ELSEVIER

Statistics & Probability Letters 44 (1999) 29–45

STATISTICS & PROBABILITY LETTERS

www.elsevier.nl/locate/stapro

On nonparametric estimation in nonlinear AR(1)-models

Marc Hoffmann *

CNRS-UMR 7599, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris VI et VII, couloir 45-55, 5e étage, 2 place Jussieu, 75251 Paris Cedex 05, France

Received May 1997; received in revised form October 1998

Abstract

We estimate the mean function and the conditional variance (the volatility function) of a nonlinear first-order autoregressive model nonparametrically. Minimax rates of convergence are established over a scale of Besov bodies B_{spq} and a range of global $L_{p'}$ error measurements, for $1 \leqslant p' < \infty$. We propose an estimating procedure based on a martingale regression approximation scheme. This enables us to implement wavelet thresholding and obtain adaptation results with respect to an unknown degree of smoothness. © 1999 Published by Elsevier Science B.V. All rights reserved

Keywords: Minimax estimation; Adaptive estimation; Weak dependence; Time series; Nonparametric regression; Wavelet thresholding

1. Introduction

1.1. Motivation

A vast literature is devoted to the study of nonlinear time series models, especially for financial economics purposes. From a statistical point of view, a nonparametric approach seems appropriate for estimating the conditional mean and variance (the volatility function) of nonlinear AR(1) models. This allows to consider a wide range of models, specified by smoothness properties on the coefficients only. Several authors dealts with the estimation of the mean function using Nadaraya–Watson estimators (Robinson, 1983; Tjøstheim, 1994; Masry and Tjøstheim, 1995). In Härdle and Tsybakov (1995), the pointwise estimation of the volatility function was proposed by local polynomial fits. The method they proposed consider the case when both the mean and the variance belong to the same smoothness class.

On the other side, nonparametric estimation in signal analysis has known over the last few years a significant development in the so-called *adaptive estimation* (among many others: Efromovitch, 1985; Lepski, 1990). The introduction of *wavelet thresholding* in the work of Donoho et al. (1995, 1996) – DJKP for abbreviation – has provided with computationaly fast adaptive procedures in density estimation and nonparametric regression.

* Tel.: +33 44 27 79 74; fax: +33 4427 7674; e-mail: hoffmann@math.jussieu.fr.

PII: S0167-7152(98)00289-2

In this paper, we study the global estimation of the mean and the variance function, when both parameters are subject to different functional constraints, within a range of Besov smoothness classes. Our study is taken from a theoretical angle, and we compute the minimax rates of convergence for $L_{p'}$ losses ($1 \leqslant p' < \infty$), when the unknown parameter has a smoothness s measured in L_p norm (without assuming $p = p'$). This naturally leads to the use of nonlinear procedures (see DJKP 1995, 1996). By thresholding the empirical wavelets coefficients, we investigate *adaptive estimation*, in the sense that our estimators achieve the optimal rates of convergence (to within a logarithmic term in some cases) without the need to specify s or p. Thus, the smoothness properties of the unknown parameter can be unknown, which is more realistic if practical considerations are taken into account.

To our knowledge, our study provides with the first adaptation results in this area (in the sense defined above), in parallel to a recent work of Barraud et al. (1997) for the case of the mean function via a model selection approach. See also Dahlhaus et al. (1995) for time-varying autoregressive processes.

We make use of a systematic *time evolving signal plus noise* analogy. This emphasises the closeness of AR(1) models to more tractable models such as nonparametric regression (see also Neumann, 1996; Neumann and Kreiss, 1996).

1.2. Outline

We consider the observation $X^{(n)} = (X_1, \ldots, X_n)$ defined by the one-dimensional first-order autoregressive process

$$X_{t+1} = m(X_t) + \sigma(X_t)\varepsilon_{t+1}, \quad X_0 = x_0, \ t = 0, \ldots, n-1, \tag{1.1}$$

where the mean function $m(\cdot)$ and the variance $\sigma^2(\cdot)$ are unknown parameters belonging to some Besov smoothness class (see Assumption A1 in Section 2 below). The innovation terms (ε_t, $t = 1, \ldots, n$) are i.i.d. variables, with a common (unknown) density g such that

$$E(\varepsilon_t) = 0 \quad \text{and} \quad E(\varepsilon_t^2) = 1.$$

We assume that X is ergodic in a strong sense (Assumption A2 below). Note that the initial condition $X_0 = x_0$ is not restrictive, and can be replaced by an arbitrary initial law with some additional technical assumptions.

Let $f = m$ or σ^2 be the unknown parameter of interest. Consider a compactly supported pair (φ, ψ) of scaling function and wavelet. For a function h and integers (j,k), define $h_{jk}(x) = 2^{j/2}h(2^j x - k)$. The classical wavelet threshold estimator $\hat{f}_n$ of f has the form

$$\hat{f}_n(x) = \sum_k \hat{\alpha}_{j_0 k}\varphi_{j_0 k}(x) + \sum_{j=j_0}^{j_1} \hat{\beta}_{jk} 1_{|\hat{\beta}_{jk}| \geqslant \lambda_j} \psi_{jk}(x), \tag{1.2}$$

where $\lambda_j = \lambda_j(n)$ is the threshold level and $\hat{\alpha}_{j_0 k} = \hat{\alpha}_{j_0 k}(X^{(n)})$ and $\hat{\beta}_{jk} = \hat{\beta}_{jk}(X^{(n)})$ are estimates of the wavelet coefficients $\alpha_{j_0 k} = \langle f, \varphi_{j_0 k}\rangle$ and $\beta_{jk} = \langle f, \psi_{jk}\rangle$ for the usual L_2 inner product. Thus, by specifying $j_0 = j_0(n)$ and $j_1(n)$, we estimate f by a low-frequency approximation at level j_0 (in a dyadic scale) and add relevant details $\hat{\beta}_{jk}$ for $j = j_0$ to j_0 only if they exceed the threshold level λ_j. The theoretical and practical advantages of the multiscale structure along with thresholding has been extensively discussed in the literature (e.g. DJKP, 1995). Going back to time series, our statistical problems reduce to the estimation of the integrals $\alpha_{j_0 k}$ and β_{jk}. For this, we use a martingale regression approach, in which we embed the two estimation problems (i.e. $f = m$ and σ^2). For the mean function m, we consider the model

$$Y_t = m(X_t) + \xi_t, \quad t = 0, \ldots, n-1, \tag{1.3}$$

where $Y_t = X_{t+1}$, and $\xi_t = \sigma(X_t)\varepsilon_{t+1}$ can be viewed as a noise term. Considering the process (X_t) subsampled at proper stopping times enables us to get rid of the spatial inhomogeneity of the design process (X_t),

while preserving the martingale structure of the noise (ξ_t). For the variance function σ^2, we apply a similar algorithm to

$$Y_t = \sigma^2(X_t) + \eta_t + \xi_t, \quad t = 0, \dots, n-1, \tag{1.4}$$

where $Y_t = \{X_{t+1} - \hat{m}_t(X_t)\}^2$ and $\hat{m}_t(x)$ is a preliminary estimator of $m(x)$ constructed using only data up to time t (see Section 3.2). The term $\eta_t = \{X_{t+1} - \hat{m}_t(X_t)\}^2$ is a *small-noise* component. The martingale noise term is now

$$\xi_t = 2\{m(X_t) - \hat{m}_t(X_t)\}\sigma(X_t)\varepsilon_{t+1} + \sigma^2(X_t)(\varepsilon_{t+1}^2 - 1).$$

Our method will prove to be optimal in the minimax sense w.r.t. rates of convergence (up to a logarithmic factor in some cases). From a practical point of view however, it has the drawback of discarding data (via the homogeneization subsampling procedure, see below). This can be healed by using a correction algorithm, and the estimators proposed here should be viewed as *pilot estimators* in a first step procedure. Since we mainly focus on asymptotical results here, we will no longer consider the problem of practical implementation.

