{f}(z)\varphi(z)\}/f(z))$, hence in the decomposition above the second term is negligible in comparison to the first term. Thereby bounding the first term of the decomposition we obtain the result. By using Lemma A.4 and A.5 and noting that $q_{FG} = q_F \wedge q_G$, we obtain Theorem 4.2. □

Proof of Corollary 4.4 Under the assumption on the bandwidth the result is obtained by balancing the terms in the bound given in Theorem 4.2. □

Lemma A.6 Suppose the stationary vector time series $\{(X_t, Z_t)\}$ is 2-mixing with size $\mathfrak{v}$ and Assumption 4.6 is fulfilled for some $q, q_G \in (0,1)$. If $1 \leq t, \tau \leq T$, then

$$|\operatorname{cov}\{X_t K_b(Z_t - z), X_\tau K_b(Z_\tau - z)\}| \lesssim \min\Big(b^{-(1-q_G)}, b^{-(1+q)}|t-\tau|^{-q\mathfrak{v}}\Big). \tag{A25}$$

Proof. Under Assumption 4.6 (ii) the first bound in (A25) follows from (A7) in Lemma A7. On the other hand, under Assumption 4.1 (i,iii) the function

$[|X_1|^p|Z_1]\cdot f$ is uniformly bounded and $\|K\|_p$ is finite for some $p=2/(1-q)>2$, therefore we have $[\ |X_1K_b(Z_1-z)|^p]^{2/p}\lesssim b^{-(q+1)}$. Using the 2-mixing property of $\{Z_t\}$ together with Hall and Heyde [27], Theorem A.6, we obtain the second bound in (A25). □

Lemma A.7 Suppose the stationary vector time series $\{(X_t,Z_t)\}$ is 2-mixing with size $\mathfrak{v}$ and Assumption 4.6 is fulfilled for some $q,q_G\in(0,1)$. In addition assume, that the function $g=\varphi\cdot f$ belongs to $\mathfrak{G}^s_\triangle$ for $s,\triangle>0$. Let $\widehat{g}$ be defined as in (11), where the kernel is of order s. Then we have

$$|\widehat{g}(z)-g(z)|\lesssim b^{2s}+b^{-1}T^{-1}+b^{-(1+q+q_G(1-[(q\mathfrak{v})\vee 1]))}T^{-[(q\mathfrak{v})\wedge 1]} \tag{A26}$$

Proof. Under the stated assumptions using Lemma A.6 the proof is very similar to the proof of Lemma A.4 and we omit the details. □

Proof of Theorem 4.7. Using Lemma A.5 and A.5 we obtain the result using a similar proof as Theorem 4.2. □

Proof of Corollary 4.8 Under the assumption on the bandwidth the result is obtained by balancing the terms in the bound given in Theorem 4.2. □

A.3. *Covariances and 2-mixing rates for linear processes*

We use the results derived in this section in Section 3.1, where we compared the rates of convergence for linear processes with the rates in the general 2-mixing case.

Let us suppose $\{Z_t\}$ satisfies the linear process representation in (6). By placing some additional conditions on the innovations we have the following lemma, which is due to Giraitis et al. [22], Lemma 1 and 2.

Lemma A.8 Giraitis et al. [22] Suppose $\{Z_t\}$ is a linear process which satisfies (6), and $\mathrm{cov}(Z_0,Z_t)\leq Ct^{-\theta}$. Let f be the density of Z_t and let f_t denote the joint density Z_0,Z_t. If $(|\varepsilon_t^3|)<\infty$, and for all $u\in\mathbb{R}$ suppose the characteristic function satisfies $|\ [\exp(-iu\varepsilon_1)]|\leq\frac{1}{(1+|u|)^\delta}$ for some $\delta>0$, then the joint density satisfies the relation

$$f_t(x,y)=f(x)f(y)+r(t)f'(x)f'(y)+O(t^{-\theta-d}),$$

where $f'\in L_1(\mathbb{R})$ and $r(t)=\mathrm{cov}(Z_0,Z_t)$, for some $0<d<\min(\frac{\theta}{7},\frac{1-\theta}{12})$.

Using the result above the MSE of the kernel estimator with observations from a linear process can be derived.

