Ann. Inst. Statist. Math.
Vol. 57, No. 2, 235–253 (2005)

ESTIMATION OF NONLINEAR AUTOREGRESSIVE MODELS USING DESIGN-ADAPTED WAVELETS*

VÉRONIQUE DELOUILLE[1] AND RAINER VON SACHS[2]

[1]*Solar Physics Department, Royal Observatory of Belgium, Avenue Circulaire 3, B-1180 Brussels, Belgium,* e-mail: veronique.delouille@oma.be

[2]*Institut de statistique, Université catholique de Louvain, Voie du Roman Pays, 20, B-1348 Louvain-la-Neuve, Belgium,* e-mail: vonsachs@stat.ucl.ac.be

(Received January 28, 2003; revised June 11, 2004)

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.

Key words and phrases: Autoregressive design, β-mixing conditions, ARCH models, 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 (1978). Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{X_t\}$ a random process associated to it. We observe the time series $X_0, X_1, \ldots, X_n$ that follow the nonlinear autoregression model

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

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, e.g. Franke *et al.* (2002*b*), Härdle and Tsybakov (1997), Masry and Tjøstheim (1995), Hafner (1998*b*), Robinson (1983), Tjøstheim (1994), Bühlmann and McNeil (2002). Very little is known however about *wavelet* estimators for autoregressive designs. The only existing method of Hoffmann (1999) that treats autoregressive models using a wavelet estimator is concerned with asymptotical results only, and does not provide an efficient algorithm in practice. It is well known (Donoho and Johnstone (1998)) that, in the case of fixed, equispaced designs, wavelet methods enjoy minimax properties

*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.

for functions with inhomogeneous smoothness, like Besov or piecewise smooth functions. Our goal here is to provide a wavelet method for *autoregressive* designs that permits to denoise data that follows (1.1) even if the function m is only piecewise continuous.

Such type of nonlinear autoregressive model, where m might be discontinuous, may arise when modelling econometric and financial data. Volatility in financial models of autoregressive type for exchange or stock index data often shows some asymmetries (Hafner (1998*a*)) or even a behaviour that calls for threshold models (Gouriéroux and Monfort (1992), Gouriéroux (1997)). Chen *et al.* (2004) show that using nonlinear models proves more flexible for multi-step ahead prediction than traditional linear autoregressive models.

The aim of this paper is to derive theoretical properties of a newly proposed algorithm for treating models of the form (1.1). This algorithm is based on the design-adapted wavelet estimators proposed in Delouille *et al.* (2001, 2004).

In Delouille *et al.* (2001), we derive a so-called 'design-adapted' wavelet estimator based on the unbalanced Haar basis in the stochastic regression context. In Delouille *et al.* (2004), we build with the *lifting scheme* (Carnicer *et al.* (1996), Sweldens (1997)) a *biorthogonal* design-adapted wavelet basis with a higher regularity than the unbalanced Haar basis, that still takes into account the irregularity of the design, i.e. no preprocessing of the data is necessary. Moreover, this wavelet transform handles non-dyadic sample sizes in a natural way, and no particular treatment of the boundary is necessary.

Other methods for treating fixed nonequispaced designs with wavelets have been proposed in the literature, but almost all of them use some preliminary step (such as binning, interpolating, or projecting) to get back to the equidistant design situation and then apply the traditional discrete orthogonal wavelet transform algorithm (Cai and Brown (1998), Hall and Turlach (1997), Kovac and Silverman (2000), Delyon and Juditsky (1997), Antoniadis and Pham (1998), Antoniadis and Fan (2001)). None of these methods have been transposed in theory or in practice to the autoregressive setting. As we face in time series a tendency to more extreme values and thus inhomogeneous designs, a preprocessing step such as binning would not give satisfactory results at all.

In this paper, we use the wavelet-type algorithm developed in Delouille *et al.* (2004) and show two properties of the resulting design-adapted wavelet estimators for autoregressive models such as (1.1):

1. We derive a thresholding scheme, tailored to the time series setting, that permits one to remove with a probability tending asymptotically to one, all the noise in the wavelet coefficients. This threshold is of order $\log n$, as opposed to the order $\sqrt{\log n}$ for the threshold obtained by Donobo (1995) in the classical setting of fixed, equispaced designs.

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.

With these theoretical results, we also cover the stochastic regression model we investigated in Delouille *et al.* (2001, 2004) without deriving any theory on it.

The above two results concern the behaviour of the wavelet coefficients. Since the lifting scheme produces biorthogonal (as opposed to orthogonal) wavelet transforms, a transfer of these results to the estimator in the time domain is not directly possible. However, having controlled the l_2-risk in the wavelet coefficient domain gives sufficient insight about the good denoising properties of the resulting estimator, in particular for a finite sample size (Simoens and Vandewalle (2003)).

This paper is organized as follows. Section 2 presents our main methodology, together with practical applications on real and simulated data sets, including data sets with conditional heteroscedasticity, that is *dependent* errors. Section 3 contains our theoretical achievements: we give the assumptions on our model and derive the two above-mentioned properties of our threshold wavelet estimator. Finally, Section 4 contains the outline of our proofs. More details on those can be found in Delouille and von Sachs (2004).

2. Main methodology and practical application

2.1 *Design-adapted wavelet transform*

An orthogonal wavelet basis consists of a set $\{\varphi_{j_0,k}\}_{k=1}^{2^{j_0}}$ of scaling functions, and of a set of wavelet functions $\{\psi_{j,k}\}_{j\geq j_0,k=1,\ldots,2^j}$, where at each scale $j \geq j_0$, $\psi_{j,k}$ is orthogonal to $\varphi_{j,k'}$, for all k, k'. A biorthogonal basis consists of two parts: a dual basis $\{\tilde{\varphi}_{j_0,k}\}_k \cup \{\tilde{\psi}_{j,k}\}_{j\geq j_0,k}$ used for analysis (decomposition), and a primal basis $\{\varphi_{j_0,k}\}_k \cup \{\psi_{j,k}\}_{j\geq j_0,k}$ that reconstructs the function. The dual and primal basis are linked through biorthogonal relationships (Cohen *et al.* (1992)). For any function $m \in L_2(\mathbb{R})$, we have:

$$m(x) = \sum_k \langle m, \tilde{\varphi}_{j_0,k}\rangle \varphi_{j_0,k}(x) + \sum_{j=j_0}^{\infty}\sum_k \langle m, \tilde{\psi}_{jk}\rangle \psi_{jk}(x).$$

In this paper, we use particular *design-adapted wavelets* (DAW), i.e. wavelets that automatically adapt to an irregular design. Let F_x denote the stationary distribution of the time series $\{X_t\}$ and let $\hat{F}_n$ be its empirical distribution function. DAW are wavelets which are (bi-)orthogonal in the space $\boldsymbol{L}_2(d\hat{F}_n)$ weighted by the empirical measure $\hat{F}_n$.