1.3. Contents

In Section 2, we discuss our assumptions on the model. Section 3 is devoted to the construction of estimators. We present a general framework of a *time-evolving signal plus noise model* which embodies AR(1) models. The results and proofs are given in Sections 4 and 5, respectively.

2. Assumptions

We write P_{x_0} for the law of the chain (X_t) with initial condition $X_0 = x_0$. Let D be a compact interval. We denote by K a real-valued Lipschitz continuous function satisfying $\int K(x)\,dx = 1$ and set $K_h(x) = h^{-1}K(h^{-1}x)$. Let B_{spq} denotes the Besov space $B_{\mathrm{spq}}(\mathbb{R})$ restricted to D, with norm $\|\cdot\|_{\mathrm{spq}}$ (e.g. Meyer, 1990), for $s > 0$, $1 \leqslant p, q < \infty$. Given $M > 0$, put

$$B_{\mathrm{spq}}(M) = \{f \in B_{\mathrm{spq}} : \|f\|_{\mathrm{spq}} \leqslant M\}.$$

We consider the statistical model defined by (1.1) and make the following assumptions:

A1. *local assumptions on $m(\cdot)$ and $\sigma(\cdot)$*

$$(m, \sigma^2) \in B_{s_1 p_1 q_1}(M_1) \times B_{s_2 p_2 q_2}(M_2).$$

A2. *global assumptions on $m(\cdot)$ and $\sigma(\cdot)$*. The process $(X_t,\ t \geqslant 0)$ has a unique stationary measure with density μ w.r.t. the Lebesgue measure on D, satisfying

$$\forall x \in D : \mu(x) \geqslant v > 0 \tag{2.5}$$

for an explicitly computable $v > 0$. Moreover, the density μ can be estimated at some polynomial rate uniformly on D: there exists $\Gamma > 0$ s.t.

$$\forall \gamma \geqslant 1 : \sup_{x \in D} E\{|\mu_n(x) - \mu(x)|^\gamma\} \leqslant M_3(\gamma) n^{-\Gamma\gamma}, \tag{2.6}$$

where

$$\mu_n(x) = \frac{1}{n+1} \sum_{t=0}^{n} K_{h_n}(X_t - x) \tag{2.7}$$

for a suitable bandwidth h_n.

A3. *assumptions on the innovation terms*. For some constant M_4 and for all $\gamma \in [1, \infty)$

$$\int_{-\infty}^{+\infty} |x|^\gamma g(x)\,\mathrm{d}x \leqslant M_4^\gamma \gamma^{\gamma/2}. \tag{2.8}$$

Remarks.

1. The constant Γ needs not be known and can be arbitrarily small.
2. The moment condition A3 is satisfied for bounded Gaussian-type errors. The exponential bound can actually be relaxed to some polynomial moment growth, at the cost of a (significantly more technical) improvement of Lemma 7 below, thanks to Fuk and Nagaev inequality (Fuk and Nagaev, 1971).
3. The Besov indices need not be equal for m and σ^2. In particular, the smoothness of σ^2 may differ from that of m to within the range $(1, \infty)$.
4. Whereas A1 is classical in the minimax theory as well as A3 in the framework of time series, we need to elaborate on assumption A2. It describes the minimal features needed to perform our estimating procedure. Denote by $\pi(x, y)$ the transition density of the chain (X_t), given by

$$\pi(x, y) = \frac{1}{\sigma(x)} g\left(\frac{y - m(x)}{\sigma(x)}\right)$$

and set, for any test function f

$$\pi f(x) = \int \pi(x, y) f(y)\,\mathrm{d}y.$$

Consider the following assumptions:

B1. The function $V(x) = |x|$ is (C_0, C_1)- Lyapunov for π, with $C_0 < 1$, i.e. $\pi V(x) \geqslant C_0|x| + C_1$ for all real x.
B2. $|m(x)| \vee |\sigma(x)| \leqslant C_2(1 + |x|)$ for all real x.
B3. $\inf_x \sigma(x) \geqslant C_3 > 0$.
B4. For all compact K: $\inf_{x \in K} g(x) > 0$.

Remarks.

1. Assumption B1 plays the role of a (the so-called) drift condition, enabling the strong ergodicity of the process (X_t).
2. Assumption B4 is satisfied for standard Gaussian errors.
3. Let us give classical examples of functions m and σ satisfying B1 and B2. Consider the space of Lipschitz continuous functions

$$\mathrm{Lip}(M) = \{f : |f(x) - f(y)| \leqslant M|x - y|\}.$$

Provided $m \in \mathrm{Lip}(M_1)$ and $\sigma \in \mathrm{Lip}(M_2)$, B1 holds if

$$M_1 + M_2 \int |x| g(x)\,\mathrm{d}x \leqslant C_0$$

and B2 holds if $|m(0)| \vee |\sigma(0)| \leqslant M_3$, with $M_1 \vee M_2 \vee M_3 \leqslant C_1$.

Lemma 1. *Assumptions* B1, B2, B3 *and* B4 *imply* A2.

Proof of Lemma 1. Under B1, B2, B3 and B4, the chain is geometrically ergodic, with exponentially fast decay mixing coefficients (Doukhan, 1995, pp. 105–106 and the references therein). This implies inequality

(2.6) by classical density kernel estimation for weakly dependent variables (see for instances Neumann, 1996). We now compute an explicit lower bound for μ on D. The invariant density satisfies

$$\mu(x) = \int \frac{1}{\sigma(y)} g\left(\frac{x - m(y)}{\sigma(y)}\right) \mu(y)\,\mathrm{d}y.$$

We assume, without loss of generality, that $D = [-c_0, c_0]$ for some $c_0 > 0$. Let $c_1 > 0$. For any $(x, y) \in [-c_0, c_0] \times [-c_1, c_1]$, we have, from assumptions B1 and B3

$$\left|\frac{x - m(y)}{\sigma(y)}\right| \leqslant C_3^{-1}(c_0 + C_2(1 + c_1)) = c_2,$$

say. Hence

$$\mu(x) \geqslant \inf_{|t| \leqslant c_2} g(t) C_2^{-1}(1 + c_1)^{-1} \int_{-c_1}^{c_1} \mu(y)\,\mathrm{d}y.$$

Applying Chebyshev's inequality yields

$$\int_{-c_1}^{c_1} \mu(y)\,\mathrm{d}y \geqslant 1 - c_1^{-1} \int V(x)\mu(y)\,\mathrm{d}y \geqslant 1 - c_1^{-1} C_1/(1 - C_0).$$

The last inequality is obtained by assumption B1 (see for instance Duflo, 1990). Finally, we can take

$$v = \inf_{|t| \leqslant c_2} g(t) C_2^{-1}(1 + c_1)^{-1}(1 - c_1 C_1/(1 - C_0)).$$

We check that $v > 0$ by taking c_1 large enough. This proves Lemma 1. □

In the following, we will denote by $\Sigma = \Sigma(M)$ for $M = (M_3, M_4)$ the functional constraint imposed by assumptions A2 and A3. Note that $\Sigma(M)$ can be replaced by $\Sigma(C, M_4)$, $C = (C_0, \ldots, C_3)$ if one prefers to consider the more familiar framework of assumptions B1–B4. We denote by

$$\Sigma_0 = \Sigma \cap B_{s_1 p_1 q_1}(M_1) \times B_{s_2 p_2 q_2}(M_2), \tag{2.9}$$

the global functional constraint on the parameter (m, σ^2).