For most processes, there isn't a direct correspondence between the 2-mixing and the covariance size. However for Gaussian processes both sizes are linked by the inequality

$$\frac{|\mathrm{cov}(Z_0,Z_t)|}{\mathrm{var}(Z_0)}\leq\sup_{A\in\sigma(Z_0),B\in\sigma(Z_t)}|P(A\cap B)-P(A)P(B)|\leq 2\pi\frac{|\mathrm{cov}(Z_0,Z_t)|}{\mathrm{var}(Z_0)} \tag{A27}$$

(see Doukhan [17], Section 2.1), thus the covariance and the 2-mixing sizes are the same. Suppose that $\{Z_t\}$ satisfies (6), where the innovations are Gaussian and $|a_j| \lesssim j^{-\theta}$. Then we have

$$\left.\begin{array}{c} \frac{|\mathrm{cov}(Z_0,Z_t)|}{\mathrm{var}(Z_0)} \\ \text{and} \\ \sup_{A\in\sigma(Z_0)B\in\sigma(Z_t)} |P(A\cap B)-P(A)P(B)| \end{array}\right\} \lesssim \begin{cases} t^{-(2\theta-1)}, & \text{if } 1/2<\theta\leq 1; \\ t^{-\theta}, & \text{if } \theta>1. \end{cases} \quad \text{(A28)}$$

We now consider more general linear processes, which are not necessarily Gaussian. Then the covariance size does not immediately give the 2-mixing size. However, if the density of the innovations satisfies certain smoothness conditions then we can obtain the following bound. We note that the supremum in the below is taken over a larger sigma-algebra than the one used in the definition of 2-mixing.

Lemma A.9 Suppose $\{Z_t\}$ is a linear process which satisfies the representation $Z_t = \sum_{j=0}^{\infty} a_j\varepsilon_{t-j}$, the function $\frac{1}{\sum_{j=0}^{\infty} a_j z^j} = \sum_{j=0}^{\infty} b_j z^j$ is analytic inside and on the unit circle. Let f_ε be the density of the innovation ε_t, which satisfies $\int |f_\varepsilon(x+a) - f_\varepsilon(x)|dx \leq C|a|$. If $\mathbb{E}(|\varepsilon_t|^\ell) < \infty$,

$$\sup_{A\in\sigma(Z_0,Z_{-1}),B\in\sigma(Z_t)} |P(A\cap B)-P(A)P(B)| \leq C(\sum_{j=t+1}^{\infty} a_j^2)^{\ell/(2(\ell+1))}$$

where C is an arbitrary constant.

PROOF. The following proof is motivated by Chanda [8], Gorodetskii [24] and Davidson [16] (Theorem 14.9). To prove the result, we use the decomposition $Z_t = a_0\varepsilon_t + \sum_{j=1}^{t} a_j\varepsilon_{t-j} + V_t$, where $V_t = \sum_{j=t+1}^{\infty} a_j\varepsilon_{t-j}$. We first note that under the stated assumptions, $\{Z_t\}$, is an invertible time series that is $\varepsilon_t = \sum_{j=0}^{\infty} b_j Z_{t-j}$, hence $\sigma(Z_t, Z_{t-1}, \ldots) = \sigma(\varepsilon_t, \varepsilon_{t-1}, \ldots)$, and the event $\{V_t \leq \eta\} \in \sigma(Z_0, Z_{-1}, \ldots)$. Now by slightly adapting Proposition 2.1, in Fryzlewicz and Subba Rao [20], we obtain

$$\begin{aligned} &\sup_{\substack{A\in\sigma(Z_t)\\ B\in\sigma(Z_0,Z_{-1},\ldots)}} |P(A\cap B)-P(A)P(B)| \quad \text{(A29)} \\ &\leq 2\sup_{|v|\leq\eta}\int f_{\underline{\varepsilon}_{t-1}}(\underline{\varepsilon}_{t-1})\left\{\int \left|f_{Z_t|\underline{\varepsilon}_{t-1},V_t}(x|\underline{\varepsilon}_{t-1},v) - f_{Z_t|\underline{\varepsilon}_{t-1},V_t}(x|\underline{\varepsilon}_{t-1},0)\right|dx\right\} d\underline{\varepsilon}_{t-1} + 3P(|V_t|\geq\eta), \end{aligned}$$

where $f_{Z_t|\underline{\varepsilon}_{t-1},V_t}$ is the conditional density of Z_t given $\underline{\varepsilon}_{t-1} = (\varepsilon_{t-1}, \ldots, \varepsilon_1)$ and V_t. We now bound each of the terms in (A29).

