Estimators of Long-Memory: Fourier versus Wavelets

Gilles Fay, Éric Moulines, François Roueff, Murad S. Taqqu

ESTIMATORS OF LONG-MEMORY: FOURIER VERSUS WAVELETS

GILLES FAŸ, ERIC MOULINES, FRANÇOIS ROUEFF, AND MURAD S. TAQQU

Abstract. There have been a number of papers written on semi-parametric estimation methods of the long-memory exponent of a time series, some applied, others theoretical. Some using Fourier methods, others using a wavelet-based technique. In this paper, we compare the Fourier and wavelet approaches to the local regression method and to the local Whittle method. We provide an overview of these methods, describe what has been done, indicate the available results and the conditions under which they hold. We discuss their relative strengths and weaknesses both from a practical and a theoretical perspective. We also include a simulation-based comparison. The software written to support this work is available on demand and we illustrate its use at the end of the paper.

Date: January 28, 2008.

1991 *Mathematics Subject Classification.* Primary 62M10, 62M15, 62G05 Secondary: 60G18.

Key words and phrases. Wavelet analysis, long range dependence, semi-parametric estimation.

Murad S. Taqqu would like to thank l'École Normale Supérieure des Télecommunications in Paris for their hospitality. This research was partially supported by the NSF Grants DMS–0505747 and DMS–0706786 at Boston University.

Contents

1. Introduction

We study here finite variance stochastic processes $\{X_k\}_{k\geq 1}$, whose spectral density $f(\lambda)$, $\lambda \in (-\pi, \pi)$ behaves like a power function at low frequencies, that is as $|\lambda|^{-2d}$ as the frequency $\lambda \to 0+$. The case $d > 0$ corresponds to *long-memory*, $d = 0$ to *short-memory* and $d < 0$ is often referred to as *negative dependence*. For $X_k, k \in \mathbb{Z}$ to be stationary it is necessary that $\int_{-\pi}^{\pi} f(\lambda)d\lambda < \infty$ and hence that $d < 1/2$. We relax these restrictions in a number of ways. We shall allow the process to be non-stationary, requiring only that it becomes stationary after it is differenced a number of times. We also suppose that the spectral density (of the differenced process) behaves not merely like $|\lambda|^{-2d}$ but as $|\lambda|^{-2d} f^\star(\lambda)$, where $f^\star$ is regarded as a short-range density function.

Our goal is to estimate d in the presence of $f^\star$. We shall not assume that the nuisance function $f^\star$ is known, nor that it is characterized by a finite number of unknown parameters, but merely that $f^\star(\lambda)$ is "smooth" in the neighborhood of $\lambda = 0$, so that if one focuses only on frequencies λ that are sufficiently low, then the spectral density $f(\lambda)$ behaves essentially like $|\lambda|^{-2d}$. What frequency cut-off should one choose will clearly become an important issue.

The estimation framework is *semi-parametric*: we must estimate the unknown parameter d while viewing the presence of $f^\star$ as a nuisance, albeit one which complicates matters. The estimation method will also be *local*, in that, it is necessary to focus only on frequencies λ that are close enough to the origin, where the influence of $f^\star(\lambda)$ can be neglected.

In this paper we provide an overview and comparison of four semi-parametric estimation methods of the parameter d which have all proven to be very effective. Two of them are Fourier-based, the other two are based on wavelets. The methods are:

- Geweke-Porter Hudak (GPH): Regression / Fourier,
- Local Whittle Fourier (LWF): Whittle / Fourier,
- Local Regression Wavelets (LRW): Regression / Wavelets,
- Local Whittle Wavelets (LWW): Whittle / Wavelets.

The Fourier methods are older and better known. They have essentially been developed by Peter Robinson in a number of fundamental papers Robinson (1995b), Robinson (1995a). If we ignore for the moment the presence of the nuisance function $f^\star$, then one has $f(\lambda) = |\lambda|^{-2d}$, that is $\log f(\lambda) \approx -2d \log |\lambda|$. Therefore, d can be estimated by linear regression on the periodogram. This is the Fourier-based regression method considered in Geweke and Porter-Hudak (1983) in a parametric setting. The semi-parametric setting was suggested by Künsch (1987) and developed by Robinson (1995b). The Fourier-based Whittle method is a pseudo-maximum likelihood method developed by Fox and Taqqu (1986) in a parametric setting and extended in a semi-parametric setting by Robinson (1995a).

The papers of Moulines et al. (2007b), Moulines et al. (2007a), Moulines et al. (2007c) and Roueff and Taqqu (2007) recast the preceding Fourier-based methods in a wavelet setting. Wavelets have a number of advantages. They allow differencing implicitly and therefore they can be used without problems when $d > 1/2$. They also automatically discount polynomial trends. The local wavelet-based regression method was first developed by Abry and Veitch (1998) under the simplifying assumption that the wavelet coefficients are uncorrelated; see also Veitch and Abry (1999) and the review articles Abry et al. (2000) and Abry et al. (2003). In addition, see Veitch et al. (2003) for the automatic selection of the cut-off frequency point and Veitch et al. (2000) for the choice of the "scale function". Bardet (2000) and Bardet (2002) provides asymptotic result for the LRW estimator in a parametric context. Bardet et al. (2000) is a first attempt to analyze the behavior of LRW in a semi-parametric context by assuming continuous-time observations. The Local Whittle wavelet method is developed in Moulines et al. (2007c).

The paper is structured as follow. In Section 2, we formalise our assumptions on $\{X_k\}$ by defining an M(d) process, that is, a process with memory parameter d. The standard ARIMA and fractional Gaussian noise examples are introduced in Section 3. The wavelet-based semi-parametric estimators are defined in Section 4 and the Fourier estimators in Section 5. The semi-parametric setting is discussed in Section 6. The asymptotic properties of the wavelet and Fourier semi-parametric estimators are described in Sections 7 and 8, respectively. Their properties are discussed further in Section 9. Section 10 contains the Monte-Carlo study which compares the effectiveness of the four methods. In Section 11, we illustrate the use of the software written in support of this work. This software may be obtained from the authors. Section 12 contains concluding remarks.

2. Definition of an M(d) process

Let $X \stackrel{\text{def}}{=} \{X_k\}_{k\in\mathbb{Z}}$ be a real-valued process, not necessarily stationary. Its first order difference is

$$[\mathbf{\Delta} X]_n \stackrel{\text{def}}{=} X_n - X_{n-1}, \quad n \in \mathbb{Z}$$

Its K-th order difference $\mathbf{\Delta}^K X$ is defined recursively. We suppose that the process X has *memory parameter* d, $d \in \mathbb{R}$, in short, is an M(d) process. We shall first define this notion for a stationary process X, where $d < 1/2$, and then provide a general definition for $d \in \mathbb{R}$.

Let f^* be a non-negative even function continuous and positive at the origin. A stationary process X is said to have memory parameter d, $-\infty < d < 1/2$, and short-range density function f^*, if its spectral density is given by

$$f_X(\lambda) \stackrel{\text{def}}{=} |1 - \mathrm{e}^{-\mathrm{i}\lambda}|^{-2d} f^*(\lambda), \quad \lambda \in (-\pi, \pi), \tag{1}$$

To allow $d > 1/2$, we consider non–stationary processes X and extend the preceding definition, valid for stationary processes, in the following way.

Definition 1. *We say that X has* memory parameter d, $d \in \mathbb{R}$ *(in short, an M(d) process), and* short-range density function f^*, *if f^* is continuous and positive at the origin and, for any integer $K > d - 1/2$, its K-th order difference $\mathbf{\Delta}^K X$ is stationary with spectral density function*

$$f_{\mathbf{\Delta}^K X}(\lambda) = |1 - \mathrm{e}^{-\mathrm{i}\lambda}|^{2(K-d)} f^*(\lambda), \quad \lambda \in (-\pi, \pi) . \tag{2}$$

Observe that $f_{\mathbf{\Delta}^K X}(\lambda)$ in (2) is integrable since $-(K - d) < 1/2$. Observe also that if the process X is as in Definition 1, then while $\mathbf{\Delta}^K X$ is stationary, the process X itself is stationary only when $d < 1/2$. Nevertheless, one can associate to X the *generalized spectral density function*

$$f_X(\lambda) = |1 - \mathrm{e}^{-\mathrm{i}\lambda}|^{-2d} f^*(\lambda) \tag{3}$$

Remark 1. This definition of M(d) processes was proposed by Hurvich and Ray (1995). It has the advantage that $\mathbf{\Delta}^K X$ is stationary, but it introduces a discontinuity at the fractional points $d = 1/2, 3/2, \ldots$ since $f_{\mathbf{\Delta}^K X}$ is quite different at these values of d. In empirical work, there are typically no inherent restrictions on the value of the memory parameter d, and this may cause a problems if the degree of integer differencing required to achieve stationarity must be guessed in advance. An alternative definition of M(d) process has been introduced by Robinson (1994) and later used by some authors (see Tanaka (1999), Shimotsu and Phillips (2005), Shimotsu and Phillips (2006)).

The memory parameter d plays a central role in the definition of M(d) processes because it characterizes the behavior of the generalized spectral density $f_X(\lambda)$ at low frequencies. Indeed, assuming that f^* is continuous at zero, then (3) implies $f_X(\lambda) \sim |\lambda|^{-2d} f^*(0)$ as $\lambda \to 0$. Allowing d to take non integer values produces a fundamental change in the correlation structure of a fractional process, as compared to the correlation structure of a standard time-series model, such as an ARMA(p, q) process.

The study of M(d) processes has recently attracted attention amongst theorists and empirical researchers. In applied econometric work, M(d) processes with $d > 0$ provide sensible models for certain macroeconomic time series (inflation, interest rates, ...) as well as certain financial time series (volatility of financial asset returns, forward exchange market premia,...). M(d) models encompass both stationary and nonstationary processes depending on the value of the memory parameter and include both short-memory series M(0) and unit-root M(1) processes as special cases when the memory parameter takes on the values zero and unity.

3. Examples

Stationarity of the increments is commonly assumed in time-series analysis. In ARIMA models, for example, (2) holds with $d = K$ integer and with f^* equal to the spectral density of an autoregressive moving average short-memory process. If $d \in \mathbb{R}$ and $f^* \equiv \sigma^2$ in (3), one gets the so-called fractionally integrated white noise process, ARFIMA(0,d,0). The choice $d \in \mathbb{R}$ and

$$f^*_{\mathrm{ARMA}}(\lambda) = \sigma^2 \frac{\left|1 - \sum_{k=1}^q \theta_k \mathrm{e}^{-\mathrm{i}\lambda k}\right|^2}{\left|1 - \sum_{k=1}^p \phi_k \mathrm{e}^{-\mathrm{i}\lambda k}\right|^2}, \quad \lambda \in (-\pi, \pi)\ , \tag{4}$$

with $1 - \sum_{k=1}^p \phi_k z^k \neq 0$ for $|z| = 1$ and $1 - \sum_{k=1}^p \theta_k \neq 0$ (so that $f^*_{\mathrm{ARMA}}(0) \neq 0$) leads to the class of ARFIMA(p, d, q) processes.

Another example is $\{B_H(k)\}_{k\in\mathbb{Z}}$, a discrete-time version of fractional Brownian motion (FBM) $\{B_H(t), t \in \mathbb{R}\}$ with Hurst index $H \in (0, 1)$. The latter is a centered Gaussian process with covariance

$$R_H(t, s) \stackrel{\mathrm{def}}{=} \mathbb{E}[B_H(t)B_H(s)] = \frac{1}{2}\left\{|t|^{2H} + |s|^{2H} - |t - s|^{2H}\right\}\ .$$

The process $\{B_H(k)\}_{k\in\mathbb{Z}}$ is increment stationary ($K = 1$) and its generalized spectral density is given up to a multiplicative constant (see Samorodnitsky and Taqqu (1994)) by

$$f_{\mathrm{FBM}}(\lambda) \stackrel{\mathrm{def}}{=} \sum_{k=-\infty}^{\infty} |\lambda + 2k\pi|^{-2H-1}\ , \quad \lambda \in (-\pi, \pi)\ .$$

We can express it in the form (3),

$$f_{\mathrm{FBM}}(\lambda) = |1 - \mathrm{e}^{-\mathrm{i}\lambda}|^{-2d} f^*_{\mathrm{FBM}}(\lambda)\ , \tag{5}$$

by setting $d = H + 1/2 \in (1/2, 3/2)$ and

$$f^*_{\mathrm{FBM}}(\lambda) = \left|\frac{2\sin(\lambda/2)}{\lambda}\right|^{2H+1} + |2\sin(\lambda/2)|^{2H+1} \sum_{k\neq 0} |\lambda + 2k\pi|^{-2H-1}\ . \tag{6}$$

Observe that $f^*_{\mathrm{FBM}}(0) = 1$ and that it is bounded on $(-\pi, \pi)$.

The process $G_H = \mathbf{\Delta} B_H$ is fractional Gaussian noise (FGN). It is a stationary Gaussian process with spectral density proportional to (5), but with $d = H - 1/2 \in (-1/2, 1/2)$.

4. Wavelet semi-parametric estimators of the memory parameter

In this section, we introduce the wavelet setting and, based on heuristical arguments, proposed possible semi-parametric wavelet estimators. We start with a brief summary of the basic ideas. A wavelet $\psi(t),\ t \in \mathbb{R}$ is a function with at least one vanishing moment, that is ,$\int_{\mathbb{R}} \psi(t)\mathrm{d}t = 0$, and which is *low-pass*, in the sense that its Fourier transform $\widehat{\psi}(\xi)$ decreases as $\xi \to \infty$. We then define the scaled and translated versions of ψ, namely $\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k),\ j, k \in \mathbb{Z}$. The *scale index* j dilates ψ so that large values of j correspond to coarse scales (low frequencies), while

the *position index* k translates the function $\psi(2^{-j}t)$ to $\psi(2^{-j}t-k)$. The corresponding *wavelet coefficients* are then defined as $W_{j,k}=\int_{\mathbb{R}}X(t)\psi_{j,k}(t)\mathrm{d}t$ and are used to estimate d. Because $\widehat{\psi}$ is low-pass, $\widehat{\psi}_{j,k}$ concentrates in the low frequency region as $j\to\infty$ and $f_X(\lambda)$ "scales" at low frequencies since $|1-\mathrm{e}^{\mathrm{i}\lambda}|^{-2d}\sim|\lambda|^{-2d}$ as $|\lambda|\to 0$. In the above definition of wavelet coefficients, we supposed, for simplicity, that the process $\{X(t)\}_{t\in\mathbb{R}}$ is defined in continuous time and that the integral above is well-defined. This definition can be adapted to discrete time series $\{X_k,\ k\in\mathbb{Z}\}$ by using a *scale function* and also to finite samples $X_1,\ldots,X_n$ by merely restricting the set of scale and translation indices of *available wavelet coefficients*. This is in sharp contrast to Fourier analysis, where the definition of discrete Fourier coefficients at a given frequency changes as the sample length increases. We now turn to a more formal presentation.

4.1. **The wavelet setting.** The wavelet setting involves a *scale function* $\phi\in\mathrm{L}^2(\mathbb{R})$ and a *wavelet* $\psi\in\mathrm{L}^2(\mathbb{R})$, with associated Fourier transforms

$$\widehat{\phi}(\xi)\stackrel{\text{def}}{=}\int_{-\infty}^{\infty}\phi(t)\mathrm{e}^{-\mathrm{i}\xi t}\,\mathrm{d}t\quad\text{and}\quad\widehat{\psi}(\xi)\stackrel{\text{def}}{=}\int_{-\infty}^{\infty}\psi(t)\mathrm{e}^{-\mathrm{i}\xi t}\,\mathrm{d}t\,.$$

We assume the following:

(W-1) ϕ and ψ are compactly-supported, integrable, and $\widehat{\phi}(0)=\int_{-\infty}^{\infty}\phi(t)\,\mathrm{d}t=1$ and $\int_{-\infty}^{\infty}\psi^2(t)\,\mathrm{d}t=1$.

(W-2) There exists $\alpha>1$ such that $\sup_{\xi\in\mathbb{R}}|\widehat{\psi}(\xi)|\,(1+|\xi|)^{\alpha}<\infty$.

(W-3) The function ψ has M vanishing moments, *i.e.* $\int_{-\infty}^{\infty}t^m\psi(t)\,\mathrm{d}t=0$ for all $m=0,\ldots,M-1$

(W-4) The function $\sum_{k\in\mathbb{Z}}k^m\phi(\cdot-k)$ is a polynomial of degree m for all $m=0,\ldots,M-1$.

Condition (W-2) ensures that the Fourier transform $\widehat{\psi}$ decreases quickly to zero. Daubechies wavelets have $\alpha>1$ (see Table 1 below) except for Haar wavelet which is discontinuous and for which $\alpha=1$. Condition (W-3) it ensures that ψ oscillates and that its scalar product with continuous-time polynomials up to degree $M-1$ vanishes. It is equivalent to asserting that the first $M-1$ derivative of $\widehat{\psi}$ vanish at the origin and hence

$$|\widehat{\psi}(\lambda)|=O(|\lambda|^M)\quad\text{as}\quad\lambda\to 0. \tag{7}$$

And, by (Cohen, 2003, Theorem 2.8.1, Page 90), (W-4) is equivalent to

$$\sup_{k\neq 0}|\widehat{\phi}(\lambda+2k\pi)|=O(|\lambda|^M)\quad\text{as}\quad\lambda\to 0. \tag{8}$$

As shown below, conditions (W-4)-(W-3) imply that the wavelet transform of discrete-time polynomials of degree $M-1$ vanishes.

We now describe the computation of the wavelet coefficients. Define the family $\{\psi_{j,k}, j \in \mathbb{Z}, k \in \mathbb{Z}\}$ of translated and dilated functions

$$\psi_{j,k}(t) = 2^{-j/2}\,\psi(2^{-j}t - k), \quad j \in \mathbb{Z},\, k \in \mathbb{Z}\,. \tag{9}$$

Consider a real-valued sequence $\mathbf{x} = \{x_k,\, k \in \mathbb{Z}\}$. We need to construct a continuous–time process from a discrete–time one. Using the scaling function ϕ, we first associate to this sequence the continuous-time functions

$$\mathbf{x}_n(t) \stackrel{\text{def}}{=} \sum_{k=1}^{n} x_k\,\phi(t-k) \quad \text{and} \quad \mathbf{x}(t) \stackrel{\text{def}}{=} \sum_{k\in\mathbb{Z}} x_k\,\phi(t-k), \quad t \in \mathbb{R}\,. \tag{10}$$

The function $\mathbf{x}_n$ only requires the values of $x_1, \ldots, x_n$ while the function $\mathbf{x}$ requires the whole sequence $\{x_k,\, k \in \mathbb{Z}\}$. Without loss of generality we may suppose that the supports of the scaling function ϕ and of the wavelet function ψ are included in $[-\mathrm{T}, 0]$ and $[0, \mathrm{T}]$, respectively, for some integer $\mathrm{T} \geq 1$. This implies that $\mathbf{x}_n(t) = \mathbf{x}(t)$ for all $t \in [0, n - \mathrm{T} + 1]$ and that the support of $\psi_{j,k}$ is included in the interval $[2^j k, 2^j(k + \mathrm{T})]$. The wavelet coefficient $W^{\mathbf{x}}_{j,k}$ at scale $j \geq 0$ and location $k \in \mathbb{Z}$ is defined as

$$W^{\mathbf{x}}_{j,k} \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} \mathbf{x}(t)\psi_{j,k}(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} \mathbf{x}_n(t)\psi_{j,k}(t)\,\mathrm{d}t, \quad j \geq 0, k \in \mathbb{Z}\,. \tag{11}$$

The second equality holds when $[2^j k, 2^j(k+\mathrm{T})] \subseteq [0, n-\mathrm{T}+1]$, that is, for all $(j,k) \in \mathcal{I}_n$, where

$$\mathcal{I}_n \stackrel{\text{def}}{=} \{(j,k) :\, j \geq 0, 0 \leq k < n_j\} \quad \text{with} \quad n_j = \lfloor 2^{-j}(n - \mathrm{T} + 1) - \mathrm{T} + 1 \rfloor\,. \tag{12}$$