To build DAW, we need an adequate partitioning (Girardi and Sweldens (1997)) of the interval containing the data. When the sample size n is a power of two, this partitioning is given by the set of random intervals

$$I_{jk} = [X_{((k-1)l_j+1)}, X_{(kl_j+1)}), k = 1, \ldots, 2^j, l_j = n2^{-j}, j_0 \leq j \leq J = \log_2(n), \tag{2.1}$$

where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are the order statistics, and, as a convention, $X_{(n+1)} := X_{(n)} + (X_{(n)} - X_{(1)})/n$. When n is not a power of two, the construction of the empirical quantile partitioning $\{I_{j,k}\}$ is generalized as follows (Sweldens (1997)). If at a given level $j+1$ we have an odd number $2p+1$ of intervals, then the first $2p$ intervals will produce p intervals at the next level j. The last interval $I_{j+1,2p+1}$ will be passed unchanged to the level j. In that case, $J := \lceil \log_2(n) \rceil$.

We now give two examples of DAW.

1. The *unbalanced Haar* basis described in Delouille *et al.* (2001) is orthogonal in $\boldsymbol{L}_2(d\hat{F}_n)$. There

$$\varphi_{jk} = \tilde{\varphi}_{jk} = 2^{j/2}1_{jk}; \psi_{jk} = \tilde{\psi}_{jk} = 2^{j/2}(1_{j+1,2k+1} - 1_{j+1,2k}), \tag{2.2}$$

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

2. The *smooth biorthogonal DAW* basis is built as follows. Starting from the unbalanced Haar basis in $\boldsymbol{L}_2(d\hat{F}_n)$ given by (2.2), an average-interpolating prediction step in $\boldsymbol{L}_2(d\hat{F}_n)$ is performed (Sweldens and Schröder (1996), Delouille *et al.* (2001)). This increases the number $\tilde{N}$ of vanishing moments of the analyzing wavelet $\tilde{\psi}_{jk}$, which leads to a *smooth reconstruction*. We call this estimator 'weighted average-interpolating' (WAI).

It is then possible to add an *update step* which performs a local semi-orthogonalization (as in Delouille *et al.* (2004), Simoens and Vandewalle (2003)), the resulting estimator is denoted 'WAI-U'. The WAI analyzing wavelets are equal to:

$$\tilde{\psi}_{jk} = 2^{(j+1)/2} \sum_{l:\tilde{g}_{jlk}\neq 0} \tilde{g}_{jlk} 1_{j+1,l}, \tag{2.3}$$

where the values $\tilde{g}_{jlk}$ are the refinements coefficients, chosen to ensure that the analyzing wavelet $\tilde{\psi}_{jk}$ has $\tilde{N} > 1$ dual (analyzing) vanishing moments in $\boldsymbol{L}_2(d\hat{F}_n)$. The values $\tilde{g}_{jlk}$ in (2.3) are uniformly bounded as soon as the values X_t belong to a compact interval. By construction, the number of terms in the sum (2.3) is uniformly bounded in j, k (Delouille *et al.* (2004)):

$$\sup_{j,k} \#\{l : \tilde{g}_{jlk} \neq 0\} =: K < \infty. \tag{2.4}$$

The WAI-U wavelets are no more piecewise constant, but the number of dual vanishing moments acquired during the average-interpolation is preserved. Hence, to prove the results announced in Section 1, it is sufficient to consider the WAI estimator.

2.2 *Design-adapted wavelet estimator*

Consider the autoregressive model (1.1) and the analyzing wavelets in (2.3). The empirical wavelet coefficients $\hat{d}_{jk}$ are computed from the data $\{X_t\}_{t=0}^{n-1}$ as follows:

$$\begin{aligned} \hat{d}_{j,k} &= d_{jk} + \rho_{jk} \\ &= \frac{1}{n}\sum_{t=1}^{n} m(X_{t-1})\tilde{\psi}_{jk}(X_{t-1}) + \frac{1}{n}\sum_{t=1}^{n} \epsilon_t \tilde{\psi}_{jk}(X_{t-1}), \end{aligned} \tag{2.5}$$

where $\hat{d}_{jk}$, d_{jk}, and ρ_{jk} are functions of n. The variance of d_{jk} is due to the stochastic nature of the design points $\{X_t\}_{t=0}^{n-1}$ and is not considered when removing the noise (innovation) part by thresholding. Indeed, only the randomness that comes from the innovations parts needs to be taken into account when denoising the data.

Denoting by $\hat{d}_{jk}^t$ the thresholded detail coefficient (see equation (3.9)), the wavelet estimator is

$$\hat{m}(x) = \sum_k \hat{s}_{j_0,k}\varphi_{j_0,k}(x) + \sum_{j,k|(j,k)\in\mathcal{J}_n, j\geq j_0} \hat{d}_{jk}^t \psi_{jk}(x), \tag{2.6}$$

where $\hat{s}_{j_0,k} := n^{-1}\sum_{t=1}^{n} X_t \tilde{\varphi}_{j_0,k}(X_{t-1})$ are the empirical scaling coefficients, and following Neumann and von Sachs (1995), Delouille *et al.* (2001), only the detail coefficients whose indices belong to

$$\mathcal{J}_n = \{(j,k) \mid 2^j \leq Cn^{1-\alpha}; 1 \leq k \leq 2^j\}, \qquad 0 < \alpha < 1,$$

are taken into account. The fine detail coefficients $d_{jk} \notin \mathcal{J}_n$ are put equal to zero.

2.3 *Practical applications*

Before stating our theoretical results, we show in practice the usefulness of WAI and WAI-U methods for estimation of nonlinear autoregressive process. Let $\mathcal{F}_t = \mathcal{F}_{-\infty}^t$ denote the σ-field generated by the stationary time-homogeneous process $\{X_s, s \leq t\}$. We first introduce algorithms for estimations, and next we give some practical examples on simulated data sets and on the 1997 CAC40 index series.

2.3.1 *Autoregressive (AR) models*

Consider the general AR model:

$$X_t = m(X_{t-1}) + \sigma(X_{t-1})\epsilon_t \quad t = 0, \ldots, n, \epsilon_t \sim \text{i.i.d. with } E(\epsilon_t) = 0, E(\epsilon_t^2) = 1. \tag{2.7}$$

In case of a constant function $\sigma(X_{t-1}) = \sigma_\epsilon$ we find back the model (1.1). To estimate (2.7), we modify the three-step algorithm of Delouille *et al.* (2004) in order to deal with the large variability (or low signal-to-noise ratio) inherent to time series data.

1. Let $J := \lfloor \log_2(n) \rfloor$ be the finest resolution level. Compute a linear design-adapted wavelet estimator (based on WAI) with a cutting level j^* equal to $j^* = J - 1$. Call $\tilde{m}(x)$ the resulting estimate and let $\tilde{X}_t := \tilde{m}(X_{t-1})$.

2. From the data set $(X_{t-1}, \tilde{X}_t)_{t=1}^n$, 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 $\sigma^2(\cdot)$ from the data set (X_{t-1}, r_t^2), $t = 1, \ldots, n$ using a *robust* linear Haar or WAI/WAI-U estimators. 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. 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.

We use the threshold $t_{jk} = \hat{\sigma}_{jk}\sqrt{2(1+\delta_{j_0})\log n}$, where δ_{j_0} is a correction factor for the correlation between wavelet coefficients induced by the biorthogonality of the transform (Berkner and Wells (1998)). δ_{j_0} is computed as the maximum of the cross-correlations between the detail coefficients. Theorem 3.1 gives the factor $\log n$ for time series. Compared to regression case, there is an additional factor of $\sqrt{\log n}$, which is introduced to account for the dependence between the design points $\{X_t\}_{t=0}^{n-1}$.