3. The estimating procedure

3.1. General setting

For $n \geqslant 0$ let $(\mathscr{F}_i^n,\ 0 \leqslant i \leqslant n)$ be a triangular array of sigma-fields such that $\mathscr{F}_0^n \subseteq \mathscr{F}_1^n \subseteq \cdots \subseteq \mathscr{F}_n^n$. Suppose one wants to recover the signal f defined on a compact interval D from data $(X_i, Y_i,\ i = 0, \ldots, n-1)$ in the model

$$Y_i = f(X_i) + \eta_i^n + \xi_i, \quad i = 0, \ldots, n-1, \tag{3.10}$$

where

- the random process $(X_i,\ i = 0, \ldots, n-1)$ is $(\mathscr{F}^n)$ adapted,
- the *noise process* $\sum_{j=0}^{i-1}$ is a $(\mathscr{F}^n)$ martingale,
- the term η_i^n is a small noise component which is $(\mathscr{F}^n)$ adapted.

This model contains usual regression frameworks, including regression with random design. Note that we do not assume that the design points (X_i) are independent nor independent from the noise terms (ξ_i). Assume that the empirical measure $1/n \sum_{i \leqslant n} \delta_{X_i}$ weakly converges to a measure with density μ w.r.t. the Lebesgue

measure on D and that $\mu(x) \geqslant v > 0$ on D for an explicitly known v. The problem of estimating the mean m and the variance σ^2 of the autoregression model defined by (1.1) is embedded in this framework via the transformations (1.3) and (1.4) of Section 1.

3.2. Algorithm

Following (1.2) our problem reduces to estimate the wavelet coefficients α_{j_0k} and β_{jk} of f. Two specific difficulties appear here

1. the limiting density μ has unknown smoothness. Even worse, its smoothness may differ from that of f due to the influence of nuisance parameters.
2. the estimation procedure should be adapted to the filtration $(\mathscr{F}^n)$ so that the noise terms remain uncorrelated.

Assuming that μ is bounded below by $v > 0$, the convergence of the empirical sampling measure ensures that $\lfloor nh_n v \rfloor$ observation points (at least) lie in a neighbourhood of size h_n of any given point in D with high probability. We can then construct an estimator by subsampling $\lfloor nv \rfloor$ observation points as follows.

Assume (without loss of generality) that $D = [0,1]$. Given $h_n > 0$, we divide D into $\lfloor h_n^{-1} \rfloor$ small intervals of size h_n each, denoted by C_λ, $\lambda = 1, \ldots, \lfloor h_n^{-1} \rfloor$. Put

$$N_i^\lambda = \left(\sum_{i<j} 1_{X_j \in C_\lambda} \right) \wedge \lfloor nh_n v \rfloor,$$

$$T_1 = 0 \text{ and for } i \geqslant 2 : T_i = \inf \left\{ j > T_{i-1} : \sum_\lambda (N_j^\lambda - N_{T_{i-1}}^\lambda) \geqslant 1 \right\} \wedge n.$$

Note that the T_i are increasing stopping times of the filtration $(\mathscr{F}^n)$. We extract from $(X_0, \ldots, X_{n-1})$ the subsampling $(X_{T_1}, \ldots, X_{T_{\lfloor nv \rfloor}})$ and compute the empirical wavelet coefficients from data $(x_{T_i}, Y_{T_i},\ i = 1, \ldots, \lfloor nv \rfloor)$ on the coarse grid $(x_{T_i},\ i = 1, \ldots, \lfloor nv \rfloor)$ defined by

$$x_{T_i} = (\lambda_{T_i} - 1)h_n + l_{T_i}/\lfloor nv \rfloor,$$

where λ_{T_i} is the index of the interval C_λ in which falls the observation X_{T_i} and $l_{T_i} = \#\{X_{T_j} \in C_{\lambda_{T_j}},\ j \leqslant i\}$. Note that the x_{T_i} are a reordering of a regular grid at coarse scale $\lfloor 1/nv \rfloor$ needed for technical reason (see the proof of Theorem 2 below).

Definition 1. The (subsampled) wavelet coefficient estimates at accuracy level v are

$$\hat{\alpha}_{j_0k} = \frac{1}{\lfloor nv \rfloor} \sum_{i=1}^{\lfloor nv \rfloor} Y_{T_i} \varphi_{j_0k}(x_{T_i}), \qquad \hat{\beta}_{jk} = \frac{1}{\lfloor nv \rfloor} \sum_{i=1}^{\lfloor nv \rfloor} Y_{T_i} \psi_{jk}(x_{T_i}).$$

Definition 2. The threshold wavelet estimator $\hat{f}_n$ of f specified by $\lambda_j = \lambda_j(n)$, $j_0 = j_0(n)$, $j_1 = j_1(n)$ and v is given by the formula

$$\hat{f}_n(x) = \sum_k \hat{\alpha}_{j_0k} \varphi_{j_0k}(x) + \sum_{j=j_0}^{j_1} \sum_k \hat{\beta}_{jk} 1_{|\hat{\beta}_{jk}| \geqslant \lambda_j} \psi_{jk}(x)$$

where the $\hat{\alpha}_{j_0k}$ and $\hat{\beta}_{jk}$ are given in Definition 1.

Remarks.

1. For sake of simplicity, we omit the boundary conditions on the edges of D, provided by wavelets on the interval (Cohen et al., 1994).
2. The above subsampling method has the major drawback that it discards data after the random time $T_{\lfloor nv \rfloor}$. The reason to introduce this subsampling procedure is mainly technical, and has no effect since only rates of convergence are studied in this paper. The next level of accuracy should be to look for minimax efficiency (i.e. optimal asymptotic constants). Such a task is beyond the techniques used in this paper.

For $f = m$ of σ^2, we consider $\hat{f}_n$ of Definition 2 at accuracy level $\bar{v} = v/2$ (this will be explained in the proof of Theorem 2 below). For $f = m$, we use data

$$Y_t = X_{t+1}, \quad t = 0, \ldots, n-1$$

and for $f = \sigma^2$, we use data

$$Y_t = \{X_{t+1} - \hat{m}_t(X_t)\}^2, \quad t = 0, \ldots, n-1,$$

where $\hat{m}_t$ is the Nadaraya–Watson estimator of m, given by

$$\hat{m}_t = \mu_t(x)^{-1} \frac{1}{t+1} \sum_{i=0}^{t} K_{\rho_t}(X_i - x)$$

and μ_t is defined by (2.7) using the bandwith ρ_t. The proper choice of ρ_t will be specified in Theorem 2 below.

4. Results

We prove upper and lower bounds for the following minimax risk:

Definition 3. Let $1 \leqslant p' < \infty$ and D be a compact interval of $\mathbb{R}$. The $L_{p'}$-minimax over D of an estimator $\hat{f}_n$ of $f = m$ or σ^2 is

$$R(\hat{f}_n) = \sup_{(m,\sigma^2) \in \Sigma_0} \left\{ E_{x_0} \left(\int_D |\hat{f}_n(x) - f(x)|^{p'} \, \mathrm{d}x \right) \right\}^{1/p'}. \tag{4.11}$$

For sake of clarity, we will denote by R_1 the risk associated to m and R_2 the risk associated to σ^2, respectively.

4.1. Lower bounds

For $i = 1$ or 2, put

$$\alpha_i = \frac{s_i}{1 + 2s_i} \wedge \frac{s_i - 1/p_i + 1/p'}{1 + 2s_i - 2/p_i} \quad \text{and} \quad \varepsilon_i = s_i p_i - \frac{p' - p_i}{2}.$$

Theorem 1. *For i=1 or 2, let $s_i > 1$, $1 \leqslant p_i \leqslant p' < \infty$, $1 \leqslant q_i \leqslant \infty$. Assume the innovation terms are Gaussian, i.e. $g(x) = 1/\sqrt{2\pi} \exp(-x^2/2)$. (In particular,* A3 *is satisfied.) Under Assumptions* A1 *and* A2, *there exists*

an explicitly computable $K_1^i = K_1^i(M, s_i, p_i, q_i, p')$ *or* $K_1^i(C, M_4, s_i, p_i, q_i, p')$ *such that*

$$\inf_{\tilde{f}_n} R^i(\tilde{f}_n) \geqslant K_1^i \left(\frac{\log n}{n}\right)^{\alpha_i}, \quad \varepsilon_i \leqslant 0,$$
$$\inf_{\tilde{f}_n} R^i(\tilde{f}_n) \geqslant K_1^i n^{-\alpha_i}, \quad \varepsilon_i > 0, \tag{4.12}$$

where the infimum is taken over all estimators.