In other words $\mathcal{I}_n$ denotes the set of indices (j,k) for which the wavelet coefficients $W_{j,k}$ depend only on $x_1, \ldots, x_n$. If the sample size n increases, these wavelet coefficients remain unchanged and new ones can be computed. Thus the definition of wavelet coefficients does not depend on the sample length, in contrast to Fourier coefficients. The wavelet coefficient $W^{\mathbf{x}}_{j,k}$ can be computed explicitly by using discrete convolution and downsampling, namely,

$$W^{\mathbf{x}}_{j,k} = \sum_{l\in\mathbb{Z}} x_l\, h_{j,2^j k - l} = (h_{j,\cdot} \star \mathbf{x})_{2^j k} = (\downarrow^j [h_{j,\cdot} \star \mathbf{x}])_k, \quad j \geq 0, k \in \mathbb{Z}\,, \tag{13}$$

where, for all $j \geq 0$, the impulse response $h_{j,\cdot}$ is defined by

$$h_{j,l} \stackrel{\text{def}}{=} 2^{-j/2} \int_{-\infty}^{\infty} \phi(t+l)\psi(2^{-j}t)\,\mathrm{d}t, \quad l \in \mathbb{Z}\,, \tag{14}$$

where '$\star$' denotes the convolution of discrete sequences and $\downarrow^j$ is the j–power downsampling operator defined, for any sequence $\{c_k\}_{k\in\mathbb{Z}}$, by $(\downarrow^j c)_k = c_{k2^j}$. Since ϕ and ψ have compact support, the associated transfer function H_j is a trigonometric polynomial,

$$H_j(\lambda) = \sum_{l\in\mathbb{Z}} h_{j,l}\,\mathrm{e}^{-\mathrm{i}\lambda l} = \sum_{l=-\mathrm{T}(2^j+1)+1}^{-1} h_{j,l}\,\mathrm{e}^{-\mathrm{i}\lambda l}\,. \tag{15}$$

Under assumption (W-4), $t \mapsto \sum_{l\in\mathbb{Z}} \phi(t+l)l^m$ is a polynomial of degree m and (W-3) therefore implies that, for all $j \geq 0$ and all $m = 0, \ldots, M-1$,

$$\sum_{l\in\mathbb{Z}} h_{j,l}\, l^m = 2^{-j/2} \int_{-\infty}^{\infty} \psi(2^{-j}t) \sum_{l\in\mathbb{Z}} \phi(t+l) l^m dt = 0\ . \tag{16}$$

Now consider $P_j(x) = \sum_{l\in\mathbb{Z}} h_{j,l}\, x^l$ and observe that (16) implies $P_j(1) = 0$, $P_j'(1) = 0$, ..., $P_j^{(M-1)}(1) = 0$, and hence $H_j(\lambda) = P_j(\mathrm{e}^{-\mathrm{i}\lambda})$ factors as

$$H_j(\lambda) = (1 - \mathrm{e}^{-\mathrm{i}\lambda})^M\, \tilde{H}_j(\lambda)\ , \tag{17}$$

where $\tilde{H}_j(\lambda)$ is also a trigonometric polynomial. The wavelet coefficient (13) may therefore be computed as

$$W_{j,k}^{\mathbf{x}} = (\downarrow^j\, [\tilde{h}_{j,\cdot} \star \boldsymbol{\Delta}^M \mathbf{x}])_k \tag{18}$$

where $\{\tilde{h}_{j,l}\}_{l\in\mathbb{Z}}$ are the coefficients of the trigonometric polynomial $\tilde{H}_j$ and $\boldsymbol{\Delta}^M\mathbf{x}$ is the M-th order difference of the sequence $\mathbf{x}$. In other words, the use of a wavelet and a scaling function satisfying (W-4) and (W-3) implicitly perform a M-th order differentiation of the time-series. Therefore, we may work with an M(d) processes X beyond the stationary regime ($d > 1/2$) possibly contaminated by a polynomial trend of degree K without specific preprocessing, provided that $d - M < 1/2$ and $M \geq K + 1$. It is perhaps less known that wavelets can be used with non-invertible processes ($d \leq -1/2$) thanks to the decay property of $\widehat{\psi}$ at infinity assumed in (W-2).

4.2. **Choice of the wavelets.** In this paper, we do not assume that $\psi_{j,k}$ are orthonormal in $\mathrm{L}^2(\mathbb{R})$ nor that they are associated to a multiresolution analysis (MRA). We may therefore use other convenient choices for ϕ and ψ as long as (W-1)-(W-4) are satisfied. A simple choice is for instance, for some integer $M \geq 2$,

$$\phi(t) \stackrel{\text{def}}{=} \mathbb{1}_{[0,1]}^{\star M}(t) \quad \text{and} \quad \psi(t) \stackrel{\text{def}}{=} c_M\, \frac{\mathrm{d}^M}{\mathrm{d}t^M} \mathbb{1}_{[0,1]}^{\star 2M}(2t), \tag{19}$$

where $\mathbb{1}_A$ is the indicator function of the set A and for an integrable function f, $f^{\star M}$ denotes the M-th self-convolution of f,

$$f^{\star M} = \underbrace{f \star \cdots \star f}_{M \text{ times}}, \text{ with } (f \star g)(t) = \int_{-\infty}^{\infty} f(t-u)\, g(u)\, \mathrm{d}u\ ,$$

and c_M is a normalizing positive constant such that $\int_{-\infty}^{\infty} \psi^2(t)\, \mathrm{d}t = 1$.

Scaling and wavelet functions associated to an MRA present two important features: 1) they give raise orthonormal $\mathrm{L}^2(\mathbb{R})$ bases $\{\psi_{j,k}\}$; 2) a recursive algorithm, the so–called pyramidal algorithm, is available for performing the convolution/downsampling operations at all scale j. The complexity of this algorithm is $O(n)$ for a sample of length n, see Mallat (1998). In other

words, in an MRA, the computation (13) can be made recursively as j grows and it is not necessary to explicitly compute the filters $h_{j,\cdot}$ defined in (14).

Assumptions (W-1)-(W-4) are standard in the context of an MRA, see for instance Cohen (2003). Common wavelets are Daubechies wavelet and Coiflets (for which the scale function also has vanishing moments). What matters in the asymptotic theory of wavelet estimators presented below is the number of vanishing moments M and the decay exponent α, which both determine the frequency resolution of ψ. For standard wavelets, M is always known and (Cohen, 2003, Remark 2.7.1, Page 86) provides a sequence of lower bounds (α_k) tending to α as $k \to \infty$. Daubechies wavelet are defined by their number M of vanishing moments, for any $M \geq 1$ (the case $M = 1$ corresponds to the so called Haar wavelet). An analytic formula for their decay exponent α is available, see (Daubechies, 1992, Eq (7.1.23), Page 225 and the table on Page 226) and note that our α equals the α of Daubechies (1992) plus 1. A simpler lower bound $\alpha \geq (1-\log_2(3)/2)M$ holds for Daubechies wavelets, see Daubechies (1992). Although it is not sharp, it shows that the number of vanishing moments M and decay exponent α of Daubechies's wavelets can be made arbitrarily large at the same time. Table 1 provides some values of α for Daubechies wavelets and the lower bound α_k with $k = 10$ for Coiflets with given number of vanishing moments M ranging from 2 to 10.

M	2	3	4	5	6	7	8	9	10
α (DB)	1.3390	1.6360	1.9125	2.1766	2.4322	2.6817	2.9265	3.1676	3.4057
α_{10} (Coif.)	1.6196	N.A.	1.9814	N.A.	2.5374	N.A.	3.0648	N.A.	3.5744

TABLE 1. The decay exponent α or its lower bound α_{10} of $|\widehat{\phi}(\xi)|$ (and hence of $|\widehat{\psi}(\xi)|$) with M vanishing moments. First line: M; second line: α for Daubechies wavelet; third line: the lower bound α_{10} for the Coiflet. N.A. stands for *not available* (Coiflets are defined for M even).

In view of Table 1, one can observe that the decays of Coiflets are slightly faster than the ones of Daubechies for given M's. On the other hand the Daubechies wavelets have shorter support, since it is of length $\mathrm{T} = 2M$ while it is of length $\mathrm{T} = 3M$ for Coiflets. The support length impacts on the number of available wavelet coefficients: given a sample size n, the greater the support length T the smaller the cardinality of the set $\mathcal{I}_n$ defined in (12).

We should also mention the so-called Shannon wavelet ψ_S whose Fourier transform $\widehat{\psi}_S$ satisfies

$$|\widehat{\psi_S}(\xi)|^2 = \begin{cases} 1 & \text{for} \quad |\xi| \in [\pi, 2\pi] \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

This wavelet satisfies (W-2)–(W-4) for arbitrary large M and α but does not have compact support, hence it does not satisfy (W-1). We may therefore not choose this wavelet in our analysis. It is of interest, however, because it gives a rough idea of what happens when α and M are large since one can always construct a wavelet ψ satisfying (W-1)–(W-4) which is arbitrarily close to the Shannon wavelet.

4.3. **The local regression wavelet (LRW) estimator of d.** For any integers n, j_0 and j_1, $j_0 \leq j_1$, the set of all available wavelet coefficients from n observations $X_1, \ldots, X_n$ having scale indices between j_0 and j_1 is

$$\mathcal{I}_n(j_0, j_1) \stackrel{\text{def}}{=} \{(j,k) : j_0 \leq j \leq j_1,\ 0 \leq k < n_j\} \ , \tag{21}$$

where n_j is given in (12). Consider two integers $L < U$ satisfying

$$0 \leq L < U \leq J_n \stackrel{\text{def}}{=} \max\{j : n_j \geq 1\} \ . \tag{22}$$

The index J_n is the maximal available scale index for the sample size n; L and U will denote, respectively, the lower and upper scale indices used in the estimation. For an M(d) process, under regularity conditions on the short-memory part f^* and for appropriately chosen scale function and wavelet ϕ and ψ, it may be shown that as $j \to \infty$, $\sigma_j^2(d, f^*) \stackrel{\text{def}}{=} \mathrm{Var}[W_{j,0}^X] \asymp \sigma^2 2^{2dj}$ and the empirical variance

$$\widehat{\sigma}_j^2 \stackrel{\text{def}}{=} n_j^{-1} \sum_{k=0}^{n_j-1} \left(W_{j,k}^X\right)^2 \ , \tag{23}$$

is a consistent sequence of estimator of $\sigma_j^2(d, f^*)$ (see Proposition 2). A popular semi-parametric estimator of the memory parameter d is the *local regression wavelet* (LRW) estimator of Abry and Veitch (1998), defined as the least squares estimator in the "linear regression model"

$$\log\left[\widehat{\sigma}_j^2\right] = \log\sigma^2 + dj\{2\log(2)\} + u_j \ ,$$

where $u_j = \log[\widehat{\sigma}_j^2/\sigma^2 2^{2dj}]$. This regression problem can be solved in closed form:

$$\widehat{d}_n^{\mathrm{LRW}}(L, U, \mathbf{w}) \stackrel{\text{def}}{=} \sum_{j=L}^{U} w_{j-L} \log\left(\widehat{\sigma}_j^2\right) \ , \tag{24}$$

(in short $\widehat{d}^{\mathrm{LRW}}$) where the vector $\mathbf{w} \stackrel{\text{def}}{=} [w_0, \ldots, w_{U-L}]^T$ of weights satisfies

$$\sum_{j=0}^{U-L} w_j = 0 \quad \text{and} \quad 2\log(2) \sum_{j=0}^{U-L} j w_j = 1 \ . \tag{25}$$

For $U - L = \ell \geq 1$, one may choose, for example, $\mathbf{w}$ corresponding to the weighted least-squares regression vector, defined by

$$\mathbf{w} \stackrel{\text{def}}{=} DB(B^T DB)^{-1}\mathbf{b} \tag{26}$$

where

$$\mathbf{b} \stackrel{\text{def}}{=} \begin{bmatrix} 0 \\ (2\log(2))^{-1} \end{bmatrix}, \quad B \stackrel{\text{def}}{=} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 0 & 1 & \dots & \ell \end{bmatrix}^T \tag{27}$$

is the design matrix and D is an arbitrary positive definite matrix. We will discuss the choice of the regression weights $\mathbf{w}$ after stating Theorem 3, which provides the asymptotic variance of $\widehat{d}^{\,\mathrm{LRW}}(L, U, \mathbf{w})$.

4.4. **The local Whittle wavelet (LWW) estimator of d.** Let $\{c_{j,k},\, (j,k) \in \mathcal{I}\}$ be an array of centered independent Gaussian random variables with variance $\mathrm{Var}(c_{j,k}) = \sigma^2_{j,k}$, where $\mathcal{I}$ is a finite set. The negative of its log-likelihood is $(1/2)\sum_{(j,k)\in\mathcal{I}} \left\{ c^2_{j,k}/\sigma^2_{j,k} + \log(\sigma^2_{j,k}) \right\}$ up to a constant additive term. The *local Whittle wavelet* (LWW) estimator uses such a contrast process to estimate the memory parameter d_0 by choosing $c_{j,k} = W^X_{j,k}$ and $\mathcal{I} = \mathcal{I}_n(L, U)$ as defined in (21) for *appropriately chosen* lower and upper scale indices L and U, so that the corresponding wavelet coefficients $W^X_{j,k}$ are computed from $X_1, \dots, X_n$. The scaling property $\sigma^2_j(d, f^*) \asymp \sigma^2 2^{2dj}$ and weak dependence conditions of the wavelet coefficients then suggest the following *pseudo* negative log-likelihood

$$\widehat{\mathrm{L}}_{\mathcal{I}}(\sigma^2, d) = \frac{1}{2\sigma^2} \sum_{(j,k)\in\mathcal{I}} 2^{-2dj} (W^X_{j,k})^2 + \frac{|\mathcal{I}|}{2} \log(\sigma^2 2^{2\langle \mathcal{I} \rangle\, d}) \,,$$

where $|\mathcal{I}|$ denotes the cardinal of $\mathcal{I}$ and $\langle \mathcal{I} \rangle$ is the average scale, $\langle \mathcal{I} \rangle \stackrel{\text{def}}{=} |\mathcal{I}|^{-1} \sum_{(j,k)\in\mathcal{I}} j$. Define $\widehat{\sigma}^2_{\mathcal{I}}(d) \stackrel{\text{def}}{=} \mathrm{Argmin}_{\sigma^2>0}\, \widehat{\mathrm{L}}_{\mathcal{I}}(\sigma^2, d) = |\mathcal{I}|^{-1} \sum_{(j,k)\in\mathcal{I}} 2^{-2dj} (W^X_{j,k})^2$. The pseudo maximum likelihood estimator of the memory parameter is then equal to the minimum of the negative profile log-likelihood,

$$\widehat{d}^{\,\mathrm{LWW}}(L, U) \stackrel{\text{def}}{=} \operatorname*{Argmin}_{d\in[\Delta_1,\Delta_2]} \widehat{\mathrm{L}}_{\mathcal{I}_n(L,U)}(\widehat{\sigma}^2_{\mathcal{I}}(d), d) \tag{28}$$

where $[\Delta_1, \Delta_2]$ is an interval of admissible values for d and

$$\tilde{\mathrm{L}}_{\mathcal{I}}(d) \stackrel{\text{def}}{=} \log \left(\sum_{(j,k)\in\mathcal{I}} 2^{2d(\langle \mathcal{I} \rangle - j)} (W^X_{j,k})^2 \right) . \tag{29}$$

This estimator has been proposed for analyzing noisy data in Wornell and Oppenheim (1992), and was then considered by several authors, mostly in a parametric context, see *e.g.* Kaplan and Kuo (1993) and McCoy and Walden (1996). If $\mathcal{I}$ contains at least two different scales then $\tilde{\mathrm{L}}_{\mathcal{I}}(d) \to \infty$ as $d \to \pm\infty$, and thus $\widehat{d}$ is finite. This contrast is strictly convex, and the minimum is unique: it can be found using any one-dimensional convex optimization procedure. In contrast to the the local regression wavelet estimator, the definition of LWWE does not relies on particular weights. An important issue for both estimators is the choice of the scale indices L and U. The asymptotic

theory developed for these estimators in a semi-parametric context sheds some light on the role played by these quantities, as will be explained in Section 7.

5. Fourier semi-parametric estimators of the memory parameter

5.1. **The periodogram.** Given n observations $X_1, \cdots, X_n$, the *discrete Fourier transform* (DFT) and the *periodogram* are respectively defined as

$$D^X(\lambda) \stackrel{\text{def}}{=} (2\pi n)^{-1/2} \sum_{t=1}^{n} X_t \mathrm{e}^{\mathrm{i}t\lambda} , \quad I^X(\lambda) \stackrel{\text{def}}{=} |D^X(\lambda)|^2 . \tag{30}$$

These quantities are computed at the Fourier frequencies

$$\lambda_j \stackrel{\text{def}}{=} 2\pi j/n , \quad \text{for } k \in \{1, \ldots, \tilde{n}\}, \text{ where } \tilde{n} = \lfloor (n-1)/2 \rfloor . \tag{31}$$

For stationary and invertible $M(d)$ processes (see for example Lahiri (2003)), the DFT coefficients at Fourier frequencies are known to be *approximately* asymptotically independent outside a shrinking neighborhood of zero. Thus the Fourier transform performs a *whitening* of the data, and, as a consequence, Fourier methods have neat asymptotic statistical properties.

5.1.1. *Differencing and Tapering.* To overcome the presence of polynomial or smooth trends (see Hurvich et al. (2005a)), or to estimate the memory parameter of an $M(d)$ process beyond the stationary regime ($d > 1/2$), some adjustments are necessary. For instance, it has been suggested to apply a data taper either to the time-series X or to its δ-th order difference $\mathbf{\Delta}^\delta X$. A taper is a non-random weight function (with certain desired properties) that is multiplied to the time-series (or its difference) prior to Fourier transformation. Tapering was originally used in nonparametric spectral analysis of short memory ($d = 0$) time series in order to reduce bias due to frequency domain leakage, where part of the spectrum "leaks" into adjacent frequencies. The leakage is due to the discontinuity caused by the finiteness of the sample and is reduced by using tapers which smooth this discontinuity. But such a bias reduction inflates the variance as will be seen later in a special case in the context of long memory semi-parametric estimation (see section 8).

The idea of applying taper directly to the observations X was proposed by Velasco (1999a,b); Velasco and Robinson (2000), who considered several tapering schemes such as the cosine bell and the Zurbenko-Kolmogorov tapers (Žurbenko, 1979). These tapers h_t, $t = 1, \ldots, n$ have the property of being orthogonal to polynomials up to a given order, for a subset of Fourier frequencies,

$$\sum_{t=1}^{n} (1 + t + \cdots + t^{\delta-1}) h_t \mathrm{e}^{\mathrm{i}t\lambda_j} = 0 , \quad j \in \mathcal{J}_{\delta,n} \subset \{1, \ldots, \tilde{n}\} , \tag{32}$$

where λ_j and $\tilde{n}$ are defined in (31). A problem with this approach is that the efficiency loss due to these tapers may be quite substantial, because the set $\mathcal{J}_{\delta,n}$ can be fairly small when δ is large.

In this contribution, we rather focus on the construction suggested in Hurvich and Ray (1995) and later developed in Hurvich and Chen (2000), which consists in differencing before tapering. Differencing is a very widely used technique for detrending and inducing stationarity. The δ-th order difference will convert the memory parameter of a $M(d)$ process to $d-\delta$, and will completely remove a polynomial trend of degree $\delta-1$. To apply this technique, an upper bound to the memory parameter (or to the degree of the polynomial trend) should be known in advance. But if only an upper bound is known, δ may be chosen too large and consequently the δ-th order difference may be non-invertible, that is one may have $d-\delta \le -1/2$. This situation, referred to as *over-differencing* which may cause difficulties in spectral inference (see Hurvich and Ray (1995)). As was suggested by these authors, the use of a data taper can alleviate the detrimental effect of overdifferencing. A main drawback with this approach is that tapering inflates the variance of the estimator. To minimize this effect, the tapers should be chosen carefully.

Hurvich and Chen (2000) have defined a family of data taper depending on a single parameter τ, referred to as the *taper order*. Set $h_t = 1-\mathrm{e}^{2\mathrm{i}\pi t/n}$ and, for any integer $\tau \ge 0$, define the tapered DFT of order τ of the sequence $\mathbf{x} = \{x_k, k \in \mathbb{Z}\}$ as follows

$$D_\tau^{\mathbf{x}}(\lambda) \stackrel{\mathrm{def}}{=} (2\pi n a_\tau)^{-1/2} \sum_{t=1}^{n} h_t^\tau x_t \mathrm{e}^{\mathrm{i}t\lambda}, \quad I_\tau^{\mathbf{x}}(\lambda) \stackrel{\mathrm{def}}{=} |D_\tau^{\mathbf{x}}(\lambda)|^2 \tag{33}$$

where the subscript τ denotes the taper order and $a_\tau \stackrel{\mathrm{def}}{=} n^{-1}\sum_{t=1}^{n} |h_t|^{2\tau}$ is a normalization factor. As shown in (Hurvich and Chen, 2000, Lemma 0), the decay of the discrete Fourier transform of the taper of order τ is given by

$$\left| (2\pi n a_\tau)^{-1/2} \sum_{t=1}^{n} h_t^\tau \mathrm{e}^{\mathrm{i}t\lambda} \right| \le C \frac{n}{(1+n|\lambda|)^\tau}, \quad \lambda \in (-\pi, \pi) \ .$$

This property means that higher-order tapers control the leakage more effectively.