We first obtain an expression the conditional density $f_{Z_t|\underline{\varepsilon}_{t-1},V_t}$ in terms of the innovation density f_ε. Since $\varepsilon_t = a_0^{-1}(Z_t - \sum_{j=1}^{t} a_j\varepsilon_{t-j} - V_t)$ we have $f_{Z_t|\underline{\varepsilon}_{t-1},V_t}(x|\underline{\varepsilon}_{t-1},v) = a_0^{-1}f_Z(a_0^{-1}(x_t - \sum_{j=1}^{t} a_j\varepsilon_{t-j} - z))$. Substituting this into (A29) and by using the stated assumptions we have

$$\int f_{\underline{\varepsilon}_{t-1}}(\underline{\varepsilon})\left\{\int \left|f_{Z_t|\underline{\varepsilon}_{t-1},V_t}(x|\underline{\varepsilon}_{t-1},v) - f_{Z_t|\underline{\varepsilon}_{t-1},V_t}(x|\underline{\varepsilon}_{t-1},0)\right|dx\right\} d\underline{\varepsilon}_{t-1} \leq K|\xi| \quad \text{(A30)}$$

This means that for all $\eta > 0$

$$\sup_{\substack{A\in\sigma(Z_t)\\ B\in\sigma(Z_0,Z_{-1},\ldots)}} |P(A\cap B) - P(A)P(B)| \leq 2K|\eta| + 3P(|V_t| \geq \eta), \qquad \text{(A31)}$$

where K is a constant independent of η. We now bound $P(|V_t| \geq \eta)$. Using the Markov and Burkhölder inequalities we have

$$P(|V_t| \geq \eta) \leq \frac{\mathbb{E}|V_t^\ell|}{\eta^\ell} \leq \frac{2^{ell-1}(\sum_{j=t+1}^\infty a_j^2)^{\ell/2}\mathbb{E}(|\varepsilon_t^\ell|)}{\eta^\ell}.$$

Substituting the above bound into (A31) gives

$$\sup_{\substack{A\in\sigma(Z_t)\\ B\in\sigma(Z_0,Z_{-1},\ldots)}} |P(A\cap B) - P(A)P(B)| \leq 2K\left[|\eta| + \frac{(\sum_{j=t+1}^\infty a_j^2)^{\ell/2})}{\eta^\ell}\right].$$

The minimum of the right hand side of the above is obtained when $\eta = (\sum_{j=t+1}^\infty a_j^2)^{\ell/(2(\ell+1))}$, giving

$$\sup_{\substack{A\in\sigma(Z_t)\\ B\in\sigma(Z_0,Z_{-1},\ldots)}} |P(A\cap B) - P(A)P(B)| \leq K(\sum_{j=t+1}^\infty a_j^2)^{\ell/(2(\ell+1))}.$$

Thus yielding the desired result. □

Remark A.10

(i) Let us suppose the parameters in the linear process satisfy $|a_j| \leq Cj^{-\theta}$ (with $\theta > 1/2$ and $\mathbb{E}(|\varepsilon_t|^\ell) < \infty$). Then we have

$$\sup_{A\in\sigma(Z_0,Z_{-1},\ldots),B\in\sigma(Z_t)} |P(A\cap B) - P(A)P(B)| \leq Cj^{(-2\theta+1)\frac{\ell}{2(\ell+1)}}$$

where C is an arbitrary constant.