2.3.2 *AutoRegressive Conditionally Heteroscedastic (ARCH) models*

The ARCH(1) model with no trend is traditionally written as:

$$X_t = \sigma_t \epsilon_t; \qquad \sigma_t^2 = s(X_{t-1}), \tag{2.8}$$

where ϵ_t are i.i.d. with $E(\epsilon_t) = 0$, $E(\epsilon_t^2) = 1$, and $\sigma_t := \mathrm{Var}(X_t \mid \mathcal{F}_{t-1})$. When a trend is present in the data, we call the model AR-ARCH:

$$X_t = m(X_{t-1}) + u_t; \qquad u_t = \sigma_t \epsilon_t$$

where ϵ_t are as above, and $E(u_t^2 \mid \mathcal{F}_{t-1}) = \sigma_t^2 = s(u_{t-1})$; $u_{t-1} := X_{t-1} - m(X_{t-2})$. The aim here is to estimate both the trend $m(\cdot)$ and the variance function $s(\cdot)$.

Since the wavelet denoising works only for additive noise, we need to rewrite the multiplicative ARCH model (2.8) as

$$X_t^2 = s(X_{t-1}) + V_t; \qquad V_t = s(X_{t-1})(\epsilon_t^2 - 1). \tag{2.9}$$

Proposition 5.1 in Delouille and von Sachs (2004) shows that $E(V_t) = 0$ and $\mathrm{Cov}(V_s, V_t) = 0$. This entails

$$E[X_t^2 \mid \mathcal{F}_{t-1}] = s(X_{t-1}); \mathrm{Var}[X_t^2 \mid \mathcal{F}_{t-1}] = s^2(X_{t-1})(E[\epsilon_t^4] - 1).$$

In case of an ARCH model, we thus obtain a *nonlinear* wavelet estimate of s from the data set (X_t, X_t^2) by using the four step algorithm described above.

In case of an AR-ARCH model, $s(\cdot)$ is function of the lagged residuals $u_{t-1} = X_{t-1} - m(X_{t-2})$. For this reason, we modify steps 3 and 4 in the algorithm of Subsection 2.3.1 as follows:

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

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

2.3.3 *Simulated examples*

We now show on some simulated examples the performance of the above described algorithms. We use as a performance criterion:

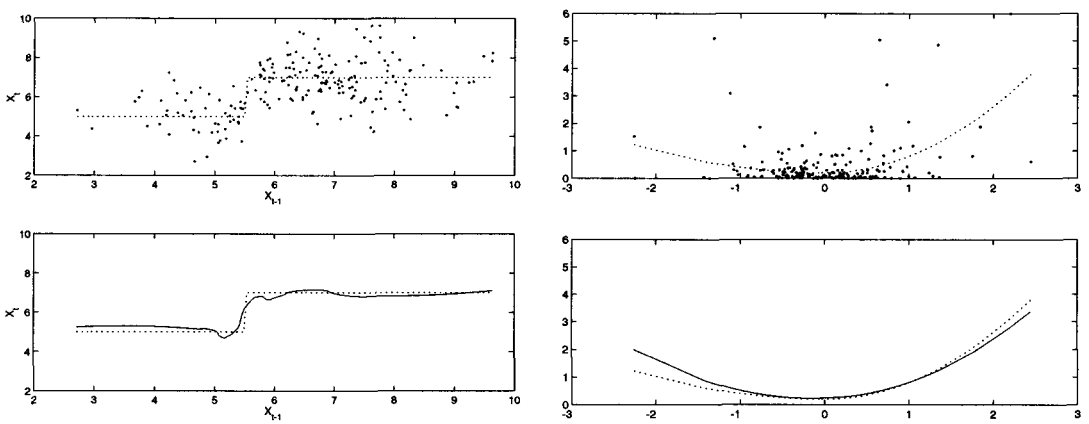
$$\text{RMSE} = \sqrt{\frac{1}{n-1}\sum_{t=2}^{n}(\hat{f}(X_{t-1}) - f(X_{t-1}))^2},$$

where the function f represents either the conditional expectation (trend) or the variance function. To simulate stationary data, we use a burn-in period of 100 observations. After the burn-in period, we take a sample of size $n = 200$. We use a hard thresholding rule and take a primary resolution level $j_0 := 2$.

AR model. Consider the piecewise constant model

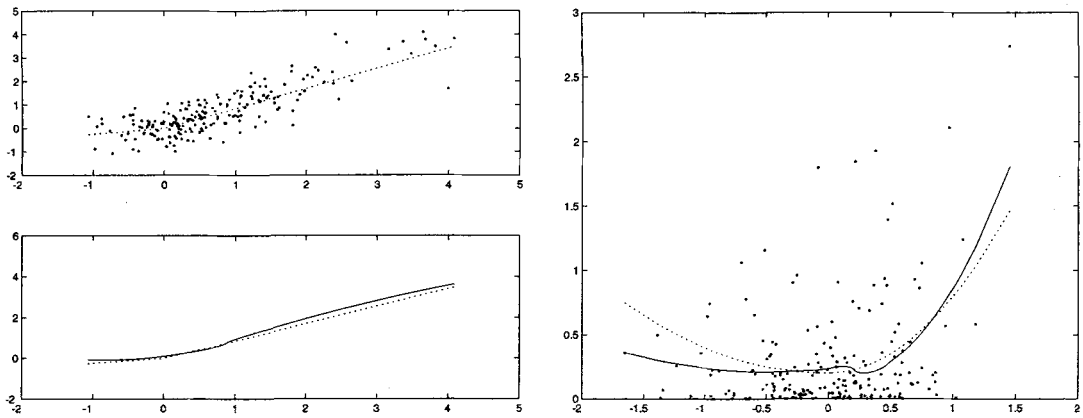
$$X_t = 5 1_{\{X_{t-1} \le 5.5\}}(X_{t-1}) + 7 1_{\{X_{t-1} > 5.5\}}(X_{t-1}) + \epsilon_t, \qquad \epsilon_t \sim N(0,1), \tag{2.10}$$

which is a type of so-called *threshold* models, see for example Gouriéroux and Monfort (1992), Tong (1983). Figure 1(a) shows an estimation of the model (2.10) using the WAI-U estimator in the algorithm of Subsection 2.3.1.



(a) (b)

Fig. 1. Data are represented by dots ($n = 200$), the underlying signal by a dotted line and the estimator by a plain line. (a) Estimation of the piecewise constant AR model using the WAI-U estimator, RMSE = 0.2249. (b) Estimation of the variance function in an ARCH model with the WAI estimator, RMSE = 0.1042.



(a) (b)

Fig. 2. 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.