Note that the Fourier transform of the taper may be expressed as a finite sum of shifted Dirichlet kernels,

$$\sum_{t=1}^{n} h_t^\tau \mathrm{e}^{\mathrm{i}t\lambda} = \sum_{k=0}^{\tau} \left\{ \sum_{t=1}^{n} \mathrm{e}^{\mathrm{i}t(\lambda+\lambda_k)} \right\} \tag{34}$$

Since $\sum_{t=1}^{n} \mathrm{e}^{\mathrm{i}t\lambda_k} = 0$, this relation implies that for $j \in \{1, \ldots, \tilde{n}-\tau\}$, $\sum_{t=1}^{n} h_t^\tau \mathrm{e}^{\mathrm{i}t\lambda_j} = 0$ so that the tapered Fourier transform (evaluated at Fourier frequencies) is invariant to shift in the mean. This shift-invariance is achieved without restricting attention to a coarse grid of Fourier frequencies, as is necessary for the Zhurbenko-Kolmogorov taper (Velasco (1999a)).

However, the construction of theoretically-justified memory estimators using the tapered periodogram may require dropping some Fourier frequencies as will be seen in the definition of the pooled periodogram in (35). This is related to the following observation. For $\tau = 0$ (no taper), the DFT coefficients $D_\tau^Z(\lambda_j)$ of a white noise $\{Z_t\}$ at Fourier frequencies λ_k, λ_j, $k \ne j \in \{1, \ldots, \tilde{n}\}$

are uncorrelated. This property is lost by tapering. For $\tau \geq 1$, the correlation $\mathbb{E}[D_\tau^Z(\lambda_j)\overline{D_\tau^Z(\lambda_k)}]$ is equal to 0 if $|k-j| > \tau$, and $(2\pi a_\tau)^{-1}(-1)^k \binom{2\tau}{\tau+k}$ if $|k| \leq \tau$.

5.1.2. *Pooling.* Let $I_\tau^{\mathbf{x}}(\lambda)$ be the tapered periodogram introduced in (33). When considering non-linear transformations of the periodogram such as taking logarithm, variance reduction can be obtained by pooling groups of finitely many, say p, consecutive $I_\tau^{\mathbf{x}}(\lambda_j)$, which results in a *pooled periodogram* (see Hannan and Nicholls (1977) and Robinson (1995b)). To understand why pooling may be helpful, recall that if Z is a white Gaussian noise, then the variance of the running mean of the log-periodogram $p^{-1/2}\sum_{k=1}^p \log I^Z(\lambda_{j+k})$ is $\psi'(1)$ whereas the variance of the logarithm of the running mean $p^{-1/2}\log\left(\sum_{k=1}^p I^Z(\lambda_{j+k})\right)$ is $p\psi'(p)$, where $\psi(z) = \Gamma'(z)/\Gamma(z)$ is the digamma function (see for instance Johnson and Kotz (1970)). The quantity $p\psi'(p)$ decreases from $\pi^2/6$ to 1 as p goes from 1 to ∞. For $p = 3$, $p\psi'(p)$ is 1.1848 and its value changes slowly thereafter as $p\psi'(p) = 1 + 1/(2p) + O(p^{-2})$. Nonetheless, this shows that the variance of the local average of the log-periodogram is larger than the variance of the logarithm of local average and explains why typical values of p are $p = 3, 4$.

As seen above for the tapered DFT coefficients of a white noise in a non-asymptotic context, in order to guarantee asymptotic independence of the *tapered* periodogram ordinates, if p successive values of the periodogram are pooled, then, at the end of the block, τ DFT coefficients are dropped, where τ is the taper order. More precisely, set $K(p,\tau) = [(n-1)/2(p+\tau)]$ and for $k \in \{1, \cdots, K(p,\tau)\}$, define the *pooled periodogram* as follows

$$\bar{I}_{p,\tau}^{\mathbf{x}}(\tilde{\lambda}_k) \stackrel{\text{def}}{=} \sum_{j=(p+\tau)(k-1)+1}^{(p+\tau)(k-1)+p} I_\tau^{\mathbf{x}}(\lambda_j). \tag{35}$$

where $\tilde{\lambda}_k \stackrel{\text{def}}{=} p^{-1}\sum_{j=(p+\tau)(k-1)+1}^{(p+\tau)k} \lambda_j = (2(p+\tau)(k-1) + p + \tau + 1)\,\pi/n$. The definition of the central frequency $\tilde{\lambda}_k$ seems somehow arbitrary. Our choice is motivated by the fact that the Chen and Hurvich's taper actually mixes together τ adjacent periodogram ordinates, so that $(p+\tau)$ Fourier frequencies are mixed in each pooled and tapered periodogram ordinate. Note that the bias of the GPH estimator defined below is very sensitive to the definition of this central frequency.

5.2. **The Geweke–Porter–Hudak (**GPH**) estimator of d.** Assume that the differencing order δ induces stationarity, *i.e.* $d < \delta + 1/2$, and the taper order τ is larger than δ. Then, for certain sequences $\{\ell_n\}$ and $\{m_n\}$ which increase slowly with n, and under smoothness conditions on the short memory component f^*, the ratios of the pooled periodogram of the δ-th order difference of $Y \stackrel{\text{def}}{=} \mathbf{\Delta}^\delta X$ divided by its spectral density $\bar{I}_{p,\tau}^Y(\tilde{\lambda}_k)/f_Y(\tilde{\lambda}_k)$, $\ell_n \leq k \leq m_n$, can be regarded as approximately independent and identically distributed (i.i.d.) in a sense that can be rigorously characterized; Robinson (1995b) and Lahiri (2003).

Based on this heuristics, a popular semiparametric estimate of d is the log-periodogram estimate of Geweke and Porter-Hudak (1983), defined here [in the manner of Robinson (1995b)] as the least squares estimate in the *linear regression model*

$$\log\left[\bar{I}^Y_{p,\tau}(\tilde{\lambda}_k)\right] = \log f^*(0) + (d-\delta)g(\tilde{\lambda}_k) + u_k \ , \quad 1 \le k \le m \ , \tag{36}$$

where $g(\lambda) \stackrel{\text{def}}{=} -2\log|1-\mathrm{e}^{\mathrm{i}\lambda}|$ and u_k is "approximately" equal to $\log\left[\bar{I}^Y_{p,\tau}(\tilde{\lambda}_k)/f_X(\tilde{\lambda}_k)\right]$. This regression equation can be solved in closed form :

$$\widehat{d}^{\,\mathrm{GPH}}(m) = \sum_{k=1}^{m} \frac{\left\{g(\tilde{\lambda}_k) - m^{-1}\sum_{k=1}^m g(\tilde{\lambda}_k)\right\}}{\sum_{k=1}^m \left\{g(\tilde{\lambda}_k) - m^{-1}\sum_{k=1}^m g(\tilde{\lambda}_k)\right\}^2} \log[\bar{I}^Y_{p,\tau}(\tilde{\lambda}_k)] + \delta \ . \tag{37}$$

To simplify the notations, we have made the dependence in the differencing, tapering and pooling orders implicit. The choice of these orders δ, τ and p will be discussed in Section 8. This estimator has been introduced by Geweke and Porter-Hudak (1983) and was later used in many empirical works.

Remark 2. In the definition of the GPH estimator in Robinson (1995b), the first ℓ_n DFT coefficients are eliminated. ℓ_n is referred to as the *trimming number*. Trimming is sometimes required to eliminate deterministic trend (Hurvich et al., 2005a), or to deal with non-Gaussian processes (Velasco, 2000).

5.3. **The local Whittle Fourier (**LWF**) estimator of** d**.** Since, as mentioned above, $I^Y_\tau(\lambda_k)/f_Y(\lambda_k)$ can be regarded as approximately i.i.d. and $f_Y(\lambda) \approx C|1-\mathrm{e}^{\mathrm{i}\lambda}|^{-2(d-\delta)}$ in the neighborhood of zero, using the same arguments as in Section 4.4, we may approximate the negated likelihood as follows:

$$\widehat{L}_{\tau,m}(C,d) = m^{-1}\sum_{k=1}^{m}\left\{\log(C|1-\mathrm{e}^{\mathrm{i}\lambda_k}|^{-2(d-\delta)}) + \frac{I^Y_\tau(\lambda_k)}{C|1-\mathrm{e}^{\mathrm{i}\lambda_k}|^{-2(d-\delta)}}\right\} \ . \tag{38}$$

Note that pooling is here irrelevant because non non-linear transformation is involved. This estimator was originally proposed by Künsch (1987) and later studied in Robinson (1995a). After eliminating C by maximizing the contrast (38), we get $\tilde{L}^{\mathrm{LWF}}_{\tau,m}$ the *profile likelihood*, defined as

$$\tilde{L}^{\mathrm{LWF}}_{\tau,m}(d) \stackrel{\text{def}}{=} \log\left(m^{-1}\sum_{k=1}^{m} I^Y_\tau(\lambda_k)|1-\mathrm{e}^{\mathrm{i}\lambda_k}|^{2(d-\delta)}\right) - 2(d-\delta)m^{-1}\sum_{k=1}^{m}\log(|1-\mathrm{e}^{\mathrm{i}\lambda_k}|) \ , \tag{39}$$

and we define the *Local Whittle Fourier* estimator (LWF) as the minimum

$$\widehat{d}^{\,\mathrm{LWF}}(m) \stackrel{\text{def}}{=} \operatorname*{Argmin}_{\bar{d}\in\mathbb{R}} \tilde{L}_{\tau,m}(\bar{d}) + \delta \ . \tag{40}$$

The function $d \to \tilde{L}_{\tau,m}(d)$ is convex and thus admit a single global minimum, which can be obtained numerically by using a standard one-dimensional convex optimization algorithm. In Robinson (1995a), the minimization in (40) is performed over a closed interval which was supposed

to include the true value of the parameter. But, because the contrast is strictly convex, there is in fact no need to impose such a restriction.

5.4. **The exact and the non-stationary extended local Whittle estimators.** To conclude this section, let us mention two recent works on the estimation of the memory parameter for non-stationary $M(d)$ processes, $d \geq 1/2$, which are not covered in details in this contribution because they are derive under slightly different conditions and henceforth do not compare well with wavelet estimators.

Shimotsu and Phillips (2005) introduced an *exact local Whittle* estimator. It is applicable when the $M(d)$ series is generated by a linear process and when the domain of d is not wider than $9/2$. Their estimator is based on fractional differencing of the data and the complexity of their algorithm is of the order n^2, where n is the number of observations. In contrast, the complexity of the estimators we consider is of the order of $n \log_2(n)$; see the discussion in Moulines et al. (2007c). Note also that the model considered by Shimotsu and Phillips (2005) is not an $M(d)$ process in the sense given above and is not time-shift invariant, see their Eq. (1). In addition, their estimator is not invariant upon addition of a constant in the data, a drawback which is not easily dealt with, see their Remark 2.

Abadir et al. (2007) propose to extend the local Whittle estimator to $d \in (-3/2, \infty)$ calling it the fully extended local Whittle estimator. This estimator is based on an extended definition of the DFT $\tilde{D}^X(\lambda_j, d)$, which include correction terms, *i.e.* $\tilde{D}^X(\lambda_j, d) \stackrel{\text{def}}{=} D^X(\lambda_j) + k^X(\lambda_j, d)$. The correction term $k^X(\lambda_j, d)$, which takes constant values on the intervals $d \in [p - 1/2, p + 1/2)$, $p = 0, 1, \dots$ is defined as $k^X(\lambda_j, d) = 0$ if $d \in (-1/2, 1/2)$ and

$$k^X(\lambda_j, d) = \mathrm{e}^{-\mathrm{i}\lambda_j} \sum_{r=1}^{p} (1 - \mathrm{e}^{-\mathrm{i}\lambda_j})^{-r} Z_{n,r} \ , \quad d \in [p - 1/2, p + 1/2) \ , \quad p = 1, 2, \dots \ , \tag{41}$$

where $Z_{n,r} \stackrel{\text{def}}{=} (2\pi n)^{-1/2} (\boldsymbol{\Delta}^{r-1} X_n - \boldsymbol{\Delta}^{r-1} X_0)$, $r = 1, \dots, p$. The corresponding corrected periodogram $\tilde{I}^X(\lambda_j, d)$ is given by $\tilde{I}^X(\lambda_j, d) = |\tilde{D}^X(\lambda_j, d)|^2$. The extended local Whittle estimator is then defined as $\widehat{d}^{\mathrm{LWF}} = \mathrm{Argmin}_{\bar{d} \in [d_{\min}, d_{\max}]} L_m^{\mathrm{LWF}}(\bar{d})$ where $L^{\mathrm{LWF}}(d)$ is

$$L_m^{\mathrm{LWF}}(d) \stackrel{\text{def}}{=} \log \left(m^{-1} \sum_{k=1}^{m} \tilde{I}^X(\lambda_k, d) |1 - \mathrm{e}^{\mathrm{i}\lambda_k}|^{2d} \right) - 2dm^{-1} \sum_{k=1}^{m} \log(|1 - \mathrm{e}^{\mathrm{i}\lambda_k}|) \ .$$

Note that to compute $k^X(\lambda_j; d)$, we have to involve additional observations $X_{-p+1}, \dots, X_n$, where $p \stackrel{\text{def}}{=} 0 \vee \lfloor d_{\max} - 1/2 \rfloor$. Compared to the Shimotsu and Phillips (2005) estimator, this estimator is easy to evaluate numerically, but the approximation of the extended local Whittle function is not continuous at $d = 1/2, 3/2, \dots$, which does not allow one to obtain limit theorems at these points (and, in the finite sample case, causes disturbances in the neighborhood of these values).

In addition, this estimator is not robust to the presence of polynomial trends in the data (if d is the memory parameter, the method tolerate a polynomial trend of degree at most $\lfloor d + 1/2 \rfloor$).

6. The semi-parametric estimation setting

The theory of semi-parametric Fourier estimators was developed in two fundamental papers by Robinson, Robinson (1995b) and Robinson (1995a), which establish, under suitable conditions, the asymptotic normality of the GPH and the LWE estimators in the stationary case. These results were later extended to non-stationary M(d) processes for different versions of the memory estimator and under various sets of assumptions. The theory of semi-parametric wavelet estimators was developed much more recently in Moulines et al. (2007b) and Moulines et al. (2007c) (some preliminary results are in Bardet et al. (2000) and Bardet (2002)). To allow for comparison the wavelet and the Fourier approaches, the asymptotic properties of the estimators are presented under a common set of assumptions. Because the theory of wavelet estimators is much less developed than the theory of Fourier estimators, these assumptions can often be relaxed in the context of Fourier estimators.

There are two types of additional assumptions that enter into play in an asymptotic theory. First, the semi-parametric rates of convergence depends on the smoothness of the short-memory component in a neighborhood of zero frequency. The most common assumption, introduced in (Robinson, 1995b), is a Hölder condition on the short-memory component of the spectral density f^* in (1).

Definition 2. *For any $0 < \beta \leq 2$, $\gamma > 0$ and $\varepsilon \in (0, \pi]$, $\mathcal{H}(\beta, \gamma, \varepsilon)$ is the set of all non-negative and even function g that satisfies $g(0) > 0$ and for all $\lambda \in (-\varepsilon, \varepsilon)$*

$$|g(\lambda) - g(0)| \leq \gamma\, g(0)\, |\lambda|^{\beta} \,. \tag{42}$$

The larger the value of β, the smoother the function at the origin. Observe that if f^* is even – as assumed – and if it is in addition infinitely differentiable, then $f^{*'}(0) = 0$ and hence, by a Taylor expansion, (42) holds with $\beta = 2$, that is, in this case, one has $f^* \in \mathcal{H}(2, \gamma, \varepsilon)$. Andrews and Guggenberger (2003) extend this definition to the case $\beta > 2$ by considering even functions satisfying

$$\left| g(\lambda) - g(0) - \sum_{k=1}^{\lfloor \beta/2 \rfloor} \varphi_k \lambda^{2k}/(2k!) \right| \leq \gamma g(0) |\lambda|^{\beta - 2\lfloor \beta/2 \rfloor} \,, \tag{43}$$

with $|\varphi_k| \leq \gamma g(0)$, for any $k \in \{1, \ldots, \lfloor \beta/2 \rfloor\}$. To take advantage of this more refined smoothness assumption when $\beta > 2$, bias reduction techniques must be applied (see for example Andrews and Guggenberger (2003); Robinson and Henry (2003); Andrews and Sun (2004)). We will only consider $\beta \leq 2$ for sake of brevity.

Second, the definition of an M(d) process accounts only for the spectral (or equivalently covariance) structure, which specifies the distribution of the process if X is Gaussian. To extend the results in the non-Gaussian context, it is necessary to specify the distribution of the process beyond its second-order properties. The most complete asymptotic theory has been developed so far for linear processes.

Definition 3. *We say that X is a* strong linear M(d) process *if there exist a $\mathrm{L}^2(\mathbb{Z})$-sequence $\{a_s\}_{s\in\mathbb{Z}}$ and an i.i.d. sequence $\{Z_s\}_{s\in\mathbb{Z}}$ satisfying $\mathbb{E}[Z_0]=0$, $\mathbb{E}[Z_0^4]<\infty$, and for any $t\in\mathbb{Z}$,*

$$(\boldsymbol{\Delta}^K X)_t = \sum_{s\in\mathbb{Z}} a_s Z_{t-s}\ , \quad K \stackrel{\text{def}}{=} \lfloor d+1/2 \rfloor\ . \tag{44}$$

Remark 3. According to the standard terminology, *strong* here refers to the fact that $\{Z_t\}$ is i.i.d. (or a strong white noise). This assumption can often be relaxed by supposing that $\{Z_t\}$ is a martingale difference sequence ($\mathrm{E}\,[\,Z_t\,|\,\mathcal{F}_{t-1}]=0$ where $\mathcal{F}_t=\sigma(Z_s, s\le t)$ is the natural filtration of the process) satisfying various additional conditional moment assumptions. For example, Robinson (1995a), and many authors after that, assume that $\{Z_t^2-\mathbb{E}[Z_t^2]\}$ is a square integrable martingale difference.

The following theorem, established in Giraitis et al. (1997), provides a lower bound for the estimation error.

Theorem 1. *Let $d_{\min}<d_{\max}$ in $\mathbb{R}$, $\varepsilon\in(0,\pi]$, $\beta\in(0,2]$ and $\gamma>0$. There exists a constant $c>0$ such that,*

$$\liminf_{n\to\infty} \inf_{\widehat{d}_n} \sup_{d_{\min}\le d\le d_{\max}} \sup_{f^*\in\mathcal{H}(\beta,\gamma,\varepsilon)} \mathbb{P}_{d,f^*}\left(n^{\beta/(2\beta+1)}|\widehat{d}_n-d|\ge c\right)>0, \tag{45}$$

where the infimum $\inf_{\widehat{d}_n}$ is taken over all possible estimators based on $\{X_1,\cdots,X_n\}$ and $\mathbb{P}_{d,f^*}$ denotes the distribution of a Gaussian M(d) process $\{X_t\}_{t\in\mathbb{Z}}$ with generalized spectral density of the form (3).

We shall see in the sequel that the best possible rate $n^{\beta/(2\beta+1)}$ is achieved by both wavelet and Fourier estimators when X is a strong linear M(d) process.

The theory of the semi-parametric estimation of d for several *non-linear* M(d) processes, used in particular in financial econometric, and in teletrafic modeling, have also been investigated in the literature (see the recent surveys Deo et al. (2006a) and Teyssière and Abry (2007) and the references therein). The stochastic volatility model (a special instance of the signal plus noise model) has been considered in Hurvich et al. (2005b); Deo et al. (2006b), which establish consistency, rate of convergence and asymptotic normality of an appropriately modified LWF estimator. Dalla et al. (2006) provide general conditions under which the LWF estimator of the memory parameter of a stationary process is consistent and examines its rate of convergence.

This class of processes include, among others, signal plus noise processes, nonlinear transforms of a Gaussian process, and exponential generalized autoregressive, conditionally heteroscedastic (EGARCH) models, etc... Faÿ et al. (2007) provide the consistency and rate of convergence of the LWW estimator for the infinite source Poisson process. The results are in general not as complete as in the linear case, and the required assumptions are specific to each considered model (abstract assumptions like in (Dalla et al., 2006, Eq. (8)) or (Moulines et al., 2007c) can be used, but checking these still require model-dependent conditions). In addition, the asymptotic behavior of the estimators are model-dependent and might depart significantly from the results obtained in the linear case.