Remark. Under assumption A3 Theorem 1 can be extended to the case of nonGaussian innovation terms, under additional technical conditions on the smoothness of g.

4.2. Upper bounds

Define

$$\tau_n(s, p, q, p') = (\log n)^{(1-\varepsilon/sp)\alpha} n^{-\alpha}, \quad \varepsilon > 0,$$
$$\tau_n(s, p, q, p') = (\log n)^{(1/2-p/qp')_+} (\log n/n)^{\alpha}, \quad \varepsilon = 0,$$
$$\tau_n(s, p, q, p') = (\log n/n)^{\alpha}, \quad \varepsilon < 0,$$

where $x_+ = \max(x, 0)$. In the following, for two real-valued sequences u_n and v_n, we will write $u_n \asymp v_n$ if there exists $C > 0$ independent of n such that $C^{-1} u_n \leqslant v_n \leqslant C u_n$.

Theorem 2. *Assume* A1, A2 *and* A3. *For* $i = 1$ *or* 2, *let* $s_i > 1 + 1/p_i$, $1 \leqslant p_i < \infty$, $1 \leqslant q_i \leqslant \infty$. *If the parameters of the threshold wavelet estimator* $\hat{f}_n$ *are specified by*

$$h_n \asymp n^{-s_0/(1+2s_0)}, \quad s_i \leqslant s_0 < \infty, \quad \lambda_j(n) = K_0 \sqrt{j/n},$$

$$2^{j_0(n)} \asymp (n(\log n)^{(p'-p_i)/(p_i) 1_{\varepsilon_i \leqslant 0}})^{1_2 \alpha_i}, \quad 2^{j_1(n)} \asymp (n/\log n)^{\alpha_i/s_i - 1/p_i},$$

and for the case $f = \sigma^2$,

$$\rho_t \asymp t^{-1/3},$$

then, there exist explicitly computable constants K_0 *and* $K_2^i = K_2^i(M, s_i, p_i, q_i, p')$ *or* $K_2^i(C, M_4, s_i, p_i, q_i, p')$ *such that*

$$R_i(\hat{f}_n) \leqslant K_2^i \tau_n(s_i, p_i, q_i, p'). \tag{4.13}$$

Remarks.

1. The rates obtained are the same as in density estimation or nonparametric regression, with fixed design (see DJKP, 1995, 1996), and, in view of Theorem 1, are sharp up to a logarithmic factor in some cases.
2. The rates of convergence for m do not depend on the characteristics of σ^2 an vice versa. Note that the considered smoothness classes contain at least the Lipschitz class (since $s_i > 1 + 1/p_i$, see assumption A2). This avoids the effect of estimating the mean for recovering the variance (see Hall and Caroll, 1989, for the specific case of nonparametric regression).

We now show how a slight modification of $\hat{f}_n$ makes it *adaptive* w.r.t. an unknown degree of smoothness, in the sense that the rates of Theorem 2 are achieved (possibly to within a logarithmic term in some cases)

without the requirement to specify (s_i, p_i, q_i). In particular, the smoothness of the estimated function can be unknown. Fix an integer $r_0 \geqslant 1$ and define

$$\mathscr{S} = \{(s,p,q):\ 1 + 1/p < s \leqslant r_0, 1 \leqslant p < \infty, 1 \leqslant q \leqslant \infty\}.$$

Definition 4. Let D be a given compact interval of $\mathbb{R}$. Let $\Sigma_0 = \Sigma_0(s,p,q)$ be the smoothness class defined in Section 2. The estimator $\hat{f}_n^*$ of $f = m$ or σ^2 is called adaptive w.r.t. $\mathscr{S}$ is there exists K_3 such that

$$\forall (s,p,q) \in \mathscr{S}: \qquad \sup_{(m,\sigma^2)\in\Sigma_0} \left\{ E_{x_0} \left(\int_D |\hat{f}_n^*(x) - f(x)|^{p'} \, \mathrm{d}x \right) \right\}^{1/p'} \leqslant K_3 \tau_n(s,p,q).$$

For $f = m$ or σ^2, we construct an estimator $\hat{f}_n^*$ from the threshold wavelet estimator by specifying the following parameters: the pair (φ, ψ) generates a r_0-regular and compactly supported multiresolution analysis and

$$\lambda_j(n) = K_0\sqrt{j/n}, \ \ 2^{j_0(n)} \asymp n^{1/(1+2r_0)}, \ \ 2^{j_1(n)} \asymp n/\log n, \ \ h_n \asymp n^{-s_0/(1+2s_0)} \quad \text{and } s_0 > 2r_0 + \tfrac{1}{2}.$$

Theorem 3. *Assume that $f = m$ or σ^2 belongs to some class Σ_0, as specified in Section 2. Then $\hat{f}_n^*$ is adaptive over $\mathscr{S}$, up to logarithmic terms.*

5. Proofs

5.1. Proof of Theorem 1

We go along a classical route, following classical ideas (Bretagnolle and Huber, 1979; Keryacharian and Picard, 1992; Neumann and Spokoiny, 1995). For a review of the likelihood method we use here, see Korostelev and Tsybakov (1993). We break the proof in two parts, the so-called sparse case ($\varepsilon_i < 0$) and the dense case ($\varepsilon_i \geqslant 0$). In the following, P_{m,σ^2} will denote the law on $\mathbb{R}^n$ of the vector $X^{(n)} = (X_1, \ldots, X_n)$ driven by the parameters (m, σ^2), with initial condition x_0.

5.1.1. The dense case $\varepsilon_i \geqslant 0$

For the mean function m, we evaluate a minimax lower bound over $\Sigma_1 = \mathscr{C}_{j_n} \times \{\sigma_0^2\}$, where $\sigma_0^2(x) = 1$ for all $x \in \mathbb{R}$ and

$$\mathscr{C}_{j_n} = \left\{ f(x) = \gamma_n \sum_{k\in K_{j_n}} v_k \psi_{j_n k}(x), v_k = \pm 1 \right\}. \tag{5.14}$$

The function ψ is a wavelet of regularity $r > s_1 \vee s_2$, with compact support in $[-A, A]$, where A is an integer (for instance a Daubechies wavelet), and

$$K_{j_n} = \{-(2^{-j_n} - 1)A + 2lA, l = 0, \ldots, 2^{j_n} - 1\},$$

so that $\#\mathscr{C}_{j_n} = 2^{2^{j_n}}$ and the functions $\psi_{j_n k}$ and $\psi_{j_n k'}$ have disjoint support for $k \neq k'$. We impose $2^{j_n} \asymp n^{1/(1+2s_1)}$ and $\gamma_n \asymp 1/\sqrt{n}$ so that $\Sigma_1 \subset \Sigma_0$.

For the variance function σ^2, we evaluate a lower bound over $\Sigma_2 = \{m_0\} \times \mathscr{C}_{j_n}$, where $m_0 = 0$ and $\mathscr{C}_{j_n}$ is defined following (5.14). The choice $2^{j_n} \asymp n^{1/(1+2s_2)}$ and $\gamma_n \asymp 1/\sqrt{n}$ ensures that $\Sigma_2 \subset \Sigma_0$.