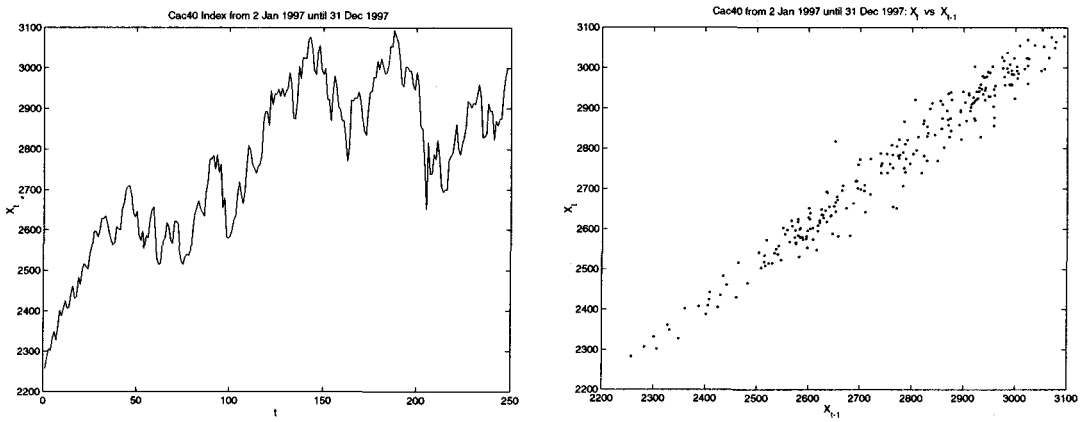
ARCH-type models. We consider the complete AR-ARCH model

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

as well as the ARCH model with no trend. Figure 1(b) gives the estimation of the variance function in an ARCH model with no trend, whereas Fig. 2 represents the estimation of the trend and volatility function in the full AR-ARCH model. 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.

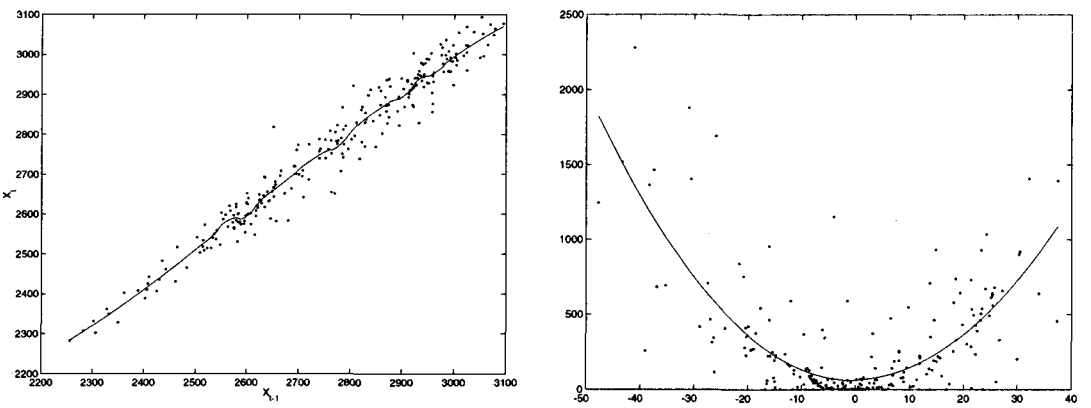
Application on the CAC 40 index. We now apply the proposed methodology on the index (called 'CAC40') of the Paris' Stock Exchange market. Figure 3(a) represents the series $\{X_t\}$ of the CAC40 from January 2, 1997 until December 31, 1997, containing 250 observations. Since we want to model $\{X_t\}$ by a nonlinear AR-process of order one, we first represent X_t as a function of X_{t-1} in Fig. 3(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 (see, e.g., Engle (1982), Gouriéroux (1997), Hafner (1998*b*)). Hence we applied the algorithm for AR-ARCH models on these data, using the WAI-U wavelet estimator. Figure 4(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 4(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 (Gouriéroux (1997), Hafner (1998*b*)).



(a) (b)

Fig. 3. 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)

Fig. 4. 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 $\hat{u}_{t-1} = X_{t-1} - \hat{m}_0(X_{t-2})$.

3. Theoretical results

3.1 *Assumptions and methodology*

Consider the autoregressive model (1.1), where the errors are i.i.d. but not necessarily Gaussian. In order to generalize nonparametric regression results for dependent data, we employ 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, we use the β-mixing

condition.

DEFINITION 3.1. Let $(\Omega, \mathcal{F}, P)$ be a probability space, with $\mathcal{G}$ and $\mathcal{H}$ two σ-subfields of $\mathcal{F}$. The β-mixing coefficients for the process $\{X_t\}$ are defined as follows (Bosq (1996), Davidson (1994)):

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

where $P(H \mid \mathcal{G})$ is the probability of the event H conditioned on the σ-subfield of $\mathcal{G}$.

For a sequence $\{X_t\}_{-\infty}^{+\infty}$, let $\mathcal{F}_{-\infty}^{t} := \sigma(X_{-\infty}, \ldots, X_{t-2}, X_{t-1}, X_t)$ be the σ-field generated by $(X_{-\infty}, \ldots, X_{t-2}, X_{t-1}, X_t)$ 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 *β-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 four assumptions on the time series $\{X_t\}_{t=0}^{n}$ coming from the model (1.1).

(A1) $\{X_t\}$ 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, and to have a density defined on a compact interval. Furthermore, we assume absolute regularity (that is, β-mixing) for $\{X_t\}$ and that the β-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 Neumann and Kreiss (1998), Franke *et al.* (2002*c*), 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

$$P\left(\sup_{t=1,\ldots,n} |\epsilon_t| > c_\lambda n^{\delta'}\right) \leq c n^{-\lambda} \qquad \forall \delta' \in (0, \delta) \tag{3.1}$$

where c_λ depends on λ, and c represents, here and in the sequel, a generic constant.

(A3) The set $\mathcal{J}_n$ used in the summation in equation (2.6) must be such that

$$\mathcal{J}_n = \{(j,k) \mid 2^j \leq C n^{1-\alpha}; 1 \leq k \leq 2^j\} \quad \text{with} \quad 4\delta < \alpha < 1, \tag{3.2}$$

where δ is defined in (3.1).

The theoretical difficulty that arises when using design-adapted wavelets to treat autoregressive models is that the intervals used to build the wavelet functions are random, and in the time series context it is not possible to condition on the values of the regressors. To circumvent this problem we follow Neumann and Kreiss (1998) and approximate a random interval $\mathcal{I} := I_{jk}$ by a *union of fixed intervals* coming from the theoretical quantile partitioning $\boldsymbol{I}^\circ = \{I_{jk}^\circ\}_{(j,k)}$ where

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

Beginning from the coarsest scale, we approximate $\mathcal{I}$ by a union of as large as possible intervals belonging to $\boldsymbol{I}^\circ$. There exist indices $(j_1, k_1), \ldots, (j_d, k_d)$ such that

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

where $j_1 < \cdots < j_{d-1} \leq j_d \leq j^*$. 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, d will increase. To emphasize these facts, we use the notation that $d := d_n(\mathcal{I})$.

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:

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

Finally, we need to relate the scale j of a random interval $\mathcal{I} \equiv I_{jk}$ to the scale j_l of the fixed intervals $I^\circ_{j_l,k_l}$ present in the approximation (3.4).

