I N S T I T U T D E
S T A T I S T I Q U E

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

0225

PROPERTIES OF DESIGN-ADAPTED WAVELETS TRANSFORMS OF NONLINEAR AUTOREGRESSION MODELS

V. DELOUILLE and R. von SACHS

http://www.stat.ucl.ac.be

Properties of Design-adapted Wavelet Transforms of Nonlinear Autoregression Models

V. Delouille[1,3] R. von Sachs[2,3]

August 27, 2002

Abstract

We estimate nonlinear autoregressive models using a design-adapted wavelet estimator. We show two properties of the wavelet transform adapted to an autoregressive design. First, in an asymptotic setup, we derive the order of the threshold that removes all the noise with a probability tending to one asymptotically. Second, with this threshold, we estimate the detail coefficients by soft-thresholding the empirical detail coefficients. We show an upper bound on the l_2-risk of these soft-thresholded detail coefficients. Finally, we illustrate the behavior of this design-adapted wavelet estimator on simulated and real data sets.

Keywords: autoregressive design, $\beta-$mixing conditions, biorthogonal wavelet transform, l_2-risk of the wavelet coefficients.

1 Introduction

Autoregressive models form an important class of processes in time series. A nonparametric version of these models was first introduced by Jones [27]. In this paper, we consider the nonlinear autoregression model

$$X_t = m(X_{t-1}) + \epsilon_t,\ t = 1, \ldots, n\ . \tag{1.1}$$

For theoretical purposes, the innovations ϵ_t are supposed to be i.i.d. with $E(\epsilon_t) = 0; E(\epsilon_t^2) = \sigma_\epsilon^2$, but are not necessarily Gaussian. In practice, an algorithm that allows for heteroscedasticity is proposed. Several authors dealt with the problem of estimating the autoregression function m nonparametrically [24, 28, 32, 35]. Very little is known however about *wavelet* estimators for autoregressive designs. The only existing method [25] that treats autoregressive models using a wavelet estimator does not allow for an efficient implementation.

The aim of this paper is to investigate some properties of the design-adapted wavelet estimator [10, 11] in the case of the autoregression model (1.1). In [10], we investigated the properties of the so-called design-adapted wavelet estimator based on the unbalanced Haar

[1]Research Assistant of the National Fund for Scientific Research – Wallonie, Belgium (F.N.R.S.). Université catholique de Louvain, Institut de Statistique, Louvain-la-Neuve, Belgium

[2]Professor. Université catholique de Louvain, Institut de Statistique, Louvain-la-Neuve, Belgium

[3]Financial support from the contract ‘Projet d’Actions de Recherche Concertées’ nr. 98/03–217 from the Belgian government, and from the IAP research network nr. P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs) is gratefully acknowledged.

basis in the stochastic regression context. Results on the near-optimal rate of convergence of the minimax L_2-risk were derived for β-Hölder functions with regularity parameter $1/2 < \beta < 1$. The Haar basis provides only piecewise-constant estimators, hence in a second step it must be improved. In [11], a mean to improve the Haar basis is proposed. Indeed, within the framework of the lifting scheme [6, 34], it is possible to obtain a wavelet basis with a higher regularity. When using the lifting scheme, orthogonality of the Haar basis must be relaxed to biorthogonality.

In the present paper, we show two properties of the design-adapted wavelet transform applied on autoregressive designs.

1. We derive a thresholding scheme that allows to remove with a probability tending to one all the noise in the wavelet coefficients. This first result parallels what Donoho [12] obtained in the classical setting of fixed, equispaced design.

2. With the threshold obtained in the first result, we estimate the detail coefficients by soft-thresholding the empirical coefficients. We then show an upper bound on the l_2-risk of these detail coefficients, that is, the risk of the estimator in the wavelet domain. This upper bound can be related to a similar upper bound in the case of a nonparametric regression with Gaussian noise.

To prove these two results, we make the assumption that the time series $\{X_t : t \geq 0\}$ is a strictly stationary time-homogeneous Markov chain having a common cumulative distribution function denoted by F_x. Moreover, this series $\{X_t, t \geq 0\}$ is assumed to be geometrically $\beta-$mixing.

This paper is organized as follows. After having briefly explained the idea of design-adapted wavelets, we give in Section 3 the assumptions on the model and the methodology to prove our main results. These are presented in Section 4. Section 5 proposes a design-adapted wavelet algorithm that accounts for the large variability and the heteroscedasticity inherent to financial time series data. We show on simulated and real data sets the usefulness of our proposed estimator for the estimation of nonlinear autoregressive designs. Finally, the proofs of the theoretical results are deferred to Section 6.

2 Design-adapted wavelet transform

The first element needed to build design-adapted wavelets is a partitioning [18] of the interval we are working on. Let F_x denote the common cumulative distribution function of the time series $\{X_t\}$. Design-adapted wavelets are wavelets which are (bi-)orthogonal in the space $L_2(d\hat{F}_n)$, that is, in the space weighted by the empirical quantile measure. Hence an adequate partitioning is given by

$$I_{jk} = [X_{((k-1)l_j+1)},\ X_{(kl_j+1)}),\ k = 1, \ldots, 2^j\ ,\ l_j = n2^{-j}, j_0 \leq j \leq J = \log_2(n)\ .$$

In this paper, we consider the Haar basis in $\mathbf{L}_2(d\hat{F}_n)$ [10] and a smoother design-adapted wavelet basis built through an average-interpolating subdivision scheme [11].

The Haar basis of [10] is orthogonal in $\mathbf{L}_2(d\hat{F}_n)$, hence the analyzing wavelets $\tilde{\psi}_{jk}$ and the synthesis wavelets ψ_{jk} are both equal to

$$\tilde{\psi}_{jk} = \psi_{jk} = 2^{j/2}\left(1_{j+1,2k+1} - 1_{j+1,2k}\right)\ , \tag{2.1}$$

where $1_{j,k}$ denotes the indicator function on the random interval I_{jk}.

If we want to improve this basis with the lifting scheme [34], we proceed as follows. An average-interpolating prediction step is performed. This increases the number of vanishing moments of the analyzing wavelet $\tilde{\psi}_{jk}$. It is then also possible to add an update step, which performs a local semi-orthogonalisation [11]. However, to prove the theoretical results, we consider only the basis built from the Haar basis followed by an average-interpolating prediction step. In this case, the dual (analyzing) wavelets after the prediction step are still piecewise constant:

$$\tilde{\psi}_{jk} = 2^{(j+1)/2} \sum_{l:\tilde{g}_{jlk}\neq 0} \tilde{g}_{jlk} 1_{j+1,l} \ , \tag{2.2}$$

where the values $\tilde{g}_{jlk}$ are the refinements coefficients, chosen in such a way that the analyzing wavelet $\tilde{\psi}_{jk}$ has $\tilde{N}$ vanishing moments in $\mathbf{L}_2(d\hat{F}_n)$. Having a piecewise constant function $\tilde{\psi}_{jk}$ allows us to easily generalize the properties of the Haar wavelets shown in this paper, cfr. Remark 4.3.

Consider the autoregressive model (1.1). When performing a forward wavelet transform on this model using design-adapted wavelets, we obtain the empirical wavelet coefficients $\hat{d}_{jk}$:

$$\begin{aligned} \hat{d}_{j,k} &= d_{jk} + \rho_{jk} \\ &= \frac{1}{n}\sum_{t=1}^{n} m(X_{t-1})\tilde{\psi}_{jk}(X_{t-1}) + \frac{1}{n}\sum_{t=1}^{n} \epsilon_t \tilde{\psi}_{jk}(X_{t-1}) \ . \end{aligned} \tag{2.3}$$

In this decomposition, we consider that d_{jk} is the theoretical wavelet coefficient and that ρ_{jk} is the noise term. Indeed, for thresholding we do not need to consider the variance of d_{jk}, which is only due to the stochastic nature of the design.

Similarly to [31, 10], in the multiscale decomposition, we only take into account the wavelet coefficients of indices (j, k) belonging to a certain set $\mathcal{J}_n$, where

$$\mathcal{J}_n = \{(j,k)|2^j \leq C\, n^{1-\alpha} \ ; 1 \leq k \leq 2^j\} \text{ with } 0 < \alpha < 1. \tag{2.4}$$

This definition of the set $\mathcal{J}_n$ involves that the wavelet decomposition begins at a lower scale than the scale given by the data. This is equivalent to performing first a linear wavelet denoising of the data, that removes the detail coefficients which belong to a scale j° not included in $\mathcal{J}_n$.

We now present in more detail the assumptions of our model and the main tool used to prove the results.

3 Assumptions and methodology

3.1 Assumptions on the model

In the autoregressive model $X_t = m(X_{t-1})+\epsilon_t$, the innovations ϵ_t are supposed to be i.i.d. with mean zero and finite variance, $E(\epsilon_t) = 0,\ \mathrm{Var}(\epsilon_t) = \sigma_\epsilon^2,\ \forall t$.

In order to generalize nonparametric regression results for dependent data, it is necessary to define some 'degrees' of dependence between the observations. The mathematical concept is that of *mixing conditions*, which can be interpreted as the speed at which the series becomes asymptotically independent. More precisely, the $\beta-$mixing condition used in this paper is defined as follows.

Definition 3.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, with $\mathcal{G}$ and $\mathcal{H}$ two $\sigma-$subfields of $\mathcal{F}$. The $\beta-$mixing coefficients for the process $\{X_t\}$ are defined as follows [4, 9]:*

$$\beta(\mathcal{G}, \mathcal{H}) = E \sup_{H \in \mathcal{H}} |P(H) - P(H|\mathcal{G})|$$

For a sequence $\{X_t\}_{-\infty}^{+\infty}$, let $\mathcal{F}_{-\infty}^{t} := \sigma(\ldots, X_{t-2}, X_{t-1}, X_t)$ be the $\sigma-$field generated by $(\ldots, X_{t-2}, X_{t-1}, X_t)$ and similarly define $\mathcal{F}_{t+m}^{+\infty} := \sigma(X_{t+m}, X_{t+m+1}, \ldots)$. The sequence $\{X_t\}_{-\infty}^{+\infty}$ is said to be *absolutely regular* or *$\beta-$mixing* if $\lim_{m\to\infty} \beta_m = 0$ where

$$\beta_m := \sup_t \beta(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{+\infty})$$

We now make the following assumptions on the time series $\{X_t\}_{t=0}^{n}$ of the model (1.1).

(A1) $\{X_t,\ t \geq 0\}$ is a strictly stationary time-homogeneous Markov chain. We denote by F_x the common cumulative distribution function of the X_t, which is assumed to be continuous. Furthermore, we assume absolute regularity (that is, $\beta-$mixing) for $\{X_t\}$ and that the $\beta-$mixing coefficients decay at a geometric rate, that is $\beta_m = O(a^{-m})$ with $a > 1$ a constant.

(A2) For all $M < \infty$, there exists finite constants C_M such that $E|\epsilon_t|^M \leq C_M$.

The assumption (A1) ensures the recurrence in the autoregression model. In [30, 17], some conditions on the regression function $m(x)$ are given in order to generate a time series $\{X_t\}$ that fulfills Assumption (A1).

Here and in the following, we make the convention that δ denotes a positive but arbitrarily small, and λ an arbitrarily large constant. With this convention, assumption (A2) tells us that there exist $\delta > 0,\ \lambda > 0$ such that

$$P\left(\sup_{t=1,\ldots,n} |\epsilon_t| > c_\lambda n^{\delta'}\right) \leq cn^{-\lambda} \tag{3.1}$$

holds for any $\delta' \in (0, \delta)$, where c_λ depends on λ and c is a generic constant. In other words, we can define a probability space Ω_n such that

$$P(\Omega_n) = P(\omega : \sup_t |\epsilon_t(\omega)| \leq n^{\delta'}) \geq 1 - O(n^{-\lambda}) \ \text{ for } \delta' \in (0, \delta)\ . \tag{3.2}$$

In order to prove the results, we assume that we are working on this probability space Ω_n. Thus, for a given λ, it is always possible to find δ such that all the $|\epsilon_t|$ are bounded with probability $1 - O(n^{-\lambda})$ by $n^{\delta'}$ for $\delta' \in (0, \delta)$.

3.2 Main tool for proving the theoretical results

The theoretical difficulty that arises when dealing with design-adapted wavelets is that the intervals used to define the wavelet functions are random. To circumvent this problem, we approximate a given random interval $\mathcal{I} := I_{jk}$ by a union of fixed intervals coming from the theoretical quantile partitioning $\mathbf{I}^\circ = \{I_{jk}^\circ\}_{(j,k)}$, defined by

$$I_{jk}^\circ = \left[F_x^{-1}\left(\frac{k-1}{2^j}\right),\ F_x^{-1}\left(\frac{k}{2^j}\right)\right),\ 0 < j < j^*,\ 1 \leq k \leq 2^j,\ 2^{j^*} \leq n < 2^{j^*+1}. \tag{3.3}$$

Note that the set $\mathbf{I}^\circ$ contains intervals with indices at a finer scale than the indices contained in the index set $\mathcal{J}_n$.

Following the proof of Corollary 2.1 in [30], we treat the interval $\mathcal{I}$ as follows. Consider the fixed partitioning $\mathbf{I}^\circ$. Beginning from the coarsest scale, we approximate $\mathcal{I}$ by a union of as large as possible intervals belonging to $\mathbf{I}^\circ$. There exist indices $(j_1, k_1), \ldots, (j_d, k_d)$ such that

$$I^\circ_{j_1,k_1} \cup \ldots \cup I^\circ_{j_d,k_d} \subseteq \mathcal{I} \tag{3.4}$$

where $j_1 < \ldots < j_{d-1} \leq j_d \leq j^*$.

The number d of terms needed in the above approximation is a function of the random interval $\mathcal{I}$. Moreover, as n tends to infinity, the number of intervals needed will grow. To emphasize these facts, in the sequel we use the convention that $d = d(\mathcal{I}) = d_n$.

Now, if we add two suitable intervals from the finest scale, $I^\circ_{j^*,l_1}$ and $I^\circ_{j^*,l_2}$, say, then we can approximate $\mathcal{I}$ from above:

$$\mathcal{I} \subseteq I^\circ_{j_1,k_1} \cup \ldots \cup I^\circ_{j_d,k_d} \cup I^\circ_{j^*,l_1} \cup I^\circ_{j^*,l_2} . \tag{3.5}$$

Note that the fine scale indices (j^*, l) do not belong to $\mathcal{J}_n$. What we want is to recover a random interval $\mathcal{I} = I_{jk}$ for $(j,k) \in \mathcal{J}_n$ using the elements in the partitioning $\mathbf{I}^\circ$. Hence we need intervals in $\mathbf{I}^\circ$ that are finer than the finest interval that is going to be approximated.

This approximation with fixed intervals will be used to evaluate the noise term ρ_{jk} in the wavelet decomposition (2.3). More precisely, the indicator functions on the random intervals present in the definition of the Haar wavelet (2.1) will be approximated by sums of indicator functions on fixed intervals. Then, usual results on sums of random variables can be applied.

In our first result, we find the order of the threshold t_{jk} such that ρ_{jk} does not exceed t_{jk} with a high probability. We formalize this fact by introducing the following notation [30].