(ii) It is interesting to compare the 2-mixing sizes derived in Lemma A.9 with the strong α-mixing results for MA(∞) processes. Under the same set of conditions, but with the additional restriction that $\theta > 3/2$, we have that

$$\sup_{A\in\sigma(Z_0,Z_{-1},\ldots),B\in\sigma(Z_t,Z_{t+1},\ldots)} |P(A\cap B) - P(A)P(B)| \lesssim |t|^{(-2\theta+1)\frac{\ell}{2(\ell+1)+1}}.$$

In other words, the 2-mixing size is larger than the α-mixing size. This is because, by definition, the σ-algebras involved in the definition of α-mixing is far larger than the σ-algebras in the definition of 2-mixing, thus allowing more extreme cases. □

Comparing Lemma A.9 with the covariance size given in (A28) we see when the Gaussianity assumption is relaxed the covariance and 2-mixing sizes no longer necessarily coincide. However by using Lemma A.9 and Hall and Heyde [27], Theorem

A.5, we have the upper and lower bounds

$$j^{(-2\theta+1)\frac{\ell}{(\ell-2)}} \lesssim \sup_{A\in\sigma(Z_0),B\in\sigma(Z_t)} |P(A\cap B)-P(A)P(B)| \lesssim j^{(-2\theta+1)\frac{\ell}{2(\ell+1)}}.$$

Therefore the 2-mixing size $\mathfrak{v}$ of the linear process $\{Z_t\}$ is bounded by

$$(2\theta-1)\frac{\ell}{2(\ell+1)} \leq \mathfrak{v} \leq (2\theta-1)\frac{\ell}{(\ell-2)}. \tag{A32}$$

References

[1] K. B. Athreya and S. Pantula. Mixing properties of Harris chains and autoregressive processes. *J. Appl. Probab.*, 23:880–892, 1986.

[2] B. Basrak, R. Davis, and T. Mikosch. Regular variation of GARCH processes. *Stochastic Processes and their Applications*, 99:95–115, 2002.

[3] D. Bosq. *Nonparametric Statistics for Stochastic Processes*. Springer, New York, 1998.

[4] F. Bousamma. *Ergodicité, mélange et estimation dans les modèles GARCH*. PhD thesis, Paris 7, 1998.

[5] R. C. Bradley. *Introduction to Strong Mixing Conditions Volumes 1,2 and 3*. Kendrick Press, 2007.

[6] R. C. Bradley. A covariance inequality under a two-part dependence assumption. *Statistics and Probability Letters*, 30:287–293, 1996.

[7] A. Bryk and J. Mielniczuk. Asymptotic properties of density estimates for linear processes: applications of linear projection method. *J. Nonparametric Statistics*, 17:121–133, 2005.

[8] K. C. Chanda. Strong mixing properties of linear stochastic processes. *J. Appl. Prob.*, 11:401–408, 1974.

[9] B. Cheng and P. M. Robinson. Density estimation in strongly dependent non-linear time series. *Statistica Sinica*, 1:335–359, 1991.

[10] B. Cheng and P. M. Robinson. Semiparametric estimation from time series with long-range dependence. *Journal of Econometrics*, 94:335–353, 1994.

[11] G. Claesken and P. Hall. effect of dependence of stochastic measures of accuracy of density estimators. *Ann. Statist.*, 30:431–454, 2002.

[12] D. Cline and H. Pu. Geometric ergodicity of nonlinear time series. *Statistica Sinica*, 9:1103–118, 1999.

[13] S. Csörgö and J. Mielniczuk. Nonparametric regression under long-range dependent normal errors. *Ann. Statist.*, 23:1000–1014, 1995.

[14] S. Csörgö and J. Mielniczuk. The smoothing dichotomoy in random-design regression with long-memory based on moving averages. *Statistica Sinica*, 10:771–787, 2001.

[15] S. Csörgö and J. Mielniczuk. Random-design regression under long range dependence errors. *Bernoulli*, 5:209–224, 1999.

[16] J. Davidson. *Stochastic Limit Theory*. Oxford University Press, Oxford, 1994.

[17] P. Doukhan. *Mixing, Properties and Examples*. Springer, New York, 1994.

[18] G. Estevas and P. Vieu. Nonparametric estimation under long memory dependence. *Nonparametric Statistics*, 15:535–551, 2003.

[19] J. Fan and Q. Yao. *Nonlinear Time Series: nonparametric and parametric models*. Springer, New York, 2005.