(A4) $\forall (j,k) \in \mathcal{J}_n$, there exist finite constants c_1, c_2 such that, if $I^\circ_{j_l,k_l} \subset I_{j,k}$ for a given $l \in \{1, 2, \ldots, d_n(I_{j,k})\}$, then $c_2 \leq 2^j/2^{j_l} \leq c_1$, or, in short, $2^j \sim 2^{j_l}$.

3.2 *Threshold that removes the noise with a high probability*

Consider the expression of the noise term ρ_{jk} for a WAI estimator, where the values $|\tilde{g}_{jlk}|$ are uniformly bounded by Assumption (A1):

$$\rho_{jk} = \frac{2^{(j+1)/2}}{n} \left(\sum_{t=1}^{n} \left(\sum_{l:\tilde{g}_{jlk} \neq 0} \tilde{g}_{jlk} \epsilon_t 1_{j+1,l}(X_{t-1}) \right) \right). \tag{3.6}$$

There is a correlation between the ρ_{jk} induced by the correlation between the regressors $\{X_t\}_{t=0}^{n-1}$. In this case, the classical universal threshold of order $O(\sqrt{\log(n)})$ does not guarantee that the empirical wavelet coefficients become asymptotically noise free. Instead, we need a threshold of order $O(\log n)$.

THEOREM 3.1. *Assume* (A1)–(A4). *Consider the WAI detail coefficient* $\hat{d}_{jk} = d_{jk} + \rho_{jk}$ *as in* (2.5), *where* $(j,k) \in \mathcal{J}_n$ *as defined in* (3.2). *For each* l *such that* $\tilde{g}_{jlk} \neq 0$, *let* $I^\circ_{j_l,k_l}$ *be the largest interval that is included in the random interval* $I_{j+1,l}$ *in a decomposition similar to* (3.5), *and let* $d_n(I_{j+1,l})$ *be the number of coarse-scale intervals in the decomposition* (3.5) *of the interval* $I_{j+1,l}$. *Assume that these* $d_n(I_{j+1,l})$ *are* $O(\log n)$. *Then we have*

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

with

$$t_{jk} = c_\lambda \sum_{l:\tilde{g}_{jlk} \neq 0} \left(2^{(j-j_l)/2} \frac{\sigma_\epsilon}{\sqrt{n}} \log n \right), \tag{3.7}$$

where the constant c_λ *depends only on* λ.

In the situation where all the j_l, $l = 1, \ldots, K$ are equal to j in (3.7), the threshold t_{jk} is of order $\frac{\sigma_\epsilon}{\sqrt{n}} \log n$ as opposed to $t = \frac{\sigma_\epsilon}{\sqrt{n}} \sqrt{2 \log n}$ in the regression with fixed, equispaced design case (Donoho (1995), Donoho and Johnstone (1994)).

Constructing t_{jk} for all $(j,k) \in \mathcal{J}_n$ and defining

$$t(n) := \max_{(j,k)\in\mathcal{J}_n} t_{jk}, \tag{3.8}$$

we have, using the Bonferroni inequality, that

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

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

Recall now the definition of the soft-threshold nonlinearity:

$$\eta_S(d,t) = \text{sign}(d)(|d| - t)_+, \tag{3.9}$$

and let $\hat{d}^t_{jk} := \eta_S(\hat{d}_{jk}, t(n))$ with $t(n)$ given by (3.7)–(3.8). By construction of $t(n)$, we have:

$$\pi_n \equiv P\left(\max_{(j,k)}|\rho_{jk}| \leq t(n)\right) \to 1 \quad \text{as} \quad n \to \infty. \tag{3.10}$$

Theorem 4.1 in Donobo (1995) allows us to deduce from (3.10) that for WAI, WAI-U and Haar transforms we have:

$$P(|\hat{d}^t_{jk}| \leq |d_{jk}| \ \forall\ (j,k) \in \mathcal{J}_n) \geq \pi_n. \tag{3.11}$$

In addition, for the orthonormal Haar basis, the above inequality implies

$$P(\|\hat{m}\|_{L_2(d\hat{F}_n)} \leq \|m\|_{L_2(d\hat{F}_n)}) \geq \pi_n. \tag{3.12}$$

As said in the Introduction, the results for the WAI basis only pertain to the wavelet coefficient domain, hence (3.12) is not valid for a WAI basis.

The proof of Theorem 3.1 is given in Subsection 4.2.1. We now check that the derived threshold is not too large by looking at the l_2-risk of the detail coefficients.

3.3 *Upper bound on the l_2-risk of the detail coefficient*

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

$$\hat{d}_{jk} = d_{jk} + \rho_{jk}, \qquad (j,k) \in \mathcal{J}_n \tag{3.13}$$

as before. Using the location and scale dependent threshold $\{t_{jk}\}$ derived in Subsection 3.2, we obtain the following risk inequality.

THEOREM 3.2. *In the autoregressive model given by* (1.1) *and* (3.13), *assume* (A1)–(A4). *In the decomposition* (3.5) *of a random interval $I_{j,k}$, where $(j,k) \in \mathcal{J}_n$ as defined in* (3.2), *assume that the number $d_n(I_{j,k})$ of intervals at coarse scales is of order $d_n(I_{j,k}) = O(\log((\log n)^{1/r}))$ for some $r > 1$. Consider the WAI detail coefficients $\hat{d}_{jk} = d_{jk} + \rho_{jk}$ as in equation* (2.5). *Let $\hat{d}^t_{jk} := \eta_S(\hat{d}_{jk}, t_{jk})$ be the soft-thresholded coefficients, where t_{jk} is given in* (3.7). *Then, for some p, $1 < p < 2$, and almost every*

realization of the time series as in model (1.1), *an upper bound on the* l_2*-risk of the thresholded coefficients is given by*:

$$\sum_{j,k|(j,k)\in\mathcal{J}_n} E|\hat{d}^t_{jk} - d_{jk}|^2 \leq \sum_{j,k|(j,k)\in\mathcal{J}_n} \left(\frac{c\sigma_\epsilon^2 (\log n)^p}{n} + \min(8d_{jk}^2, 4t_{jk}^2) \right), \tag{3.14}$$

where c *is a finite constant not depending on* n.

This upper bound goes to zero as n tends to infinity by definition of the set $\mathcal{J}_n$ in (3.2). Note that the noise is not assumed to be Gaussian. Since we use *biorthogonal* wavelet transforms, a direct transfer of the l_2-risk of the wavelet coefficients to the L_2-risk of the estimator $\hat{m}$ in (2.6) is not possible. Indeed, Parseval's inequality is not valid anymore. However, simulation studies in Simoens and Vandewalle (2003) show that the design-adapted wavelets form a Riesz basis where $\|\hat{m} - m\|_{L_2} \leq C(\sum_{j,k|(j,k)\in\mathcal{J}_n} E|\hat{d}^t_{jk} - d_{jk}|^2)$, with C a constant independent of n. This means that the risk in the wavelet coefficient and estimator domains are comparable; having controlled the l_2-risk as in (3.14) gives thus sufficient insight about the L_2-risk of the estimator $\hat{m}$ itself.