Lemma 2. *For given $k \in K_{j_n}$, denote by $f_+ = f_+(k)$ and $f_- = f_-(k)$ any pair of functions in $\mathscr{C}_{j_n}$ such that*

$$f_+(x) - f_-(x) = 2\gamma_n \psi_{j_n k}(x).$$

Under the assumptions of Theorem 1, *there exist* $\lambda_i > 0$ *and* $z_i > 0$, $i = 1, 2$, *independent of* n *such that for large enough* n

$$P_{f_-,\sigma_0^2}(\Lambda_1(f_+, f_-, X^{(n)}) \geqslant \mathrm{e}^{-\lambda_1}) \geqslant z_1 > 0 \quad \textit{and} \quad P_{m_0,f_-}(\Lambda_2(f_+, f_-, X^{(n)}) \geqslant \mathrm{e}^{-\lambda_2}) \geqslant z_2 > 0,$$

where

$$\Lambda_1(f_+, f_-, X^{(n)}) = \frac{\mathrm{d}P_{f_+,\sigma_0^2}}{\mathrm{d}P_{f_-,\sigma_0^2}}(X^{(n)}) \quad \textit{and} \quad \Lambda_2(f_+, f_-, X^{(n)}) = \frac{\mathrm{d}P_{m_0,f_+}}{\mathrm{d}P_{m_0,f_-}}(X^{(n)}).$$

Proof of Theorem 1, dense case. The lower bound is a consequence of Lemma 2, as follows from Korostelev and Tsybakov (1993, Ch. 2). □

5.1.2. The sparse case $\varepsilon_i < 0$

For the mean function m, we now consider the parametric family $\Sigma_1 = \mathscr{P}_{j_n} \times \{\sigma_0^2\}$, where

$$\mathscr{P}_{j_n} = \{f_0(x), f_{j_n k}(x) = f_0(x) + \gamma_n \psi_{j_n k}(x),\ k \in K_{j_n}\} \tag{5.15}$$

with the same notation as for the dense case. Here, $f_0 = \sigma_0^2$ and we choose $2^{j_n} \asymp (n/\log n)^{1/(1+2s_1-2/p_1)}$ and $\gamma_n \asymp (\log n/n)^{1/2}$ so that $\Sigma_1 \subset \Sigma_0$.

For the variance function σ^2, we consider the subfamily $\Sigma_2 = \{m_0\} \times \mathscr{P}_{j_n}$ with $\mathscr{P}_{j_n}$ defined by (5.15), with now $f_0 = m_0 = 0$. Again, we choose $2^{j_n} \asymp (n/\log n)^{1/(1+2s_2-2/p_2)}$ and $\gamma_n \asymp (\log n/n)^{1/2}$ so that $\Sigma_2 \subset \Sigma_0$.

Lemma 3. *With the notation of Lemma* 2, *for any* $k \in K_{j_n}$, *the following representation holds*:

$$\Lambda_i(f_0, f_{j_n k}, X^{(x)}) = \exp(\kappa_k^{(i)} - \lambda_k^{(i)} \log 2^{j_n}) \tag{5.16}$$

where $\lambda_k^{(i)} \leqslant \lambda_*^{(i)} < 1$ *and* $(\kappa_k^{(i)}, k \in K_{j_n})$ *are random variables such that for large enough* n

$$P_{f_{j_n k},\sigma_0^2}(\kappa_k^{(1)} \geqslant -\lambda^{(1)}) \geqslant z^{(1)} > 0 \quad \textit{and} \quad P_{m_0,f_{j_n k}}(\kappa_k^{(2)} \geqslant -\lambda^{(2)}) \geqslant z^{(2)} > 0$$

for some $\lambda^{(i)}$ *and* $z^{(i)} > 0$ *independent of* n.

Proof of Theorem 1, sparse case. The lower bound follows from Lemma 3, as it follows from Korostelev and Tsybakov (1993, Ch. 2). □

5.1.3. Proof of Lemma 2

The mean function m. For clarity, we abbreviate P_{m_-,σ_0^2} by P_{m_-} and we substitute $f_\pm$ by $m_\pm$. For $m_\pm$, the transition density of the chain is given by

$$\pi_\pm(x, y) = \frac{1}{\sqrt{2\pi}} \exp[-\tfrac{1}{2}(y - m_\pm(x))^2].$$

Under P_{m_-}, direct computation shows that

$$\Lambda_1(m_+, m_-, X^{(n)}) = \exp\left(2 \sum_{i=0}^{n} \{\gamma_n \psi_{j_n k}(X_i)\varepsilon_{i+1} - \gamma_n^2 \psi_{j_n k}^2(X_i)\}\right).$$

Assuming $\gamma_n = 1/\sqrt{n}$ without any loss of generality, we obtain the following inclusion:

$$(\Lambda_1(m_+, m_-, X^{(n)}) \geqslant \mathrm{e}^{-\lambda_1}) \supseteq \left(\left| \frac{1}{\sqrt{n}} \sum_{i=0}^{n} \psi_{j_n k}(X_i)\varepsilon_{i+1} - \frac{1}{n} \sum_{i=0}^{n} \psi_{j_n k}^2(X_i) \right| \leqslant \lambda_1\right).$$

Using Chebyshev and Schwarz inequalities together with the fact that the $\psi_{j_nk}(X_i)\varepsilon_{i+1}$ are zero-mean uncorrelated, we derive that the probability (under P_{m_-}) of the last event is greater than

$$1-\frac{1}{\lambda_1}\left(E_{m_-}\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=0}^{n}\psi_{j_nk}(X_i)\varepsilon_{i+1}\right|\right\}+E_{m_-}\left\{\frac{1}{n}\sum_{i=0}^{n}\psi_{j_nk}^2(X_i)\right\}\right)$$

$$\geqslant 1-\frac{2}{\lambda_1}\max\left(1,E_{m_-}\left\{\frac{1}{n}\sum_{i=0}^{n}\psi_{j_nk}^2(X_i)\right\}\right).$$

It remains to show that the term within the expectation is bounded, and the conclusion follows by taking λ_1 large enough. Clearly, for $i\geqslant 1$

$$E_{m_-}[\psi_{j_nk}^2(X_i)]$$
$$=\int_{\mathbb{R}^{i-1}}\pi_-(x_0,x_1)\cdots\pi_-(x_{i-2},x_{i-1})\left(\int\pi_-(x_{i-1},x_i)\psi_{j_nk}^2(x_i)\,\mathrm{d}x_i\right)\mathrm{d}x_1\cdots\mathrm{d}x_{i-1}\leqslant\frac{1}{\sqrt{2\pi}},$$

since ψ is orthonormal in L_2 and $\sup_{x,y}\pi_-(x,y)\leqslant 1/\sqrt{2\pi}$. For $i=0$, we simply use the fact that $n^{-1}\psi_{j_nk}^2(x_0)\leqslant\sup_x|\psi^2(x)|2^{j_n}n^{-1}$, which converges to 0 as $n\to\infty$. The proof for the mean function is complete. □

The variance function σ^2. For clarity, we abbreviate P_{m_0,σ_-^2} by $P_{\sigma_-^2}$ and substitute $f_\pm$ by $\sigma_\pm^2$. For $\sigma_\pm^2$, the transition density of the chain over Σ_1 is now given by

$$\pi_\pm(x,y)=\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_\pm^2(x)}\exp-\frac{1}{2}\frac{y^2}{\sigma_\pm^2(x)}.$$

Straightforward computation shows that, under $P_{\sigma_-^2}$

$$\Lambda_2(\sigma_+^2,\sigma_-^2,X^{(n)})=\prod\frac{\sigma_-}{\sigma_+}(X_i)\exp-\frac{1}{2}\sum_{i=0}^{n}\left(\frac{\sigma_-^2}{\sigma_+^2}(X_i)-1\right)\varepsilon_{i+1}^2.$$