7. Asymptotic properties of the wavelet estimators LRW and LWW

7.1. **The between-scale process.** Before stating the main known results on the asymptotic behavior of LWW and LRW estimators, some additional definitions related to the spectral density of wavelet coefficients are required.

If the process X is an $\mathrm{M}(d)$ process, as defined in Section 2, and $M > d - 1/2$, then $\mathbf{\Delta}^M X$ is weakly stationary. It follows that the process $\{W^X_{j,k}\}_{k\in\mathbb{Z}}$ of wavelet coefficients at scale $j \geq 0$ is weakly stationary in k. However the two–dimensional process $\{[W^X_{j,k},\, W^X_{j',k}]^T\}_{k\in\mathbb{Z}}$ of wavelet coefficients at two different scales j and j', with $j \neq j'$, is not weakly stationary. This is why we consider the *between-scale* process $\{[W^X_{j,k},\, \mathbf{W}^X_{j,k}(j-j')^T]^T\}_{k\in\mathbb{Z}}$, where

$$\mathbf{W}^X_{j,k}(u) \stackrel{\text{def}}{=} \left[W^X_{j-u,2^u k},\, W^X_{j-u,2^u k+1},\, \ldots,\, W^X_{j-u,2^u k+2^u-1}\right]^T . \tag{46}$$

The index u in (46) denotes the scale difference $j - j' \geq 0$ between the finest scale j' and the coarsest scale j. If $u = 0$, that is $j = j'$, then $\mathbf{W}^X_{j,k}(0)$ is the scalar $W^X_{j,k}$. For $u > 0$, the second component of the between–scale process is a vector whose entries are the wavelet coefficients with scale index $j - u = j'$ and translation indices $2^{j-j'}k, 2^{j-j'}k + 1, \ldots, 2^{j-j'}(k+1) - 1$. Using (13), it follows that

$$\begin{aligned}\mathbf{W}^X_{j,k}(u) &= [(h_{j-u,\cdot} \star \mathbf{x})_{2^j k + l\, 2^{j-u}},\ l = 0, \ldots, 2^u - 1]^T \\ &= \left[(\downarrow^j \circ\, \mathbf{B}^{-l\, 2^{j-u}}[h_{j-u,\cdot} \star \mathbf{x}])_k,\ l = 0, \ldots, 2^u - 1\right]^T ,\end{aligned}$$

where $\mathbf{B}$ is the shift operator defined, for any sequence $\{c_k\}_{k\in\mathbb{Z}}$, by $(\mathbf{B}c)_k = c_{k-1}$. The between–scale process is stationary in k because all its entries can be expressed with the same downsampling operator $\downarrow^j$ applied to jointly stationary time series, namely, $h_{j,\cdot} \star \mathbf{x}$ and $\mathbf{B}^{-l\, 2^{j-u}}[h_{j-u,\cdot} \star \mathbf{x}] = [\mathbf{B}^{-l\, 2^{j-u}} h_{j-u,\cdot}] \star \mathbf{x},\ l = 0, \ldots, 2^u - 1$. One can therefore write, for all $0 \leq u \leq j$,

$$\mathrm{Cov}_f(W^X_{j,k}, \mathbf{W}^X_{j,k'}(u)) = \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}\lambda(k-k')}\, \mathbf{D}_{j,u}(\lambda; f)\, \mathrm{d}\lambda ,$$

where $\mathbf{D}_{j,u}(\lambda; f)$ is the cross-spectral density function of the between-scale process. The case $u = 0$ corresponds to the spectral density of the *within-scale* process $\{W_{j,k}^X\}_{k\in\mathbb{Z}}$. When X is an M(d) process with spectral density function (3), we often denote $\mathbf{D}_{j,u}(\lambda; f)$ by $\mathbf{D}_{j,u}(\lambda; d, f^*)$.

7.2. **Generalized fractional Brownian motion.** We shall approximate the within- and between-scale spectral densities $\mathbf{D}_{j,u}(\lambda; d, f^*)$ of the process X with memory parameter $d \in \mathbb{R}$ by the corresponding spectral densities of the *generalized fractional Brownian motion* $B_{(d)}$. This process is parametrized by a family $\Theta_{(d)}$ of smooth test functions $\theta(t)$, $t \in \mathbb{R}$ and is defined as follows: $\{B_{(d)}(\theta),\, \theta \in \Theta_{(d)}\}$ is a mean zero Gaussian process with covariance

$$\mathrm{Cov}\left(B_{(d)}(\theta_1),\, B_{(d)}(\theta_2)\right) = \int_{\mathbb{R}} |\lambda|^{-2d}\, \widehat{\theta}_1(\lambda)\, \overline{\widehat{\theta_2}(\lambda)}\, \mathrm{d}\lambda\ . \tag{47}$$

The finiteness of the integral $\int_{\mathbb{R}} |\lambda|^{-2d}\, |\widehat{\theta}(\lambda)|^2\, \mathrm{d}\lambda$ provides a constraint on the family $\Theta_{(d)}$. For instance, when $d > 1/2$, this condition requires that $\widehat{\theta}(\lambda)$ decays sufficiently quickly at the origin and, when $d < 0$, it requires that $\widehat{\theta}(\lambda)$ decreases sufficiently rapidly at infinity. Hence, under (W-1)-(W-4), θ can be a wavelet ψ if $d \in (1/2 - \alpha, M + 1/2)$. The discrete wavelet transform of $B_{(d)}$ is defined as

$$W_{j,k}^{(d)} \stackrel{\mathrm{def}}{=} B_{(d)}(\psi_{j,k}), \quad (j,k) \in \mathbb{Z}\times\mathbb{Z}\ . \tag{48}$$

As shown in (Moulines et al., 2007b, Relation (35)), for all j, k and k' in $\mathbb{Z}$, and $u \geq 0$, one has

$$\mathrm{Cov}\left(W_{j,k}^{(d)},\, \mathbf{W}_{j,k'}^{(d)}(u)\right) = 2^{2dj} \int_{-\pi}^{\pi} \mathbf{D}_{\infty,u}(\lambda; d)\, \mathrm{e}^{\mathrm{i}\lambda(k-k')}\, \mathrm{d}\lambda\ , \tag{49}$$

where $\mathbf{D}_{\infty,u}(\lambda; d)$ does not involve j and is given by

$$\mathbf{D}_{\infty,u}(\lambda; d) = \sum_{l\in\mathbb{Z}} |\lambda + 2l\pi|^{-2d}\, \mathbf{e}_u(\lambda + 2l\pi)\, \overline{\widehat{\psi}(\lambda + 2l\pi)}\widehat{\psi}(2^{-u}(\lambda + 2l\pi))\ , \tag{50}$$

where, for all $u \geq 0$, $\mathbf{e}_u(\xi) \stackrel{\mathrm{def}}{=} 2^{-u/2}\, [1, \mathrm{e}^{-\mathrm{i}2^{-u}\xi}, \ldots, \mathrm{e}^{-\mathrm{i}(2^u-1)2^{-u}\xi}]^T$, $\quad \xi \in \mathbb{R}$. As mentioned above, $W_{j,k}^{(d)}$ is well defined under (W-1)-(W-4) when $d \in (1/2 - \alpha, M + 1/2)$. Under the same condition, using the decay condition of $\widehat{\psi}$ in (W-2) and (7), one easily gets that for any $u \geq 0$, $\mathbf{D}_{\infty,u}(\lambda; d)$ is continuous $(-\pi, \pi) \setminus \{0\}$ and $|\mathbf{D}_{\infty,u}(\lambda; d)| = O(|\lambda|^{2(M-d)})$ as $\lambda \to 0$. If moreover $d \leq M$, one has (see Relation (72) in Moulines et al. (2007c))

$$\sup_{u\geq 0} 2^{u(2d-1/2)} \int_{-\pi}^{\pi} |\mathbf{D}_{\infty,u}(\lambda; d)|^2\ \mathrm{d}\lambda < \infty\ . \tag{51}$$

The variance $\sigma_j^2(d, f^*) \stackrel{\mathrm{def}}{=} \mathrm{Var}[W_{j,0}^X]$ can be interpreted as a *scale spectrum*, in words, the *power* at scale j. It is approximated by $f^*(0)\mathrm{K}(d)\, 2^{2jd}$ where the constant $\mathrm{K}(d)$ is given by

$$\mathrm{K}(d) \stackrel{\mathrm{def}}{=} \int_{-\infty}^{\infty} |\xi|^{-2d}\, |\widehat{\psi}(\xi)|^2\, d\xi\ . \tag{52}$$

7.3. **Uniform bounds.** Using (Moulines et al., 2007b, Theorem 1) and (Moulines et al., 2007c, Theorem 1), the following result holds.

Proposition 2. *Let X be an M(d) process with $d \in \mathbb{R}$ and $f^* \in \mathcal{H}(\beta, \gamma, \varepsilon)$ for some $\beta, \gamma > 0$ and $\varepsilon \in (0, \pi]$. Assume (W-1)-(W-4) with $(1+\beta)/2 - \alpha < d < M + 1/2$,*

$$\left| \sigma_j^2(d, f^*) - f^*(0)\, \mathrm{K}(d)\, 2^{2jd} \right| \le C\, f^*(0)\, L\, 2^{(2d-\beta)j} \tag{53}$$

If moreover $\varepsilon = \pi$ and $d \le M$, then, for all $\lambda \in (-\pi, \pi)$, $j \ge u \ge 0$,

$$\left| \mathbf{D}_{j,u}(\lambda; d, f^*) - f^*(0)\, \mathbf{D}_{\infty,u}(\lambda; d)\, 2^{2jd} \right| \le C\, f^*(0)\, L\, 2^{(2d-\beta)j} \tag{54}$$

where $|y|$ denotes the Euclidean norm of any vector y.

In (53) and (54), the constant C only depends on d, β and on the wavelets ψ and ϕ. It can be made independent of d on any compact set included in $((1+\beta)/2 - \alpha, M + 1/2)$ for (53) and in $((1+\beta)/2 - \alpha, M]$ for (54). This proposition shows that the covariance properties of the wavelet coefficients of an $\mathrm{M}(d)$ process resemble those of the generalized FBM B_d at large scales. The latter are not, in general, decorrelated as sometimes heuristically assumed (see Abry and Veitch (1998) and Veitch and Abry (1999) for example). Exact decorrelation occurs but in very specific cases: if $\{\psi_{j,k}\}$ is an orthonormal basis of $\mathrm{L}^2(\mathbb{R})$ and $d = 0$, see (Moulines et al., 2007b, Remark 7). Due to its very specific spectral property, the *ideal* Shannon wavelet coefficients (defined in (20)) of B_d satisfy partial independence for $d \neq 0$. Indeed, applying (20) in (50), we get, for all $\lambda \in (-\pi, \pi)$,

$$\mathbf{D}_{\infty,u}(\lambda; d) = \begin{cases} 0 & \text{for} \quad u \ge 1 \\ (2\pi - |\lambda|)^{-2d} & \text{otherwise,} \end{cases} \tag{55}$$

implying that $W_{j,k}^{(d)}$ and $W_{j',k'}^{(d)}$ are uncorrelated for $j \neq j'$ and that $W_{j,k}^{(d)}$ and $W_{j,k'}^{(d)}$ are uncorrelated only if $d = 0$.

The asymptotic behavior of wavelet estimators $\widehat{d}^{\,\mathrm{LRW}}(L_n, U_n, \mathbf{w}_n)$ and $\widehat{d}^{\,\mathrm{LWW}}(L_n, U_n)$ defined in (24) and (28) will be derived for specific lower and upper scale sequences (L_n) and (U_n). In the semiparametric framework, the lower scale sequence (L_n) governs the rate of convergence of the memory estimator. There are two possible settings as far as the upper scale sequence (U_n) is concerned:

(S-1) $U_n - L_n$ is fixed, equal to $\ell > 0$

(S-2) $U_n \le J_n$ for all n and $U_n - L_n \to \infty$ as $n \to \infty$, where J_n is the largest available scale defined in (22).

7.4. **Asymptotic properties of the** LRW **estimator.** We will use the following definition, for all $i, j \geq 0$,

$$\mathbf{V}_{i,j}(d) \stackrel{\text{def}}{=} \frac{4\pi 2^{2d|i-j|} 2^{i\wedge j}}{\mathrm{K}(d)^2} \int_{-\pi}^{\pi} \left|\mathbf{D}_{\infty,|i-j|}(\lambda; d)\right|^2 \, d\lambda \, , \tag{56}$$

with $\mathbf{D}_{\infty,u}(\lambda; d) \in \mathbb{R}^{2^u}$ defined by (50). The following result, adapted from Moulines et al. (2007a), applies to Gaussian M(d) processes; it has recently been extended to strong linear M(d) processes in Roueff and Taqqu (2007).

Theorem 3. *Let X be a Gaussian M(d) process with generalized spectral density given by* (3) *with $d \in \mathbb{R}$ and $f^* \in \mathcal{H}(\beta, \gamma, \varepsilon)$ for some $\gamma > 0$, $\beta \in (0, 2]$ and $\varepsilon \in (0, \pi]$. Assume (W-1)-(W-4) with*

$$(1 + \beta)/2 - \alpha < d \leq M \, . \tag{57}$$

Let (L_n) be a sequence satisfying

$$\lim_{n\to\infty} \left\{ n 2^{-(1+2\beta)L_n} + (n 2^{-L_n})^{-1} \right\} = 0 \, . \tag{58}$$

and $\mathbf{w}$ be a weight of length $\ell + 1$ satisfying (25)*. Then, as $n \to \infty$,*

$$\sqrt{n 2^{-L_n}} \left(\widehat{d}^{\,\mathrm{LRW}}(L_n, L_n + \ell, \mathbf{w}) - d \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{i,j=0}^{\ell} \mathbf{w}_i \mathbf{V}_{i,j}(d) \mathbf{w}_j \right) \, . \tag{59}$$

Remark 4. This result is stated in Moulines et al. (2007a) with $\varepsilon = \pi$. The case $\varepsilon < \pi$ can be obtained by using (Moulines et al., 2007c, Corollary 2).

Remark 5. The larger the value of β, the smaller the size of the allowed range for d_0 in (57) for a given decay exponent α and number M of vanishing moments. Indeed the range in (57) has been chosen so as to obtain a bound on the bias which corresponds to the best possible rate under the condition $f^* \in \mathcal{H}(\beta, \gamma, \varepsilon)$. If (57) is replaced by the weakest condition $d_0 \in (1/2 - \alpha, M]$, which does not depend on β, the same CLT (57) holds but β in Condition (58) must be replaced by $\beta' \in (0, \beta]$. This β' must satisfy $1/2 - \alpha < (1 + \beta')/2 - \alpha < d_0$, that is $0 < \beta' < 2(d_0 + \alpha) - 1$. When $\beta' < \beta$ one gets a slower achievable rate in (59).

Remark 6. The condition $n 2^{-(1+2\beta)L_n} \to 0$ guarantees that the bias is negligible in the limit. The optimal rate $n^{\beta/(1+2\beta)}$ given by Theorem 1 is obtained with $n 2^{-(1+2\beta)L_n} \asymp 1$, in which case the squared bias and the variance are of the same order of magnitude, see Theorem 3 in Moulines et al. (2007b) where uniform bounds of the mean square error and an exact equivalent of the variance are given. The asymptotic equivalent of the variance is $(n 2^{-L_n})^{-1} \sum_{i,j=0}^{\ell} \mathbf{w}_i \mathbf{V}_{i,j}(d) \mathbf{w}_j$ as can be expected from (59).

Remark 7. Theorem 3 applies only to the setting (S-1) since $U_n = L_n + \ell$. Under the setting (S-2), $U_n - L_n \to \infty$ as $n \to \infty$, one has to replace the weights $\mathbf{w}$ by a sequence of weights $(\mathbf{w}_n)$ of lengths $U_n - L_n + 1$. Using (Moulines et al., 2007b, Proposition 4), if $\varepsilon = \pi$, it is possible to extend the Theorem 3 to this setting, provided that $w_{n,i} \to w_{\infty,i}$ as $n \to \infty$ for all i and

$$\lim_{\ell\to\infty} \sum_{i>\ell} \sup_n |w_{n,i}| 2^{i/2} = 0 \,, \tag{60}$$

in which case one has

$$\sqrt{n2^{-L_n}} \left(\widehat{d}^{\mathrm{LRW}}(L_n, U_n, \mathbf{w}_n) - d \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{i,j=0}^{\infty} w_{\infty,i} \mathbf{V}_{i,j}(d) w_{\infty,j} \right) \,.$$

The variance in the right-hand side of the last display is finite as a consequence of (51) and (60).

The standard theory of linear regression shows that, for any fixed $\ell \geq 1$, the optimal design matrix is $D = \mathbf{V}^{-1}(d,\ell)$, where $\mathbf{V}(d,\ell) = [\mathbf{V}_{i,j}(d)]_{0\leq i,j,\leq \ell}$ is a $(\ell+1)\times(\ell+1)$ matrix. By (26), the corresponding weights read

$$\mathbf{w}^{\mathrm{opt}}(d,\ell) \stackrel{\mathrm{def}}{=} \mathbf{V}^{-1}(d,\ell) B (B^T \mathbf{V}^{-1}(d,\ell) B)^{-1} \mathbf{b} \,, \tag{61}$$

where B and $\mathbf{b}$ are defined by (27) and the associated limiting variance is

$$\rho^2_{\mathrm{opt}}(d,\ell) \stackrel{\mathrm{def}}{=} \mathbf{w}^{\mathrm{opt}}(d,\ell)^T \mathbf{V}(d,\ell) \mathbf{w}^{\mathrm{opt}}(d,\ell) = \mathbf{b}^T (B^T \mathbf{V}^{-1}(d,\ell) B)^{-1} \mathbf{b} \,. \tag{62}$$

Since the regression vector weights w of length ℓ can be viewed as regression vector weights of length $\ell+1$ with zero as last coordinate, we have that $\rho^2_{\mathrm{opt}}(d,\ell)$ decreases as ℓ increases and we will denote its limit by $\rho^2_{\mathrm{opt}}(d,\infty)$. Figure 3 shows that the limit is approximately attained for $\ell \geq 7$ for a standard choice of wavelet.

The optimal regression vector $\mathbf{w}^{\mathrm{opt}}(d,\ell)$ cannot be used directly since it depends on unknown the memory parameter d, but a plug-in method can be used as suggested by Bardet (2002) in a similar context: a preliminary consistent estimator, say $\widehat{d}^{(1)}$, is used to estimate the optimal weights $\widehat{\mathbf{w}} = \mathbf{w}^{\mathrm{opt}}(\widehat{d}^{(1)},\ell)$ and then the estimator $\widehat{d}^{\mathrm{LRW}}(L,U,\widehat{\mathbf{w}})$ is applied.

A different choice of weights is suggested by Abry and Veitch (1998) (in a parametric context). This choice relies on the approximation that $\mathbf{D}_{\infty,u}(\lambda;d)/\mathrm{K}(d)$ is nearly zero for $u > 0$ and nearly constant equal to $(2\pi)^{-1}$ for $u = 0$. This provides a diagonal approximation of $\mathbf{V}^{-1}(d,\ell)$ yielding a diagonal design matrix D with diagonal entries $D_{i,i} = 2^{-i}$, $i = 0,\ldots,\ell$, up to a multiplicative constant. By straightforward computations, this design matrix define the following *Abry–Veitch* weights.

$$\mathbf{w}^{\mathrm{AV}}_i(\ell) \stackrel{\mathrm{def}}{=} \frac{(i-\eta_\ell)2^{-i}}{2\log(2)\kappa_\ell(2-2^{-\ell})}, \quad i = 0,\ldots,\ell \,, \tag{63}$$

where

$$\eta_\ell \stackrel{\text{def}}{=} \sum_{j=0}^{\ell} j \frac{2^{-j}}{2-2^{-\ell}} \quad \text{and} \quad \kappa_\ell \stackrel{\text{def}}{=} \sum_{j=0}^{\ell} (j-\eta_\ell)^2 \frac{2^{-j}}{2-2^{-\ell}} \, . \tag{64}$$

This choice, while not optimal, is closed to it in practice, at least for not too large values of d and with a standard choice of wavelets, see Figure 2. One advantage of this choice of regression vector weights stems from the fact that it does not require the use of a pilot estimator, since the weights do not depend on the unknown parameter d.