Theorem 3.2 also covers the stochastic regression model we investigate in Delouille *et al.* (2004) without deriving any theory on it. An outline of the proof is given in Section 4 and details can be found in Delouille and von Sachs (2004). Note that due to the random nature of the wavelets (and not only the wavelet coefficients) the proofs need different techniques and are not just a straightforward modification of the original results of Donoho and co-authors, e.g., Donoho and Johnstone (1994).

3.4 *Dependent error models*

The theoretical results of Subsections 4.2–4.3 were proven under the assumption of i.i.d. errors. In case of an ARCH model, when rewriting the model as in (2.9), the errors V_t become dependent. It is possible to transfer an ARCH model into an AR model with independent errors, if one considers the slightly less general ARCH model:

$$X_t = \sigma_t \epsilon_t \quad \text{and} \quad \sigma_t^2 = s(X_{t-1}^2) \tag{3.15}$$

where no asymmetric effect can be represented. With the logarithmic squared data $Y_t := \log(X_t^2)$; $Z_t := \log(\epsilon_t^2)$, we have, see Franke *et al.* (2002*a*):

$$\begin{aligned} Y_t &= \log(X_t^2) = \log(\sigma_t^2) + \log(\epsilon_t^2) = \log(s(X_{t-1}^2)) + Z_t \\ &= \log(s_1(\log(X_{t-1}^2))) + Z_t \quad \text{with} \quad s_1(x) = s(e^x) \\ &= f_1(Y_{t-1}) + Z_t \quad \text{with} \quad f_1(\cdot) = \log(s_1(\cdot)), \end{aligned}$$

and we recover a model with additive and independent errors Z_t. It is a subject for future research to prove the analog of Theorems 3.1 and 3.2 in the general case of dependent errors.

4. Outline of proofs

We give here an outline of the proofs for Theorems 3.1 and 3.2. The complete proofs are available in Delouille and von Sachs (2004). We begin by some remarks and definitions.

4.1 *Definitions*

Note that with the inequality (3.1), we can define a probability space Ω_n such that

$$P(\Omega_n) = P\left(\omega : \sup_t |\epsilon_t(\omega)| \le n^{\delta'}\right) \ge 1 - O(n^{-\lambda}) \text{ for } \delta' \in (0, \delta). \tag{4.1}$$

To prove the results, we work on this probability space Ω_n. We will use the following definition, which is stronger than the usual O_P.

DEFINITION 4.1. Let Z_n be a sequence of random variables and let $\{\alpha_n\}$ and $\{\beta_n\}$ be sequences of positive reals. We write

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

if

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

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

Assumption (A4) can be formulated more precisely as follows.

(A4) For a given $\lambda > 0$, there exists a constant $c_\lambda < \infty$ such that

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

4.2 *Order of the threshold needed to remove the noise*

To prove Theorem 3.1, we use Lemma 2.1 in Neumann and Kreiss (1998) that treats sums of geometrically β-mixing random variables.

LEMMA 4.1. (Neumann and Kreiss (1998), Lemma 2.1.) *Suppose that $\{Z_t\}_{t=1}^n$ is a geometrically β-mixing process with $E(Z_t) = 0$. If for all $M < \infty$, there exists a constant $C_M < \infty$ such that $E|Z_t|^M \le C_M$, we have*

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

4.2.1 *Proof of Theorem* 3.1.

PROOF. We need to find the smallest threshold t_{jk} such that, with λ arbitrarily large, we have

$$\begin{aligned} P(|\rho_{jk}| > t_{jk}) &\le P\left(\frac{2^{j+1/2}}{n}\sum_{l:\tilde{g}_{jlk}\neq 0}|\tilde{g}_{jlk}|\left|\sum_{t=1}^n \epsilon_t 1_{j+1,l}(X_{t+1})\right| > t_{jk}\right) \\ &\le \sum_{l:\tilde{g}_{jlk}\neq 0} P\left(\frac{2^{j+1/2}}{n}|\tilde{g}_{jlk}|\left|\sum_{t=1}^n \epsilon_t 1_{j+1,l}(X_{t+1})\right| > t_{jk}\right) \\ &\le Kcn^{-\lambda}, \end{aligned} \tag{4.2}$$

where K is defined in (2.4). Let $t_{jk} := \sum_{l:\tilde{g}_{jlk}\neq 0} a^{(\lambda)}_{j,k,l}$. It is then sufficient to find the order of the term

$$P\left(\frac{2^{(j+1)/2}}{n}|\tilde{g}_{jlk}|\left|\sum_{t=1}^{n}\epsilon_t 1_{j+1,l}(X_{t-1})\right| > a^{(\lambda)}_{j,k,l}\right), \tag{4.3}$$

for a given l. Under Assumption (A1), we have that $\tilde{G} := \sup_{j,l,k}|\tilde{g}_{jlk}| < \infty$. Hence the probability (4.3) is bounded from above by

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

Let $\mathcal{I} := I_{j+1,l}$ be the random interval of interest. With $I^{\circ}_{j_1,k_1}$ the largest coarse-scale interval included in $I_{j+1,l}$ in (3.5), we need to prove that

$$P\left(\left|\sum_t \epsilon_t 1_{\mathcal{I}}(X_{t-1})\right| > c_\lambda \sigma_\epsilon \sqrt{n2^{-j_1}}\log n\right) = O(n^{-\lambda}). \tag{4.4}$$

To obtain the result we then multiply $c_\lambda\sigma_\epsilon\sqrt{n2^{-j_1}}\log n$ by $2^{(j+1)/2}\tilde{G}/n$.

To prove (4.4), we approximate $\mathcal{I}$ by a union of fixed intervals as in (3.5). Then for arbitrary c we have:

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

Lemmas 4.2 and 4.3 below state that

$$\begin{aligned} &P\left(\sum_{i=1}^{d_n(\mathcal{I})}\left|\sum_{t=1}^{n}\epsilon_t 1^{\circ}_{j_i,k_i}(X_{t-1})\right| > c_\lambda(\sigma_\epsilon\sqrt{n2^{-j_1}}\log n + d_n(\mathcal{I})n^{\delta})\right) \\ &\quad\le O(n^{-\lambda}). \end{aligned} \tag{4.5}$$

$$\begin{aligned} &P\left(\sum_{t=1}^{n}|\epsilon_t| 1_{I^{\circ}_{j^*,l_1}\cup I^{\circ}_{j^*,l_2}}(X_{t-1}) > c_\lambda(n^{\delta}[\sqrt{n2^{-j^*}}\log n + n2^{-j^*} + n^{\delta}])\right) \\ &\quad\le O(n^{-\lambda}). \end{aligned} \tag{4.6}$$

Putting (4.5) and (4.6) together, we obtain: $\sum_t \epsilon_t 1_{\mathcal{I}}(X_{t-1}) = \tilde{O}(t_\lambda, n^{-\lambda})$, where

$$t_\lambda = c_\lambda(\sigma_\epsilon\sqrt{n2^{-j_1}}\log n + d_n(\mathcal{I})n^{\delta} + n^{\delta}(\sqrt{n2^{-j^*}}\log n + n2^{-j^*} + n^{\delta})). \tag{4.7}$$

It is easy to see that $\sigma_\epsilon\sqrt{n2^{-j_1}}\log n$ is the dominating term as $n \to \infty$ as soon as Assumption (A3) is satisfied. □