Assuming $\gamma_n=1/\sqrt{n}$ with no loss of generality, taking logarithm and using a second-order Taylor expansion, it is easily checked that

$$\log\Lambda_2(\sigma_+^2,\sigma_-^2,X^{(n)})=\frac{1}{\sqrt{n}}\sum_{i=0}^{n}\psi_{j_nk}(X_i)(\varepsilon_{i+1}^2-1)-\frac{1}{n}\sum_{i=0}^{n}\psi_{j_nk}^2(X_i)\varepsilon_{i+1}+R_n,$$

where R_n is a remainder term, uniformly bounded thanks to the fact that $2^{3j_n/2}/\sqrt{n}$ is bounded (from the choice of j_n). To complete the proof, we use similar arguments as for the mean function, using now that ε_{i+1}^2-1 is zero-mean. The proof of Lemma 2 is complete. □

5.1.4. Proof of Lemma 3

The mean function m. We abbreviate $P_{m_{j_n},\sigma_0^2}$ by P_{j_nk}. By elementary computation, we have, under P_{j_nk}

$$\Lambda_1(m_0,m_{j_nk},X^{(n)})=\exp\left(\gamma_n\sum_{i=0}^{n}\psi_{j_nk}(X_i)\varepsilon_{i+1}-\frac{1}{2}\gamma_n^2\sum_{i=0}^{n}\psi_{j_nk}^2(X_i)\right).$$

Set, for $L_0>0$

$$\tilde{\kappa}_k^{(1)}=L_0\frac{\log n}{\sqrt{n}}\sum_{i=0}^{n}\psi_{j_nk}(X_i)\varepsilon_{i+1},Z_n=\int\psi_{j_nk}^2(x)\mu(x)\,\mathrm{d}x-\frac{1}{n}\sum_{i=0}^{n}\psi_{j_nk}^2(X_i)$$

and

$$\kappa_k^{(1)} = \tilde{\kappa}_k^{(1)} + \frac{L_0^2}{2}(\log n)^2 Z_n.$$

Assume for simplicity that $\gamma_n = L_0 \log n/\sqrt{n}$. Denote by r_n a real-valued sequence converging to 0. We write $\log 2^{j_n} = \beta \log n(1 + r_n)$, for some $\beta > 0$, say. Define

$$\lambda_k^{(1)} = (1 + r_n)\frac{L_0}{2}\beta^{-1} \int \psi_{j_n k}^2(x)\mu(x)\,\mathrm{d}x.$$

Thus we formally obtain the decomposition required by (5.16). Since the choice of L_0 is free and $|\int \psi_{j_n k}^2(x)\mu(x)\,\mathrm{d}x| \leqslant \sup_{x\in D}|\mu(x)|$, it suffices to take n large enough so that

$$(1 + r_n)\frac{L_0}{2}\beta^{-1} \int \psi_{j_n k}^2(x)\mu(x)\,\mathrm{d}x < 1,$$

to have the condition on the $\lambda_k^{(1)}$ fulfilled. It remains to prove that $P_{j_n k}(\kappa_k^{(1)} \geqslant -\lambda^{(1)}) \geqslant z_0^{(1)} > 0$. From Chebyshev's inequality

$$P_{j_n k}(\kappa_k^{(1)} \geqslant -\lambda^{(1)}) \geqslant P_{j_n k}(\tilde{\kappa}_k^{(1)} \geqslant 0) - \frac{L_0^2}{2\lambda^{(1)}}(\log n)^2 E_{j_n k}(|Z_n|).$$

The function ψ^2 satisfies the kernel condition of assumption A2. It follows from the choice of j_n that

$$(\log n)^2 E_{j_n k}(|Z_n|) \leqslant L_1 (\log n)^2 n^{-\Gamma}$$

for some constant L_1 and is asymptotically negligible. Therefore, it is enough to prove that $P_{j_n k}(\tilde{\kappa}_k^{(1)} \geqslant 0) \geqslant z^{(1)}$, or, equivalently

$$P_{j_n k}\left\{\frac{1}{\sqrt{n}}\sum_{i=0}^{n} \psi_{j_n k}(X_i)\varepsilon_{i+1} \geqslant 0\right\} \geqslant z^{(1)} > 0 \tag{5.17}$$

for some $z^{(1)}$ which does not depend on n. For this, we use the following elementary lemma (proof of which we omit).

Lemma 4. *Let U_n be a sequence of random variables such that $E(U_n) = 0$, $E(U_n)^2 \geqslant z_0 > 0$ for large enough n and such that U_n^2 is uniformly integrable. Then, there exists $z_1 > 0$ such that for large enough n*

$$P(U_n \geqslant 0) \geqslant z_1 > 0. \tag{5.18}$$

We turn to (5.18). We apply Lemma 4 to $U_n = 1/\sqrt{n}\sum_{i=0}^{n} \psi_{j_n k}(X_i)\varepsilon_{i+1}$. Clearly, $E_{j_n k}(U_n) = 0$. To prove that $E_{j_n k}(U_n^2)$ is bounded from below, we use assumption A2 and the fact that $\inf_{x\in D}\mu(x) \geqslant v > 0$. Likewise, one easily checks the uniform integrability of U_n^2. This completes the case of the mean function m.

The variance function σ^2. We abbreviate $P_{m_0,\sigma_{n^k}}^2$ by $P_{j_n k}$. Routine computation yields, under $P_{j_n k}$

$$\Lambda_2(\sigma_0^2, \sigma_{j_n k}^2, X^{(n)}) = \prod_{i=0}^{n} \sqrt{1 + \gamma_n \psi_{j_n k}(X_i)}\exp - \frac{1}{2}\sum_{i=0}^{n} \gamma_n \psi_{j_n k}(X_i)\varepsilon_{i+1}.$$

Taking the logarithm and using a Taylor expansion, we derive

$$\log \Lambda_2(\sigma_0^2, \sigma_{j_n k}^2, X^{(n)}) = \frac{\gamma_n}{2}\sum_{i=0}^{n} \psi_{j_n k}(X_i)(1 - \varepsilon_{i+1}^2) - \frac{\gamma_n^2}{4}\sum_{i=0}^{n} \psi_{j_n k}^2(X_i) + R_n,$$

where R_n is a remainder term. We procede analogously as for the mean function and we obtain the same conclusion. We omit the details of the computations, which are similar to the case of m. The proof of Lemma 2 is complete. □

5.2. Proof of Theorems 2 and 3

Theorems 2 and 3 follow from moment bounds and moderate deviation inequalities. Let $g \in L_m(\mathbb{R})$ be continuous, bounded, compactly supported, and such that $\int g^2(x)\,dx = 1$. For $f = m$ or σ^2, define

$$\gamma_{jk} = \int f(x) g_{jk}(x)\,dx \quad \text{and} \quad \hat{\gamma}_{jk} = \frac{1}{\lfloor n\bar{v} \rfloor} \sum_{i=0}^{\lfloor n\bar{v} \rfloor} g_{jk}(x_{T_i}) Y_{T_i}.$$

The Y_{T_i} are the transformations on the observation scheme $(X_t, t = 1, \ldots, n)$, following the regression approximation defined by (1.3) for m and (1.4) for σ^2.

Lemma 5 (moment bounds). *Let $2^j \leqslant n$. Under the assumptions of Theorem 2, for all $r \geqslant 2$, there exists an explicitly computable $K_4(r)$ such that*

$$E_{x_0}\{|\hat{\gamma}_{jk} - \gamma_{jk}|^r\} \leqslant K_4(r) n^{-r/2}. \tag{5.19}$$

Lemma 6 (moderate deviations). *Let $2^j \leqslant n$. Under the assumptions of Theorem 2, for all $r \geqslant 2$, there exist explicitly computable $K_5(r)$ and $K_0(r)$ such that*

$$P_{x_0}\{|\hat{\beta}_{jk} - \beta_{jk}| \geqslant K_0(r)\sqrt{j/n}\} \leqslant K_5(r) 2^{-jr}. \tag{5.20}$$

Proof of Theorems 2 and 3. Using Lemmas 5 and 6, we readily follow the proof of Theorems 3 and 4 in DJKP (1996).