Let us denote, for all $u \geq 0$ (see Moulines et al. (2007c)),

$$\mathrm{I}_u(d) \stackrel{\text{def}}{=} \int_{-\pi}^{\pi} |\mathbf{D}_{\infty,u}(\lambda;d)|^2 \, \mathrm{d}\lambda = (2\pi)^{-1} \sum_{\tau\in\mathbb{Z}} \mathrm{Cov}^2 \left(W_{0,0}^{(d)}, W_{-u,\tau}^{(d)} \right) \, , \tag{65}$$

with $\mathbf{D}_{\infty,u}(\lambda;d) \in \mathbb{R}^{2^u}$ defined by (50). In the case where the weights $\mathbf{w}$ are chosen as proposed by Abry and Veitch (1998) given by (63), the asymptotic variance in the right-hand side of (59) reads

$$\rho^2(d,\ell) \stackrel{\text{def}}{=} \sum_{i,j=0}^{\ell} \mathbf{w}_i^{\mathrm{AV}}(d,\ell) \mathbf{V}_{i,j}(d,\ell) \mathbf{w}_j^{\mathrm{AV}}(d,\ell) \, ,$$

where $\ell+1$ is the number of scales used in the regression. Inserting (63) and (65), we get, for any $\ell \geq 1$,

$$\rho^2(d,\ell) = \frac{\pi}{(2-2^{-\ell})\kappa_\ell (\log(2)\mathrm{K}(d))^2} \times \left\{ \mathrm{I}_0(d) + \frac{2}{\kappa_\ell} \sum_{u=1}^{\ell} \mathrm{I}_u(d) \, 2^{(2d-1)u} \sum_{i=0}^{\ell-u} \frac{2^{-i}}{2-2^{-\ell}} (i-\eta_\ell)(i+u-\eta_\ell) \right\} \, , \tag{66}$$

where $\mathrm{K}(d)$ is defined in (52), and η_ℓ and κ_ℓ in (64). When ℓ is large, the last display can be approximated by its limit as $\ell \to \infty$, namely,

$$\rho^2(d,\infty) \stackrel{\text{def}}{=} \frac{\pi}{[2\log(2)\mathrm{K}(d)]^2} \left\{ \mathrm{I}_0(d) + 2 \sum_{u=1}^{\infty} \mathrm{I}_u(d) \, 2^{(2d-1)u} \right\} \, . \tag{67}$$

7.5. **Asymptotic properties of the** LWW **estimator.** Let us now consider the LWW estimator $\widehat{d}^{\mathrm{LWW}}(L_n, U_n)$ defined in (28). The following results was first proved for a Gaussian $\mathrm{M}(d)$ process in Moulines et al. (2007c) and then extended to linear processes in Roueff and Taqqu (2007).

Theorem 4. *Let X be a strong linear M(d) process with generalized spectral density given by* (3)*. with $d \in \mathbb{R}$ and $f^* \in \mathcal{H}(\beta,\gamma,\varepsilon)$ for some $\gamma > 0$, $\beta \in (0,2]$ and $\varepsilon \in (0,\pi]$. Assume (W-1)-(W-4) with Condition (57). Let (L_n) be a sequence satisfying*

$$\lim_{n\to\infty} \left\{ n 2^{-(1+2\beta)L_n} + L_n^2 (n 2^{-L_n})^{-1/4} \right\} = 0 \tag{68}$$

and (U_n) *be a sequence such that either (S-1) or (S-2) holds. Then, as* $n \to \infty$,

$$(n2^{-L_n})^{1/2} \left(\widehat{d}^{\mathrm{LWW}}(L_n, U_n) - d\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left[0, \rho^2(d, \ell)\right] , \tag{69}$$

where $\ell = \lim_{n\to\infty}(U_n - L_n) \in \{1, 2, \ldots, \infty\}$ *and where* $\rho^2(d, \ell)$ *is given by* (66) *for* $l < \infty$ *and* (67) *for* $l = \infty$.

Remark 8. As in Theorem 3, the condition $n2^{-(1+2\beta)L_n} \to 0$ guarantees that the bias is negligible in the limit by imposing a sufficiently fast growth for L_n. The condition $L_n^2(n2^{-L_n})^{-1/4} \to 0$ means that $n2^{-L_n}$ has to grow faster than L_n^8 which is at most of order $\log^8(n)$ and hence always holds in the typical regime where $n2^{-L_n} \asymp n^\gamma$ with $\gamma \in (0, 1)$. Relation (69) is an asymptotic normality result. If we are interested merely in the rate of convergence of the LWW estimator $\widehat{d}^{\mathrm{LWW}}$, then we can relax condition (68). It follows from (Moulines et al., 2007c, Theorems 1 and 3) that if under the assumptions of Theorem 4, (68) is replaced by

$$\lim_{n\to\infty} \left\{ L_n^{-1} + L_n^2 (n2^{-L_n})^{-1/4} \right\} = 0 ,$$

then

$$\widehat{d}^{\mathrm{LWW}}(L_n, U_n) = d + O_{\mathbb{P}}\left((n2^{-L_n})^{-1/2} + 2^{-\beta L_n} \right) .$$

Hence, for $2^{L_n} \asymp n^{1/(1+2\beta)}$, we get the optimal rate $n^{\beta/(1+2\beta)}$, stated in Theorem 1. As shown in Moulines et al. (2007c), this result holds for a class of "weak" linear M(d) processes (see remark 3).

The following observations, which follow directly from Theorem 4 seems to have be unkwnown so far:

Corollary 5. *The* LWW *estimator has the same asymptotic variance as the* LRW *estimator with Abry–Veitch weights* (63).

Corollary 6. *Among all wavelet estimators of the memory parameter* d *presented in this paper, for a given choice of wavelet and scales involved in the estimates, the estimator with optimal asymptotic variance is the* LRW *estimator using the optimal weights defined in* (61).

As explained above, the optimal LRW in Corollary 6 estimator requires plugging a preliminary consistent estimator of d.

7.6. **Asymptotic variances.** The asymptotic variances of both the LRW and the LWW estimators depend on true value of the memory parameter d and on the wavelet ψ, as they are all expressed in terms of $\mathbf{D}_{\infty,u}(\lambda; d)$, defined in (50). In practice, one estimates the limiting variance $\rho^2(d, \ell)$ by $\rho^2(\widehat{d}, \ell)$ in order to construct asymptotic confidence intervals. The continuity of $\rho^2(\cdot, \ell)$ and the consistency of $\widehat{d}$ justify this procedure.

A comparison of the asymptotic variances $\rho^2(d,\ell)$ for several wavelets can be found in Moulines et al. (2007c), (see also Figure 1) In particular, as Figure 1 in Moulines et al. (2007c) indicates, the choice of wavelets does not matter much (provided that $(1+\beta)/2-\alpha < d \leq M$ holds) and a sensible approximation can be obtained by using the Shannon wavelet, for which a simple expression of the asymptotic variance can be obtained thanks to (55). Using the Shannon wavelet in (50), we get, for all $\lambda \in (-\pi,\pi)$, $\mathbf{D}_{\infty,u}(\lambda;d)=0$ for $u \geq 1$ and $\mathbf{D}_{\infty,0}(\lambda;d)=(2\pi-|\lambda|)^{-2d}$ so that, for all $d \in \mathbb{R}$, (66) becomes

$$\rho^2(d,\ell) = \frac{\pi\, g(-4d)}{2(2-2^{-\ell})\kappa_\ell \log^2(2)\, g^2(-2d)} \quad \text{where} \quad g(x) = \int_\pi^{2\pi} \lambda^x \,\mathrm{d}\lambda\,. \tag{70}$$

8. Asymptotic properties of the Fourier estimators GPH and LWF

8.1. **Asymptotic properties of the** GPH **estimator.** The consistency and asymptotic normality of the GPH have been established by Robinson (1995b) for stationary invertible Gaussian $M(d)$ process $-1/2 < d < 1/2$ with no data taper ($\tau=0$) and no differencing ($\delta=0$). As shown by Velasco (1999b), the GPH estimator with ($\tau=0$ and $\delta=0$) exhibits non-standard behavior when $d>1/2$. Although it is consistent for $d \in (1/2,1]$ and asymptotically normally distributed for $d \in (1/2,3/4)$, the GPH estimator has a non-normal limit distribution for $d \in [3/4,1]$, and for $d>1$, it converges to 1 in probability and is inconsistent. Hence, the interest in applying the GPH estimator under differencing $\delta>0$ and tapering $\tau>0$.

The following result is adapted from Moulines and Soulier (2003). To state the results, some additional notations are required. If $\{Z_t\}$ is a Gaussian white noise, $\bar{I}^Z_{p,\tau}(\tilde{\lambda}_k)$ is distributed as $\|G_{p,\tau}\|^2/2$ where $G_{p,\tau} = [G^{(1)}_{p,\tau},\ldots,G^{(2p)}_{p,\tau}]$ is a $2p$-dimensional zero-mean Gaussian vector with covariance matrix $\Sigma_{p,\tau}$, whose expression is given in Hurvich et al. (2002). Define

$$\gamma_{p,\tau} = \mathbb{E}\left[\log(\|W_{p,\tau}\|^2/2)\right]\,, \quad \sigma^2_{p,\tau} = \mathrm{Var}\left[\log(\|W_{p,\tau}\|^2/2)\right]\,. \tag{71}$$

Numerical expressions for these quantities are given in Hurvich et al. (2002).

Theorem 7. *Assume that X is a Gaussian $M(d)$ process and $f^* \in \mathcal{H}^*(\beta,\gamma,\varepsilon)$ for some $\beta \in (0,2]$, $\gamma>0$, $\varepsilon \in (0,\pi]$. Let $\delta \geq 0$ be the differencing order, $\tau \geq 0$ be the tapering order, and $p \geq 1$ be the pooling order. Let m_n be a non-decreasing sequence of integers such that*

$$\lim_{n\to\infty}(m_n^{-1} + m_n^{2\beta+1} n^{-2\beta}) = 0. \tag{72}$$

Then, for any d satisfying

$$\delta - \tau - 1/2 < d < \delta + 1/2\,, \tag{73}$$

the GPH *estimator defined in* (37) *satisfies,*

$$\sqrt{m_n}(\widehat{d}^{\mathrm{GPH}}(m_n) - d) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2_{p,\tau}/4)\,. \tag{74}$$

where $\sigma^2_{p,\tau}$ is defined in (71).

Compared to the wavelet estimators, the size of the confidence intervals does not depend on d, which may be seen as a significant advantage. On the other hand, there is an inflation of the variance over all the interval $(\delta-\tau-1/2, \delta+1/2)$ and it can be greater than the limiting variance obtained using by the wavelet estimator, at least for certain values of the memory parameter.

The definition of the pooled periodogram (35) implies that the number of Fourier frequencies used to evaluate $\widehat{d}^{\mathrm{GPH}}(m_n)$ is equal to $(p+\tau)m_n$. Hence the efficiency ratio between two GPH estimators using *same* number of Fourier frequencies but two *different* pooling numbers, say p and p', may be expressed as $(p+\tau)\sigma^2_{p,\tau}/(p'+\tau)\sigma^2_{p',\tau}$. As shown in (Hurvich et al., 2002, Theorem 1), the function $p \mapsto (p+\tau)\sigma^2_{p,\tau}$ is decreasing showing that pooling increases asymptotic efficiency. In addition, for any fixed τ, $\lim_{p\to\infty}(p+\tau)\sigma^2_{p,\tau} = \Phi(\tau)$, where

$$\Phi(\tau) = \frac{\Gamma(4\tau+1)\Gamma^4(\tau+1)}{\Gamma^4(2\tau+1)}\,. \tag{75}$$

As seen below, this efficiency bound is achieved by the local Whittle estimator. Therefore, the order of pooling can be made arbitrarily large, and at least asymptotically, an increase in the pooling order will result in a decrease of the asymptotic variance; see Hannan and Nicholls (1977) and Theorem 7. In practice, of course, this is not possible and since the improvements in asymptotic efficiency happen quickly, there is no need to use a very large p; $p = 3, 4, 5$ are typical values.

Remark 9. As observed in (73), the differencing order δ and the taper order τ control the range of values of the memory parameter d which can be inferred. The number of differentiation δ controls the upper bound for d, while the taper order τ controls the range. These two parameters are independent, by choosing the number of differentiations δ and the taper order τ we can therefore cover any intervals of admissible values for d (the same comment apply to the LWF estimator). If $\tau = 0$, the interval over which the GPH estimator is consistent and asymptotically normal is $(\delta - 1/2, \delta + 1/2)$. If $\tau > 0$, the range is $[\delta - \tau - 1/2, \delta + 1/2]$, as indicated in Theorem 7. Note that these ranges may not be optimal: for $\tau = 0$, using the sharpened results from Velasco (1999b), the range over which the memory parameter is asymptotically normal can be shown to be $(\delta - 3/4, \delta + 3/4)$.

Remark 10. It is possible to replace the Gaussian assumption by the weaker assumption that the process X is a strong linear $M(d)$ process by adding moment and regularity conditions on the distribution of the driving noise in the definition (44). In this case however, the estimator $\widehat{d}^{\mathrm{GPH}}(m_n)$ should be slightly modified to avoid a number of Fourier frequencies near 0. In addition, tapering and pooling are then required, even if the process X is stationary and invertible; see (Velasco, 2000, Theorem 3) and Faÿ et al. (2004).

8.2. **Asymptotic properties of the** LWF **estimator.** The consistency and asymptotic normality of the LWF estimator have been established by Robinson (1995a) for stationary invertible linear $M(d)$ process $-1/2 < d < 1/2$ with no data taper ($\tau = 0$) and no differencing ($\delta = 0$) (under the weaker assumption that $\{Z_t\}$ in (44) are martingale differences, whose squares, centered at their expectation, are also weakly stationary martingale differences). Velasco (1999a) has shown that the LWF estimator with $\tau = 0$ and $\delta = 0$ was consistent for $d \in (-1/2, 1]$ and asymptotically $\mathcal{N}(0, 1/4)$ for $d \in (-1/2, 3/4)$ under the same assumptions than Robinson (1995a). (Hurvich and Chen, 2000, Theorem 2) established Theorem 8 for $\tau = 1$ and $\delta = 1$. This result was later extended in Moulines and Soulier (2003) to general τ and δ. The consistency of the LWF was established (with $\delta = 0$ and $\tau = 0$) for $-1/2 < d < 1/2$ for a general class of non-linear processes in Dalla et al. (2006).

Theorem 8. *Assume that X is a strong linear $M(d)$ process for some $d \in \mathbb{R}$ and $f^* \in \mathcal{H}^*(\beta, \gamma, \varepsilon)$ for some $\beta \in (0, 2]$, $\gamma > 0$, $\varepsilon \in (0, \pi]$. Let δ be the differencing order and τ be the taper order. Let m_n be a non decreasing sequence of integers such that*

$$\lim_{n\to\infty} \left(m_n^{-1} + m_n^{2\beta+1} n^{-2\beta} \right) = 0 \; . \tag{76}$$

Then, the LWF *estimator defined in* (40) *satisfies, for any d satisfying*

$$\delta - \tau - 1/2 < d < \delta + 1/2 \; , \tag{77}$$

$$\sqrt{m_n}(\widehat{d}^{\mathrm{LWF}}(m_n) - d) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Phi(\tau)/4) \; , \tag{78}$$

where $\Phi(\tau)$ is defined in (75)

The quantity $\Phi(\tau)$ quantifies the loss of efficiency due to tapering. As the taper order increases, the limiting variance inflates: $\Phi(0) = 1$ (no tapering), $\Phi(1) = 1.5$, $\Phi(2) = 35/18$, etc. Since the LWF estimator is based on linear functionals of the periodogram, pooling is irrelevant. Recall that, in the definition (40), the standard periodogram is used and not the pooled one. This is why the pooling order does not appear in the conditions of Theorem 8.

Remark 11. The condition on the bandwidth (76) is slightly weaker than Assumption A4′ in Robinson (1995a) and Hurvich and Chen (2000). It seems that the $\log^2(m_n)$ term in these assumptions is superfluous (see (Andrews and Sun, 2004, Comments of Assumption 4) for the required adaptation of the proof).

9. Discussion

The Fourier and wavelet estimators of the memory parameter present similar characteristics and distinctive advantages. In Table 2, we summarize the main features of the estimators considered in this paper.

	Fourier	Wavelets
Non-stationarity (d large)	Pre-apply $(I-B)^\delta$ with $\delta > d - 1/2$	take $M \geq d$ (a sufficient number of vanishing moments)
Polynomial trends of degree K	same as above with $\delta \geq K+1$	same as above with $M \geq K+1$
Leakage (d small)	Use taper $(1-\mathrm{e}^{2\mathrm{i}\pi k/n})^p$ with taper order $p > 1/2 - d$	take $\alpha > (1+\beta)/2 - d$ (sufficiently smooth wavelet).
Rate of convergence of $\widehat{d}$ to d	$n^{\beta/(1+2\beta)}$ $(\beta \leq 2)$ with $m_n \asymp n^{\beta/(1+2\beta)}$	$n^{\beta/(1+2\beta)}$ $(\beta \leq 2)$ with $2^{L_n} \asymp n^{1/(1+2\beta)}$
Asymptotic variance	depends on taper order p only; GPH's $\downarrow$ LWF's, as pooling order $\to \infty$.	depends on d and ψ; LRW's with $\mathbf{w}^{\mathrm{AV}} =$ LWW's $\geq$ LRW's with $\mathbf{w}^{\mathrm{opt}}(\widehat{d}^{(1)}, \ell)$.

Table 2. Fourier VS wavelets: trends, non-stationarity, non-invertibility. In the *Wavelets* column, M and α are defined in (W-1)-(W-4).

To allow comparison between wavelet and Fourier estimators, we must first link the normalization factors, $(n2^{-L_n})^{1/2}$ for wavelet estimators and $m_n^{1/2}$ for Fourier estimators. A Fourier estimator with bandwidth m_n projects the observations $[X_1 \ \dots \ X_n]^T$ on the space generated by the vectors $\{\cos(2\pi k \cdot /n), \sin(2\pi k \cdot /n)\}$, $k = 1, \dots, m_n$, whose dimension is $2m_n$; on the other hand, the wavelet coefficients $\{W_{j,k}, j \geq L, k = 0, \dots, n_j - 1\}$ used in the wavelet estimator correspond to a projection on a space whose dimension is at most $\sum_{j=L_n}^{J_n} n_j \sim \sum_{j=L_n}^{\infty} n2^{-j} \sim 2n2^{-L_n}$. Hence, for m_n or $n2^{-L_n}$ large, it makes sense to consider $n2^{-L_n}$ as an analog of the bandwidth parameter m_n.

We shall now compare the asymptotic variances in the CLT's in Theorems 3, 4, 7 and 8. While the asymptotic variance of the Fourier estimators is a constant, the variance of the wavelet estimators is a function of the memory parameter, which can be numerically computed. For the Fourier estimators, the allowed range of the memory parameter d is given by (see Theorems 7

and 8)

$$\delta - \tau - 1/2 < d < \delta + 1/2 \; . \tag{79}$$

The length of this range equals $\tau + 1$, while the differentiation order δ allows to shift it towards large values of d. For instance, if one wishes to shift the upper boundary of the range towards large values of d while keeping the lower boundary unchanged, one has to increase both τ and δ by the same factor. As shown in Theorem 8, increasing τ inflates the asymptotic variance of the estimator. For wavelet estimators, the allowed range of the memory parameter d is given by

$$1/2 - \alpha < d \leq M \; , \tag{80}$$

[see Theorems 3 and 4; here we took β arbitrarily small, since we focus on the asymptotic variance in this discussion]. Of course one may still shift this range by a factor δ to the right by differentiating the series X at the order δ before processing the wavelet transform. This will also eliminate polynomial trends up to degree $M + \delta - 1$.

Observe that the higher the α in (80), the more negative d is allowed to be. This is because the higher the α, the smoother the wavelet $\psi_{j,k}$ and hence the better the spectral resolution of the wavelet. This matters particularly when $d < 0$ because $f(\lambda)$ is then very small around the origin making it harder to estimate d. In Fourier (see (79), it is τ that plays a role similar to α.

It is important to note that, for a given wavelet family such as Daubechies and Coiflets, increasing M yields a larger α, so that the allowed range is effectively increased by increasing M. In contrast to Fourier methods, by increasing the number of vanishing moments M, say of a Daubechies wavelet, the asymptotic variance converges to the asymptotic variance obtained with the Shannon wavelet, presented in (70). Thus, for a given d, there is asymptotically no price to pay for increasing the number of vanishing moments M and the number of available scales. This is an argument in favor of wavelet estimators as compared to tapered Fourier estimator. This should, however, be interpreted with care. For a given sample size n, an increase of M causes an increase of the support of the wavelet and a decrease in the number of available scales. While this does not influence the asymptotic variances, it affects the performance on finite samples[1].