4.2.2 *Proof of auxiliary lemmas*

LEMMA 4.2. *Let $\mathcal{I}$ be a random interval approximated as in expression* (3.4). *Under the assumptions* (A1)–(A2) *on the autoregressive model* (1.1), *we have*

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

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

$$P\left(\left|\sum_{t=1}^{n}\epsilon_t 1^{\circ}_{j_i,k_i}(X_{t-1})\right| > c_{1,\lambda}\right) \le O(n^{-\lambda})$$

since the result follows by summation over $i = 1,\dots,d_n(\mathcal{I})$. The random variables $Z_t := \epsilon_t 1^{\circ}_{j_i,k_i}(X_{t-1})$ are geometrically β-mixing, with $E(Z_t) = 0$ and $\mathrm{Var}(Z_t) = \sigma_\epsilon^2 2^{-j_i}$ by assumptions. Applying Lemma 4.1 on $Z = \sum_{t=1}^{n} Z_t$ we obtain

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

$\sum_{i=1}^{d_n(\mathcal{I})} 2^{-j_i/2} = O(2^{-j_1/2})$ where j_1 is the coarsest level used in the approximation, hence the result. □

LEMMA 4.3. *Let $\mathcal{I}$ be a random interval approximated as in expression* (3.5), *with* $2^{j^*} \le n \le 2^{j^*+1}$. *Assume* (A1)–(A2). *Then*

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

PROOF. On Ω_n defined in (4.1), we have $|\epsilon_t| \le n^\delta$. We have:

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

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,

$$P\left(\sum_t n^\delta 1^{\circ}_{j^*,l}(X_{t-1}) > c_\lambda\right) \le cn^{-\lambda}, \tag{4.8}$$

then the lemma is proven, provided we take $\lambda > 1$ sufficiently large. If we find a_λ such that $P(\sum_t 1^\circ_{j^*,l}(X_{t-1}) > a_\lambda) \le cn^{-\lambda}$, this means that $c_\lambda := n^\delta a_\lambda$ will satisfy (4.8). Introducing b_λ such that $a_\lambda := b_\lambda + n2^{-j^*}$, we have:

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

Defining $W_t := 1^\circ_{j^*,l}(X_{t-1}) - 2^{-j^*}$ and applying Lemma 4.1 to $W := \sum_t W_t$ we obtain:

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

With $\tilde{c}_\lambda$ depending only on λ, the quantity c_λ that satisfies (4.8) is equal to

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

4.3 *Upper bound on the l_2-risk of wavelet coefficients*

PROOF OF THEOREM 3.2. We show the following inequalities:

$$E((\hat{d}^t_{jk} - d_{jk})^2 1_{\{d^2_{jk} \le t^2_{jk}/2\}}) \le O\left(\frac{(\log n)^p}{n}\right) + 8|d_{jk}|^2 \quad \text{a.s.} \tag{4.9}$$

$$E((\hat{d}^t_{jk} - d_{jk})^2 1_{\{d^2_{jk} > t^2_{jk}/2\}}) \le 4t^2_{jk} \quad \text{a.s.}, \tag{4.10}$$

where the expectation is made w.r.t. the innovation ϵ_t, and the inequalities are valid for almost every realization of the process X_t in the model (1.1). The inequality (3.14) then results from the minimum of these two upper bounds and by summation over $(j,k) \in \mathcal{J}_n$.

PROOF OF (4.9). We decompose the space where $\{d^2_{jk} \le t^2_{jk}/2\}$ into two parts. On the subspace where $|\rho_{jk}| \le \frac{t_{jk}}{2}$, $|\hat{d}_{jk}| = |d_{jk} + \rho_{jk}| \le t_{jk}$. Thus the thresholded coefficients $\hat{d}^t_{jk} = 0$ and $E[(\hat{d}^t_{jk} - d_{jk})^2] = d^2_{jk}$. If $|\rho_{jk}| > \frac{t_{jk}}{2}$, we decompose $|\hat{d}^t_{jk} - d_{jk}|$ as

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

Hence

$$E((\hat{d}^t_{jk} - d_{jk})^2 1_{\{d^2_{jk} \le t^2_{jk}/2\}}) \le 2E[\eta_S(\rho_{jk}, t_{jk})^2] + 8|d_{jk}|^2. \tag{4.11}$$

Finally one can prove that (for details of this rather technical proof, see Lemma 6.4 in Delouille and von Sachs (2004))

$$E[\eta_S(\rho_{jk}, t_{jk})^2] = O\left(\frac{(\log n)^p}{n}\right) \quad \text{for some} \quad p \text{ s.t. } 1 < p < 2. \tag{4.12}$$

And this completes the proof of (4.9). □

PROOF OF (4.10). On the space where $\{d_{jk}^2 > t_{jk}^2/2\}$, we perform the decomposition

$$E|\hat{d}_{jk}^t - d_{jk}|^2 \le 2E|\hat{d}_{jk}^t - \hat{d}_{jk}|^2 + 2E|\hat{d}_{jk} - d_{jk}|^2 \le 2t_{jk}^2 + 2E|\rho_{jk}|^2.$$

By construction, $P(|\rho_{jk}| > t_{jk}) \le cn^{-\lambda}$ with t_{jk} given by (3.7). We now show that $E[\rho_{jk}^2] \le t_{jk}^2$. Indeed, with ρ_{jk} as in (3.6), we have:

$$E[\rho_{jk}^2] \le \frac{\tilde{G}^2 2^{j+1}}{n^2} \sum_{l:\tilde{g}_{jlk}\neq 0} E\left(\left|\sum_{t=1}^n \epsilon_t 1_{j+1,l}(X_{t-1})\right|^2\right),$$

since the random intervals $I_{j+1,l}$ are disjoint. Using the approximation (3.5), we decompose $E|\sum_{t=1}^n \epsilon_t 1_{\mathcal{I}}(X_{t-1})|^2$ as

$$\begin{aligned}(4.13)\qquad \left|\sum_{t=1}^n \epsilon_t 1_{\mathcal{I}}(X_{t-1})\right|^2 &\le 2\left(\sum_{i=1}^{d_n(\mathcal{I})}\left|\sum_{t=1}^n \epsilon_t 1_{j_i,k_i}^{\circ}(X_{t-1})\right|\right)^2 \\ &\quad + 2\left(\sum_t |\epsilon_t| 1_{j^*,l_1}^{\circ}(X_{t-1}) + \sum_t |\epsilon_t| 1_{j^*,l_2}^{\circ}(X_{t-1})\right)^2.\end{aligned}$$

Since the intervals I_{j_i,k_i}° are disjoint by construction, we have:

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

Considering the method of independent blocks as in Neumann and Kreiss (1998), Doukhan *et al.* (1995), we have:

$$E\left(\left|\sum_{t=1}^n \epsilon_t 1_{j_i,k_i}^{\circ}(X_{t-1})\right|^2\right) = \operatorname{Var}\left(\sum_{t=1}^n \epsilon_t 1_{j_i,k_i}^{\circ}(X_{t-1})\right) \le c\log(n)n2^{-j_i};$$

$$\begin{aligned}E\left(\left|\sum_t |\epsilon_t| 1_{j^*,l}^{\circ}(X_{t-1})\right|^2\right) &\le n^{2\delta}\operatorname{Var}\left(\sum_t 1_{j^*,l}^{\circ}(X_{t-1})\right) + n^{2\delta}\left(E\left[\sum_t 1_{j^*,l}^{\circ}(X_{t-1})\right]\right)^2 \\ &\le cn^{2\delta}\log(n)n2^{-j^*}.\end{aligned}$$