Definition 3.2. *Let Z_n be a sequence of random variables and let $\{\alpha_n\}$ and $\{\beta_n\}$ be sequence of positive reals. We write*

$$Z_n = \tilde{O}(\alpha_n, \beta_n)$$

if

$$P(|Z_n| > C_1\alpha_n) \leq C_2\beta_n,$$

holds for $n \geq 1$ and some finite constants C_1, C_2.

Obviously, this definition is stronger than the usual O_P. We are now able to present the two main results of this paper.

4 Main theoretical results

4.1 Threshold that removes the noise with a high probability

Here, we are concerned with the order of the threshold that removes the noise with a high probability. The obtained threshold will be conservative. However, it will provide us with a benchmark as to which order of the threshold to use in case of an autoregressive design, as compared to the regression setting.

Consider first one detail coefficient $\hat{d}_{jk}$ and its corresponding noise term ρ_{jk}. In case of the Haar transform, we have:

$$\rho_{jk} = \frac{2^{j/2}}{n} \left(\sum_{t=1}^{n} \epsilon_t 1_{j+1,2k+1}(X_{t-1}) - \sum_{t=1}^{n} \epsilon_t 1_{j+1,2k}(X_{t-1}) \right) . \tag{4.1}$$

Our present aim is to find the smallest quantities $a^{(\lambda)}_{j+1,2k}$ and $a^{(\lambda)}_{j+1,2k+1}$, both depending on a constant λ and such that

$$\begin{aligned} P(\frac{2^{j/2}}{n}\left|\sum_{t=1}^{n}\epsilon_t 1_{j+1,2k}(X_{t-1})\right| &> a^{(\lambda)}_{j+1,2k}) \leq cn^{-\lambda} \\ P(\frac{2^{j/2}}{n}\left|\sum_{t=1}^{n}\epsilon_t 1_{j+1,2k+1}(X_{t-1})\right| &> a^{(\lambda)}_{j+1,2k+1}) \leq cn^{-\lambda} \end{aligned}$$

with λ arbitrarily large, and where c represents, here and in the sequel, a generic constant. Then with

$$t_{jk} := a^{(\lambda)}_{j+1,2k} + a^{(\lambda)}_{j+1,2k+1} \ , \tag{4.2}$$

we obtain

$$\begin{aligned} P(|\rho_{jk}| > t_{jk}) &\leq P(\frac{2^{j/2}}{n}\left|\sum_{t=1}^{n}\epsilon_t 1_{j+1,2k}(X_{t-1})\right| + \frac{2^{j/2}}{n}\left|\sum_{t=1}^{n}\epsilon_t 1_{j+1,2k+1}(X_{t-1})\right| > t_{jk}) \\ &\leq P\left(\frac{2^{j/2}}{n}\left|\sum_{t=1}^{n}\epsilon_t 1_{j+1,2k}(X_{t-1})\right| > a^{(\lambda)}_{j+1,2k}\right) \\ &+ P\left(\frac{2^{j/2}}{n}\left|\sum_{t=1}^{n}\epsilon_t 1_{j+1,2k+1}(X_{t-1})\right| > a^{(\lambda)}_{j+1,2k+1}\right) \\ &\leq cn^{-\lambda} \ . \end{aligned} \tag{4.3}$$

Repeating this operation for all $(j,k) \in \mathcal{J}_n$ and defining

$$t(n) := \max_{(j,k)\in\mathcal{J}_n} t_{jk} \ , \tag{4.4}$$

we have, using the Bonferroni inequality, that

$$\begin{aligned} P(\max_{jk}|\rho_{jk}| > t(n)) &= P(\cup_{jk}(|\rho_{jk}| > t(n))) \\ &\leq \sum_{(j,k)\in\mathcal{J}_n} P(|\rho_{jk}| > t(n)) = O(n^{1-\alpha-\lambda}) \end{aligned}$$

with $0 < \alpha < 1$ and λ arbitrarily large.

To find an adequate threshold t_{jk} for the average-interpolating wavelet basis, we proceed in a completely similar manner. The only difference is that there will be (approximately) $2\tilde{N}$ terms in (4.2) instead of only two with the Haar basis. The *order* of the final threshold however will still be the same, the difference with the Haar case will only show up in the constant factor. Hence, for sake of simplicity, we consider only the Haar wavelet transform to present the two main results of this paper.

To prove the following theorem, the set $\mathcal{J}_n$ must be such that

$$\mathcal{J}_n = \{(j,k) | 2^j \leq C\, n^{1-\alpha} \ ; 1 \leq k \leq 2^j\} \text{ with } 4\delta \leq \alpha < 1 \ , \tag{4.5}$$

where δ is defined in (3.1).

In the proof of the theorems, we need to relate the scale j of a random interval $\mathcal{I} \equiv I_{jk}$ to the scale j_l of the intervals $I^{\circ}_{j_l,k_l}$ which are included in I_{jk} and which are used in the approximation (3.4). More precisely, the ratio $2^j/2^{j_l}$ must be bounded away from zero and infinity, independently of n. We formalize this fact with the following assumption, made for each $(j,k) \in \mathcal{J}_n$.

(A3) For a given $\lambda > 0$, there exists a constant $c_\lambda < \infty$ such that

$$\forall\,(j,k) \in \mathcal{J}_n,\ \exists\, k_1,\ \exists\, c \in \mathbb{Z} \text{ such that } |c| \le c_\lambda,\ I^\circ_{j+c,k_1} \subset I_{j,k} \text{ and}$$
$$P(|I^\circ_{j+c,k_1}| \le |I_{jk}|) \ge 1 - O(n^{-\lambda})$$

In particular, if $I^\circ_{j_1,k_1}$ is the largest interval $\in \mathbf{I}^\circ$ that is included in $I_{j,k}$, then we have finite constants c_1, c_2 such that $c_2 \le 2^j/2^{j_1} \le c_1$, or, in short, $2^j \sim 2^{j_1}$.

Theorem 4.1. *Assume (A1)-(A3). Consider the Haar detail coefficient $\hat{d}_{jk} = d_{jk} + \rho_{jk}$ as in equation* (2.3)*, where $(j,k) \in \mathcal{J}_n$ as defined in* (4.5)*. Let $I^\circ_{j_1,k_1}$ (resp. $I^\circ_{j_1',k_1'}$) be the largest interval that is included in the random interval $I_{j+1,2k}$ (resp. $I_{j+1,2k+1}$) in a decomposition similar to expression* (3.5)*. Let d_n (resp. d_n') be the number of coarse-scale intervals in the decomposition* (3.5) *of the interval $I_{j+1,2k}$ (resp. $I_{j+1,2k+1}$). Assume that d_n and d_n' are* $O(\log n)$*. Then we have*

$$P(|\rho_{jk}| \ge t_{jk}) \le O(n^{-\lambda})$$

with

$$t_{jk} = c_\lambda \left(2^{(j-j_1)/2} \frac{\sigma_\epsilon}{\sqrt{n}} \log n + 2^{(j-j_1')/2} \frac{\sigma_\epsilon}{\sqrt{n}} \log n \right) , \tag{4.6}$$

where the constant c_λ depends only on λ.

Remark 4.2. Note that the assumption requiring that d_n and d_n' are $O(\log n)$, as well as the condition (4.5) on the set $\mathcal{J}_n$ are merely technical for this first result. If we do not make these assumptions, the first term in (4.6) is simply replaced by

$$\frac{2^{j/2}}{n} \left(\sigma_\epsilon \sqrt{n 2^{-j_1}} \log n + d_n n^\delta + n^{2\delta} \right) ,$$

and the second term in (4.6) is modified in a similar way.

Remark 4.3. In the case of an average-interpolating basis with $\tilde{\psi}_{jk}$ given by (2.2), the noise term ρ_{jk} is equal to

$$\rho_{jk} = \frac{2^{(j+1)/2}}{n} \left(\sum_{t=1}^{n} \left(\sum_{l:\tilde{g}_{jlk} \neq 0} \tilde{g}_{jlk} 1_{j+1,l}(X_{t-1}) \right) \right) \tag{4.7}$$

instead of (4.1). Suppose now that the values $\tilde{g}_{jlk}$ are uniformly bounded in j (this is the case for example if the density associated to F_x is bounded away from zero on a compact interval). Then the order of the threshold in Theorem 4.1 is the same when ρ_{jk} defined in (4.1) is replaced by (4.7); only the number of terms in (4.6) differs: in case of an average-interpolating basis it is equal to $K_{jk} := \#\{l : \tilde{g}_{jlk} \neq 0\}$, which is uniformly bounded in j and k by construction.

Remark 4.4. Consider for a moment the situation where $j = j_1 = j_1'$ in (4.6). The threshold t_{jk} is then of order $\frac{\sigma_\epsilon}{\sqrt{n}} \log n$. This order has to be compared with the so-called Universal or VisuShrink threshold [12, 13], given by $t_{\text{univ}} = \frac{\sigma_\epsilon}{\sqrt{n}} \sqrt{2 \log n}$. The fact that the threshold needs to be increased when there is some correlation between the wavelet coefficients is well-known [2, 1]. An intuitive explanation is easy: in presence of correlation, a detail coefficient can be large, not because it conveys some important information about the underlying function, but only because of its correlation with another, possibly large, detail coefficient.

Recall now the definition of the soft-threshold nonlinearity:

$$\eta_S(d,t) = \mathrm{sign}(d)(|d| - t)_+ \ ,$$

and let $\hat{d}^t_{jk} := \eta_S(\hat{d}_{jk}, t(n))$ with $t(n)$ given by (4.4)-(4.6). By construction of $t(n)$, we have that

$$\pi_n \equiv P(\max_{(j,k)} |\rho_{jk}| \le t(n)) \to 1 \text{ as } n \to \infty \ . \tag{4.8}$$

Theorem 4.1 of Donoho given in [12] allows us to deduce from (4.8) that

$$P\left(|\hat{d}^t_{jk}| \le |d_{jk}| \ \forall \ (j,k) \in \mathcal{J}_n\right) \ge \pi_n \tag{4.9}$$

By orthonormality of the unbalanced Haar basis in $\mathbf{L}_2(d\hat{F}_n)$, this leads directly to

$$P\left(\|\hat{m}\|_{\mathbf{L}_2(d\hat{F}_n)} \le \|m\|_{\mathbf{L}_2(d\hat{F}_n)}\right) \ge \pi_n \ . \tag{4.10}$$

As it has been noted in the Remark 4.3 above, in case of an AI (average-interpolating) basis, it is still possible to find a threshold that satisfies (4.8), and hence an inequality similar to (4.9) is still valid. However, since we have to give up orthonormality when building smoother design-adapted wavelets, the relationship (4.10) does not exist anymore for an AI basis.

In the regression setting on a fixed design, Donoho [12] derived from (4.9) a result on the smoothness of the estimator of the regression function. More precisely, the estimated values $\hat{m}(i/n), i = 0, \ldots, n-1$, are first obtained by applying an interval-adapted wavelet transform [7] and a soft-thresholding nonlinearity on the empirical detail coefficients before inverting the transform. The threshold is chosen so that the property (4.9) is satisfied.

From the estimated values $\hat{m}(i/n)$, Donoho constructed a special smooth interpolation of these values, producing a continuous estimator $\hat{m}^*(x)$ on $[0,1]$. This estimator is, with probability tending to one, at least as smooth as the underlying function m, see Theorem 1.1 in [12].

In our case, it is not possible to reproduce the construction of Theorem 1.1 given by Donoho in [12]. Indeed, the only orthonormal basis in $\mathbf{L}_2(d\hat{F}_n)$ is the Haar basis, which is not smooth enough to allow for the characterization of the wavelet coefficients in a Besov space $B^s_{p,q}$, having a smoothness parameter s such that $1/p < s < \min(R,N)$. In this notation, R is the regularity of the functions belonging to the space spanned by φ_{jk}, and N is the number of vanishing moments of the wavelets. Indeed, in case of the Haar wavelet, we have $R = 0, N = 1$.

The proof of Theorem 4.1 is given in Section 6.1.2. We now have a thresholding scheme that removes the noise in the empirical wavelet coefficients with a high probability. However, it is necessary to check that the threshold chosen is not too large, otherwise it might remove the underlying signal as well as the noise term. Hence, in a second step, it is necessary to derive an upper bound on the l_2-risk of the detail coefficients (i.e. the risk of the estimator in the wavelet domain). This is the object of the next section.

4.2 Upper bound on the l_2-risk of the detail coefficient

Consider the l_2-risk

$$\mathcal{R}(\hat{d}^t, d) = E\sum_{(j,k)} (\hat{d}^t_{jk} - d_{jk})^2 = E\left\|\hat{d}^t - d\right\|^2_{l_2} \ ,$$

where $\hat{d}^t$ is the vector containing the soft-thresholded wavelet coefficients. We use the threshold t_{jk} given by (4.6), which is level and location dependent since $j_1 = j_1(k)$. To measure the quality of the estimator $\hat{d}^t$, one can define an *ideal risk* [13]. The ideal risk is the risk one would obtain if an oracle told us which detail coefficients to keep or to kill.

Oracle inequalities indicate how far is the l_2-risk $\mathcal{R}(\hat{d}^t, d)$ from the ideal risk. In [13, 26, 37], the authors derived oracle inequalities for the regression model with i.i.d., correlated and locally stationary errors, respectively.

In this section, we begin by recalling the oracle inequality derived by Donoho and Johnstone [13] in the case of Gaussian i.i.d. errors, and by giving another upper bound on $\mathcal{R}(\hat{d}^t, d)$. This second upper bound will be our benchmark to compare the upper bounds obtained in the regression and in the autoregression setting.

4.2.1 Regression model with Gaussian noise

Consider for a moment the regression model

$$Y_t = m(x_t) + \epsilon_t, \tag{4.11}$$

where the design is fixed and the errors are Gaussian i.i.d., $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. In the wavelet domain, the above model can be written as:

$$\hat{d}_{jk} = d_{jk} + \sigma_W z_{jk} \tag{4.12}$$

where $z_{jk} \sim N(0,1)$ and σ_W is the standard deviation of the wavelet coefficients. In case of an orthonormal transform, σ_W is equal to $\sigma_\epsilon/\sqrt{n}$. In the model (4.12), ideally we would keep the detail coefficients $\{\hat{d}_{jk}\}$ for which $\{|d_{jk}| > \sigma_W\}$. This yields ideal thresholded coefficients $\hat{d}^t_{jk} := \hat{d}_{jk} 1_{\{|d_{jk}|>\sigma_W\}}$ whose corresponding risk is equal to

$$\begin{aligned}
\mathcal{R}_{\text{ideal}} &= E\left(\sum_{(j,k)} (\hat{d}_{jk} 1_{\{|d_{jk}|>\sigma_W\}} - d_{jk})^2\right) \\
&= E\left(\sum_{(j,k)} ((d_{jk} + \sigma_W z_{jk}) 1_{\{|d_{jk}|>\sigma_W\}} - d_{jk})^2\right) \\
&= E\left(\sum_{(j,k)} (\sigma_W z_{jk} 1_{\{|d_{jk}|>\sigma_W\}} - d_{jk} 1_{\{|d_{jk}|\le\sigma_W\}})^2\right) \\
&= \sum_{(j,k)} \min(\sigma_W^2, d_{jk}^2) \ .
\end{aligned}$$

In Theorem 4.5, we first recall the oracle-type inequality obtained by Donoho and Johnstone in [13]. We then give a second upper bound, that will be compared in the next section with a similar upper bound obtained in the case of autoregressive designs.

