İSTATİSTİK: JOURNAL OF THE TURKISH STATISTICAL ASSOCIATION

İSTATİSTİK

Vol. 6, No. 1, January 2013, pp. 27–41
ISSN 1300-4077 | 13 | 1 | 27 | 41

SEMIPARAMETRIC INFERENCE AND BANDWIDTH CHOICE UNDER LONG MEMORY: EXPERIMENTAL EVIDENCE*

Uwe Hassler† and Maya Olivares
Goethe University Frankfurt

Abstract: The most widely used semiparametric estimators under fractional integration are variants of the local Whittle [LW] estimator. They are consistent for the long memory parameter d and follow a limiting normal distribution. Such properties require the bandwidth m to satisfy certain restrictions for the estimators to be "local" or semiparametric in large samples. Optimal rates for m are known and data-driven selection procedures have been proposed. A Monte Carlo study is conducted to compare the performance of the LW and the so-called exact LW estimators both in terms of experimental size when testing hypotheses about d and in terms of root mean squared error. In particular, the choice of the bandwidth is addressed. Further, competing approximations to limiting normality are compared.

Key words: Fractional integration, approximate normality, bandwidth selection.
History: Submitted: 11 June 2012; Revised: 14 September 2012; Accepted: 25 October 2012

1. Introduction

Persistence in the sense of slowly decaying autocorrelations is a stylized fact with many economic and financial time series, see Henry and Zaffaroni [12] for a survey. Such persistence is often called long memory. It can be captured by models that are fractionally integrated of order d, $I(d)$, with $0 < d < 1$, which extends the classical $I(0)/I(1)$ paradigm.

Two popular procedures to analyze long memory are the log-periodogram regression (GPH after Geweke and Porter-Hudak [4]), and the local Whittle estimator [LW], proposed by Künsch [18] and investigated by Robinson [22]. Both estimators are asymptotically normally distributed and consistent for $d \in (-0.5, 0.75)$, but the LW estimator is more efficient asymptotically. Still, LW is inconsistent for $d \geq 1$ and lacks asymptotic normality for $d > 3/4$, see Velasco [30]. Therefore, diverse methods have been proposed to improve its statistical properties. Well-known and easily implemented extensions include data differencing and periodogram tapering, see for example Hurvich and Ray [17], Hurvich and Chen [15] and Velasco [30]. However, Shimotsu and Phillips [28] argue that the first alternative requires prior knowledge of the degree of differencing, while the second one leads to an increase in the variance of the estimator. They propose a computationally more demanding variant of LW called Exact Local Whittle estimator [ELW]. It is consistent and follows the same limiting distribution as the LW estimator, however, it does so for a much larger parameter space. Alternatively, Abadir, Distaso and Giraitis [1] introduce a fully extended (or: nonstationarity-extended) version of the LW estimator. A thorough discussion of these LW-type estimators, and also of tapered LW estimators as well as wavelet-based competitors can be found in Faÿ, Moulines, Roueff and Taqqu [3].

* This paper is a revised and shortened version of a chapter from the PhD thesis of the second author. We thank an anonymous referee, Jörg Breitung, Dieter Nautz and Katsumi Shimotsu for helpful comments and suggestions. Financial support by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) through HA-3306/4-1 is gratefully acknowledged.

† Corresponding author: Goethe University Frankfurt, Statistics and Econometric Methods, Grüneburgplatz 1, D-60323 Frankfurt, Germany. Phone: +49 69 798 34762, Email: hassler@wiwi.uni-frankfurt.de.