Putting the above results together, we obtain

$$(4.14)\qquad E\left|\sum_{t=1}^n \epsilon_t 1_{\mathcal{I}}(X_{t-1})\right|^2 \le c(\log(n)2^{-j_1}n + n^{2\delta}\log(n)n2^{-j^*}).$$

In (4.14), the term $\log(n)2^{-j_1}n$ is dominating by Assumptions (A3)–(A4). We obtain

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

and thus

$$E((\hat{d}^t_{jk} - d_{jk})^2 1_{\{d^2_{jk} > t^2_{jk}/2\}}) \leq 4t^2_{jk} \leq 4t^2_{jk} + O\left(\frac{(\log n)^p}{n}\right). \tag{4.15}$$

Putting (4.11), (4.12), and (4.15) together, and summing over $(j,k) \in \mathcal{J}_n$, we obtain the desired result. □

Acknowledgements

The authors would like to sincerely thank M. Neumann for the many insightful discussions we had with him. Furthermore, we thank the three anonymous referees for their comments that helped greatly to improve the presentation of this paper. This research was done while the first author was Research assistant of the National Fund for Scientific Research – Wallonie, Belgium (F.N.R.S.).

References

Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussion), *Journal of the American Statistical Association*, **96**, 939–967.

Antoniadis, A. and Pham, D. (1998). Wavelet regression for random or irregular design, *Computational Statistics and Data Analysis*, **28**, 353–369.

Berkner, K. and Wells, R. (1998). A correlation-dependent model for denoising via nonorthogonal wavelet transforms, Tech. Report No. 98-07, Computational Mathematics Laboratory, Rice University, Houston, Texas.

Bosq, D. (1996). *Nonparametric Statistics for Stochastic Processes*, Springer-Verlag, New York.

Bühlmann, P. and McNeil, A. (2002). An algorithm for nonparametric GARCH modelling, *Computational Statistics and Data Analysis*, **40**, 665–683.

Cai, T. and Brown, L. (1998). Wavelet shrinkage for nonequispaced samples, *Annals of Statistics*, **26**(5), 1783–1799.

Carnicer, J., Dahmen, W. and Peña, J. (1996). Local decomposition of refinable spaces and wavelets, *Applied and Computional Harmonic Analysis*, **3**, 127–153.

Chen, R., Yang, L. and Hafner, C. (2004). Nonparametric multi-step ahead prediction in time series analysis, *Journal of the Royal Statistical Society, Series B*, **66**, 669–686.

Cohen, A., Daubechies, I. and Feauveau, J. (1992). Bi-orthogonal bases of compactly supported wavelets, *Communications in Pure and Applied Mathematics*, **45**, 485–560.

Davidson, J. (1994). *Stochastic Limit Theory*, Oxford University Press.

Delouille, V. and von Sachs, R. (2004). Estimation of nonlinear autoregressive models using design-adapted wavelets. (Available at www.stat.ucl.ac.be/rvs/dp0225rev.pdf)

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

Delouille, V., Simoens, J. and von Sachs, R. (2004). Smooth design-adapted wavelets for nonparametric stochastic regression, *Journal of the American Statistical Association*, **99**, 643–658.

Delyon, B. and Juditsky, A. (1997). On the computation of wavelet coefficients, *Journal of Approximation Theory*, **88**, 47–79.

Donoho, D. (1995). De-noising via soft-thresholding, *IEEE Transactions on Information Theory*, **41**, 613–627.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, **81**, 425–455.

Donoho, D. and Johnstone, I. (1998). Minimax estimation via wavelet shrinkage, *Annals of Statistics*, **26**, 879–921.

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

Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation, *Econometrica*, **50**, 987–1008.

Franke, J., Holzberger, H. and Müller, M. (2002*a*). Nonparametric estimation of ARMA- and GARCH-processes, *Applied Quantitative Finance*, 367–388, Springer-Verlag, Berlin.

Franke, J., Kreiss, J.-P. and Mammen, E. (2002*b*). Bootstrap of kernel smoothing in nonlinear time series, *Bernoulli*, **8**, 1–37.

Franke, J., Kreiss, J.-P., Mammen, E. and Neumann, M. (2002*c*). Properties of the nonparametric autoregressive bootstrap, *Journal of Time Series Analysis*, **23**, 555–585.

Girardi, M. and Sweldens, W. (1997). A new class of unbalanced Haar wavelets that form an unconditional basis for L_p on general measure spaces, *Journal of Fourier Analysis and Applications*, **3**(4), 457–474.

Gouriéroux, C. (1997). *ARCH Models and Financial Applications*, Springer-Verlag, New York.

Gouriéroux, C. and Monfort, A. (1992). Qualitative threshold ARCH models, *Journal of Econometrics*, **52**, 159–199.

Hafner, C. (1998*a*). Estimating high frequency foreign exchange rate volatility with nonparametric ARCH models, *Journal of Statistical Planning and Inference*, **68**, 247–269.

Hafner, C. (1998*b*). *Nonlinear Time Series Analysis with Applications to Foreign Exchange Rate Volatility*, Physica-Verlag, Heidelberg.

Hall, P. and Turlach, B. (1997). Interpolation methods for nonlinear wavelet regression with irregularly spaced design, *Annals of Statistics*, **25**(5), 1912–1925.

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

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

Jones, D. (1978). Nonlinear regressive processes, *Proceedings of the Royal Society London, Series A*, **360**, 71–95.

Kovac, A. and Silverman, B. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding, *Journal of the American Statistical Association*, **95**, 172–183.

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

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

Neumann, M. and von Sachs, R. (1995). Wavelet thresholding: Beyond the Gaussian i.i.d. situation, *Wavelets and Statistics*, Lecture notes in statistics (eds. A. Antoniadis and G. Oppenheim), **103**, 301–329, Springer-Verlag, New York.

Robinson, P. (1983). Nonparametric estimators for time series, *Journal of Time Series Analysis*, **4**, 185–207.

Simoens, J. and Vandewalle, S. (2003). A stabilized lifting construction of wavelets on irregular meshes on the interval, *SIAM Journal of Scientific Computing*, **24**(4), 1356–1378.

Sweldens, W. (1997). The lifting scheme: A construction of second generation wavelets, *SIAM Journal of Mathematical Analysis*, **29**(2), 511–546.

Sweldens, W. and Schröder, P. (1996). Building your own wavelets at home, *Wavelets in Computer Graphics*, 15–87, ACM SIGGRAPH Course Notes.

Tjøstheim, D. (1994). Non-linear time series: A selective review, *Scandinavian Journal of Statistics*, **21**, 97–130.

Tong, H. (1983). *Threshold Models in Nonlinear Time Series Analysis*, Springer-Verlag, Berlin.