Theorem 4.5 ([13], Theorem 1). *Assume model* (4.11)-(4.12)*. The soft-thresholded detail coefficients*

$$\hat{d}^t_{jk} = \eta_S(\hat{d}_{jk}, t), \quad \textit{where } t = \sigma_W (2\log n)^{1/2} \ ,$$

satisfy

$$E \sum_{(j,k)} \left| \hat{d}^t_{jk} - d_{jk} \right|^2 \leq (2 \log n + 1) \left(\frac{\sigma_\epsilon^2}{n} + \sum_{(j,k)} \min(d_{jk}^2, \frac{\sigma_\epsilon^2}{n}) \right) \tag{4.13}$$

$$\leq \sum_{(j,k)} \frac{2\sigma_\epsilon^2}{n}(1 + 2\log n) + \min(8d_{jk}^2, 4t^2) \; . \tag{4.14}$$

4.2.2 Autoregressive model with non-Gaussian noise

Consider now the autoregressive model (1.1). In the wavelet domain, we have

$$\hat{d}_{jk} = d_{jk} + \rho_{jk}, \; (j,k) \in \mathcal{J}_n \tag{4.15}$$

as before. Using the location and scale dependent threshold $\{t_{jk}\}$ derived in Section 4.1, we obtain the following risk inequality.

Theorem 4.6. *In the autoregressive model given by* (1.1) *and* (4.15)*, assume (A1)-(A3). In the decomposition* (3.5) *of a random interval $I_{j,k}$, $(j,k) \in \mathcal{J}_n$, assume that the number d_n of intervals at coarse scales is of order $d_n = O(\log((\log n)^{1/r}))$ for some $r > 1$. Consider the Haar detail coefficients $\hat{d}_{jk} = d_{jk} + \rho_{jk}$ as in equation* (2.3)*. Let $\hat{d}^t_{jk} := \eta_S(\hat{d}_{jk}, t_{jk})$ be the soft-thresholded coefficients, where t_{jk} is given in* (4.6)*. Then, for some p, $1 < p < 2$, an upper bound on the l_2-risk of the thresholded coefficients is given by:*

$$\sum_{(j,k)\in\mathcal{J}_n} E \left| \hat{d}^t_{jk} - d_{jk} \right|^2 \leq \sum_{(j,k)\in\mathcal{J}_n} \frac{c\sigma_\epsilon^2 (\log n)^p}{n} + \min\left(8d_{jk}^2, 4t_{jk}^2\right) \; , \tag{4.16}$$

where the set $\mathcal{J}_n$ is defined in (4.5) *and c is a finite constant not depending on n.*

The similarity between the expression (4.14), obtained in the regression setting, and the upper bound given by (4.16) in case of an autoregressive design is clear. The order of these two upper bounds is not the same however, since in case of dependent data, the threshold has an additional factor of order $\sqrt{\log n}$.

5 Estimation of nonlinear autoregressive models in practice

In the previous sections, we saw some properties of the design-adapted wavelets in case of the homoscedastic model (1.1). We now present an algorithm to estimate in practice nonlinear autoregression models of order one using design-adapted wavelets. This algorithm is then applied on a few simulated data sets and on the 1997 CAC40 index series.

5.1 Nonlinear autoregressive models

The literature on time series models has produced in the past decades a huge collection of linear (ARMA) or nonlinear (ARCH, GARCH) parametric models [16, 3]. More recently, nonparametric estimators have also been proposed [24, 28, 32, 35, 5] in order to capture the nonlinear behavior of the time series. Let $\mathcal{F}_t = \mathcal{F}^t_{-\infty}$ denote the σ-field generated by $\{X_s, s \leq t\}$. With our design-adapted wavelet estimator, it is possible to estimate the following four types of models.

1. The AR(1) model : $X_t = m(X_{t-1}) + \epsilon_t \, , \epsilon_t \sim N(0, \sigma_\epsilon^2)$.
2. The Conditional Heteroscedastic Autoregressive Nonlinear (CHARN) model, where an heteroscedastic variance is allowed. $X_t = m(X_{t-1}) + \sigma(X_{t-1})\epsilon_t \, , \epsilon_t \sim N(0,1)$. This model has been studied in [24, 21].
3. The Autoregressive Conditionally Heteroscedastic ARCH(1) model with no trend:
$$X_t = \sigma_t \epsilon_t \ , \tag{5.1}$$
where $\{\epsilon_t\}$ is an i.i.d sequence with zero mean and unit variance: $E(\epsilon_t) = 0, E(\epsilon_t^2) = 1$. The conditional variance $\sigma_t := \mathrm{Var}(X_t|\mathcal{F}_{t-1})$ evolves according to
$$\sigma_t^2 = s(X_{t-1}) \ . \tag{5.2}$$
4. If a trend is present in the data, we call the model AR-ARCH:
$$\begin{aligned} X_t &= m(X_{t-1}) + u_t \\ u_t &= \sigma_t \epsilon_t \end{aligned}$$
where again the $\{\epsilon_t\}$ is an i.i.d sequence with $E(\epsilon_t) = 0, E(\epsilon_t^2) = 1$. The conditional expectation of u_t^2 is a function of u_{t-1}.
$$E(u_t^2|\mathcal{F}_t) = \sigma_t^2 = s(u_{t-1}) \ .$$
Thus the variance function σ_t^2 is a function of the lagged residuals $u_{t-1} := Y_{t-1} - m(Y_{t-2})$. The purpose here is to estimate both the trend $m(.)$ and the variance function $s(.)$.

The parametric ARCH-like models are commonly used in finance, see for example [3, 16, 22] Our aim here is to estimate the trend and/or the volatility function in the above models in a nonparametric way, thus without making restrictives assumptions on the form or on the smoothness of the functions to be estimated.

5.2 Estimation of AR and CHARN models

We now present an algorithm tailored to the AR and CHARN models, which is a robust version of the three-step algorithm proposed in [11] for the estimation of heteroscedastic regression models. Indeed, in practice data sets $(X_{t-1}, X_t)_{t=1}^n$ coming from a stationary time series have a low signal-to-noise ratio, that is, a large variability. This is the reason why we add a first step to the algorithm presented in [11].

1. Let $J := \lfloor \log_2(n) \rfloor$ be the finest resolution level. Compute a linear design-adapted wavelet estimator (based on Haar or on weighted AI) with a cutting level j^* equal to $j^* = J - 1$ in case of AR and AR-ARCH models, and to $j^* = J - 2$ in case of CHARN and ARCH models. Call $\tilde{m}(x)$ the resulting estimate and let $\tilde{X}_t := \tilde{m}(X_{t-1})$.
2. Consider the data set given by $(X_{t-1}, \tilde{X}_t)_{t=1}^n$. From this data set, obtain a pilot estimator $\widehat{m}_0(x)$ of $m(x)$ using a linear design-adapted wavelet estimator, where the cut-off scale j_1 is chosen as : $j_1 := \lfloor \log_2(n)/2 \rfloor$.

3. Take the residuals $r_t := \tilde{X}_t - \widehat{m}_0(X_{t-1})$ and estimate the variance function $\sigma^2(.)$ from the data set $(X_{t-1}, r_t^2), t = 1, \ldots, n$ using a *robust* linear Haar or average-interpolating estimator. That is, we choose a cutting level $j_2 := \lfloor \log_2(n)/3 \rfloor$, and we discard the 1% of the residuals having the highest absolute value. Call this estimate $\hat{\sigma}^2(x)$.
4. Finally, estimate m from the data set $(X_{t-1}, \tilde{X}_t)_{t=1}^n$ by a non-linear thresholding wavelet estimator, using $\widehat{\sigma}(x)$ to estimate the variance σ_{jk}^2 of the wavelet coefficients.

The threshold t_{jk} is chosen as a modified version of the universal threshold of Donoho [12]:

$$t_{jk} = \hat{\sigma}_{jk} \sqrt{2(1 + \delta_{j_0})} \log n \ ,$$

where δ_{j_0} is computed as the maximum of the cross-correlations between the detail coefficients. This factor takes into account the correlation between wavelet coefficients induced by the biorthogonal nature of the transform [2]. Compared to the traditional universal threshold, the additional factor $\sqrt{\log n}$ takes into account the dependance between the observed design points.

5.3 Estimation of ARCH-like models

As it stands, the ARCH model (5.1)-(5.2) contains multiplicative errors. The denoising wavelet method as initiated by the work of Donoho and Johnstone [13, 14] is only suitable for additive noise. Hence, using wavelets, it is not possible to estimate the volatility function σ_t directly from model (5.1)-(5.2). Clearly, we can rewrite the ARCH model (5.1)-(5.2) as

$$X_t^2 = s(X_{t-1}) + V_t \tag{5.3}$$

$$V_t = s(X_{t-1})(\epsilon_t^2 - 1) \ . \tag{5.4}$$

Proposition 5.1 below shows that the random variable V_t is of zero-mean and that V_t, V_k are uncorrelated for $k < t$. This entails that

$$\begin{aligned} E[X_t^2 | \mathcal{F}_{t-1}] &= s(X_{t-1}) \\ \mathrm{Var}[X_t^2 | \mathcal{F}_{t-1}] &= s^2(X_{t-1})(E[\epsilon_t^4] - 1) \ . \end{aligned}$$

This suggests to estimate the function s by regressing X_t^2 on the lagged variables X_{t-1} and use the four steps algorithm described above.

Proposition 5.1. *Assume model* (5.3)-(5.4)*, where the time series* $\{X_t\}$ *is strictly stationary and where the innovations* ϵ_t *are i.i.d. Then we have*

1. $E(V_t) = E(V_t | \mathcal{F}_t) = 0$ *and thus* V_t *is a martingale difference series.*
2. $Cov(V_k, V_{k'}) = Cov(V_k, V_{k'} | \mathcal{F}_t) = 0, k' < k.$

Proof. 1. $E(V_t | \mathcal{F}_t) = s(X_{t-1})(E(\epsilon_t^2) - 1) = 0$

2. We have:

$$\begin{aligned} \mathrm{Cov}(V_t, V_k) &= E(V_t V_k) \\ &= E(E(s(X_{t-1})(\epsilon_t^2 - 1)s(X_{k-1})(\epsilon_k^2 - 1) | \mathcal{F}_t)) \\ &= E(s(X_{t-1})s(X_{k-1})E[(\epsilon_t^2 - 1)(\epsilon_k^2 - 1) | \mathcal{F}_t]) \end{aligned}$$

and

$$E[(\epsilon_t^2 - 1)(\epsilon_k^2 - 1) | \mathcal{F}_t] = E(\epsilon_t^2 \epsilon_k^2 - \epsilon_t^2 - \epsilon_k^2 + 1) = 0.$$

□

In case of an AR-ARCH model, the variance function $s(\cdot)$ has as argument the lagged residuals $u_t := X_t - m(X_{t-1})$. For this reason, we modify the steps 3 and 4 in the algorithm of Section 5.2 as follows.

3b. Estimate the residuals by $\hat{u}_t = X_t - \hat{m}_0(x)$. From the data set $(\hat{u}_{t-1}, \hat{u}_t^2)$, obtain an estimator $\hat{s}(x)$ of $s(x)$ by using a linear design-adapted wavelet estimator. The cutting level is chosen equal to $j_2 := \lfloor \log_2(n)/3 \rfloor$.

4b. Knowing $\hat{s}(\hat{u}_{t-1})$, re-estimate the trend m using a nonlinear wavelet estimator.

In case of an ARCH model (without trend) the variance function is estimated with a nonlinear wavelet estimator. In contrast, in an AR-ARCH model, at first the variance σ_t^2 is considered as a nuisance parameter, and is estimated with a linear wavelet estimator from the residuals, see Step 3b above. This estimation of σ_t^2 allows then to compute the threshold t_{jk} needed for the nonlinear denoising of the trend. If desired, in a second step one may substract the estimated trend from the data and come back to the case of an ARCH model, where the variance function can then be estimated with a nonlinear procedure.

5.4 Simulated examples

As mentioned in the theoretical part of this paper, we have at our disposal two types of design-adapted wavelet estimators. First, the unbalanced Haar wavelet estimator was studied in [10]. For practical implementations, an interval-adapted translation-invariant [8, 29] version of this Haar transform was used. We call this estimator Haar-TI. Secondly, in [11], the authors proposed smooth-design adapted wavelets based on the lifting scheme. Here we call 'WAI-U' the estimator obtained with a weighted average interpolating step followed by a stabilizing update step, while the abbreviation 'WAI' denotes the estimator with no stabilizing update step. We show on some examples the performance of the WAI and WAI-U estimators.

To measure the performance of an estimator, we record the square root of the MSE:

$$\text{RMSE} = \sqrt{\frac{1}{n-1}\sum_{t=2}^{n}(\hat{f}(X_{t-1}) - f(X_{t-1}))}\ ,$$

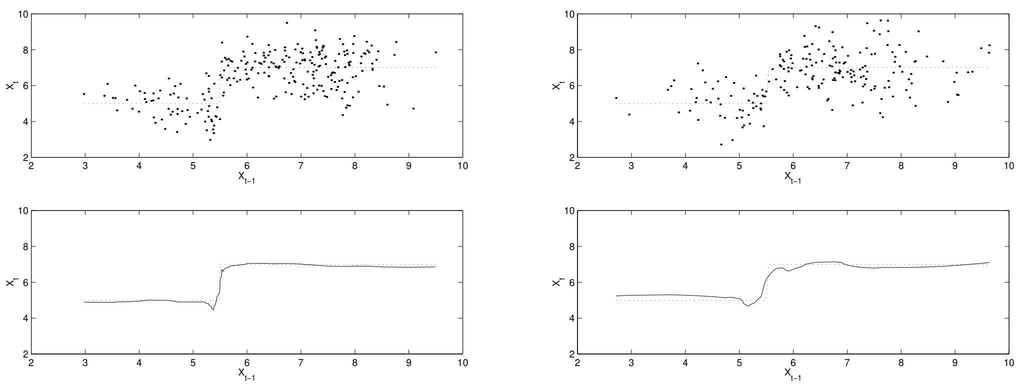
where the function f represents either the conditional expectation (trend) or the variance function. In order to simulate stationary data, we use a burn-in period of 100 observations. After this burn-in period, we take a sample of size $n = 200$. We use hard thresholding and choose for the primary resolution level the value $j_0 = 2$. We now illustrate how the wavelet-type estimator performs on some examples of nonlinear autoregression models.

5.4.1 Example of AR models

Consider the piecewise constant model

$$X_t = 5 1_{\{X_{t-1} \leq 5.5\}}(X_{t-1}) + 7 1_{\{X_{t-1} > 5.5\}}(X_{t-1}) + \epsilon_t,\ \epsilon_t \sim N(0,1)\ . \tag{5.5}$$

The above model is a type of so-called *threshold* models, see for example [20, 36]. Figure 1 shows an estimation of the model (5.5) using Haar-TI and WAI-U estimator. For such piecewise constant models, the Haar basis is optimal, but the WAI-U estimator performs also quite well.