In this paper, we focus on two variants of the LW procedure and compare the original LW and ELW estimators in an extensive Monte Carlo study. Several papers have studied the selection of the number of periodogram ordinates (bandwidth m) in relation to estimation bias and root mean squared error [RMSE], see for instance Robinson and Henry [24], Henry [10], and Nielsen and Frederiksen [20]. But it seems that size properties when testing for the true parameter value building on approximate normality have hardly been reported (an exception being the limited evidence in Hauser [9] and Hurvich and Chen [15, Table II] for just one bandwidth choice). For that reason we provide on the one hand experimental evidence not only on RMSE, but in particular on distortions of the nominal size (or coverage) as a function of the chosen bandwidth. Additionally, we compare deterministic with data-driven bandwidth selection rules. The latter ones have been proposed by Robinson [21], Delgado and Robinson [2] and Henry and Robinson [11]. On the other hand, we compare an alternative approximation of the test statistic ($\mathcal{R}^*$ from eq. (3.6)) with the usual asymptotic version ($\mathcal{R}$ from (3.4)), motivated by earlier findings in Hurvich and Chen [15]. We arrive at three relevant conclusions for empirical work that are summarized at the end of the paper.

The rest of the paper is organized as follows. Section 2 introduces the long memory model of fractional integration and the (E)LW estimator. In Section 3, we discuss approximations to the limiting normal distribution used for testing hypotheses about d, while simulation evidence is contained in Section 4. The final section summarizes our main findings.

2. Model and estimation

2.1. Fractional integration

The most widely used model to capture long memory is a fractionally integrated process $\{y_t, t \in \mathbb{Z}\}$, given by

$$(1-L)^d y_t = x_t \,, \quad -1 < d < 0.5 \,, \tag{2.1}$$

where $(1-L)^d$ with the usual lag operator L is given by binomial expansion,

$$(1-L)^d = \sum_{j=0}^{\infty} \pi_{j,d} L^j \,, \quad \pi_{0,d} = 1, \; \pi_{j,d} = \frac{j-1-d}{j} \pi_{j-1,d} \,, \quad j \geq 1 \,, \tag{2.2}$$

and $\{x_t\}$ is a purely stochastic, stationary process with short memory. More precisely, we assume that the spectral density of $\{x_t\}$ is bounded and bounded away from zero at frequency zero, such that the process is $I(0)$. In case of fractional integration one allows for non-integer values of $d \in \mathbb{R}$. Equation (2.1) defines a stationary process if and only if $d < 0.5$, see e.g. Granger and Joyeux [6] and Hosking [13].

In the time domain, the persistence of a fractionally integrated process is reflected by the behaviour of a hyperbolically decaying autocovariance sequence. For $0 < d < 0.5$, the autocovariances $\gamma_h = E(y_t y_{t+h})$ decay for a constant C_d depending on d so slowly,

$$\gamma_h \sim C_d \, h^{2d-1}, \quad h \to \infty \,, \tag{2.3}$$

that they are not summable, which characterizes long memory in the time domain:

$$\sum_{h=0}^{H} |\, \gamma_h \,| \to \infty, \quad H \to \infty \,.$$

Often, it is assumed that $\{x_t\}$ is an autoregressive moving-average process [ARMA], which we do not require here. For a fairly general sufficient condition on the short memory component $\{x_t\}$ that guarantees (2.3), see Hassler and Kokoszka [7, Coro. 2.1].

In the frequency domain, long memory translates into unboundedness of the spectral density at frequency zero. Particularly, it holds for $\{y_t\}$ with spectral density $f_y(\lambda)$ that

$$f_y(\lambda) \sim \lambda^{-2d} f_x(0), \quad \lambda \to 0\,, \tag{2.4}$$

where $f_x(\lambda)$ stands for the spectral density of the short memory component $\{x_t\}$. Hence, f_y is integrable for $d < 0.5$ notwithstanding the singularity at frequency zero.

Nonstationary fractionally integrated processes can be defined in terms of integer differences $(\Delta = 1 - L)$ for $0.5 \leq d < 1.5$,

$$\Delta y_t = z_t \ \sim I(d-1)\,,$$

where $\{z_t,\, t \in \mathbb{Z}\}$ is $I(d-1)$ as defined in (2.1). Consequently,

$$y_t = y_0 + \sum_{j=1}^{t} z_j\,, \quad t = 1, \ldots, T\,, \tag{2.5}$$

is integrated of order d. Such processes have been labelled "type I" by Marinucci and Robinson [19], see also Robinson [23] for a discussion. Alternatively, many people work under the assumption of "type II" processes defined as

$$y_t = \mu + \sum_{j=0}^{t-1} \psi_{j,d}\, x_{t-j}\,, \quad t = 1, \ldots, T\,, \tag{2.6}$$

where $\psi_{j,d}$ are from the truncated expansion of Δ^{-d}, i.e. $\psi_{j,d} = \frac{j-1+d}{j}\psi_{j-1,d}$, and a constant μ is added to the process. The model from (2.6) can also be used for $d < 0.5$, although the process becomes stationary only asymptotically.