The plots in Figure 1 indicate that the asymptotic variance of the LWW estimator is lower than the one obtained using the tapered version of LWF estimator, for most values of the memory parameter and this difference increases as τ increases in order to adapt to larger ranges for d. The asymptotic variance $\rho^2(d, \ell)$ in (66) of the LWW estimator is displayed for $\ell = 7$ using the Daubechies wavelets with $M = 2$ (Left) to $M = 4$ (Right). For these choices of wavelets, the corresponding α's are 1.34 and 1.91, and the allowed intervals for d are $[-0.84, 2]$ and $[-1.41, 4]$, respectively. The asymptotic variances $\rho^2(d, \infty)$ in (67) for $M = 2$ and 4 can also be compared

[1]The same remark apply to the pooling order for the GPH estimator: the asymptotic variance decreases as $p \to \infty$ but in practice, one takes small values, *e.g.* $p = 3, 4$.

to the one of the Shannon wavelet on this plot. The asymptotic variance $\Phi(\tau)$ in (75) of the LWF estimator is constant in d but increases when the taper order τ increases from $\tau = 2$ to $\tau = 4$, these values corresponding to intervals lengths for d close to those of the $M = 2, 4$ wavelet estimators (a bit larger for the former: 3 versus 2.84, and smaller for the latter: 5 versus 5.41).

Using wavelet to estimate the memory parameter has several additional benefits compared to using Fourier estimators. The wavelets present a rich time/frequency representation of the process, which can be more informative than that of the classical Fourier analysis, as discussed in Serroukh et al. (2000), Stoev et al. (2006) and Percival and Walden (2006). Wavelets can be used to detect the presence of outliers or jumps in the mean. The short-range dependence of the wavelet coefficients suggests construction of bootstrap confidence intervals for functionals of the wavelet coefficients, a procedure referred to as *wavestrapping*. This technique, which still is not rigorously justified, may be used to construct bootstrapped confidence interval for the memory parameter; see for example Percival et al. (2000).

10. A Monte-Carlo study

In this section, we present some Monte-Carlo simulation results that compare the root-mean square error performance of our four estimators for finite samples. The four estimators are denoted GPH (Geweke-Porter-Hudak), LWF (local Whittle Fourier), LWW (local Whittle wavelet) and LRW (local regression wavelet). We consider three models and several parameter combinations for each model:

(1) The ARFIMA models, introduced by Granger and Joyeux (1980), and generalized here to any value of the memory parameter d. We considered the ARFIMA(0,d,0) and ARFIMA(1,d,0) with d in $\{-1.2, 0, 0.3, 1.5, 2.5, 3.5\}$ and sizable lag 1 AR coefficient equal to 0.8. The innovation is assumed to be Gaussian. The short-memory component f^* of the spectral density satisfies $f^* \in \mathcal{H}(2, \gamma, \pi)$, where $\mathcal{H}$ is defined in Definition 2.

(2) The DARFIMA models, defined in Andrews and Sun (2004), is an ARFIMA-like process that has a discontinuity in its spectral density at frequency $\lambda = \lambda_0$. The DARFIMA(1,d,0) has the spectral density of an ARFIMA(1,d,0) on the interval $[-\lambda_0, \lambda_0]$ and is zero for $|\lambda| \in [\lambda_0, \pi]$. It is obtained by low-pass filtering of an ARFIMA(1,d,0) trajectory by a truncated *sinc* function in the time domain. We chose $\lambda_0 = \pi/2$ and Gaussian innovations.

(3) The third model is a non-linear function of a Gaussian sequence: $X_t = G(Y_t)$ where $\{Y_t\}$ is a stationary Gaussian sequence with zero-mean and variance 1 and $G : \mathbb{R} \to \mathbb{R}$ is a measurable function such that $\mathbb{E}[G^2(Y_0)] < \infty$. Then, X_t may be expressed as the sum $X_t = c_0 + \sum_{k=k_0}^{\infty} (c_k/k!) H_k(Y_t)$, where $H_k(\cdot)$ is the k-th Hermite polynomial and $c_k = \mathbb{E}[G(Y_t)H_k(Y_t)]$. The minimal integer $k_0 \geq 1$ such that $c_{k_0} \neq 0$ is called the Hermite rank of G. If Y is a $M(d)$ process with memory parameter $d_Y \leq 1/2$, then X is also an

$M(d)$ process with memory parameter $d_X = \frac{1}{2}(1 - k_0(1 - 2d_Y))$ (see (Dalla et al., 2006, p. 229, Eq. (55)) for details). In simulations, we use $G(x) = \exp(x)$ (for which $k_0 = 1$) and $G(x) = H_2(x) = x^2 - 1$ (for which $k_0 = 2$) and denote those models SUBORD1 and SUBORD2, respectively.

For the estimators LWW and LWF, we have used a convex minimization procedure of the contrast functions (29) and (39). In all cases, 1000 simulation runs for each value of d are used. This produces simulation standard errors that are roughly 3%.

The tuning parameters of each estimation procedure have been chosen to allow a fair comparison of those methods in a realistic setting, where the order of magnitude of the memory parameter d is only loosely known and where one may be in the presence of high-order polynomial trends. In order to cover all the values of d above ($-1.2 \leq d \leq 3.5$), we have used a Daubechies wavelet with $M = 4$ vanishing moments for the wavelet estimators (hence $\alpha \approx 1.91$, see Table 1) and we have differenced the series $\delta = 4$ times and have used a taper order $\tau = 5$ for the Fourier estimators. The corresponding admissible ranges are (see (80) and (79)) $(-1.41, 4]$ and $(-1.5, 4.5)$, respectively. For the GPH estimator, we took $p = 4$ in Relation (35) defining the pooled periodogram. This reduces the number of frequencies by a factor $\tau + p = 9$ (see Relation (35)). In the case of the LRW estimator, the computation of the optimal weights in the least-square criterion is numerically quite involved so we ran the simulations using the weights suggested by Abry and Veitch (1998). The difference in the results becomes significant only when d gets close to the boundaries of the admissible range, see Figure 2 for the asymptotic variance. The remaining free parameters are the number of frequencies (LWF) or blocks of frequencies (GPH with pooling) denoted m_n, and the minimal (*i.e.* finest) dyadic wavelet scale L_n (for the LRW and the LWW estimators); the highest (*i.e.* coarsest) scale is chosen to be the highest available ($U_n = J_n$).

The box and whisker plots of the estimators are displayed in Figures 4 and 5 for different values of the bandwidth (Fourier methods) and the finest scale (wavelet methods). In Figure 4, the model is an ARFIMA(1,d,0) with $d = 1.5$. In Figure 5, the model is an DARFIMA(1,d,0) with $d = 0.3$. The AR coefficient is 0.8 in both cases. These figures illustrate the bias-variance trade-off inherent to semi-parametric methods (the variance decreases as the bandwidth or the number of scales increases, but then the bias increases). In general, the standard deviation of the 1000 runs of the wavelet methods is comparable to that of the Fourier methods.

Tables 3 and 4 give the bias, variance, and RMSE (root mean square error) for models 1,2,3 for sample sizes 512 and 4096, respectively. Those quantities are computed for the optimal bandwidth m_n (Fourier methods) or the optimal finest scale L_n (wavelet methods) in the RMSE-sense, whose values are displayed in the fourth column. For each model, the lowest RMSE among the four methods appears in boldface. Note that all the possible values of finest scale L_n have been considered, but only a subset of the many possible values of the bandwidth m_n. The

standard deviations of the LWW estimator are lower than those of the LRW estimator, which is consistent with our theoretical findings. Also, the standard deviations of the Fourier methods remain approximately constant for the different values of the memory parameter, whereas the variance of the wavelet methods increase with $|d|$. Also, as predicted by the expressions of the limiting variance, the variance of the wavelet methods are lower than those of the Fourier methods, especially when d is small. The reported values of the standard deviations agree with our theoretical findings for the sample size $n = 4096$.

For the non-linear processes, the results suggest that the wavelet estimator remains consistent. However, the presence of non-linearity worsens the behavior of the estimator at a given finite sample and a larger sample size is required to achieve a prescribed accuracy.

The root mean square error is shown in some particular cases in Figure 6. The MSE of the Fourier criteria is plotted against the value of the bandwidth, that is, m_n for the LWF estimator and $m_n \times (p + \tau)$ for the GPH estimator. For the wavelet methods, the somehow arbitrary "equivalent bandwidth" abscissa is half the number of wavelet coefficients used by the estimators: $m_n^{\text{equiv}} = \frac{1}{2}\sum_{j=L_n}^{U_n} n_j$ (see the discussion on the comparison of the asymptotic variances in the previous section).

11. Software

The software used to perform the estimation of the long-memory parameter was written in MATLAB/OCTAVE and may be obtained from the authors. It includes the four estimators (LWF, GPH, LWWand LRW), basic random processes generators and some other utilities such as the pyramidal algorithm for computing wavelet coefficient.

Basic installation. After downloading the tar archive (`toolboxLRD.tar`) and expanding it in *e.g.* `/home/user/octave`, one has to add the directory to the search path:

```
addpath(genpath('/home/user/octave/ToolboxLRD'));
```

This line can be added to the `.octave` or `.matlab` file. Some demos are available in `ToolboxLRD/Examples`.

Loading the data. Use the `load` command to load a data set (time series) into some vector, say `x`. One can also synthesize some trajectories using one's personal generator or the one present in the `Utils` subdirectory. For instance, a 4096 long trajectory of the ARFIMA model $(1 - B)^d(X_t - \alpha X_{t-1}) = Z_t$ with $d = 1.4$, $\alpha = 0.8$ and Gaussian i.i.d sequence Z can be obtained by setting:

```
n = 4096;
x = randARFIMA(1.4,[0.8],[],n);
```

The argument `[]` above means that the MA part of the generated ARFIMA is a weak white noise (MA(0)). This example is used in the following to describe the package.

Estimating the long-memory parameter. We shall now obtain the LRW, LWW, GPH and LWF estimators of the memory parameter d of the series as well as an estimate of their standard deviation using the asymptotic variance given in Theorems 3, 4, 7, and 8. The standard deviations can be used to build asymptotic confidence intervals.

1. The *Geweke-Porter Hudak* (GPH) estimator is obtained as follows:
```
param.taper=5; param.pooling= 4;
param.bandwidth=[6 12 24 50]; param.differorder = 4;
[d, stds]=GPH(x,param)
```
where `param.bandwidth` is a vector giving the different values for the upper Fourier frequency m on which the regression is to be performed (Theorem 7). The taper order τ, pooling order p and differentiation order δ are specified by `param.taper`, `param.pooling` and `param.differorder`, respectively. One obtains
```
d =
    1.2744    1.2736    1.3670    1.3930
stds =
    0.1803    0.1215    0.0840    0.0576
```
Note that `d` and `stds` are vectors. In the above example, they have four components, corresponding respectively to the bandwidths $m = 6, 12, 24, 50$.
2. The *Local Whittle Fourier* (LWF) estimator is invoked in the following way:
```
param.taper=5; param.differorder=4;
param.bandwidth=[50 100 200 500];
[d, stds]=LWF(x,param,[])
```
One gets :
```
d =
    1.3114    1.2974    1.2905    1.3422
stds =
    0.1206    0.0853    0.0603    0.0381
```
Here the minimization of (39) is done over the whole real line. As for the `LWW` function, one may specify the range where to optimize the LWF contrast function (39) by replacing the third argument `[]` by an interval $[\Delta_1, \Delta_2]$. For instance, if one wants to restrict this minimization to the set (79) of admissible values of d for the CLT Theorem 8 to hold,
```
range= [ param.differorder-param.taper-0.5,  param.differorder+0.5];
[d, stds]=LWF(x,param,range)
```
In this specific case, the output is unchanged since the minimizing values of d are within the corresponding interval $[-1.5, 4.5]$.
3. The *Local Whittle Wavelets* (LWW) estimator is obtained as follows:
```
LU = [6 9; 5 9; 4 9; 3 9];
[phi, M, alpha] = scalingfilter('Daubechies',4);
[d, stds] = LWW(x,LU,phi,[])
```
where `phi` indicates the scaling function, `M` the corresponding number of vanishing moments, `alpha` the Fourier decay exponent (see (W-1)-(W-4)), `x` contains a finite set of observations and `LU` is a two column matrix, containing scales limits L and U in the LWW objective

function (29). If `LU` is a one column vector then it contains different values of the lower scale L and U is taken equal to the maximal available scale index J defined in (22). The argument `[]` above means that the interval, denoted $[\Delta_1, \Delta_2]$ in Definition (28), over which the contrast function (29) is minimized is $(-\infty, \infty)$. It can be replaced by an interval $[\Delta_1, \Delta_2]$, if one wants to restrict the minimization to a particular range, for instance to the one given by (80) which corresponds to admissible values of d for the CLT Theorem 4 to hold. One gets :

```
d =
    1.3183    1.3366    1.4209    1.3788
stds =
    0.1138    0.0719    0.0479    0.0324
```

4. The *Local Regression Wavelet* (LRW) estimator and the standard deviation is invoked in the following way:

```
L = [6;5;4;3];
[d,stds] = LRW(x,L,phi)
```

The three first argument of `LRW` are the same as `LWW` but `LU` has been replaced by `L`, a column vector containing different choices for the lower scale L used in the regression, see Eq. (24). In this case, the upper scale is the largest scale available ($U = J_n$). If one wants to take different values for U, a two-columns matrix must replace `L`, for example, by the `LU` in the `LWW` case. Here the LRW estimator is computed using Abry–Veitch weights (see 63) and one gets the output:

```
d =
    1.4161    1.3815    1.4305    1.3882
stds =
    0.1020    0.0675    0.0461    0.0317
```

But the LRW estimator can also be obtained using the optimal weights. In fact, the following additional output variables are available :

(a) the value of a log-regression multiplicative constant c so that $\widehat{\sigma}_j^2 \approx c\, 2^{2dj}$. Equivalently, $\log \widehat{\sigma}_j^2 \approx \log c + 2dj$, where $\log c$ is the intercept of the regression line;
(b) new estimates of d using the two-step procedure based on the optimal weights (61). These weights are computed using the preliminary value of d estimated with the Abry-Veitch weights;
(c) the standard deviations of the new estimates of d;
(d) the corresponding values of the log-regression multiplicative constant.

Thus if the `LRW` call of the last example is replaced by

```
[d,stds, c, dopt,stdopt, copt] = LRW(x,L,phi)
```

the following additional output is added to the previous one :

```
c =
    0.0074    0.0105    0.0067    0.0094
dopt =
    1.4199    1.3852    1.4451    1.3834
stdopt =
    0.1011    0.0666    0.0453    0.0311
```

```
copt =
    0.0073    0.0103    0.0060    0.0097
```

Alternatively, one could also use a different preliminary estimate of d, say $d = 1.3823$, a value obtained as the first output of the `GPH` routine above :

```
[d,stds, c, dopt,stdopt, copt] = LRW(x,L,phi,1.3823)
```

The last three output values are then replaced by

```
dopt =
    1.4199    1.3852    1.4441    1.3835
stdopt =
    0.1011    0.0666    0.0451    0.0310
copt =
    0.0073    0.0103    0.0061    0.0097
```

Obtaining confidence intervals. A routine for obtaining the asymptotically normal *confidence intervals* for d at a given level has also been included. It works as follows :

```
p=0.95; d=dopt; stds=stdopt;
[I]=ConfidenceInterval(d,stds,p)
```

where `d` and `stds` are any outputs of the above procedures and `p` is the confidence level. Here we used the last displayed estimates with $p = 0.95$ and get

```
I =
    1.2217    1.6180
    1.2546    1.5158
    1.3557    1.5325
    1.3227    1.4443
```

In this particular example we can see that the true value of d, namely 1.4, belongs to the four intervals.

Obtaining the theoretical asymptotic variances. It is possible also to obtain directly the asymptotic variances $\rho^2_{\text{opt}}(d,\ell)$ and $\rho^2(d,\ell)$ defined in (62) and (66). These are the asymptotic variances of the LRW estimator (Theorem 3), when, respectively, the optimal weigths (61) are chosen or when the Abry-Veitch weights (63) are chosen. The asymptotic variance of the LWW estimator is also $\rho^2(d,\ell)$ (Theorem 4). The approximation of $\rho^2(d,\ell)$ given in (70) and obtained by replacing ψ by the Shannon wavelet is also available. This is how to get these asymptotic variances :

```
d=1.4; l=5;
[v, vs, vopt, wopt]= AsymptoticVarianceLRW(phi,d,l)
```

Here $\rho^2(d,\ell)$ and $\rho^2_{\text{opt}}(d,\ell)$ are computed for $d = 1.4$ and $\ell = 5$ and stacked in the output `v` and `vopt` respectively. The output `vs` corresponds to the Shannon approximation (70) and `wopt` to the optimal weights (61), of length $\ell + 1 = 6$. For these values one gets

```
v =
    0.5848
vs =
    0.4949
vopt =
    0.5698
```

```
wopt =
   -0.2693    0.0546    0.0827    0.0587    0.0410    0.0322
```

We observe that for this value of d the optimal variance (0.5698) is sensitively lower than the one obtained with Abry-Veitch weights (0.5848), but still larger than the Shannon approximation (0.4949). Since this approximation gets sharper as the number of vanishing moments of the Daubechies wavelet increases, it indicates that one could get a better variance by increasing M, here $M = 2$. Notice, however, that the length of the wavelet filters would also increase and thus the number of available wavelet coefficients decrease for a finite n, an effect which is not considered in the asymptotic variance, see Section 9.

12. Conclusion

We have compared four semi-parametric methods for the estimation of the long-memory parameter d in times series, two Fourier-based and two wavelet-based. These are the Geweke-Porter Hudak (GPH) [Regression/Fourier], Local Whittle Fourier (LRW) [Whittle/Fourier], Local Regression Wavelet (LRW) [Regression/Wavelets] and Local Whittle Wavelet (LWW) [Whittle/Wavelets]. We have discussed issues related to differencing, tapering and pooling in the case of Fourier-based estimators and choices of wavelets in the case of wavelet-based estimators. Conditions for the asymptotic normality of the estimators are specified in Theorems 3, 4, 7 and 8.

We have undertaken a Monte Carlo comparison. In the Monte Carlo study, we have focused on ARFIMA(0,d,0) and ARFIMA(1,d,0) models with an AR(1) parameter equal to 0.8, a relatively high value, as well as on DARFIMA and subordinated models defined in Section 10. All four methods appear to work well with similar performances at the optimal bandwidth lower scale index. We have also developed a software package for the benefit of the practitioner which computes the corresponding estimates of the long-memory parameter d and provides confidence intervals, based on the asymptotically normal distribution of the estimators.

We noted that the LRW estimator with Abry-Veitch weights (63) has the same asymptotic variance as the LWW estimator. This means that the LRW estimator, when used with the optimal weights (61), has smaller asymptotic variance than the LWW estimator.

References

Abadir, K., Distaso, W. and Giraitis, L. (2007). Nonstationarity-extended local whittle estimation. *J. of Econometrics* **141** 1353–1384.

Abry, P., Flandrin, P., Taqqu, M. S. and Veitch, D. (2000). Wavelets for the analysis, estimation and synthesis of scaling data. In *Self-Similar Network Traffic and Performance Evaluation* (K. Park and W. Willinger, eds.). Wiley (Interscience Division), New York.

Abry, P., Flandrin, P., Taqqu, M. S. and Veitch, D. (2003). Self-similarity and long-range dependence through the wavelet lens. In *Theory and Applications of Long-range Dependence* (P. Doukhan, G. Oppenheim and M. S. Taqqu, eds.). Birkhäuser, 527–556.

Abry, P. and Veitch, D. (1998). Wavelet analysis of long-range-dependent traffic. *IEEE Trans. Inform. Theory* **44** 2–15.

Andrews, D. W. K. and Guggenberger, P. (2003). A bias-reduced log-periodogram regression estimator for the long-memory parameter. *Econometrica* **71** 675–712.

Andrews, D. W. K. and Sun, Y. (2004). Adaptive local polynomial Whittle estimation of long-range dependence. *Econometrica* **72** 569–614.

Bardet, J.-M. (2000). Testing for the presence of self-similarity of Gaussian time series having stationary increments. *Journal of Time Series Analysis* **21** 497–515.

Bardet, J.-M. (2002). Statistical study of the wavelet analysis of fractional Brownian motion. *IEEE Trans. Inform. Theory* **48** 991–999.

Bardet, J.-M., Lang, G., Moulines, E. and Soulier, P. (2000). Wavelet estimator of long-range dependent processes. *Stat. Inference Stoch. Process.* **3** 85–99. 19th "Rencontres Franco-Belges de Statisticiens" (Marseille, 1998).

Cohen, A. (2003). *Numerical analysis of wavelet methods*, vol. 32 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam.