(a) (b)

Figure 1: Two examples of the estimation of the piecewise constant AR model. In (a), the Haar-TI estimator is used on $n = 256$ data and in (b) the WAI-U estimator has been applied on $n = 200$ data. Data are represented by dots, the underlying signal by a dotted line and the estimators by a plain line. The RMSE values are equal to: (a) 0.1466 for the Haar-TI method and (b) 0.2249 for the WAI-U method.

5.4.2 Example of CHARN models

Consider the following process

$$\begin{aligned} X_t &= 0.25X_{t-1}1_{\{X_{t-1}\leq 0\}} + 0.85X_{t-1}1_{\{X_{t-1}>0\}} \\ &+ (0.07 + 0.2(X_{t-1} - 0.4)^2 1_{\{X_{t-1}\leq 0.4\}})\epsilon_t,\ \epsilon_t \sim N(0,1)\ , \end{aligned}$$

where the trend follows a piecewise linear model, and the variance function is quadratic. Figure 2 represents a typical realization of this process together with an estimation of the trend using WAI wavelet estimator (without stabilizing update). We see that the estimator without update performs well.

5.4.3 Example of ARCH-type models

We consider the complete AR-ARCH model

$$\begin{aligned} X_t &= 0.25X_{t-1}1_{\{X_{t-1}\leq 0\}} + 0.85X_{t-1}1_{\{X_{t-1}>0\}} + u_t \\ u_t &= \sigma_t\epsilon_t\ \epsilon_t \sim N(0,1)\ , \\ \sigma_t^2 &= 0.2 + 0.2u_{t-1}^2 1_{\{u_{t-1}<0\}} + 0.6u_{t-1}^2 1_{\{u_{t-1}\geq 0\}}\ , \end{aligned}$$

as well as the ARCH model with no trend. In case of an AR-ARCH model, we use the WAI-U wavelet estimator, whereas for the ARCH model with no trend we use the WAI estimator.

As mentioned in Section 5.3, the nature of the variance function estimator is different in case of an AR-ARCH and ARCH models. It is a *linear* wavelet estimator for the AR-ARCH models (the trend in this model is estimated by a thresholding procedure), whereas in case of the ARCH model with no trend, the algorithm produces a thresholding estimator.

Figure 3 represents the estimation of the trend and volatility function in the full model, whereas Figure 4 gives an example of an estimator of the above ARCH model when no trend

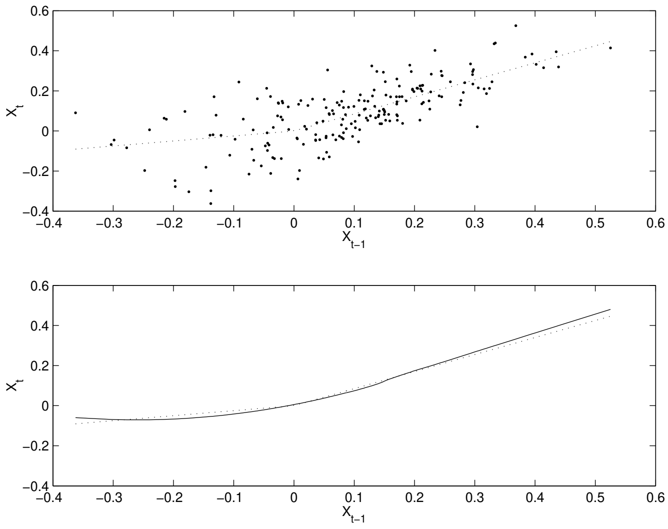


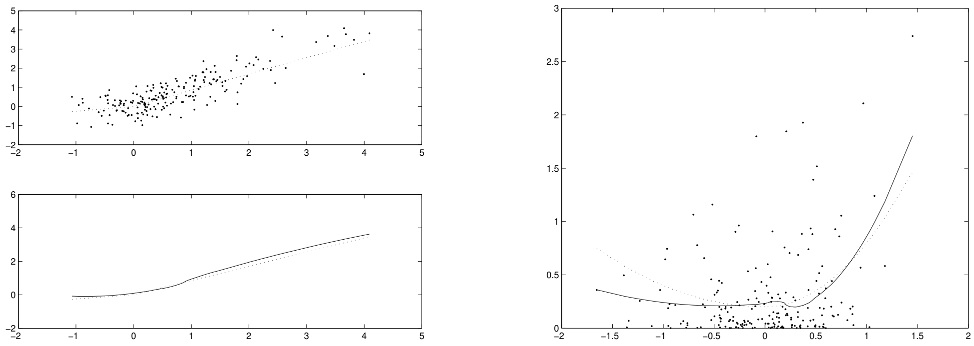
Figure 2: Estimation of the piecewise linear CHARN model using the WAI method with $n = 200$ data. Data are represented by dots, the underlying signal by a dotted line and the estimators by a plain line. The RMSE value is equal to 0.0105.

are present in the data. Note how the procedure easily handles the presence of asymmetry in the volatility function. Also, the trend and volatility function can be estimated all together in a single procedure.

5.5 Application on real data

We now apply the proposed methodology on the index (called 'CAC40') of Paris' Stock Exchange. Figure 5(a) represents the series $\{X_t\}$ of the CAC40 from January, 2, 1997 until December 31, 1997. This data set contains 250 observations. Since we want to model the series $\{X_t\}$ by a nonlinear autoregressive process of order one, we first represent the X_t as a function of the lagged residuals X_{t-1} in Figure 5(b). After dealing with repeated data, the effective sample size was reduced to $n = 241$ observations.

It is a common practice to model such financial time series by means of ARCH-like models, [16, 19, 21]. Hence we applied the algorithm for AR-ARCH models on these data, using the WAI-U wavelet estimator. We took $j^* = J - 1$ in the first step of the algorithm presented in Section 5.2. Figure 6(a) shows the pilot estimator $\hat{m}_0(x)$ of the trend. Indeed, with this data set, the nonlinear thresholding procedure did not improve the estimation. Figure 6(b) represents the linear estimation of the variance function. This figure clearly shows the presence of a quadratic component in the variance function, as it is often the case for such financial time series [19, 21].



(a) (b)

Figure 3: Estimation of the AR-ARCH model using the WAI-U estimator. Data are represented by dots ($n = 200$), the underlying signal by a dotted line and the estimators by a plain line. In (a), an example of the estimation of the trend is given, whereas in (b) the estimation of the variance function σ_t^2 is given. The x-axis in (b) represents the lagged estimated residuals $\hat{u}_t = X_t - \hat{m}(X_{t-1})$. The corresponding RMSE values are equal to (a) 0.1226 and (b) 0.0758.

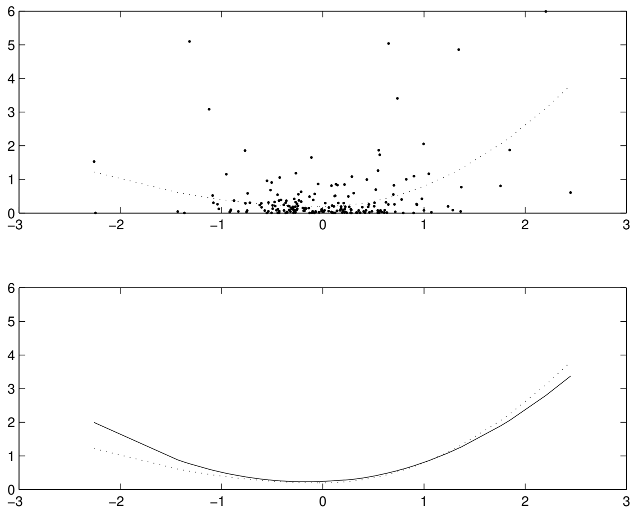
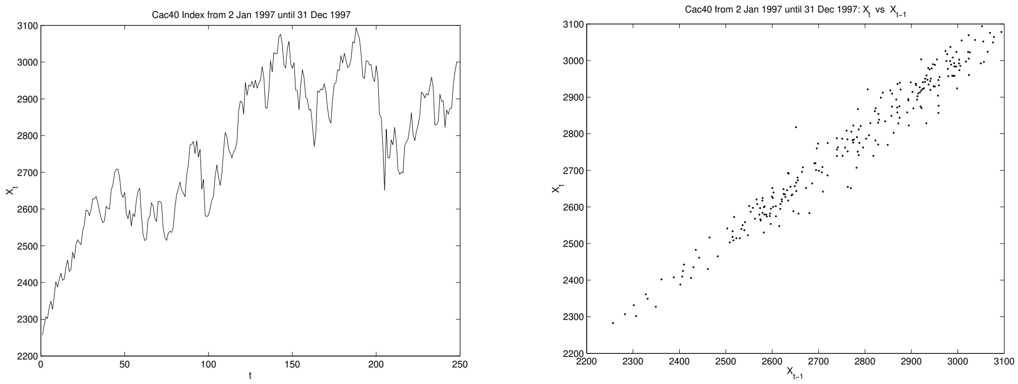
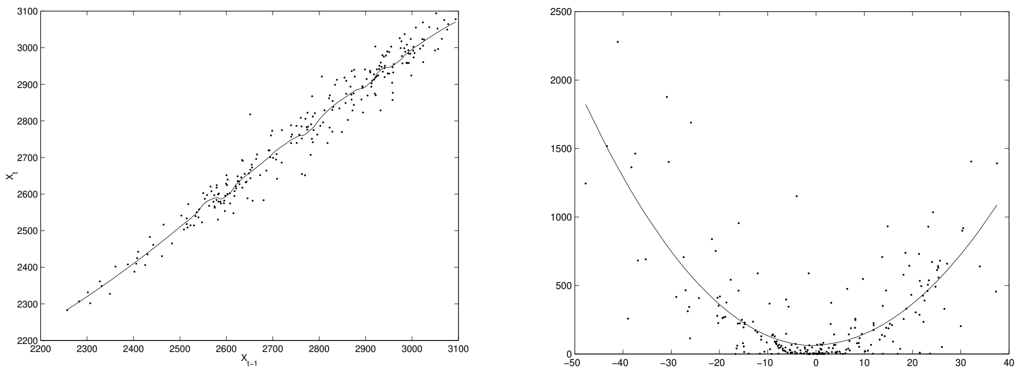


Figure 4: Estimation of the variance function in an ARCH model with the WAI estimator. Data are represented by dots ($n = 200$), the underlying signal by a dotted line and the estimator by a plain line. The corresponding RMSE value is equal to RMSE = 0.1042



(a) (b)

Figure 5: CAC40 index series of Paris' Stock Exchange, from January 2, 1997 until December 31, 1997. In (a) the time series itself is represented, in (b), the value at time t, X_t is given as a function of the lagged value X_{t-1}



(a) (b)

Figure 6: Estimation, using a nonparametric AR-ARCH model, of the CAC40 index series of Paris' Stock Exchange, from January 2, 1997 until December 31, 1997. In (a) the estimation $\hat{m}_0(x)$ of the trend is represented. In (b), the variance function σ_t^2 is estimated as a function of the lagged estimated residuals $u_t = X_t - \hat{m}_0(X_{t-1})$.

6 Proof of the main theorems

6.1 Order of the threshold needed to remove the noise

6.1.1 Preliminaries

To prove the result of Theorem 4.1, it is sufficient to find the order of the term

$$P\left(\frac{2^{j/2}}{n}\left|\sum_{t=1}^{n}\epsilon_t 1_{j+1,2k}(X_{t-1})\right| > a^{(\lambda)}_{j+1,2k}\right)$$

in (4.3) since the other term will be treated in a completely similar manner. Changing notation from $(j+1,2k)$ to (j,k), we let $\mathcal{I} := I_{j,k}$ be the random interval of interest and $a^{(\lambda)}_{j,k} := a^{(\lambda)}_{j+1,2k}$.

As explained in Section 3.2, to prove the results we approximate $\mathcal{I}$ by a union of fixed intervals, as in expression (3.5). This means that, for a given ω in the probability space Ω_n:

$$\begin{aligned}\left|\sum_t \epsilon_t 1_{\mathcal{I}}(X_{t-1}(\omega))\right| &\leq \sum_{i=1}^{d(\mathcal{I})}\left|\sum_t \epsilon_t 1^{\circ}_{j_i,k_i}(X_{t-1}(\omega))\right| \\ &+ \left(\sum_{t=1}^{n}|\epsilon_t|\,1^{\circ}_{j^*,l_1}(X_{t-1}(\omega)) + \sum_{t=1}^{n}|\epsilon_t|\,1^{\circ}_{j^*,l_2}(X_{t-1}(\omega))\right) \\ &= A+B\ . \end{aligned} \tag{6.1}$$

The term A represents the approximation of $\mathcal{I}$ using coarse scale intervals, which are all included in the interval $\mathcal{I}$. The term B contains the indicator on fine scale intervals used to recover $\mathcal{I}$. These intervals $I^{\circ}_{j^*,l}$ are not necessarily included in $\mathcal{I}$, and hence it is easier to take the absolute values $|\epsilon_t|$ in the expression of the fine scale term B, see Lemma 6.3. Under the assumptions (A1)-(A2), the terms A and B are sums of geometrically $\beta-$mixing random variables. To treat these sums, a central lemma is the Lemma 2.1 given by Neumann and Kreiss in [30]. Hence it is useful to recall this lemma. For sake of convenience of the reader, its proof is given in Section 6.1.3.

Lemma 6.1 ([30], Lemma 2.1). *Suppose that $\{Z_t\}_{t=1}^n$ is a geometrically $\beta-$mixing process with $E(Z_t)=0$. If for all $M<\infty$, there exists a constant $C_M<\infty$ such that $E|Z_t|^M \leq C_M$, we have*

$$\sum_{t=1}^{n} Z_t = \tilde{O}\left(\sqrt{\sum_t \mathrm{Var}(Z_t)}\log n + n^{\delta}, n^{-\lambda}\right)\ .$$

We now give the proof of Theorem 4.1.