2.2. (Exact) Local Whittle [LW] estimation

Whittle [31, 32] suggested for stationary processes an approximation of the likelihood function in the frequency domain, which relies on the periodogram. Consider the discrete Fourier transform [DFT] $w_y(\cdot)$ of $\{y_t\}$ $(t = 1 \ldots T)$,

$$w_y(\lambda_j) = \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} y_t \exp\{i\lambda_j t\}\,, \quad i^2 = -1\,,$$

at the jth harmonic frequency $\lambda_j = \frac{2\pi j}{T}$. Then the periodogram simply is

$$I_y(\lambda_j) = |w_y(\lambda_j)|^2\,, \quad j = 1, \ldots, M = \left\lfloor \frac{T-1}{2} \right\rfloor, \tag{2.7}$$

where $\lfloor \cdot \rfloor$ stands for the floor operator. The log likelihood approximation in the frequency domain becomes

$$l_M(d) = -\sum_{j=1}^{M} \log f_y(\lambda_j) - \sum_{j=1}^{M} \frac{I_y(\lambda_j)}{f_y(\lambda_j)}\,.$$

Now, let us assume model (2.5). The LW estimator maximizes the log likelihood *locally* over a vicinity close to frequency 0, where the slope of f_y varies with d alone, see (2.4). To that end M is replaced by the bandwidth m. A crucial condition for the consistency of the estimator in the presence of short memory in $\{x_t\}$ is that the number of harmonic frequencies m used in the estimation must diverge more slowly than the sample size T:

$$\frac{1}{m} + \frac{m}{T} \to 0\,, \quad T \to \infty\,. \tag{2.8}$$

Replacing $f_y(\lambda_j) \sim G\lambda_j^{-2d}$ where $G = f_x(0)$, the negative local log likelihood becomes

$$-l_m(d) \approx \sum_{j=1}^{m} \left(\log G - 2d \log(\lambda_j) + \frac{I_y(\lambda_j)}{G\lambda_j^{-2d}} \right) =: Q_m(G,d) \, .$$

Concentrating G out yields $\widehat{G} = \frac{1}{m}\sum_{j=1}^{m} \lambda_j^{2d} I_y(\lambda_j)$. Hence, the estimation of d requires minimizing

$$R(m,d) := \log\left\{ \frac{1}{m}\sum_{j=1}^{m} \lambda_j^{2d} I_y(\lambda_j) \right\} - \frac{2d}{m}\sum_{j=1}^{m} \log(\lambda_j) \, , \tag{2.9}$$

being the LW estimator thus defined as

$$\widehat{d}_{LW} = \arg\min R(m,d) \, .$$

Obviously, $\widehat{d}_{LW}$ crucially hinges on m as well as on the slope of $f_x(\lambda)$ at the origin, which is surpressed for notational convenience. Robinson [22] establishes limiting normality under (2.8) and further assumptions,

$$\sqrt{m}\,(\widehat{d}_{LW} - d) \Rightarrow \mathcal{N}\left(0, \frac{1}{4}\right) \, , \tag{2.10}$$

for $d \in (-0.5, 0.5)$ where "$\Rightarrow$" denotes convergence in distribution. Since the periodogram is shift invariant, the LW estimator is not affected by a mean μ different from zero and does not require an estimation of μ. Notice that the limiting variance is smaller than that of the famous semiparametric competitor, the log-periodogram regression by Geweke and Porter-Hudak [4]. Moreover, LW has been recommended since it is robust with respect to heteroskedasticity of a certain degree, see Robinson and Henry [24] and Shao and Wu [25]. Finally, Velasco [30] extended the results by Robinson [22] showing that the LW estimator is consistent for $d \in (-0.5, 1)$ and asymptotically normal for $d \in (-0.5, 0.75)$.