Dalla, V., Giraitis, L. and Hidalgo, J. (2006). Consistent estimation of the memory parameter for nonlinear time series. *J. Time Ser. Anal.* **27** 211–251.

Daubechies, I. (1992). *Ten lectures on wavelets*, vol. 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Deo, R., Hsieh, M., Hurvich, C. M. and Soulier, P. (2006a). Long memory in nonlinear processes. In *Dependence in probability and statistics*, vol. 187 of *Lecture Notes in Statist.* Springer, New York, 221–244.

Deo, R., Hurvich, C. M. and Lu, Y. (2006b). Forecasting realized volatility using a long-memory stochastic volatility model: estimation, prediction and seasonal adjustment. *J. Econometrics* **131** 29–58.

Faÿ, G., Moulines, E. and Soulier, P. (2004). Edgeworth expansions for linear statistics of possibly long-range-dependent linear processes. *Statist. Probab. Lett.* **66** 275–288.

Faÿ, G., Roueff, F. and Soulier, P. (2007). Estimation of the memory parameter of the infinite-source Poisson process. *Bernoulli* **13** 473–491.

Fox, R. and Taqqu, M. S. (1986). Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Ann. Statist.* **14** 517–532.

Geweke, J. and Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *J. Time Ser. Anal.* **4** 221–238.

Giraitis, L., Robinson, P. M. and Samarov, A. (1997). Rate optimal semiparametric estimation of the memory parameter of the Gaussian time series with long range dependence. *J. Time Ser. Anal.* **18** 49–61.

GRANGER, C. W. J. and JOYEUX, R. (1980). An introduction to long-memory time series models and fractional differencing. *J. Time Ser. Anal.* **1** 15–29.

HANNAN, E. J. and NICHOLLS, D. F. (1977). The estimation of the prediction error variance. *J. Amer. Statist. Assoc.* **72** 834–840.

HURVICH, C. M. and CHEN, W. W. (2000). An efficient taper for potentially overdifferenced long-memory time series. *J. Time Ser. Anal.* **21** 155–180.

HURVICH, C. M., LANG, G. and SOULIER, P. (2005a). Estimation of long memory in the presence of a smooth nonparametric trend. *J. Amer. Statist. Assoc.* **100** 853–871.

HURVICH, C. M., MOULINES, E. and SOULIER, P. (2002). The FEXP estimator for potentially non-stationary linear time series. *Stoch. Proc. App.* **97** 307–340.

HURVICH, C. M., MOULINES, E. and SOULIER, P. (2005b). Estimating long memory in volatility. *Econometrica* **73** 1283–1328.

HURVICH, C. M. and RAY, B. K. (1995). Estimation of the memory parameter for nonstationary or noninvertible fractionally integrated processes. *J. Time Ser. Anal.* **16** 17–41.

JOHNSON, N. L. and KOTZ, S. (1970). *Distributions in statistics. Continuous univariate distributions. 2.* Houghton Mifflin Co., Boston, Mass.

KAPLAN, L. M. and KUO, C.-C. J. (1993). Fractal estimation from noisy data via discrete fractional Gaussian noise (DFGN) and the Haar basis. *IEEE Trans. Signal Process.* **41** 3554–3562.

KÜNSCH, H. (1987). Statistical aspects of self-similar processes. In *Proceedings of the 1st World Congress of the Bernoulli Society, Vol. 1 (Tashkent, 1986)*. VNU Sci. Press, Utrecht.

LAHIRI, S. N. (2003). A necessary and sufficient condition for asymptotic independence of discrete Fourier transforms under short- and long-range dependence. *Ann. Statist.* **31** 613–641.

MALLAT, S. (1998). *A wavelet tour of signal processing*. Academic Press Inc., San Diego, CA.

MCCOY, E. J. and WALDEN, A. T. (1996). Wavelet analysis and synthesis of stationary long-memory processes. *J. Comput. Graph. Statist.* **5** 26–56.

MOULINES, E., ROUEFF, F. and TAQQU, M. S. (2007a). Central Limit Theorem for the log-regression wavelet estimation of the memory parameter in the Gaussian semi-parametric context. To appear.

MOULINES, E., ROUEFF, F. and TAQQU, M. S. (2007b). On the spectral density of the wavelet coefficients of long memory time series with application to the log-regression estimation of the memory parameter. *J. Time Ser. Anal.* **28**.
URL `http://arxiv.org/abs/math.ST/0512635`

MOULINES, E., ROUEFF, F. and TAQQU, M. S. (2007c). A wavelet Whittle estimator of the memory parameter of a non-stationary Gaussian time series. Tech. rep., Ecole Nationale Supérieure des Télécommunications et Boston University. To appear in the Annals of Statistics.
URL `http://arxiv.org/abs/math.ST/0601070`

MOULINES, E. and SOULIER, P. (2003). Semiparametric spectral estimation for fractional processes. In *Theory and applications of long-range dependence*. Birkhäuser Boston, Boston, MA, 251–301.

PERCIVAL, D. B., SARDY, S. and DAVISON, A. C. (2000). Wavestrapping time series: adaptive wavelet-based bootstrapping. In *Nonlinear and nonstationary signal processing (Cambridge, 1998)*. Cambridge

Univ. Press, Cambridge, 442–471.

Percival, D. B. and Walden, A. T. (2006). *Wavelet methods for time series analysis*, vol. 4 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. Reprint of the 2000 original [MR1770693].

Robinson, P. M. (1994). Efficient tests of nonstationary hypotheses. *J. Amer. Statist. Assoc.* **89** 1420–1437.

Robinson, P. M. (1995a). Gaussian semiparametric estimation of long range dependence. *Ann. Statist.* **23** 1630–1661.

Robinson, P. M. (1995b). Log-periodogram regression of time series with long range dependence. *The Annals of Statistics* **23** 1048–1072.

Robinson, P. M. and Henry, M. (2003). Higher-order kernel semiparametric M-estimation of long memory. *J. Econometrics* **114** 1–27.

Roueff, F. and Taqqu, M. S. (2007). Asymptotic normality of wavelet estimators of the memory parameter: the linear case. Tech. rep.

Samorodnitsky, G. and Taqqu, M. S. (1994). *Stable non-Gaussian processes: stochastic models with infinite variance*. Chapman and Hall.

Serroukh, A., Walden, A. T. and Percival, D. B. (2000). Statistical properties and uses of the wavelet variance estimator for the scale analysis of time series. *J. Amer. Statist. Assoc.* **95** 184–196.

Shimotsu, K. and Phillips, P. C. B. (2005). Exact local Whittle estimation of fractional integration. *Ann. Statist.* **33** 1890–1933.

Shimotsu, K. and Phillips, P. C. B. (2006). Local Whittle estimation of fractional integration and some of its variants. *J. Econometrics* **130** 209–233.

Stoev, S., Taqqu, M. S., Park, C., Michailidis, G. and Marron, J. S. (2006). LASS: a tool for the local analysis of self-similarity. *Comput. Statist. Data Anal.* **50** 2447–2471.

Tanaka, K. (1999). The nonstationary fractional unit root. *Econometric Theory* **15** 549–582.

Teyssière, G. and Abry, P. (2007). Wavelet analysis of nonlinear long-range dependent processes. Applications to financial time series. In *Long memory in economics*. Springer, Berlin, 173–238.

Veitch, D. and Abry, P. (1999). A wavelet-based joint estimator of the parameters of long-range dependence. *IEEE Trans. Inform. Theory* **45** 878–897.

Veitch, D., Abry, P. and Taqqu, M. S. (2003). On the automatic selection of the onset of scaling. *Fractals* **11** 377–390.

Veitch, D., Taqqu, M. S. and Abry, P. (2000). Meaningful MRA initialisation for discrete time series. *Signal Processing* **80** 1971–1983.

Velasco, C. (1999a). Gaussian semiparametric estimation of non-stationary time series. *J. Time Ser. Anal.* **20** 87–127.

Velasco, C. (1999b). Non-stationary log-periodogram regression. *J. Econometrics* **91** 325–371.

Velasco, C. (2000). Non-Gaussian log-periodogram regression. *Econometric Theory* **16** 44–79.

Velasco, C. and Robinson, P. M. (2000). Whittle pseudo-maximum likelihood estimation for nonstationary time series. *J. Am. Statist. Assoc.* **95** 1229–1243.

Wornell, G. W. and Oppenheim, A. V. (1992). Estimation of fractal signals from noisy measurements using wavelets. *IEEE Trans. Signal Process.* **40** 611 – 623.

Žurbenko, I. (1979). On the efficiency of estimates of a spectral density. *Scand. J. Statist.* **6** 49–56.

Laboratoire Paul-Painlevé, Université Lille-1, 59655 Villeneuve-d'Ascq Cedex, France.

Current address: Laboratoire APC, Université Paris-7, Bâtiment Condorcet, 10, rue Alice Domon et Léonie Duquet, 75205 Paris Cedex 13, France.

E-mail address: gilles.fay@univ-lille1.fr

LTCI (CNRS, TELECOM ParisTech) , 46, rue Barrault, 75634 Paris Cédex 13, France.

E-mail address: moulines@tsi.enst.fr

E-mail address: roueff@tsi.enst.fr

Department of Mathematics and Statistics, Boston University Boston, MA 02215, USA.

E-mail address: murad@math.bu.edu

	GPH				LWF				LRW				LWW			
Model	bias	std	RMSE	m_n^{opt}	bias	std	RMSE	m_n^{opt}	bias	std	RMSE	L_n^{opt}	bias	std	RMSE	L_n^{opt}
ARFIMA(0,-1.2,0)	0.007	0.105	**0.105**	26	-0.108	0.071	0.129	234	0.047	0.106	0.116	2	0.105	0.083	0.134	2
ARFIMA(1,-1.2,0)	0.000	0.161	0.161	12	-0.138	0.128	0.188	72	-0.093	0.108	0.142	2	-0.048	0.083	**0.096**	2
ARFIMA(0,0.0,0)	-0.022	0.103	0.105	26	-0.099	0.073	0.123	234	-0.026	0.058	0.064	1	-0.002	0.046	**0.046**	1
ARFIMA(1,0.0,0)	-0.073	0.154	0.170	12	-0.175	0.134	0.220	72	-0.169	0.104	0.198	2	-0.058	0.143	**0.154**	3
ARFIMA(0,0.3,0)	-0.029	0.104	0.108	26	-0.094	0.071	0.118	234	-0.065	0.060	0.088	1	-0.045	0.046	**0.065**	1
ARFIMA(1,0.3,0)	-0.076	0.151	0.169	12	-0.194	0.105	0.221	108	-0.172	0.100	0.199	2	-0.060	0.143	**0.154**	3
ARFIMA(0,1.5,0)	-0.049	0.097	0.109	26	-0.077	0.072	0.105	234	-0.085	0.110	0.139	2	-0.045	0.093	**0.103**	2
ARFIMA(1,1.5,0)	-0.121	0.148	0.190	12	-0.182	0.106	0.210	108	-0.167	0.115	0.203	2	-0.135	0.091	**0.163**	2
ARFIMA(0,2.5,0)	-0.039	0.094	0.102	26	-0.050	0.072	**0.087**	234	-0.093	0.120	0.152	2	-0.047	0.097	0.108	2
ARFIMA(1,2.5,0)	-0.132	0.141	0.194	12	-0.157	0.108	0.190	108	-0.136	0.115	0.178	2	-0.101	0.098	**0.141**	2
ARFIMA(0,3.5,0)	-0.023	0.092	0.095	26	-0.023	0.070	**0.074**	234	-0.077	0.110	0.134	2	-0.037	0.089	0.097	2
ARFIMA(1,3.5,0)	-0.097	0.136	0.167	12	-0.111	0.109	0.155	108	-0.102	0.113	0.152	2	-0.063	0.097	**0.116**	2
DARFIMA(0,0.0,0)	0.064	0.275	0.282	5	0.002	0.162	0.162	72	-0.013	0.188	0.188	3	0.072	0.139	**0.157**	3
DARFIMA(1,0.0,0)	0.031	0.287	0.288	5	-0.027	0.155	0.157	72	-0.038	0.187	0.191	3	0.041	0.144	**0.149**	3
DARFIMA(0,0.3,0)	0.036	0.282	0.284	5	0.005	0.161	0.161	72	-0.038	0.190	0.193	3	0.046	0.145	**0.152**	3
DARFIMA(1,0.3,0)	0.015	0.266	0.266	5	-0.030	0.154	0.157	72	-0.064	0.182	0.193	3	0.016	0.146	**0.147**	3
SUBORD1(0,0.0,0)	-0.018	0.099	0.101	26	-0.095	0.066	0.116	234	-0.032	0.069	0.076	1	-0.003	0.057	**0.057**	1
SUBORD1(1,0.0,0)	0.121	0.114	0.166	26	0.006	0.075	**0.076**	234	-0.043	0.097	0.106	1	-0.015	0.086	0.087	1
SUBORD1(0,0.3,0)	-0.111	0.113	0.159	26	-0.179	0.087	0.199	234	-0.155	0.090	0.179	1	-0.127	0.083	**0.152**	1
SUBORD1(1,0.3,0)	-0.063	0.161	**0.173**	17	-0.133	0.157	0.206	153	-0.122	0.174	0.212	2	-0.032	0.185	0.188	2
SUBORD2(0,0.0,0)	-0.014	0.103	0.104	26	-0.095	0.067	0.117	234	-0.026	0.061	0.066	1	-0.001	0.048	**0.048**	1
SUBORD2(1,0.0,0)	0.207	0.260	0.332	5	0.022	0.183	**0.184**	45	0.160	0.212	0.266	3	0.293	0.197	0.353	3
SUBORD2(0,0.3,0)	0.014	0.115	0.116	26	-0.058	0.093	0.109	234	-0.005	0.076	0.077	1	0.025	0.067	**0.072**	1
SUBORD2(1,0.3,0)	0.157	0.209	0.262	8	0.036	0.168	**0.172**	72	0.174	0.088	0.195	1	0.214	0.085	0.230	1

TABLE 3. Length of the time series = 512. Bias, standard deviation, root mean-square error and optimal bandwidth/minimal scale for the four estimators applied to the three models of Section 10. The lowest RMSE among the four methods appears in boldface.

	GPH				LWF				LRW				LWW			
Model	bias	std	RMSE	m_n^{opt}	bias	std	RMSE	m_n^{opt}	bias	std	RMSE	L_n^{opt}	bias	std	RMSE	L_n^{opt}
ARFIMA(0,-1.2,0)	-0.012	0.032	**0.034**	224	-0.031	0.022	0.038	2016	0.020	0.037	0.043	3	0.038	0.032	0.050	3
ARFIMA(1,-1.2,0)	-0.024	0.061	0.065	54	-0.081	0.041	0.091	486	-0.046	0.023	0.051	2	-0.037	0.021	**0.043**	2
ARFIMA(0,0.0,0)	-0.016	0.031	0.035	224	-0.027	0.022	0.035	2016	-0.006	0.014	0.015	1	0.000	0.012	**0.012**	1
ARFIMA(1,0.0,0)	-0.037	0.060	0.070	54	-0.072	0.043	0.084	486	-0.042	0.034	0.055	3	-0.031	0.029	**0.043**	3
ARFIMA(0,0.3,0)	-0.017	0.031	0.036	224	-0.026	0.023	0.034	2016	-0.019	0.022	0.029	2	-0.010	0.019	**0.021**	2
ARFIMA(1,0.3,0)	-0.044	0.062	0.076	54	-0.071	0.043	0.083	486	-0.042	0.036	0.056	3	-0.029	0.030	**0.041**	3
ARFIMA(0,1.5,0)	-0.017	0.031	0.035	224	-0.021	0.021	**0.029**	2016	-0.038	0.026	0.046	2	-0.028	0.024	0.037	2
ARFIMA(1,1.5,0)	-0.051	0.057	0.077	54	-0.062	0.041	0.074	486	-0.038	0.042	0.057	3	-0.022	0.037	**0.043**	3
ARFIMA(0,2.5,0)	-0.013	0.030	0.033	224	-0.014	0.022	**0.026**	2016	-0.040	0.030	0.050	2	-0.029	0.027	0.040	2
ARFIMA(1,2.5,0)	-0.043	0.058	0.072	54	-0.047	0.042	0.063	486	-0.035	0.044	0.056	3	-0.018	0.039	**0.043**	3
ARFIMA(0,3.5,0)	-0.005	0.030	0.031	224	-0.006	0.022	**0.023**	2016	-0.033	0.028	0.043	2	-0.019	0.027	0.034	2
ARFIMA(1,3.5,0)	-0.033	0.058	0.066	54	-0.035	0.044	0.056	486	-0.029	0.044	0.053	3	-0.014	0.041	**0.043**	3
DARFIMA(0,0.0,0)	-0.023	0.060	0.064	54	-0.055	0.042	0.069	486	0.025	0.035	**0.043**	3	-0.002	0.044	0.044	4
DARFIMA(1,0.0,0)	-0.037	0.060	0.070	54	-0.072	0.043	0.084	486	0.010	0.034	**0.035**	3	0.029	0.029	0.041	3
DARFIMA(0,0.3,0)	-0.024	0.063	0.068	54	-0.053	0.044	0.069	486	0.012	0.036	**0.038**	3	0.033	0.030	0.044	3
DARFIMA(1,0.3,0)	-0.044	0.062	0.076	54	-0.071	0.043	0.083	486	-0.003	0.037	0.037	3	0.016	0.030	**0.034**	3
SUBORD1(0,0.0,0)	-0.015	0.030	0.033	224	-0.027	0.021	0.034	2016	-0.006	0.014	0.016	1	0.000	0.013	**0.013**	1
SUBORD1(1,0.0,0)	0.028	0.089	0.093	26	0.032	0.028	0.043	2016	0.034	0.026	**0.042**	1	0.044	0.026	0.051	1
SUBORD1(0,0.3,0)	-0.103	0.043	0.112	224	-0.115	0.038	0.121	2016	-0.103	0.038	0.109	2	-0.082	0.051	**0.096**	3
SUBORD1(1,0.3,0)	-0.091	0.073	0.117	110	-0.086	0.076	0.115	990	-0.093	0.057	0.109	2	-0.060	0.066	**0.089**	2
SUBORD2(0,0.0,0)	-0.018	0.031	0.036	224	-0.028	0.022	0.036	2016	-0.006	0.014	0.016	1	-0.001	0.013	**0.013**	1
SUBORD2(1,0.0,0)	0.043	0.096	0.106	26	-0.008	0.071	**0.072**	234	0.054	0.063	0.083	4	0.031	0.090	0.095	5
SUBORD2(0,0.3,0)	0.032	0.045	0.055	224	0.022	0.040	0.046	2016	0.028	0.025	**0.037**	1	0.035	0.024	0.042	1
SUBORD2(1,0.3,0)	-0.013	0.079	0.080	37	0.043	0.061	0.074	486	-0.045	0.063	0.078	4	-0.006	0.054	**0.055**	4

TABLE 4. Same as Table 3 with length of the time series = 4096.

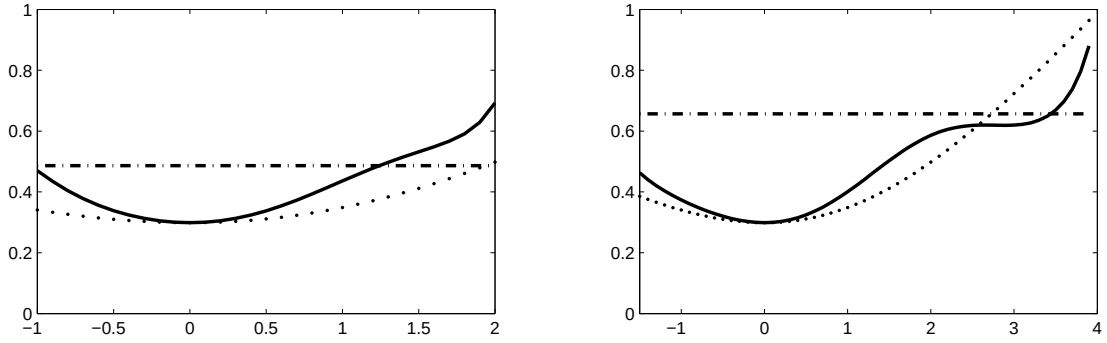


FIGURE 1. Comparison of the asymptotic variances of LWF and LWW estimators as functions of d. The dot/dash line displays the variance (75) of the LWF estimator with taper order τ; the plain curve displays the variance $\rho^2(d,\ell)$ in (66) with $\ell = 7$ of the LWW estimator using Daubechies wavelets of order M; the dotted curve displays the variance (67) of the LWW estimator using the ideal Shannon wavelet. Left panel: $\tau = 2$ for the LWF, $M = 2$ for the LWW. Right panel: $\tau = 4$ for the LWF, $M = 4$ for LWW.