6.1.2 Proof of Theorem 4.1

Proof. Let $\mathcal{I} := I_{jk}$ be one of the random interval such that $(j,k)\in\mathcal{J}_n$ with $\mathcal{J}_n$ defined in (4.5). Given the approximation $\mathcal{I}\subseteq I^{\circ}_{j_1,k_1}\cup\ldots\cup I^{\circ}_{j_d,k_d}\cup I^{\circ}_{j^*,l_1}\cup I^{\circ}_{j^*,l_2}$, where $I^{\circ}_{j_1,k_1}$ is the largest interval used to recover $\mathcal{I}$, it is sufficient to prove that

$$P\left(\left|\sum_t \epsilon_t 1_{\mathcal{I}}(X_{t-1})\right| > c_\lambda \sigma_\epsilon \sqrt{n2^{-j_1}}\log n\right) = O(n^{-\lambda})$$

Using the decomposition (6.1), we have for arbitrary c

$$P\left(\left|\sum_t \epsilon_t 1_{\mathcal{I}}(X_{t-1})\right| > c\right) \leq P\left(\sum_{i=1}^{d(\mathcal{I})}\left|\sum_t \epsilon_t 1^{\circ}_{j_i,k_i}(X_{t-1})\right| + \sum_{t=1}^{n}|\epsilon_t|\, 1_{I^{\circ}_{j^*,l_1}\cup I^{\circ}_{j^*,l_2}}(X_{t-1})) > c\right) .$$

Lemma 6.2 states that the coarse scale approximation can be bounded as follows:

$$P\left(\sum_{i=1}^{d(\mathcal{I})}\left|\sum_t \epsilon_t 1^{\circ}_{j_i,k_i}(X_{t-1})\right| > c_\lambda(\sigma_\epsilon\sqrt{n2^{-j_1}}\log n + dn^\delta)\right) \leq O(n^{-\lambda}) .$$

For the fine scale term, Lemma 6.3 shows that

$$P\left(\sum_{t=1}^{n}|\epsilon_t|\, 1_{I^{\circ}_{j^*,l_1}\cup I^{\circ}_{j^*,l_2}}(X_{t-1})) > c_\lambda(n^\delta[\sqrt{n2^{-j^*}}\log n + n2^{-j^*} + n^\delta])\right) \leq O(n^{-\lambda}) .$$

Putting both results together, we obtain

$$\sum_t \epsilon_t 1_{\mathcal{I}}(X_{t-1}) = \tilde{O}(t_\lambda, n^{-\lambda}) ,$$

where

$$t_\lambda = c_\lambda\left(\sigma_\epsilon\sqrt{n2^{-j_1}}\log n + d(n)n^\delta + n^\delta(\sqrt{n2^{-j^*}}\log n + n2^{-j^*} + n^\delta)\right) . \tag{6.2}$$

We now look for the dominating term as n tends to infinity. Recall first that $2^{j^*} \leq n < 2^{j^*+1}$, hence $n2^{-j^*} = O(1)$. Under the assumption that $d(n) = O(\log n)$, the term $\sqrt{n2^{-j_1}}\log n$ will be the dominating term in (6.2) if $n^{2\delta} = o(\sqrt{n2^{-j_1}}\log n)$. By Assumption (A3), $j_1 = j + c'$ with $|c'| \leq |c_\lambda| < \infty$, thus $2^{j_1} \sim 2^j$. This implies that, for some constants $c, c_1 < \infty$, $2^{j_1} \leq c_1 2^j \leq cn^{1-\alpha}$, since $(j,k) \in \mathcal{J}_n$.

Hence, the term $\sqrt{n2^{-j_1}}\log n$ is bounded from below as follows.

$$n^{1/2}n^{-1/2+\alpha/2}\log n < \sqrt{n2^{-j_1}}\log n .$$

Thus, we have

$$n^{2\delta} = o(n^{\alpha/2}\log n) = o(\sqrt{n2^{-j_1}}\log n)$$

as soon as $\alpha/2 > 2\delta$, that is, if $\alpha > 4\delta$, as it is given in the condition (4.5) on $\mathcal{J}_n$. It remains to multiply t_λ in (6.2) by $2^{j/2}/n$ to obtain the result. ☐

6.1.3 Proof of auxiliary lemmas

Proof of Lemma 6.1 ([30], Lemma 2.1)

Proof. The proof proceeds in two steps.

(i) Assume that $|Z_t| \leq 1$ a.s. and show that

$$\left|\sum_t Z_t\right| = \tilde{O}\left(\sqrt{\sum_t \mathrm{Var}(Z_t)}\log n + (\log n)^2, n^{-\lambda}\right) .$$

(ii) Use truncated variables $Z''_t = Z_t 1_{|Z_t|\le n^\delta}$ to show that

$$\left|\sum_t Z_t\right| = \tilde{O}\left(\sqrt{\sum_t \mathrm{Var}(Z_t)\log n} + n^\delta, n^{-\lambda}\right) .$$

Proof of (i)

The idea to handle the dependence between the $\{Z_t\}$ is to consider blocks of consecutive observations $\{Z_t, t \in \mathcal{B}_i\}$ where $\mathcal{B}_i = \{(i-1)\rho_n + 1, \dots, \min(i\rho_n, n)\}$. The quantity $\rho_n := [\tilde{c}_\lambda \log n]$ is the length of the blocks. Now, if you consider only the blocks with odd indices, then by Proposition 2 in [15], we have the following result: for geometrically $\beta-$mixing random variables there exists a sequence of independent blocks $\{Z'_t, t \in \mathcal{B}_i\}$, i odd, with the property that

$$\begin{aligned} P\left((Z_t, t\in\mathcal{B}_i) \neq (Z'_t, t\in\mathcal{B}_i)\right) \le \beta_n(\rho_{n+1}) &= a^{-(\rho_n+1)} \\ &= a^{-(\tilde{c}_\lambda \log n+1)} = O(n^{-\lambda}) . \end{aligned} \tag{6.3}$$

(Naturally, the same operation can be done on the even blocks). After this reduction to the independent case, it is possible to use Berstein's inequality, see [30], page 1596 or [33], page 855 for example. Bernstein's inequality states that, if $U_1, \dots, U_n$ are independent random variables with $E(U_i) = 0$ and $|U_i| \le K_n$ almost surely, then for $U = \sum_i U_i$:

$$\begin{aligned} P(|U| > x) &\le 2\exp\left(\frac{-x^2/2}{\mathrm{Var}(U) + K_n x/3}\right) \\ &\le 2\exp\left(\frac{-x^2}{4\,\mathrm{Var}(U)}\right) + 2\exp\left(\frac{-3x}{4K_n}\right) \end{aligned}$$

holds for arbitrary $x > 0$. Setting $x_\lambda = \sqrt{\mathrm{Var}(U)}\sqrt{4\lambda\log(n)} + (4/3)K_n\lambda\log(n)$, we get $P(|U| > x_\lambda) \le 4\exp(-\lambda\log(n)) = 4n^{-\lambda}$. That is, we have that

$$U = \tilde{O}(\sqrt{\mathrm{Var}(U)}\sqrt{\log(n)} + K_n\log(n), n^{-\lambda}) ,$$

where we used Definition 3.2. To apply Bernstein's inequality on the independent blocks, we need to compute the variance of $\{\sum_t Z'_t, t \in \mathcal{B}_i\}$. By Jensen inequality, we have

$$\mathrm{Var}\left(\sum_{t\in\mathcal{B}_i} Z'_t\right) \le \rho_n \,\mathrm{Var}\left(\sum_{t\in\mathcal{B}_i} Z'_t\right) = \rho_n \,\mathrm{Var}\left(\sum_{t\in\mathcal{B}_i} Z_t\right) .$$

The variance of the sum of all r.v. Z_t belonging to the odd blocks is then given by

$$\mathrm{Var}\left(\sum_i\sum_{t\in\mathcal{B}_i} Z'_t\right) = \sum_i \mathrm{Var}\left(\sum_{t\in\mathcal{B}_i} Z'_t\right) \le \rho_n \sum_i\sum_{t\in\mathcal{B}_i} \mathrm{Var}(Z_t) ,$$

since the blocks $\{Z'_t, t \in \mathcal{B}_i\}$ are independent. By doing the same operation on the even blocks, we finally obtain

$$\mathrm{Var}\left(\sum_{t=1}^n Z_t\right) \le O(\log n)\sum_{t=1}^n \mathrm{Var}(Z_t) . \tag{6.4}$$

Applying Bernstein's inequality on $U_i = \sum_{t\in\mathcal{B}_i} Z'_t$ with $|U_i| \le K_n = \tilde{c}_\lambda \log(n)$, we get

$$\begin{aligned}
\sum_{t=1}^{n} Z_t &= \tilde{O}\left(\sqrt{\rho_n \sum_t \mathrm{Var}(Z_t)}\sqrt{\log n} + K_n(\log n), n^{-\lambda}\right) \\
&= \tilde{O}\left(\sqrt{\sum_t \mathrm{Var}(Z_t)}\log n + (\log n)^2, n^{-\lambda}\right)
\end{aligned}$$

Proof of (ii)

Let c_λ be a generic notation for a constant depending only on λ. Under the assumption that, $\forall\, M < \infty, \exists C_M < \infty$ such that $E|Z_t|^M \le C_M$, we have that

$$P\left(\sup_{t=1,\ldots,n} |Z_t| > c_\lambda n^{\delta'}\right) \le cn^{-\lambda} \tag{6.5}$$

hold for any $\delta' \in (0,\delta)$. Define the truncated variables $Z''_t = Z_t 1_{\{-c_1 n^{\delta'} \le Z_t \le c_2 n^{\delta'}\}}$ with c_1, c_2 chosen so that $EZ''_t = 0$. Using the same construction involving independent blocks as before, we have:

$$\begin{aligned}
\left\|\sum_{t\in\mathcal{B}_i} Z''_t\right\|_\infty &\le c_\lambda \log(n) n^{\delta'} \le c_\lambda n^{\delta} \ ; \\
\mathrm{Var}(\sum_{t=1}^{n} Z''_t) &= \sum_i \mathrm{Var}\left(\sum_{t\in\mathcal{B}_i} Z''_t\right) \\
&\le (\tilde{c}_\lambda \log n)\sum_i \sum_{t\in\mathcal{B}_i} \mathrm{Var}(Z''_t) \\
&= (\tilde{c}_\lambda \log n)\sum_t (\mathrm{Var}(Z_t) + O(n^{-\lambda})) \ .
\end{aligned}$$

The last equality is valid since

$$E[Z_t^2] = E[Z_t''^2] + \int_{-\infty}^{n^{\delta'}} z^2 dF_z + \int_{n^{\delta'}}^{\infty} z_d^2 F_z = E[Z_t''^2] + O(n^{-\lambda}) \ ,$$

by assumption on the moments of Z_t. Applying Bernstein's inequality on $U_i = \sum_{t\in\mathcal{B}_i} Z''_t$ with $|U_i| \le \log(n) n^{\delta'} \le n^{\delta}$, we obtain

$$\sum_{t=1}^{n} Z_t = \tilde{O}\left(\sqrt{\sum_{t=1}^{n} \mathrm{Var}(Z_t)}\log n + n^{\delta}, n^{-\lambda}\right) \ .$$

□

Lemma 6.2. *Let $\mathcal{I}$ be a random interval such that*

$$\mathcal{I} \supseteq I^\circ_{j_1,k_1} \cup \ldots \cup I^\circ_{j_d,k_d}$$

where the interval I°_{jk} are given in (3.3) *and $j_1 < j_2 < j_3, \ldots$. Under the assumptions (A1)-(A2) on the autoregressive model* (1.1), *we have*

$$P\left(\sum_{t=1}^{d(\mathcal{I})}\left|\sum_t \epsilon_t 1^\circ_{j_i,k_i}(X_{t-1})\right| > c_\lambda(\sigma_\epsilon \sqrt{n2^{-j_1}}\log n + d(\mathcal{I})n^{\delta})\right) \le O(n^{-\lambda}) \ .$$

Proof. Consider first the problem of finding $c_{1,\lambda}$ such that

$$P\left(\left|\sum_t \epsilon_t 1^\circ_{j_i,k_i}(X_{t-1})\right| > c_{1,\lambda}\right) \leq O(n^{-\lambda})$$

(The result will then follow by summation). Since the time series $\{X_t\}$ is geometrically $\beta-$mixing, the indicator function $1^\circ_{j_i,k_i}(X_{t-1})$ is also geometrically $\beta - mixing$, (since the σ-field $\sigma(1^\circ_{j_i,k_i}(X_t)) \subset \sigma(X_t)$). Furthermore, the innovation ϵ_t is independent from X_{t-1}, thus the random variables $Z_t := \epsilon_t 1^\circ_{j_i,k_i}(X_{t-1})$ are also geometrically $\beta-$mixing. We then have

$$E(Z_t) = E(E(\epsilon_t 1^\circ_{j_i,k_i}(X_{t-1})|X_{t-1})) = E(\epsilon_t)E(1^\circ_{j_i,k_i}(X_{t-1})) = 0,$$

since ϵ_t is independent of X_{t-1} and $E(\epsilon_t) = 0$. The variance can be computed as

$$\begin{aligned} \mathrm{Var}(Z_t) &= E(E(\epsilon_t^2 1^\circ_{j_i,k_i}(X_{t-1})|X_{t-1})) = E(1^\circ_{j_i,k_i}(X_{t-1}))E(\epsilon_t^2|X_{t-1}) \\ &= \sigma_\epsilon^2 P(1^\circ_{j_i,k_i}(X_{t-1}) = 1) = \sigma_\epsilon^2 2^{-j_i}, \end{aligned}$$

since we assume a homoscedastic variance. Since $Z = \sum_{t=1}^n Z_t$ is a sum of mean-zero $\beta - mixing$ random variables, we can apply Lemma 6.1 and obtain

$$\begin{aligned} \sum_{t=1}^n Z_t &= \tilde{O}\left(\sqrt{\sum_t \mathrm{Var}(Z_t)}\log n + n^\delta, n^{-\lambda}\right) \\ &= \tilde{O}\left(\sigma_\epsilon\sqrt{n2^{-j_i}}\log n + n^\delta, n^{-\lambda}\right) . \end{aligned}$$

When summing over $i = 1, \ldots, d$, we have $\sum_{i=1}^d 2^{-j_i/2} = O(2^{-j_1/2})$ with j_1 the coarsest level used in the approximation, hence the result. □

Lemma 6.3. *Let $\mathcal{I}$ be a random interval such that*

$$\mathcal{I} \subseteq I^\circ_{j_1,k_1} \cup \ldots \cup I^\circ_{j_d,k_d} \cup I^\circ_{j^*,l_1} \cup I^\circ_{j^*,l_2} ,$$

where $I^\circ_{j,k}$ are defined in (3.3) *and $2^{j^*} \leq n \leq 2^{j^*+1}$. Assume (A1)-(A2). Then*

$$P\left(\sum_{t=1}^n |\epsilon_t| \, 1_{I^\circ_{j^*,l_1} \cup I^\circ_{j^*,l_2}}(X_{t-1})) > c_\lambda(n^\delta[\sqrt{n2^{-j^*}}\log n + n2^{-j^*} + n^\delta])\right) \leq O(n^{-\lambda}) .$$

Proof. Note first that, on the probability space Ω_n defined in (3.2), we have that $|\epsilon_t| \leq n^\delta$. We have to evaluate $P\left(\sum_t |\epsilon_t| 1^\circ_{j^*,l_1}(X_{t-1}) > c\right)$. However, since we do not know which precise intervals $I^\circ_{j^*,l}$ are needed in order to recover $\mathcal{I}$, we will take the maximum over all locations l. Moreover, a given interval $I^\circ_{j^*,l}$ is not necessarily included in $\mathcal{I}$. For these two reasons, it is easier to take an upper bound on $|\epsilon_t|$ and to work with the random variables $\{1^\circ_{j^*,l}(X_{t-1})\}_{l=1}^{2^{j^*}}$.