5.3. Proof of Lemma 5

We will follow the method we previously used in Hoffmann (1999). We recall however all the technical steps in order to give a self-containing exposition. Without loss of generality, we assume that $D = [0, 1]$. We will denote by C a generic constant, possibly varying from line to line. We will not distinguish between m and σ^2, taking advantage of the general framework defined in Section 3.1, with

$$f = m, \quad \eta_i = 0, \quad \xi_i = \sigma(X_i)\varepsilon_{i+1}$$

or

$$f = \sigma^2, \quad \eta_i = \{m(X_i) - \hat{m}_i(X_i)\}^2, \quad \xi_i = 2\{m(X_i) - \hat{m}_i(X_i)\}\sigma(X_i)\varepsilon_{i+1} + \sigma^2(X_i)(\varepsilon_{i+1}^2 - 1).$$

We use the following decomposition:

$$\hat{\gamma}_{jk} - \gamma_{jk} = Q_1 + Q_2 + Q_3$$

with

$$Q_1 = \frac{1}{\lfloor n\bar{v} \rfloor} \sum_{i=j}^{\lfloor n\bar{v} \rfloor} f(X_{T_i}) g_{jk}(x_{T_i}) - \int f(x) g_{jk}(x)\,dx,$$

$$Q_2 = \frac{1}{\lfloor n\bar{v} \rfloor} \sum_{i=j}^{\lfloor n\bar{v} \rfloor} \eta_{T_i} g_{jk}(x_{T_i}),$$

$$Q_3 = \frac{1}{\lfloor n\bar{v} \rfloor} \sum_{i=j}^{\lfloor n\bar{v} \rfloor} \xi_{T_i} g_{jk}(x_{T_i}).$$

Clearly, it is enough to prove moment bounds for each term Q_i, $i=1,2,3$. Let us first study Q_1. For $u>0$, we introduce the following *penalty event*:

$$\mathscr{A}_{jk}=\prod_{\lambda\in C_{jk}} 1_{\mu_n(x_\lambda)\geqslant u},$$

where $C_{jk}=\{\lambda:\ C_\lambda\cap[k2^{-j}),(k+1)2^{-j}\neq\}$ and x_λ is the midpoint of the interval C_λ. Recall that the empirical sampling measure defined in Assumption A2 depends on a kernel K. In the following, we will choose K such that

$$K(x)=\tilde{K}(x)(1/\tilde{K}(v)\,\mathrm{d}v)\quad\text{with}\quad 0\leqslant\tilde{K}(x)\leqslant 1_{[-1/2,1/2]}(x)\quad\text{and}\quad\int\tilde{K}(x)\,\mathrm{d}x,$$

a choice which is obviously possible. First, remark that $|Q_1|\leqslant C\sup_{x\in D}|f(x)g(x)|2^{-j/2}$. Therefore,

$$E_{x_0}(|Q_1|^r)\leqslant C[E_{x_0}(|Q_1|^r,\mathscr{A}_{jk})+C2^{-jr/2}P_{x_0}(\mathscr{A}_{jk}^c)]. \tag{5.21}$$

Next,

$$P_{x_0}(\mathscr{A}_{jk}^c)\leqslant\sum_{\lambda\in C_{jk}}P_{x_0}(\mu_n(x\lambda)<u)\leqslant Ch_n^{-1}2^{-j}\sup_{x\in D}P_{x_0}(\mu_n(x)<u).$$

The choice of $u=v/(2\int\tilde{K})$ together with the properties of K entails

$$E_{x_0}(|Q_1|^r)\leqslant C\left[E_{x_0}(|Q_1|^r,\mathscr{A}_{jk})+Ch_n^{-1}2^{-j}\sup_{x\in D}P_{x_0}\left(|\mu_n(x)-\mu(x)|>v\left(1-\frac{1}{2\int\tilde{K}}\right)\right)\right].$$

The choice of $\tilde{K}$ ensures that $1-1/(2\int\tilde{K})>0$. By Assumption A2 and Chebyshev's inequality, the last term in the previous inequality is arbitrary small in power of n hence asymptotically negligible. In order to bound the first term in the right-hand side, we use the following decomposition:

$$Q_1=Q_{11}+Q_{12}$$

with

$$Q_{11}=\frac{1}{\lfloor n\bar{v}\rfloor}\sum_{i=j}^{\lfloor n\bar{v}\rfloor}[f(X_{T_i})-f(x_{T_i}]g_{jk}(x_{T_i}) \tag{5.22}$$

and

$$Q_{12}=\frac{1}{\lfloor n\bar{v}\rfloor}\sum_{i=j}^{\lfloor n\bar{v}\rfloor}f(x_{T_i})g_{jk}(x_{T_i})-\int f(x)g_{jk}(x)\,\mathrm{d}x. \tag{5.23}$$

For $u=v/(2\int\tilde{K})$, we have the following inclusion:

$$(\mu_n(x)\geqslant u)=\left(\frac{1}{n+1}\sum_{i=0}^{n}\tilde{K}(h_n^{-1}(X_i-x))\geqslant v/2\right)\leqslant\left(\sum_{i=0}^{n}1_{|X_i-x|\leqslant h_n}\geqslant\lfloor(n+1)h_n\bar{v}\rfloor\right).$$

The condition $s_i>1+1/p_i$ implies that f is Lipschitz continuous (by classical Sobolev embeddings in Besov spaces), therefore

$$|f(X_{T_i})-f(x_{T_i})|\leqslant C|X_{T_i}-x_{T_i}|. \tag{5.24}$$

Moreover, the T_i are all distinct on the event $\mathscr{A}_{jk}$. It follows from the construction of the T_i and the x_{T_i} that

$$|X_{T_i}-x_{T_i}|\leqslant Ch_n. \tag{5.25}$$

Likewise, on $\mathscr{A}_{jk}$

$$Q_{12} = \frac{1}{\lfloor n\bar{v}\rfloor}\sum_{i=j}^{\lfloor n\bar{v}\rfloor} f(i/\lfloor n\bar{v}\rfloor)g_{jk}(i/\lfloor n\bar{v}\rfloor) - \int f(x)g_{jk}(x)\,\mathrm{d}x.$$

Using again that the number of term involved in the sums in (5.22) is $\mathcal{O}(n2^{-j})$ and using (5.24) and(5.25), we derive

$$E_{x_0}(|Q_{11}|^r, \mathscr{A}_{jk}) \leqslant C2^{-jr/2}h_n^r \tag{5.26}$$

and for Q_{12}, by Riemann's approximation

$$E_{x_0}(|Q_{12}|^r, \mathscr{A}_{jk}) \leqslant C(2^{-j/2}/n)^r. \tag{5.27}$$

From the choice of h_n and 2^j, these terms have the right order. We now turn to the term Q_2, which only appears in the case of the estimation of the variance function. We will use the following bound, for $1 \leqslant i \leqslant n$

$$\sup_{x\in D}|m(x) - \hat{m}_i(x)|^2 \leqslant \tilde{C}/\sqrt{i}, \quad \text{a.s.} \tag{5.28}$$

which is valid since m is Lipschitz continuous, for a random constant $\tilde{C}$ such that $E_{x_0}(|\tilde{C}|^r)$ is finite for all $r \geqslant 1$. This bound is a classical result for Nadaraya–Watson estimators in nonparametric regression with random design. The extension for estimating the mean function in a AR(1) model in our context is straightforward. In fact, the rate in (5.28) can be improved, but is sufficient for our purpose. Using successively that $T_i \geqslant i$, inequality (5.28) and the same arguments on j as for Q_1, we obtain

$$E_{x_0}(|Q_2|^r) \leqslant Cn^{-r/2}. \tag{5.29}$$

To bound Q_3, we use Rosenthal inequality for martingales (Hall and Heyde, 1980, p. 23). Indeed, the process $(\sum_{j\leqslant i} g_{jk}(x_{T_i})\xi_{T_i},\ i = 1,\ldots,\lfloor n\bar{v}\rfloor)$ is a $(\mathscr{F}^n_{T_i})$-martingale. By Assumption A3, straightforward computation shows that $E_{x_0}(|\xi_{T_i}|^r)$ is bounded, therefore

$$E_{x_0}(|Q_3|^r) \leqslant Cn^{-r/2} \tag{5.30}$$

and the conclusion follows. The proof of Lemma 5 is complete.