[20] P. Fryzlewicz and S. Subba Rao. On mixing properties of ARCH and time-varying ARCH processes. *Preprint*, 2007.

[21] J. Geweke and S. Porter-Hudak. The estimation and application of long memory time series models. *Journal of Time Series Analysis*, 4:221–238, 1983.

[22] L. Giraitis, H. L. Koul, and D. Surgailis. Asymptotic normality of regression estimators with long memory errors. *Statistics and Probability Letters*, 29:317–335, 1996.

[23] L. Giraitis, P. Kokoskza, and R. Leipus. Stationary ARCH models: Dependence structure and central limit theorem. *Econometric Theory*, 16:3–22, 2000.

[24] V. Gorodetskii. On the strong mixing propery for linear sequences. *Theory of Probability and its Applications*, 22:411–413, 1977.

[25] P. Hall and J. Hart. Convergence rates in density estimation for data from infinite-order moving average processes. *Probability Theory and Related Fields*, 87:253–274, 1990.

[26] P. Hall and J. Hart. Nonparametric regression with long-range dependence. *Stochastic Processes and their Applications*, 87:339–351, 1990.

[27] P. Hall and C. Heyde. *Martingale Limit Theory and its Application*. Academic Press, New York, 1980.

[28] P. Hall, S. N. Lahiri, and Y. K. Truong. On bandwidth selection choice for density estimation with dependent data. *Ann. Statist.*, 23:2241–2264, 1995.

[29] J. Hart. Efficiency of a kernel density estimator under an autoregressive dependence model. *Journal of the American Statistical Association*, 83:86–99, 1984.

[30] J. Hart. Automated kernel smoothing of dependent data using time series cross validation. *Journal of the Royal Statistical Society (B)*, 56:529–542, 1994.

[31] H. R. Künsch. Statistical aspects of self-similar processes. *Proceedings of the World Congress of the Bernoulli Society*, 1:67–74, 1987.

[32] O. B. Linton and E. Mammen. Estimating semiparametric ARCH(∞) models by kernel smoothing methods. *Econometrica*, 73:771–836, 2004.

[33] E. Masry and D. Tjøstheim. Nonparametric estimation and identification of nonlinear ARCH time series. *Econometric Theory*, 11:258–289, 1995.

[34] J. Mielniczuk. On the asymptotic mean integrated squares error of kernel density estimator for dependent data. *Statistics and Probability Letters*, 34:53–58, 1997.

[35] D. T. Pham. The mixing propery of bilinear and generalised random coefficient autorregressive models. *Stochastic Processes and their Applications*, 23:291–300, 1986.

[36] E. Rio. Covariance inequalities for strongly mixing processes. *Ann. Inst. H. Poincaré Prob. Statist.*, 29:587–597, 1993.

[37] E. Rio. About the lindeberg method for strongly mixing sequences. *ESAIM Probab. Statist.*, 1:35–61, 1997.

[38] P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56:931–954, 1988.

[39] P. M. Robinson. Nonparamatric estimates for time series. *Journal of Time Series Analysis*, 4:185–201, 1983.

[40] P. M. Robinson. Testing for strong serial correlation and dynamic conditional heteroskedasity in multiple regression. *Journal of Econometrics*, 47:67–78, 1991.

[41] P. M. Robinson. Log-periodogram regression of time series with long range dependence. *Ann. Statist.*, 23:1048–1072, 1995.

[42] M. Rosenblatt. Density estimates and markov sequences. In M. Puri, editor, *Non-parametric techniques in statistical inference*, pages 199–210. Cambridge University Press, London, 1970.

[43] D. W. Scott. *Multivariate Density Estimation*. Wiley, New York, 1992.

[44] S. Subba Rao. A note on uniform convergence of an ARCH(∞) estimator. *Sankhya*, 68:600–620, 2006.

[45] M.-C. Viano, C. Deniau, and G. Oppenheim. Long-range dependence and mixing for discrete time fractional processes. *J. Time Ser. Anal.*, 16:323–338, 1995.

[46] P. Vieu. Quadratic errors for nonparametric estimates under dependence. *Journal of Multivariate Analysis*, 39:324–47, 1991.