For an arbitrary constant c, we have

$$\begin{aligned}
P\left(\sum_t |\epsilon_t| 1^\circ_{j^*,l_1}(X_{t-1}) > c\right) &\leq P\left(\sum_t n^\delta 1^\circ_{j^*,l_1}(X_{t-1}) > c\right) \\
&\leq P\left(\max_{l=1,\ldots,2^{j^*}} \sum_t n^\delta 1^\circ_{j^*,l}(X_{t-1}) > c\right) \\
&\leq \sum_{l=1}^{2^{j^*}} P\left(n^\delta \sum_t 1^\circ_{j^*,k}(X_{t-1}) > c\right) \\
&\leq nP\left(n^\delta \sum_t 1^\circ_{j^*,k}(X_{t-1}) > c\right)
\end{aligned}$$

and similarly for the term $\sum_t |\epsilon_t| 1^\circ_{j^*,l_2}(X_{t-1})$. If we find a constant c_λ such that, for any l,

$$P\left(\sum_t n^\delta 1^\circ_{j^*,l}(X_{t-1}) > c_\lambda\right) \leq cn^{-\lambda}\ , \tag{6.6}$$

then the lemma is proven, provided we take $\lambda > 1$ sufficiently large. If we find a_λ such that $P(\sum_t 1^\circ_{j^*,l}(X_{t-1}) > a_\lambda) \leq cn^{-\lambda}$, this means that $c_\lambda := n^\delta a_\lambda$ will satisfy (6.6).

The random variable $1^\circ_{j^*,l}(X_{t-1})$ is not of zero-mean, hence we cannot apply directly Lemma 6.1. Introducing b_λ such that $a_\lambda := b_\lambda + n2^{-j^*}$, we have:

$$P\left(\sum_{t=1}^n 1^\circ_{j^*,l}(X_{t-1}) > b_\lambda + n2^{-j^*}\right) \leq P\left(\left|\sum_{t=1}^n (1^\circ_{j^*,l}(X_{t-1}) - 2^{-j^*})\right| > b_\lambda\right)$$

Define $W_t := 1^\circ_{j^*,l}(X_{t-1}) - 2^{-j^*}$. Then $E(W_t) = 0$ and $\mathrm{Var}(W_t) = 2^{-j^*}(1 - 2^{-j^*}) = O(2^{-j^*})$. We can now apply Lemma 6.1 to the sum $W := \sum_t W_t$ to obtain

$$\begin{aligned}
\sum_{t=1}^n W_t &= \tilde{O}\left(\sqrt{\sum_t \mathrm{Var}(W_t)} \log n + n^\delta, n^{-\lambda}\right) \\
&= \tilde{O}\left(\sqrt{n2^{-j^*}} \log n + n^\delta, n^{-\lambda}\right)
\end{aligned}$$

Finally, the quantity c_λ which satisfies (6.6) is equal to

$$c_\lambda = \tilde{c}_\lambda (n2^{-j^*} + \sqrt{n2^{-j^*}} \log n + n^\delta) n^\delta,$$

where $\tilde{c}_\lambda$ depends only on λ. □

6.2 Upper bound on the l_2-risk for the regression model with Gaussian noise

6.2.1 Proof of Theorem 4.5

Proof. The first inequality in Theorem 4.5 is given by the Theorem 1 in [13]. To prove (4.14), we consider only the univariate case, for the multivariate case follows by summation. Let $X = \mu + Z$ be a normal r.v. with $X \sim N(\mu, 1)$ and $\eta_t(X) := \eta_S(X, t)$. The result is shown for a threshold t of the form $t := \sqrt{2 \log n}$ with $n \geq 2$.

First, we decompose $|\eta_t(X) - \mu|$ as follows.

$$|\eta_t(X) - \mu| \le |\eta_t(X)| + |\mu| \le |\eta_t(Z)| + 2|\mu| \ ,$$

where we used the monotonicity of η_t. This entails that

$$E[(\eta_t(X) - \mu)^2] \le 2E[\eta_t(Z)^2] + 8\mu^2 \ .$$

From the proof of Theorem 1 in [13], we have

$$E[\eta_t(Z)^2] \le 2P(|Z| \le t) \le \frac{1}{n}(2\log n + 1) \ ,$$

where the second inequality follows from the normality of Z, see [13], and thus:

$$E[(\eta_t(X) - \mu)^2] \le \frac{2}{n}(2\log n + 1) + 8\mu^2 \ . \tag{6.7}$$

Second, we consider the fact that

$$\begin{aligned} E[(\eta_t(X) - \mu)^2] &\le 2E[(\eta_t(X) - X)^2] + 2E[(X - \mu)^2] \\ &\le 2t^2 + 2E[Z^2] \le 4t^2 \\ &\le 4t^2 + \frac{2}{n}(2\log n + 1) \end{aligned} \tag{6.8}$$

Now, if $\mu^2 \le t^2/2$, we use the decomposition (6.7), and if $\mu^2 > t^2/2$, we use (6.8). Putting both expression together, we obtain:

$$E[(\eta_t(X) - \mu)^2] \le \frac{2}{n}(2\log n + 1) + \min\left(8\mu^2, 4t^2\right) \ .$$

□

6.3 Upper bound on the l_2-risk for autoregression models

To prove Theorem 4.6, we consider separately the case where $|d_{jk}| \le \frac{t_{jk}}{\sqrt{2}}$ (small detail coefficients) and the case where $|d_{jk}| \ge t_{jk}/\sqrt{2}$ (large detail coefficients). In each case, we find an upper bound for $E[(\hat{d}^t_{jk} - d_{jk})^2]$. The inequality (4.16) then results from the minimum of these two upper bounds and by summation over $(j, k) \in \mathcal{J}_n$. To find these upper bounds, we approximate once more the random interval I_{jk} by a union of fixed intervals.

The difficult part in this proof is to show an upper bound for $E[(\hat{d}^t_{jk} - d_{jk})^2]$ when $|d_{jk}| \le \frac{t_{jk}}{\sqrt{2}}$ and $|\rho_{jk}| > \frac{t_{jk}}{2}$. The methodology needed in this case is developed in Lemma 6.4.

6.3.1 Proof of Theorem 4.6

Proof. First, if $|d_{jk}|^2 \le t^2_{jk}/2$, we divide the problem into two parts.

1. $|\rho_{jk}| \le \frac{t_{jk}}{2}$. In that case, $|\hat{d}_{jk}| = |d_{jk} + \rho_{jk}| \le t_{jk}$, hence the thresholded coefficient $\hat{d}^t_{jk} = 0$ and thus $E[(\hat{d}^t_{jk} - d_{jk})^2] = d^2_{jk}$.

2. $|\rho_{jk}| > \frac{t_{jk}}{2}$. Similarly to the proof of Theorem 4.5, we decompose $|\hat{d}^t_{jk} - d_{jk}|$ as follows.

$$\begin{aligned} & |\hat{d}^t_{jk} - d_{jk}| = |\eta_S(\hat{d}_{jk}, t_{jk}) - d_{jk}| \\ = \quad & |\eta_S(d_{jk} + \rho_{jk}, t_{jk}) - d_{jk}| \le |\eta_S(\rho_{jk}, t_{jk})| + 2|d_{jk}| \ . \end{aligned}$$

Hence

$$E|\hat{d}^t_{jk} - d_{jk}|^2 \le 2E[\eta_S(\rho_{jk}, t_{jk})^2] + 8|d_{jk}|^2 \ . \tag{6.9}$$

Lemma 6.4 below tells us that

$$E[\eta_S(\rho_{jk}, t_{jk})^2] = O\left(\frac{(\log n)^p}{n}\right) \tag{6.10}$$

for some p such that $1 < p < 2$.

Second, if $|d_{jk}|^2 > t^2_{jk}/2$, we perform the decomposition

$$\begin{aligned} E|\hat{d}^t_{jk} - d_{jk}|^2 & \le 2E|\hat{d}^t_{jk} - \hat{d}_{jk}|^2 + 2E|\hat{d}_{jk} - d_{jk}|^2 \\ & \le 2t^2_{jk} + 2E|\rho_{jk}|^2 \ . \end{aligned} \tag{6.11}$$

By construction, $P(|\rho_{jk}| > t_{jk}) \le cn^{-\lambda}$ with t_{jk} given by (4.6). We now show that

$$E[\rho^2_{jk}] \le t^2_{jk} \tag{6.12}$$

Indeed,

$$\begin{aligned} E[\rho^2_{jk}] & \le E\left(\frac{2^j}{n^2}\left[\left|\sum_{t=1}^n \epsilon_t 1_{j+1,2k}(X_{t-1})\right| + \left|\sum_{t=1}^n \epsilon_t 1_{j+1,2k+1}(X_{t-1})\right|\right]^2\right) \\ & \le 2(2^j n^{-2})E\left(\left|\sum_{t=1}^n \epsilon_t 1_{j+1,2k}(X_{t-1})\right|^2\right) + 2(2^j n^{-2})E\left(\left|\sum_{t=1}^n \epsilon_t 1_{j+1,2k+1}(X_{t-1})\right|^2\right) \ . \end{aligned}$$

Consider now the problem of finding an upper bound for $E\left|\sum_{t=1}^n \epsilon_t 1_{\mathcal{I}}(X_{t-1})\right|^2$, where $\mathcal{I}$ is random. As before, we decompose this sum as follows.

$$\left|\sum_{t=1}^n \epsilon_t 1_{\mathcal{I}}(X_{t-1})\right|^2 \le 2\left(\sum_{i=1}^{d_n}\left|\sum_{t=1}^n \epsilon_t 1^\circ_{j_i,k_i}(X_{t-1})\right|\right)^2 + 2\left(\sum_t |\epsilon_t| 1^\circ_{j^*,l_1}(X_{t-1}) + \sum_t |\epsilon_t| 1^\circ_{j^*,l_2}(X_{t-1})\right)^2 \ . \tag{6.13}$$

Since the intervals $I^\circ_{j_i,k_i}$ are disjoint by construction, we have:

$$\left(\sum_{i=1}^{d_n}\left|\sum_{t=1}^n \epsilon_t 1^\circ_{j_i,k_i}(X_{t-1})\right|\right)^2 = \sum_{i=1}^{d_n}\left|\sum_{t=1}^n \epsilon_t 1^\circ_{j_i,k_i}(X_{t-1})\right|^2 \ .$$

Taking expectation, we have, using the method of independent blocks as in (6.4):

$$\begin{aligned} E\left(\left|\sum_{t=1}^n \epsilon_t 1^\circ_{j_i,k_i}(X_{t-1})\right|^2\right) & = \mathrm{Var}\left(\sum_{t=1}^n \epsilon_t 1^\circ_{j_i,k_i}(X_{t-1})\right) \le c\log(n) n 2^{-j_i} \ ; \\ E\left(\left|\sum_t |\epsilon_t| 1^\circ_{j^*,l}(X_{t-1})\right|^2\right) & \le n^{2\delta}\,\mathrm{Var}(\sum_t 1^\circ_{j^*,l}(X_{t-1})) \\ & \le cn^{2\delta}\log(n) n 2^{-j^*}. \end{aligned}$$

Putting the above results together, we obtain

$$E\left|\sum_{t=1}^{n}\epsilon_t 1_{\mathcal{I}}(X_{t-1})\right|^2 \le c(\log(n)2^{-j_1}n + n^{2\delta}\log(n)n2^{-j^*}) \ . \tag{6.14}$$

In (6.14), the first term in the sum is the dominating one by the assumption (A3) and the condition (4.5) on the set $\mathcal{J}_n$. Hence we obtain

$$E[|\rho_{jk}|^2] \le c\frac{2^j}{n^2}\log(n)2^{-j_1}n = o(t_{jk}^2)$$
$$\text{since } t_{jk}^2 = c\frac{2^j}{n}2^{-j_1}\log^2(n) \text{ and } 2^j \sim 2^{j_1} \ .$$

The expression (6.11) can thus be bounded by

$$E|\hat{d}_{jk}^t - d_{jk}|^2 \le 4t_{jk}^2 + O\left(\frac{(\log n)^p}{n}\right) \tag{6.15}$$

Finally, putting (6.9),(6.10), and (6.15) together, we have

$$E|\hat{d}_{jk}^t - d_{jk}|^2 \le \frac{c(\log n)^p}{n} + \min(8d_{jk}^2, 4t_{jk}^2) \ .$$

By summation over $(j,k) \in \mathcal{J}_n$, we obtain the desired result. □

Lemma 6.4. *In the autoregressive model given by* (1.1) *and* (4.15), *assume (A1)-(A3). Under the assumption of Theorem 4.6, we have for some* $p \in (1;2)$

$$E[\eta_S(\rho_{jk}, t_{jk})^2] = O\left(\frac{(\log n)^p}{n}\right) \ ,$$

where $\eta_S(d,t) = sign(d)(|d| - t)_+$ *is the soft-threshold nonlinearity.*

Proof. Let $t := t_{jk}$ and note first that:

$$\begin{aligned} E[\eta_S(\rho_{jk}, t)^2] &= E\left[(|\rho_{jk}| - t)^2 1_{\{|\rho_{jk}|>t\}}\right] \\ &= \int_0^\infty P\left[(|\rho_{jk}| - t)1_{\{|\rho_{jk}|>t\}} > z\right] 2zdz \\ &= \int_0^\infty P\left((|\rho_{jk}| - t) > z\right) 2zdz \end{aligned}$$

This shows that in order to compute $E[\eta_S(\rho_{jk}, t_{jk})^2]$, we need to evaluate the quantity:

$$\int_0^\infty P(|\rho_{jk}| > z + t)2zdz \ . \tag{6.16}$$

In case of the Haar transform, we have

$$|\rho_{jk}| \le \frac{2^{j/2}}{n}\left\{\left|\sum_t \epsilon_t 1_{j+1,2k}(X_{t-1})\right| + \left|\sum_t \epsilon_t 1_{j+1,2k+1}(X_{t-1})\right|\right\} \ .$$

Let $\mathcal{I} := I_{j+1,2k}$. We concentrate on the term involving $I_{j+1,2k}$, since the other term will be treated in a completely similar manner. For sake of clarity, change notation from $(j+1, 2k)$

to (j,k). Similarly to Theorem 4.1, we approximate the random interval $\mathcal{I}$ by a union of fixed interval coming from the theoretical quantile partitioning.

$$\mathcal{I} \subseteq I^\circ_{j_1,k_1} \cup \ldots \cup I^\circ_{j_d,k_d} \cup I^\circ_{j^*,l_1} \cup I^\circ_{j^*,l_2} \ .$$

This gives us

$$\left|\sum_t \epsilon_t 1_{\mathcal{I}}(X_{t-1})\right| \le \sum_{t=1}^{d_n}\left|\sum_t \epsilon_t 1^\circ_{j_i,k_i}(X_{t-1})\right| + \sum_{t=1}^{n}|\epsilon_t|\, 1_{I^\circ_{j^*,l_1}\cup I^\circ_{j^*,l_2}}(X_{t-1}) \ ,$$

where the notation d_n indicates that, as n grows, the number of intervals needed in the approximation of $\mathcal{I}$ also grows. Let $z' := z+t$, $U_{jk,l} := \frac{2^{j/2}}{n}\left|\sum_t \epsilon_t 1^\circ_{j_l,k_l}(X_{t-1})\right|$ and U_{j^*,l,d_n} $= \frac{2^{j/2}}{n}\sum_{t=1}^n |\epsilon_t| 1_{j^*,l}(X_{t-1})$. The event $\{|\rho_{jk}|>z'\}$ can be decomposed as follows.

$$\begin{aligned}
\{|\rho_{jk}|>z'\} \quad &\subset \quad \left\{\left|\sum_{l=1}^{d_n} U_{j,k,l} + U_{j^*,l_1,d_n} + U_{j^*,l_2,d_n}\right| > z'\right\} \\
&\subset \quad \left\{U_{j,k,l} \ge z' 2^{-l/\tilde{k}} \text{ for at least one of the } l'\text{s}\right\} \\
&\cup \quad \left\{U_{j^*,l_1,d_n} \ge z' 2^{-d_n/\tilde{k}}\right\} \cup \left\{U_{j^*,l_2,d_n} \ge z' 2^{-d_n/\tilde{k}}\right\}
\end{aligned}$$

for some constant $\tilde{k}$ such that $\sum_{l=1}^{d_n} 2^{-l/\tilde{k}} + 2\cdot 2^{-d_n/\tilde{k}} \le 1$. Indeed, in order to have $\{|\rho_{jk}|>z'\}$, we need at least that, either $U_{j,k,1} \ge z'2^{-1/\tilde{k}}$ or that $U_{j,k,2} \ge z'2^{-2/\tilde{k}}$, etc.