5.4. Proof of Lemma 6

We use the decomposition

$$\hat{\beta}_{jk} - \beta_{jk} = Q_1 + Q_2 + Q_3$$

as in Lemma 5, replacing g by ψ. Clearly,

$$P_{x_0}(|\hat{\beta}_{jk} - \beta_{jk}| > K_0\sqrt{j/n}) \leqslant \sum_{i=2}^{3} P_{x_0}\left(|Q_i| > \frac{K_0}{2}(1 - |\bar{Q}_1|)\sqrt{j/n}\right),$$

where $\bar{Q}_1 = (K_0\sqrt{j_n})^{-1}Q_1$. Recall that on the event $(T_{\lfloor\tilde{n}v\rfloor} < n-1)$, we have

$$|Q_1| \leqslant C2^{-j/2}(h_n + n^{-1}).$$

Hence, using the same penalty argument as for the moment bounds, we derive

$$P_{x_0}(|\hat{\beta}_{jk} - \beta_{jk}| > K_0\sqrt{j/n}) \leqslant C\sum_{i=2}^{3} P_{x_0}\left(|Q_i| > \frac{K_0}{4}\sqrt{j/n}\right), \tag{5.31}$$

plus a negligible term coming from the probability of the event $\mathscr{A}_{jk}^c$, which we can insert in the constant C. For the first term in the right-hand side of (5.31), using Chebyshev's inequality and similar arguments as for Lemma 5 yield

$$P_{x_0}\left(|Q_2| > \frac{K_0}{4}\sqrt{j/n}\right) \leqslant C2^{-jr/2}. \tag{5.32}$$

For Q_3, we use a Bernstein-type inequality for unbounded martingales.

Lemma 7. *Let* $(S_n = \sum_{i=0}^n d_i, n \geqslant 0)$ *be a* $(\mathscr{F}_n)$*-martingale such that*

$$\forall r \geqslant 1: \quad E(|d_i|^r|\mathscr{F}_{i-1}) \leqslant C_0 C_1^r r^r$$

for two constants C_0 *and* C_1*. Then*

$$\forall t \geqslant 0: \quad P(|S_n| \geqslant t) \leqslant 2\exp\left[-\frac{t^2}{2(\bar{C}_0 C_1^2 n + \bar{C}_1 t)}\right]$$

with $\bar{C}_0 = 4eC_0$ *and* $\bar{C}_1 = 2eC_1$.

The proof is obtained in the same lines as the classical Bernstein inequality, so we omit it. By Assumption A3, the term ξ_{T_i} satisfies the moment condition of Lemma 7 in both cases (m and σ^2). The verification is straightforward. We then apply Lemma 7 to the $(\mathscr{F}_{T_i}^n)$-martingale

$$Q_3(k) = \frac{1}{\lfloor n\bar{v}\rfloor}\sum_{i=0}^{k} d_i$$

with

$$d_i = \psi_{jk}(x_{T_i})\xi_{T_i}.$$

Hence, for any $t \geqslant 0$

$$P_{x_0}(|Q_3| \geqslant t) \leqslant \exp\left(-\frac{nt^2}{C(1+2^{j/2}t)}\right).$$

Since $j2^j \leqslant n$ for $j_0(n) \leqslant j \leqslant 2^{j_1(n)}$, the choice of the threshold $K_0\sqrt{j/n}$ entails, for sufficiently large $K_0 = K_0(r)$

$$P_{x_0}(|Q_3| \geqslant K_0(r)\sqrt{j/n}) \leqslant 2^{-jr}. \tag{5.33}$$

Putting together (5.32) and (5.33), we obtain (5.29). The proof of Lemma 6 is complete. □

Acknowledgements

I gratefully thank M. Neumann, D. Picard and A. Tsybakov for helpful discussion and comments. The careful remarks and advices of a referee were valuable to improve a former version of this manuscript.

References

Barraud, Y., Comte, F., Viennet, G., 1997. Adaptive estimation in an autoregression and geometrical beta mixing frmaework. Preprint 98-07, CREST-INSEE.

Bretagnolle, J., Huber, C., 1979. Estimation des densités: risques minimax. Z. Wahrscheinlichkeitstheorie verw. Gebiete 47, 119–137.

Cohen, A., Daubechies, I., Vial, P., 1994. Wavelets on the interval and fast wavelet transform. Appl. Comput. Harm. Anal. 1, 54–81.

Dahlhaus, R., Neumann, M., von Sachs, R., 1995. Nonlinear wavelet estimation of time varying autoregressive processes. WIAS, preprint 150.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D., 1995. Wavelet shrinkage: asymptopia? J. Roy. Statist. Soc. Ser. B 57, 301–369. With discussion.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D., 1996. Density estimation by wavelet thresholding. Ann. Statist. 24, 508–539.

Doukhan, P., 1995. Mixing: Properties and Examples. Lecture Notes in Statistics, vol. 85. Springer, New York.

Duflo, M., 1990. Méthodes récursives aléatoires. Masson, Paris.

Efromovitch, S.Y., 1985. Nonparametric estimation of a density of unknown smoothness. Theory Probab. Appl. 30, 557–661.

Fuk, D.K., Nagaev, S.V., 1971. Probability inequalities for sums of independent random variables. Theory Probab. Appl. 4, 643–660.

Hall, P., Caroll, R.J., 1989. Variance function estimation in regression: the effect of estimating the mean. J. Roy. Statist. Soc. Ser. B 51, 3–14.

Hall, P., Heyde, C.C., 1980. Martingale Limit Theory and its Applications. Academic Press, New York.

Härdle, W., Tsybakov, A.B., 1995. Local polynomial estimators of the volatility function in nonparametric autoregression. Discussion paper 42, Humboldt University, Berlin.

Hoffmann, M., 1999. Adaptive estimation in diffusion processes. Stochastic Process. Appl. 79, 135–163.

Korostelev, A.P., Tsybakov, A.D., 1993. Minimax Theory of Image Reconstruction. Lecture Notes in Statistics, vol. 82. Springer, Berlin.

Lepski, O.V., 1990. One problem of adaptive estimation in Gaussian white noise. Theory Probab. Appl. 35, 459–470.

Masry, E., Tjøstheim, D., 1995. Nonparametric estimation and identification of nonlinear ARCH time series. Econom. Theory 11, 258–289.

Meyer, Y., 1990. Ondelettes et Opérateurs I. Hermann, Paris.

Neumann, M., 1996. Strong aproximation of density estimators from weakly dependent observations by density estimators from independent observations. WIAS, preprint 295.

Neumann, M., Spokoiny, V., 1995. On the efficiency of wavelets estimators under arbitrary errors distributions. Math. Meth. Statist. 4, 137–166.

Neumann, M., Kreiss, J.P., 1996. Bootstrap confidence bands for the autoregression function. WIAS, preprint 263.

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

Tjøstheim, D., 1994. Nonlinear time series: a selective review. Scand. J. Statist. 97–130.