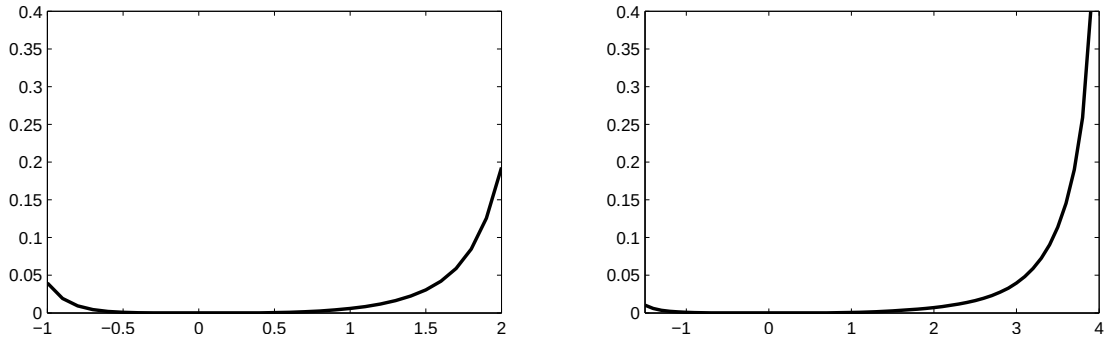


FIGURE 2. Comparison of the asymptotic variance of the LRW estimator using Abry-Veitch weights given by (63) with the LRW estimator using optimal weights given by (61). We plot $\rho^2(d,\ell) - \rho^2_{\mathrm{opt}}(d,\ell)$ as a function of d with $\ell = 7$. We used Daubechies wavelets for two different values for M. Left panel: $M = 2$. Right panel: $M = 4$.

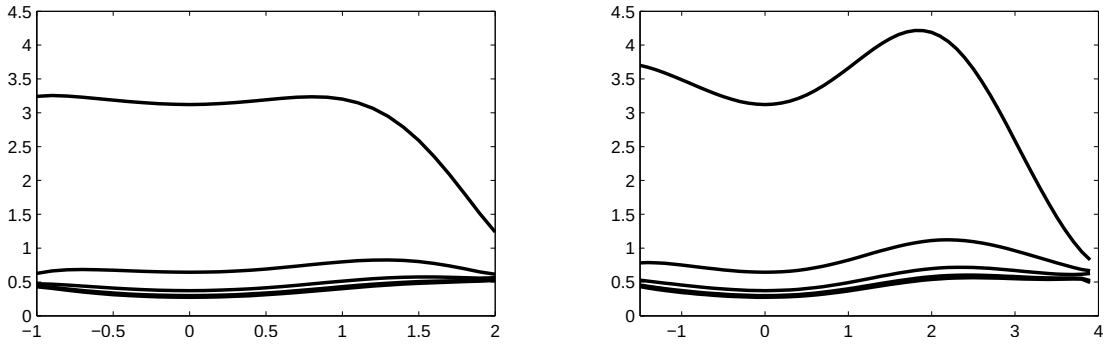
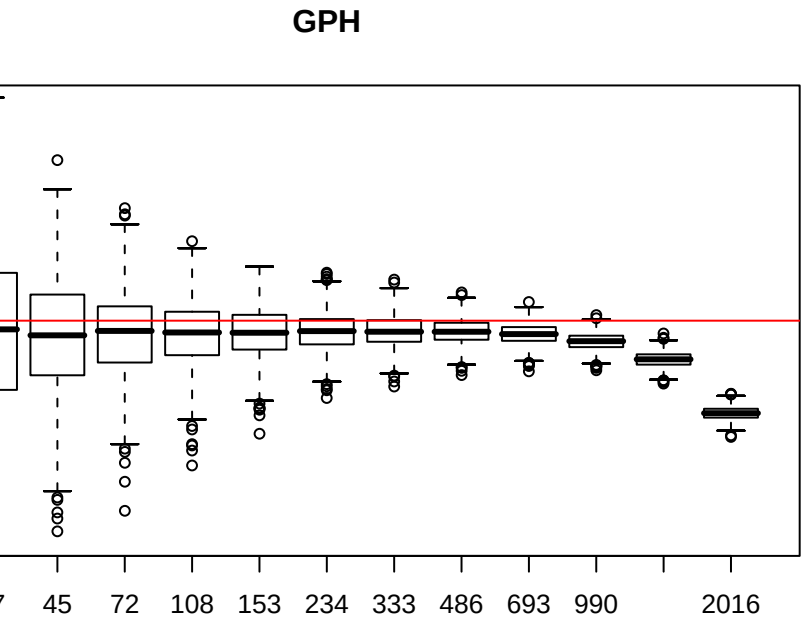
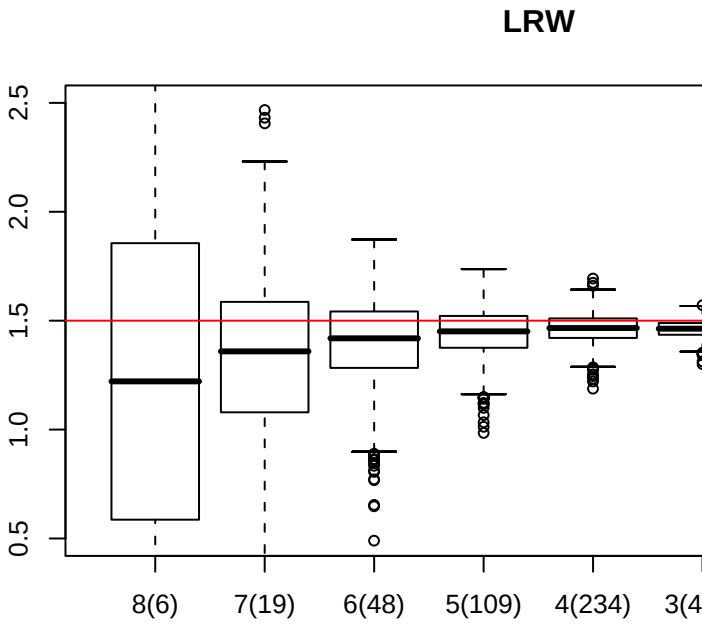
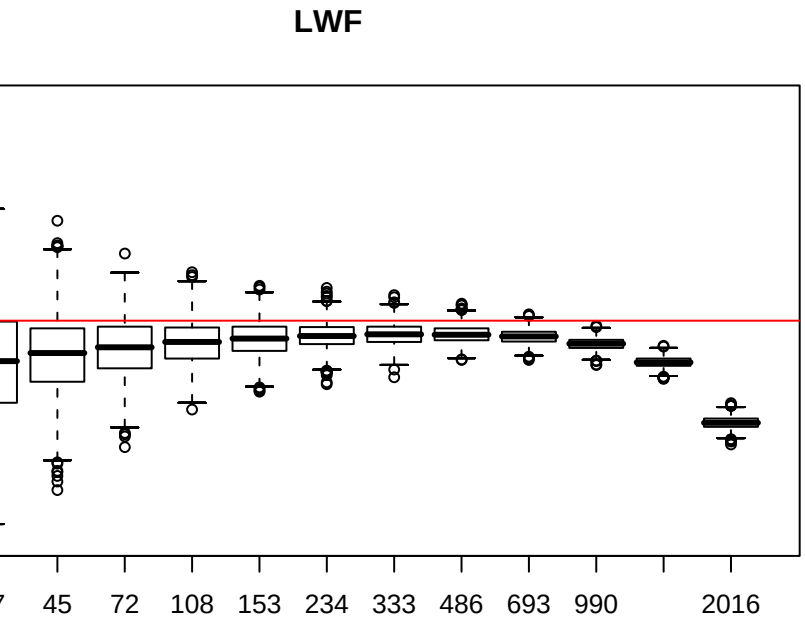
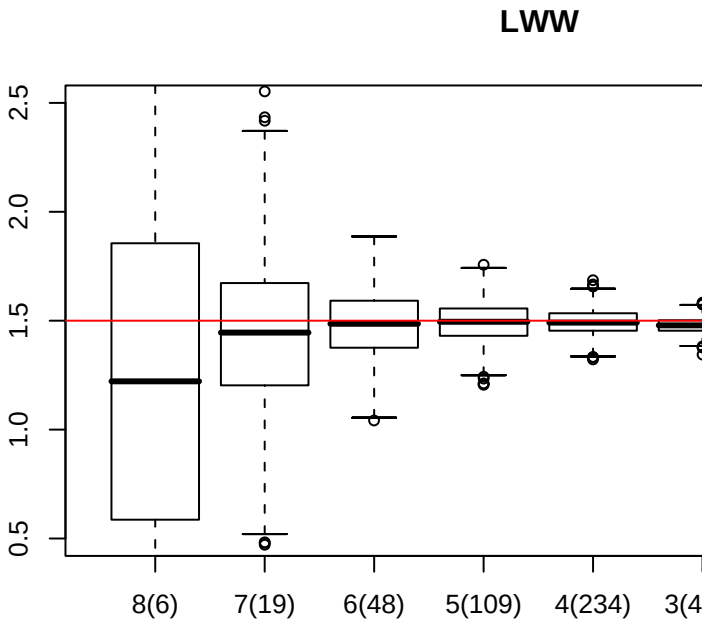


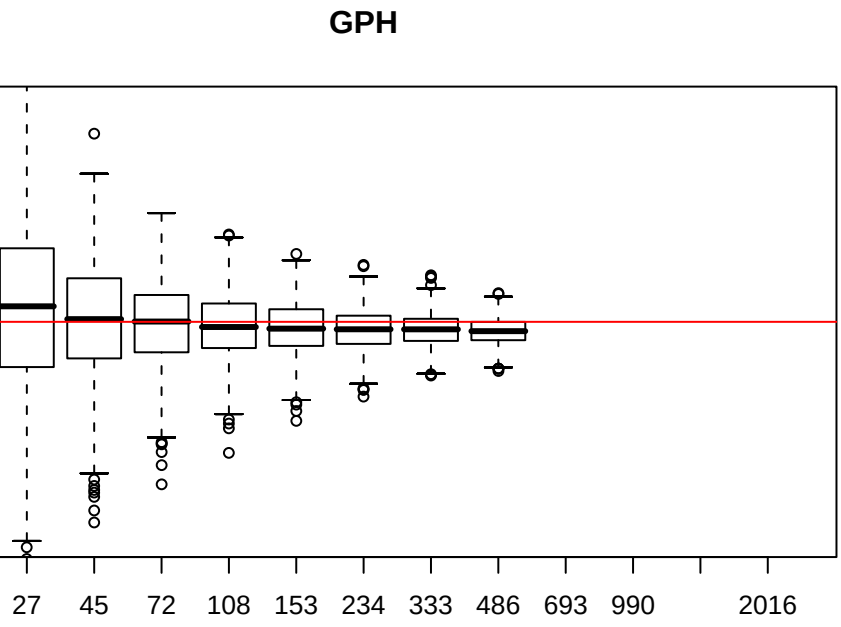
FIGURE 3. Comparison of the asymptotic variance (66) of the LRW estimator using Abry-Veitch weights given by (63) with the one of LRW estimator using optimal weights given by (61). We plot $\rho^2_{\mathrm{opt}}(d,\ell)$ as a function of d for successive values of $\ell = 1, 3, 5, 7, 9$ (from top to bottom). We used Daubechies wavelets for two different values for M. Left panel: $M = 2$. Right panel: $M = 4$.

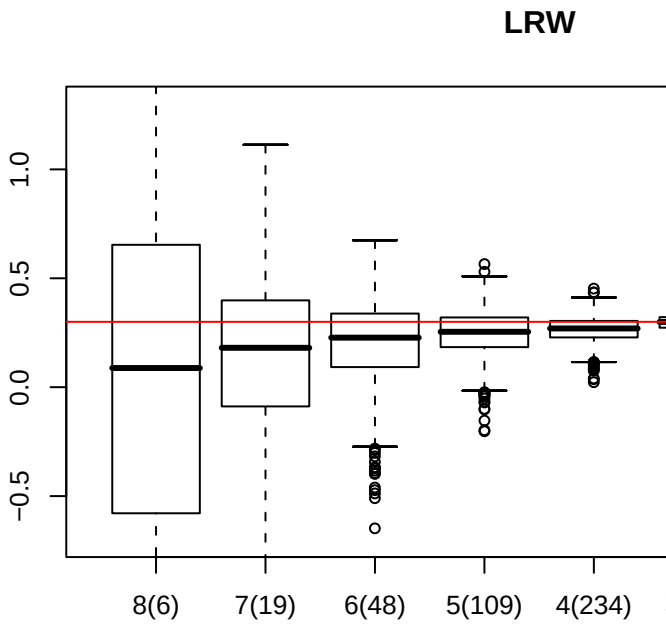

GPH
45
72
108
153
234
333
486
693
990
2016
Number of Fourier frequencies used
Taper order = 5 , pooling order = 4

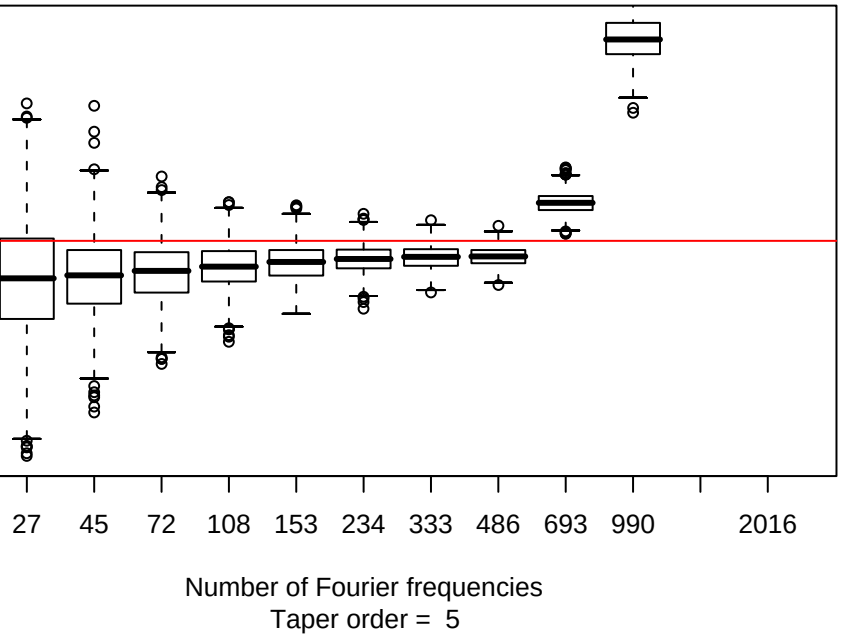

LRW
0.5
1.0
1.5
2.0
2.5
8(6)
7(19)
6(48)
5(109)
4(234)
Finest scale (equivalent number of frequ
Daubechies 8, largest scale =

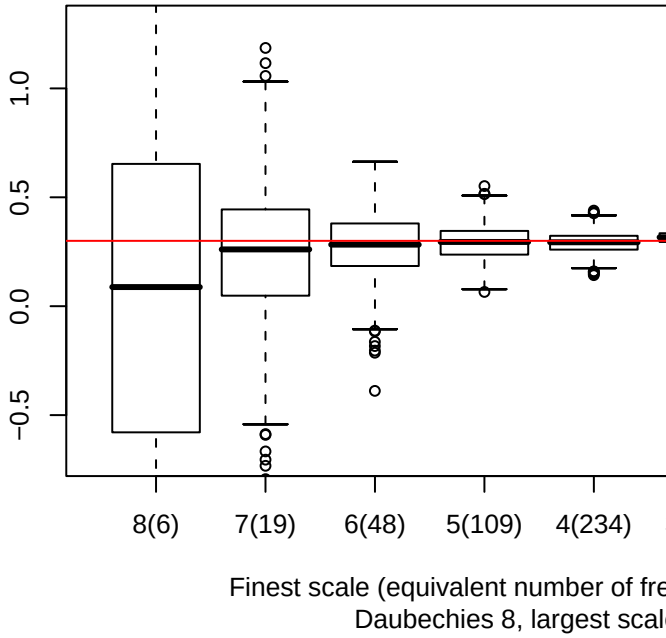

LWF
45
72
108
153
234
333
486
693
990
2016
Number of Fourier frequencies
Taper order = 5

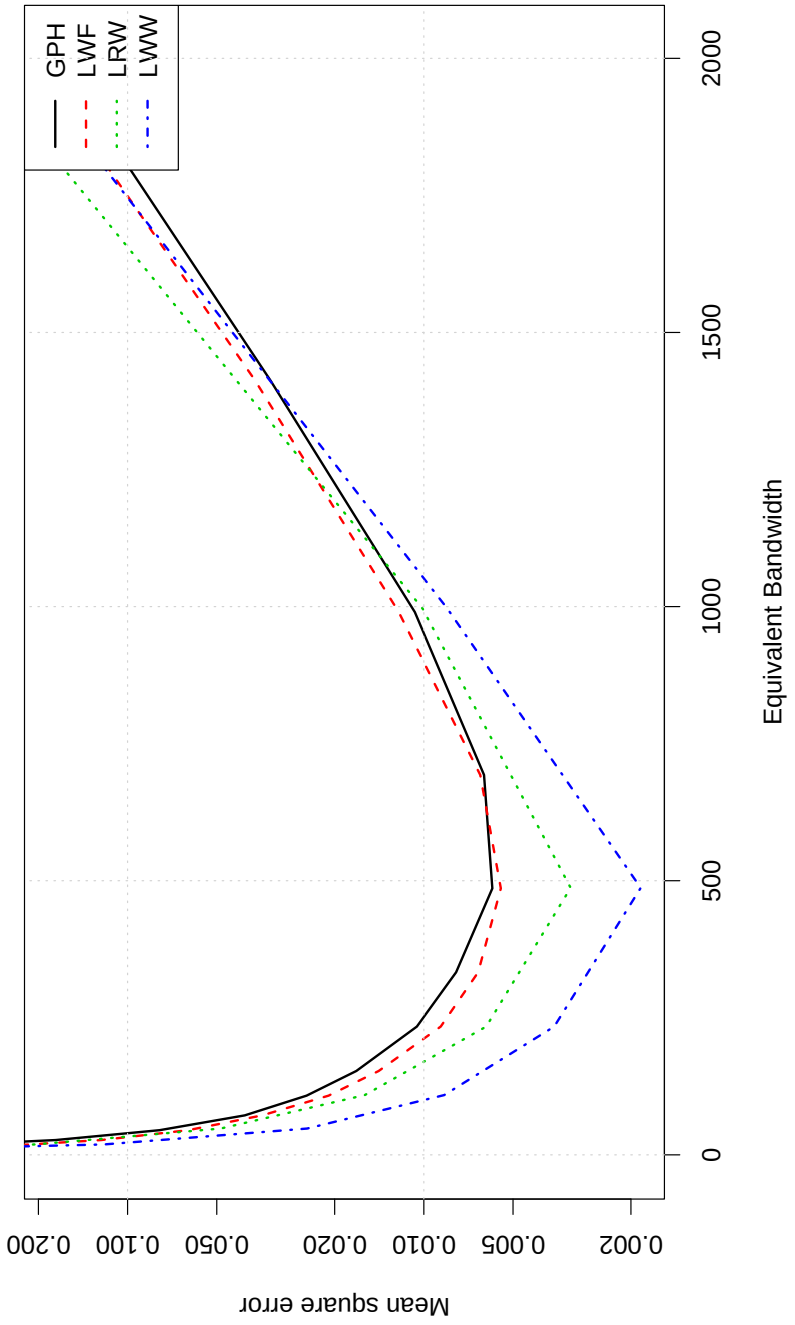

LWW
0.5
1.0
1.5
2.0
2.5
8(6)
7(19)
6(48)
5(109)
4(234)
Finest scale (equivalent number of frequ
Daubechies 8, largest scale =


GPH
27
45
72
108
153
234
333
486
693
990
2016
Number of Fourier frequencies used
Taper order = 5 , pooling order = 4


LRW
1.0
0.5
0.0
−0.5
8(6)
7(19)
6(48)
5(109)
4(234)
Finest scale (equivalent number of fre
Daubechies 8, largest scal


LWF
27
45
72
108
153
234
333
486
693
990
2016
Number of Fourier frequencies
Taper order = 5


LWW
1.0
0.5
0.0
−0.5
8(6)
7(19)
6(48)
5(109)
4(234)
Finest scale (equivalent number of fre
Daubechies 8, largest scal


GPH
LWF
LRW
LWW
Mean square error
0.002
0.005
0.010
0.020
0.050
0.100
0.200
0
500
1000
1500
2000
Equivalent Bandwidth


GPH
LWF
LRW
LWW
1500
2000