Using the Bonferroni inequality, we have:

$$P\left(|\rho_{jk}|>z'\right) \le \sum_{l=1}^{d_n} P\left(U_{j,k,l} \ge z'2^{-l/\tilde{k}}\right) + \sum_{i=1,2} P\left(U_{j^*,l_i,d_n} \ge z'2^{-d_n/\tilde{k}}\right)$$

Integrating out this last expression with respect to $2zdz$ will give us the desired result. Hence we need to show the following inequalities.

$$\text{[A1]} \qquad \int_0^\infty P\left(U_{j,k,l} \ge z'2^{-l/\tilde{k}}\right) 2zdz \le c\frac{(\log n)^p}{n}, \ 1<p<2 \tag{6.17}$$

$$\text{[A2]} \qquad \int_0^\infty P\left(U_{j,l_i,d_n} \ge z'2^{-d_n/\tilde{k}}\right) 2zdz \le c\frac{(\log n)^q}{n^{1+\alpha/2}}, \ 2<q<3, \ i=1,2 \ , \tag{6.18}$$

where α is defined by (4.5). Note that the order of $[A1]$ is the dominating one.

Proof of [A1]

We have

$$\begin{aligned}
P\left(U_{j,k,l} \ge z'2^{-l/\tilde{k}}\right) \quad &= \quad P\left(\frac{2^{j/2}}{n}\left|\sum_t \epsilon_t 1^\circ_{j_l,k_l}(X_{t-1})\right| \ge z'2^{-l/\tilde{k}}\right) \\
&= \quad P\left(\left|\sum_t \epsilon_t 1^\circ_{j_l,k_l}(X_{t-1})\right| \ge z'2^{-l/\tilde{k}} n 2^{-j/2}\right)
\end{aligned}$$

By assumption (A2), $P(\Omega_n) = P\left\{\omega : \sup_{1\le t\le n}|\epsilon_t(\omega)| \le n^{\delta'}\right\} \ge 1-cn^{-\lambda}$ for any $\delta' \in (0,\delta)$. Thus on the probability space Ω_n, the ϵ_t are bounded by $n^{\delta'}$.

Let now $Z_t := \epsilon_t 1^{\circ}_{j_l,k_l}(X_{t-1})$. Since the $\{X_t\}$ are geometrically $\beta-$mixing and ϵ_t is an i.i.d. white noise, the random variables Z_t's are also geometrically $\beta-$mixing. Similarly to the proof of Lemma 6.1, we handle the dependence between the Z_t by considering blocks of consecutive observations Z_t. Proposition 2 in Doukhan, Massart and Rio (1995) [15] allows us to replace these blocks by independent blocks of length $[c_\lambda \log n]$ of r.v. Z'_t. These independent blocks have the same distribution as the original blocks with probability $1 - cn^{-\lambda}$, see expression (6.3). It is then possible to apply Bernstein's inequality [33, 23] on the independent blocks. Let $\mathcal{B}_i$ be one of these blocks. On the space Ω_n, we have $\left|\sum_{t\in\mathcal{B}_i} Z_t\right| \leq K_n = \log(n)n^{\delta'} \leq n^{\delta}$. We have the following two inequalities.

1. In the middle of the distribution of $|\sum_t Z_t|$, that is, when $x \leq \frac{3\,\mathrm{Var}(\sum_t Z_t)}{K_n}$, we have

$$P\left(\left|\sum_t Z'_t\right| \geq x\right) = P\left(\left|\sum_t Z_t\right| \geq x\right) \leq 2\exp\left(\frac{-x^2}{4\,\mathrm{Var}(\sum_t Z_t)}\right) \tag{6.19}$$

2. In the tail of the distribution, that is, when $x > 3\,\mathrm{Var}(\sum_t Z_t)/K_n$:

$$P\left(\left|\sum_t Z_t\right| \geq x\right) \leq 2\exp\left(\frac{-3x}{4K_n}\right) \tag{6.20}$$

In our application, we have $x = z' 2^{-l/\tilde{k}} n 2^{-j/2}$, hence if $x \leq 3\,\mathrm{Var}(\sum_t Z_t)K_n^{-1}$, this means that

$$z' = z + t \leq 3\,\mathrm{Var}\left(\sum_t Z_t\right)(n^{-\delta})2^{l/\tilde{k}}n^{-1}2^{j/2} =: a_1 \text{ and } z \leq a_1 - t =: b\ .$$

The integral $\int_0^\infty P(U_{j,k,l} \geq z' 2^{-l/\tilde{k}})2zdz$ can thus be decomposed as follows.

$$\begin{aligned}\int_0^\infty P\left(U_{j,k,l} \geq z' 2^{-l/\tilde{k}}\right)2zdz &= \int_0^b P\left(\left|\sum_t Z_t\right| \geq (z+t)2^{-l/\tilde{k}}n2^{-j/2}\right)2zdz \quad (6.21)\\ &+ \int_b^\infty P\left(\left|\sum_t Z_t\right| \geq (z+t)2^{-l/\tilde{k}}n2^{-j/2}\right)2zdz \quad (6.22)\end{aligned}$$

Naturally, to compute (6.21) (resp. (6.22)), we use the upper bound (6.19)(resp. (6.20)). Let $a_2 := 2^{-l/\tilde{k}}n2^{-j/2}$ and $\sigma_Z^2 = \mathrm{Var}(\sum_t Z_t)$.

[A1]-1.*Order of expression* (6.21)

The integral in (6.21) can be bounded by

$$\int_0^{a_1-t} 4\exp\left(\frac{-((z+t)a_2)^2}{4\sigma_Z^2}\right)zdz\ . \tag{6.23}$$

Performing the change of variable $u = (z+t)/\sigma_Z$, the last expression is equal to

$$\begin{aligned}
&\int_{t/\sigma_Z}^{a_1/\sigma_Z} 4\exp\left(\frac{-u^2 a_2^2}{4}\right)(\sigma_Z u - t)\sigma_Z du \\
\leq\ & 4\sigma_Z^2 \int_{t/\sigma_Z}^{\infty} \exp\left(\frac{-u^2 a_2^2}{4}\right) u du - t\int_{t/\sigma_Z}^{a_1/\sigma_Z} 4\exp\left(\frac{-u^2 a_2^2}{4}\right)\sigma_Z du \qquad (6.24)\\
\leq\ & 4\sigma_Z^2 \int_{t/\sigma_Z}^{\infty} \left[\exp\left(\frac{-u^2 a_2^2}{4}\right)\left(-\frac{2}{a_2^2}\right)\right]' du \\
=\ & 4\sigma_Z^2 \frac{2}{a_2^2}\exp\left(-\frac{t^2 a_2^2}{4\sigma_Z^2}\right) , \qquad (6.25)
\end{aligned}$$

since the second term in (6.24) is negative and thus can be neglected.

It remains to find the order of (6.25). Let $I^\circ_{j_1,k_1}$ be the largest interval used to approximate our random interval of interest $\mathcal{I}$. Following (4.6), we use as threshold $t_{jk} = c_\lambda 2^{(j-j_1)/2}\sigma_\epsilon \log(n)/\sqrt{n}$ (indeed, we consider only the interval $\mathcal{I} = I_{j+1,2k}$, the interval $I_{j+1,2k+1}$ being treated in a completely similar manner). The expression (6.25) is equal to

$$\begin{aligned}
& 8\operatorname{Var}(\sum_t Z_t)\frac{2^{2l/\tilde{k}}}{n^2}2^j \exp\left(-\frac{(c_\lambda^2 2^{j-j_1}\sigma_\epsilon^2 \log(n)^2)2^{-2l/\tilde{k}} n^2 2^{-j}}{n\operatorname{Var}(\sum_t Z_t)}\right)\\
=\ & 8\operatorname{Var}(\sum_t Z_t)\frac{2^{2l/\tilde{k}}}{n^2}2^j \exp\left(-\frac{c_\lambda^2 2^{-j_1}\sigma_\epsilon^2 \log(n)^2 2^{-2l/\tilde{k}} n}{\operatorname{Var}(\sum_t Z_t)}\right) \qquad (6.26)
\end{aligned}$$

Recalling that $Z_t = \epsilon_t 1^\circ_{j_l,k_l}(X_{t-1})$ and that the variance $\operatorname{Var}(\sum_t Z_t)$ is bounded by $[\tilde{c}_\lambda \log(n)]\sigma_\epsilon^2 n 2^{-j_l}$, we can bound (6.26) from above by

$$\frac{\tilde{c}_\lambda \log(n)\sigma_\epsilon^2 2^{j-j_l}2^{2l/\tilde{k}}}{n}\exp\left(-c_\lambda^2 2^{j_l-j_1}\log(n)2^{-2l/\tilde{k}}\right) \qquad (6.27)$$

Note first that the exponential term is $O(1)$. By assumption (A3), we have that $2^j \sim 2^{j_l}$. In order to have the desired order, the term $2^{2l/\tilde{k}}$ must be at most of order $\log(n)^{1/r}$ for a given $r \geq 1$. This means that the number of terms at coarse scales used to approximate $\mathcal{I}$ does not grow faster than

$$d_n = O\left(\frac{\tilde{k}}{2}\log_2\left((\log n)^{1/r}\right)\right), \text{ or } d_n = O\left(\log\left((\log n)^{1/r}\right)\right) ,$$

as it is given in the assumptions of the present lemma. Finally, we have that (6.21) is bounded by

$$\frac{c_\lambda \log(n)\log(n)^{1/r}}{n}O(1) = O\left(\frac{\log(n)^{1+1/r}}{n}\right) \qquad (6.28)$$

[A1]-2. *Order of expression* (6.22)

Consider the change of variable $u := z+t$. Using the inequality (6.20), the expression (6.22) is bounded by

$$\begin{aligned}
& \int_{a_1}^{\infty} 4\exp\left(-\frac{3ua_2}{4n^\delta}\right)(u-t)du \\
=\ & \int_{a_1}^{\infty} 4\exp\left(-\frac{3ua_2}{4n^\delta}\right)u du - t\int_{a_1}^{\infty} 4\exp\left(-\frac{3ua_2}{4n^\delta}\right)du . \qquad (6.29)
\end{aligned}$$

Again, the second term in (6.29) is negative, and thus we look only at the first term. Let $a_4 := (3a_2)/4n^{\delta}$. Then the quantity given in (6.29) is smaller than

$$\int_{a_1}^{\infty} 4\exp(-a_4 u)udu = 4\frac{\exp(-a_4 a_1)}{a_4}(a_1 + a_4^{-1}) \tag{6.30}$$

Recall now that $a_1 := 3\operatorname{Var}\left(\sum_t Z_t\right)(n^{-\delta})2^{l/\tilde{k}}n^{-1}2^{j/2}$ and $a_4 = (3/4)2^{-l/\tilde{k}}n2^{-j/2}n^{-\delta}$. Then

$$a_1 a_4 = \frac{9}{4}n^{-2\delta}\operatorname{Var}(\sum_t Z_t)\ ; \quad \frac{a_1}{a_4} = 4\operatorname{Var}(\sum_t Z_t)2^{2l/\tilde{k}}2^j n^{-2}$$

and $a_4^{-2} = c2^{2l/\tilde{k}}n^{-2}2^j n^{2\delta}$. Consider the two following cases.

1. $\operatorname{Var}(\sum_t Z_t) \geq n^{2\delta} + \epsilon$ for $\epsilon > 0$. Then $\exp(-a_1 a_4) \leq \exp(-\frac{9}{4}n^{\epsilon})$, which gives an exponential decay of (6.30), hence this term is negligible compared to (6.28).
2. $\operatorname{Var}(\sum_t Z_t) \leq n^{2\delta}$. In that case, $\exp(-a_1 a_4) = O(1)$. Since $(j,k) \in \mathcal{J}_n$, $2^j \leq cn^{1-\alpha}$. Then

$$\begin{aligned} \frac{a_1}{a_4} &\leq c\operatorname{Var}(\sum_t Z_t)\log(n)^{1/r}n^{1-\alpha}n^{-2} \leq cn^{2\delta}(\log n)^{1/r}n^{-\alpha}n^{-1} \leq c\frac{\log(n)^{1/r}}{n^{1+\alpha/2}} \\ a_4^{-2} &\leq \left(\frac{4}{3}^2\right)(\log n)^{1/r}n^{-2}n^{1-\alpha}n^{2\delta} \leq c\frac{(\log n)^{1/r}}{n^{1+\alpha/2}}\ , \end{aligned}$$

by the condition (4.5) on $\mathcal{J}_n$.

It follows that the term (6.22) is of order $O(\frac{\log(n)^{1/r}}{n^{1+\alpha/2}})$, which is of smaller order than (6.28). Hence the expression (6.28) gives the result (6.17).

Proof of [A2]

We now show the inequality (6.18) involving the fine-scale terms $U_{j,l_i,d_n}, i = 1,2$. Recall first that

$$P\left(U_{j,l_1,d_n} \geq z'2^{-d_n/\tilde{k}}\right) = P\left(\frac{2^{j/2}}{n}\sum_{t=1}^{n}|\epsilon_t|1^{\circ}_{j^*,l_1}(X_{t-1}) \geq (z+t)2^{-d_n/\tilde{k}}\right) \tag{6.31}$$

On the probability space Ω_n defined in (3.2), $\sup_t|\epsilon_t| \leq n^{\delta}$. Similarly to what we did in the proof of Lemma 6.3, we introduce the random variable $W_t := 1^{\circ}_{j^*,k}(X_{t-1}) - 2^{-j^*}$, which has a zero mean and for which $\operatorname{Var}(W_t) = 2^{-j^*}(1-2^{-j^*})$. Then (6.31) is smaller or equal to

$$\begin{aligned} & P\left(\sum_{t=1}^{n}1^{\circ}_{j^*,l_1}(X_{t-1}) \geq (z+t)2^{-d_n/\tilde{k}}n2^{-j/2}n^{-\delta}\right) \\ = \ & P\left(\sum_{t=1}^{n}(1^{\circ}_{j^*,l_1}(X_{t-1}) - 2^{-j^*}) \geq (z+t)2^{-d_n/\tilde{k}}n2^{-j/2}n^{-\delta} - n2^{-j^*}\right) \\ = \ & P\left(\sum_{t=1}^{n}W_t > x\right), \end{aligned}$$

where

$$x = (z+t)2^{-d_n/\tilde{k}}n2^{-j/2}n^{-\delta} - n2^{-j^*}\ . \tag{6.32}$$

Let $W := \sum_{t=1}^{n} W_t$. Paralleling the proof of [A1], we use Proposition 2 in [15] to build independent blocks of data of length $[c_\lambda \log(n)]$ which have with high probability the same distribution as the original blocks of data. This allows us to bound $\mathrm{Var}(W)$ as $\mathrm{Var}(W) \leq c_\lambda \log(n) n 2^{-j^*}$, and also to use Bernstein's inequality on the independent blocks denoted by $\mathcal{B}_i$. Here we have $|\sum_{t\in\mathcal{B}_i} W_t| \leq K_n = O(\log n)$. With $\sum_t Z_t$ replaced by $\sum_t W_t$ and x given by (6.32), the inequalities (6.19) and (6.20) are still valid.

If $x < 3\,\mathrm{Var}(W) K_n^{-1}$, that is, if

$$\begin{aligned} &(z+t)2^{-d_n/\tilde{k}} n 2^{-j/2} n^{-\delta} - n2^{-j^*} < 3\,\mathrm{Var}(W)K_n^{-1}, \text{ then} \\ &\quad z+t < \tilde{a}_1 := (3\,\mathrm{Var}(W)K_n^{-1} + n2^{-j^*})2^{d_n/\tilde{k}} n^{-1+\delta} 2^{j/2}\ . \end{aligned} \tag{6.33}$$

In order to show (6.18), we decompose $\int_0^\infty P\left(U_{j,l_1,d_n} \geq z' 2^{-d_n/\tilde{k}}\right) 2z dz$ into two parts.

$$\begin{aligned} \int_0^\infty \ & P\left(U_{j,l_1,d_n} \geq z' 2^{-d_n/\tilde{k}}\right) 2zdz \\ \leq \ & \int_0^{\tilde{a}_1 - t} P\left(|\sum_t W_t| \geq (z+t)2^{-d_n/\tilde{k}} n 2^{-j/2} n^{-\delta} - n2^{-j^*}\right) && (6.34) \\ + \ & \int_{\tilde{a}_1 - t}^\infty P\left(|\sum_t W_t| \geq (z+t)2^{-d_n/\tilde{k}} n 2^{-j/2} n^{-\delta} - n2^{-j^*}\right) && (6.35) \end{aligned}$$

[A2]-1 Order of expression (6.34)

We use the Bernstein-type inequality (6.19) to bound (6.34) as

$$\int_0^{\tilde{a}_1 - t} \exp\left(-\frac{[(z+t)2^{-d_n/\tilde{k}} n 2^{-j/2} n^{-\delta} - n2^{-j^*}]^2}{4\,\mathrm{Var}(W)}\right) 2zdz \tag{6.36}$$

Let $\tilde{a}_2 := 2^{-d_n/\tilde{k}} n 2^{-j/2} n^{-\delta}$ and $\sigma_W^2 := \mathrm{Var}(W)$. Using the change of variable $u := [(z+t)\tilde{a}_2 - n2^{-j^*}]/\sigma_W$, the expression (6.36) above is bounded by

$$\begin{aligned} \int_{\frac{t\tilde{a}_2 - n2^{-j^*}}{\sigma_W}}^\infty 4\exp(-u^2/4)\frac{\sigma_W^2 u}{\tilde{a}_2^2} du \ + \ \int_{\frac{t\tilde{a}_2 - n2^{-j^*}}{\sigma_W}}^\infty 4\exp(-u^2/4)\left(\frac{n2^{-j^*}}{\tilde{a}_2} - t\right)\frac{\sigma_W}{\tilde{a}_2} du \\ = E \ + \ F\ . \end{aligned}$$

We first show that the term F is asymptotically negative, that is, $n2^{-j^*}\tilde{a}_2^{-1} = O(\tilde{a}_2^{-1}) = o(t)$, where $t = 2^{(j-j_1)/2}\log(n) n^{-1/2}$ and with $I^\circ_{j_1,k_1}$ the largest interval used to approximate $\mathcal{I}$. We have

$$\begin{aligned} \tilde{a}_2^{-1} &= 2^{d_n/\tilde{k}} n^{-1} 2^{j/2} n^\delta \leq \frac{2^{d_n/\tilde{k}}}{n} n^{1/2-\alpha/2} n^\delta \\ &\leq \frac{2^{d_n/\tilde{k}}}{n^{1/2}} n^{-\alpha/2} n^{\alpha/4} \leq c\frac{(\log n)^{1/(2r)}}{n^{1/2+\alpha/4}}\ , \end{aligned}$$

using the condition (4.5) on the set $\mathcal{J}_n$. Now, by assumption (A3), we have $t = c\frac{\log n}{n^{1/2}}$, and thus $\tilde{a}_2^{-1} = o(t)$.

We now show the order of the term E.

$$\begin{aligned}
\int_{\frac{t\tilde{a}_2 - n2^{-j^*}}{\sigma_W}}^{\infty} \quad & 4\exp(-u^2/4)\frac{\sigma_W^2 u}{\tilde{a}_2^2}du = \frac{\sigma_W^2}{\tilde{a}_2^2}8\exp\left(-\frac{(t\tilde{a}_2 - n2^{-j^*})^2}{4\sigma_W^2}\right) \\
\leq \quad & c\frac{\sigma_W^2}{2^{-2d_n/\tilde{k}}n^2 2^{-j}n^{-2\delta}} \leq c\frac{(\log n)2^{-j^*}n}{2^{-2d_n/\tilde{k}}n^2 2^{-j}n^{-2\delta}}
\end{aligned}$$

By the definition of the set $\mathcal{J}_n$, $2^j \leq cn^{1-\alpha}$ and $2^{j^*} \leq n < 2^{j^*+1}$, thus $2^j/2^{j^*} \leq cn^{-\alpha}$. Moreover, $n^{2\delta} \leq n^{\alpha/2}$, hence $n^{2\delta}2^{j-j^*} \leq cn^{-\alpha/2}$. The term E is thus of order $\frac{\log(n)^{1+1/r}}{n^{1+\alpha/2}}$, which gives us the order of (6.34).

[A2]-2 Order of expression (6.35)

If x, defined in (6.32) is such that $x \leq 3\,\mathrm{Var}(W)K_n^{-1}$, then (6.35) can be bounded by

$$\begin{aligned}
& \int_{\tilde{a}_1 - t}^{\infty} 4\exp\left(\frac{-3[(z+t)2^{-d_n/\tilde{k}}n^{1-\delta}2^{-j/2} - n2^{-j^*}]}{4\log n}\right) zdz \\
= & \int_{\tilde{a}_1}^{\infty} 4\exp\left(\frac{-3u2^{-d_n/\tilde{k}}n^{1-\delta}2^{-j/2}}{4\log n}\right)\exp\left(\frac{n2^{-j^*}}{4\log n}\right) udu \\
- & \ t\int_{\tilde{a}_1}^{\infty} 4\exp\left(\frac{-3(u2^{-d_n/\tilde{k}}n2^{-j/2}n^{-\delta} - n2^{-j^*})}{4\log n}\right) du \\
= & \ G + H \ , \qquad (6.37)
\end{aligned}$$

by making the change of variable $u := z + t$.

First note that, since $n2^{-j^*} = O(1)$, $\exp(n2^{-j^*}/\log n) \to 1$ as $n \to \infty$. The term H in (6.37) is negative and thus can be neglected. We now show the order of the term G. Let $\tilde{a}_4 = 3(2^{-d_n/\tilde{k}}2^{-j/2}n)/(4n^\delta \log n)$. Integrating by part, we obtain that the term G is equal to

$$\frac{\exp(-\tilde{a}_4\tilde{a}_1)}{\tilde{a}_4}(\tilde{a}_1 + \tilde{a}_4^{-1}) \ ,$$

where the value of $\tilde{a}_1$ is given in (6.33). With $K_n = \log n$, we have:

$$\begin{aligned}
\tilde{a}_4\tilde{a}_1 &= \frac{9\,\mathrm{Var}(W)}{4(\log n)^2} + \frac{3n2^{-j^*}}{4\log n} \\
\tilde{a}_1\tilde{a}_4^{-1} &= (3\,\mathrm{Var}(W) + n2^{-j^*}\log n)\frac{(4/3)2^{2d_n/\tilde{k}}2^j n^{2\delta}}{n^2} \\
\tilde{a}_4^{-2} &= \left(\frac{4}{3}\right)^2 2^{2d_n/\tilde{k}}n^{-2+2\delta}2^j\log(n)^2 \leq c(\log n)^{2+1/r}n^{-2}n^{1-\alpha}n^{2\delta} = O\left(\frac{(\log n)^{2+1/r}}{n^{1+\alpha/2}}\right)
\end{aligned} \qquad (6.38)$$

The term $\tilde{a}_4\tilde{a}_1$ tends to zero as $n \to \infty$ since $\mathrm{Var}(W) \leq (\log n)n2^{-j^*} = O(\log n)$. The term $\tilde{a}_1\tilde{a}_4^{-1}$ can be bounded by a quantity of order

$$\begin{aligned}
& \left(n2^{-j^*}\log n + n2^{-j^*}\log n\right)\frac{(\log n)^{1/r}2^j n^{2\delta}}{n^2} \\
\leq \ & 2(\log n)^{1+1/r}\frac{n^{1-\alpha}n^{2\delta}}{n^2} \leq 2\frac{(\log n)^{1+1/r}}{n^{1+\alpha/2}} \ .
\end{aligned}$$

Hence the order of (6.35) is given by (6.38). This expression (6.38) also gives the order of (6.18). □

Acknowledgements

The authors would like to sincerely thank M. Neumann for the many insightful discussions we had with him.

References

[1] K. Berkner and R. Wells. Smoothness estimates for soft-threshold denoising via translation invariant wavelet transforms. *Appl. Comput. Harmon. Anal.*, 12:1–24, 2002.

[2] K. Berkner and R.O. Wells. A correlation-dependent model for denoising via nonorthogonal wavelet transforms. Technical report, Computational Mathematics Laboratory, Rice University, 1998.

[3] T. Bollerslev, R. Chou, and K. Kroner. ARCH modeling in finance. *Journal of Econometrics*, 52:5–59, 1992.

[4] D. Bosq. *Nonparametric Statistics for Stochastic Processes.* Lecture Notes in Statistics 110. Springer-Verlag, 1996.

[5] P. Bühlmann and A.J. McNeil. Nonparametric GARCH models. Technical report, ETH, Zürich, 1999.

[6] J.M. Carnicer, W. Dahmen, and J.M. Peña. Local decomposition of refinable spaces and wavelets. *Appl. Comput. Harmon. Anal.*, 3:127–153, 1996.

[7] A. Cohen, I. Daubechies, P. Vial, and B. Jawerth. Multiresolution analysis, wavelets and fast algorithms on an interval. *Comptes Rendus Acad. Sci. Paris, Serie I*, 316:417–421, 1993.

[8] R. R. Coifman and D.L. Donoho. Translation-Invariant De-Noising. In A. Antoniadis and G. Oppenheim, editors, *Wavelets in Statistics, Lectures Notes in Statistics, Vol. 103.*, pages 125–150. Springer-Verlag, 1995.

[9] J. Davidson. *Stochastic Limit Theory.* Advanced Texts in Econometrics. Oxford University Press, 1994.

[10] V. Delouille, J. Franke, and R. von Sachs. Nonparametric stochastic regression with design-adapted wavelets. *Sankhya: The Indian Journal of Statistics. Special Issue on Wavelets, Series A*, 63(3):328–366, 2001.

[11] V. Delouille, J. Simoens, and R. von Sachs. Smooth design-adapted wavelets for nonparametric stochastic regression. Discussion Paper 0117, Institut de Statistique, UCL, Belgium, 2001.

[12] D.L. Donoho. De-noising via soft-thresholding. *IEEE Transactions on Information Theory*, 41:613–627, 1995.

[13] D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[14] D.L. Donoho and I.M. Johnstone. Adapting to unknown smoothness via wavelet shrinking. *J. Am. Statist. Assoc.*, 90:1200–1224, 1995.

[15] P. Doukhan, P. Massart, and E. Rio. Invariance principles for absolutely regular empirical processes. *Annales de l'Institut Henri Pointcaré-Probabilités et Statistiques*, 31(2):393–427, 1995.

[16] R. Engle. Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica*, 50:987–1008, 1982.

[17] J. Franke, J-P., Kreiss, E. Mammen, and M. Neumann. Properties of the nonparametric autoregressive bootstrap. To appear in *Journal of Time Series Analysis*, 2002.

[18] M. Girardi and W. Sweldens. A new class of unbalanced Haar wavelets that form an unconditional basis for L_p on general measure spaces. *J. Fourier Anal. Appl.*, 3(4):457–474, 1997.

[19] C. Gouriéroux. *ARCH Models and Financial Applications*. Springer Series in Statistics. Springer-Verlag, New York, 1997.

[20] C. Gouriéroux and A. Monfort. Qualitative threshold ARCH models. *Journal of Econometrics*, 52:159–199, 1992.

[21] C. Hafner. *Nonlinear Time Series Analysis with Applications to Foreign Exchange Rate Volatility*. Physica-Verlag, 1998.

[22] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[23] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, Approximation, and Statistical Applications*. Lecture Notes in Statistics 129. Springer-Verlag, 1998.

[24] W. Härdle and A. Tsybakov. Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics*, 81:223–242, 1997.

[25] M. Hoffmann. On nonparametric estimation in nonlinear AR(1)-models. *Statistics and Probability letters*, 44:29–45, 1999.

[26] I.M. Johnstone and B.W. Silverman. Wavelet methods for data with correlated noise. *J. Royal Statist. Soc., Series B*, 59:319–351, 1997.

[27] D.A. Jones. Nonlinear regressive processes. *Proc. Roy. Soc. London Ser. A*, 360:71–95, 1978.

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

[29] G.P. Nason and B.W. Silverman. The stationary wavelet transform and some statistical applications. In A. Antoniadis and G. Oppenheim, editors, *Wavelets in Statistics, Lectures Notes in Statistics, Vol. 103.*, pages 281–299. Springer-Verlag, 1995.

[30] M. Neumann and J.P. Kreiss. Regression-type inference in nonparametric autoregression. *Annals of Statistics*, 26:1570–1613, 1998.

[31] M.H. Neumann and R. von Sachs. Wavelet thresholding: Beyond the Gaussian i.i.d. situation. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics, Lecture Notes in Statistics 103*, pages 301–329. Springer-Verlag, New York, 1995.

[32] P.M. Robinson. Nonparametric estimators for time series. *J. Time. Ser. Anal.*, 4:185–207, 1983.

[33] G.R. Shorack and J.A. Wellner. *Empirical Processes with Applications to Statistics*. John Wiley and Sons, New York, 1986.

[34] W. Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546, 1997.

[35] D. Tjøstheim. Non-linear time series: a selective review. *Scand. J. Statist.*, pages 97–130, 21.

[36] H. Tong. *Threshold Models in Nonlinear Time Series Analysis*. Springer-Verlag, Berlin, 1983.

[37] R. von Sachs and B. MacGibbon. Nonparametric curve estimation by wavelet thresholding with locally stationary errors. *Scandinavian Journal of Statistics*, 27:475–499, 2000.