The issue of nonstationarity has been addressed more fully by Shimotsu and Phillips [28]. They propose to correct the DFT by adding a complementing term ensuring a valid approximation that holds for every value of d. This so-called exact LW procedure [ELW] implies replacing $\lambda_j^{2d} I_y(\lambda_j)$ in (2.9) by $I_{\Delta^d y}(\lambda_j)$, and it is valid if $\mu = 0$ in (2.6). For means different from zero, Shimotsu [27] suggests to demean $\{y_t\}$ with an appropriate estimator $\widehat{\mu}$, and to compute the exact LW estimator from the demeaned data. The objective function to be minimized becomes

$$R_E(m,d) := \log\left\{ \frac{1}{m}\sum_{j=1}^{m} I_{\Delta^d(y-\widehat{\mu})}(\lambda_j) \right\} - \frac{2d}{m}\sum_{j=1}^{m} \log(\lambda_j) \, , \tag{2.11}$$

where $I_{\Delta^d(y-\widehat{\mu})}(\lambda_j)$ is the periodogram of $\{\Delta^d(y_t - \widehat{\mu})\}$. To determine the fractional differences, it is assumed that $\{y_t\}$ is given by a type II process as in (2.6). It turns out that the first sample observation y_1 is a reliable mean estimator in the case of large values of d, while the usual arithmetic mean $\overline{y}$ does a good job for small parameter values of d. Hence, Shimotsu [27] puts forward the following weighted estimator:

$$\begin{aligned} \widehat{\mu}(d) &= v(d)\,\overline{y} + (1 - v(d))\, y_1 \, , \\ v(d) &= \begin{cases} 1 \, , & d \le 0.5 \\ \frac{1+\cos(4\pi d)}{2} \, , & 0.5 < d < 0.75 \\ 0 \, , & d \ge 0.75 \end{cases} \, . \end{aligned}$$

To get a feasible procedure, he considers two steps. First, one determines an estimator of $\widehat{d}$ independent of μ in order to get an estimator of the constant: $\widehat{\mu} = \widehat{\mu}(\widehat{d})$. In a second step, the slope and Hessian of $R_E(m,d)$ are used to compute the feasible estimator (a MATLAB code is available from the homepage of K. Shimotsu):

$$\widehat{d}_{2ELW} = \widehat{d} - \frac{R_E'(m,\widehat{d})}{R_E''(m,\widehat{d})}.$$

Shimotsu [27, Theo. 3] shows that the two-step ELW estimator $\widehat{d}_{2ELW}$ is consistent and has the same limiting distribution as the LW and ELW estimators under $-0.5 < d < 2$.

3. Approximate inference

One goal of semiparametric inference is hypothesis testing about the long memory parameter d. Hypotheses of interest are $d \leq 0$ (short memory) vs. $d > 0$ (long memory), or $d \geq 0.5$ (nonstationarity) vs. $d < 0.5$ (stationarity). The tests are approximate in that they rely on limiting normality (as $m \to \infty$) of the appropriately standardized estimators.

3.1. Bandwidth selection

Robinson [22] proves limiting normality for the LW estimator under

$$m = T^{\alpha}, \quad 0 < \alpha < 0.8. \tag{3.1}$$

This rate is from Robinson [22, Assumption A4'] in the case $\{x_t\}$ is ARMA (for $\beta = 2$ in his notation); see Shimotsu and Phillips [28] for a slightly stronger assumption in the case of ELW estimation. In practice, the choice of the bandwidth m will be crucial for reliable inference. It balances a trade-off between variance and bias, see Henry and Robinson [11]. Thus, if m is chosen too large or too small, the outcome of the estimation may wrongly suggest a certain degree of persistence. To avoid such pitfalls, many empirical researchers typically opt for choosing a grid of bandwidth values and then plot the estimates against different values of m, see also Taqqu and Teverovsky [29] for graphical bandwidth selection. Ideally, one observes three regimes: With small values of m the estimates will display high variability, then the plot of the estimates should become approximately flat, while with further growing values of m the estimates may start to fall or to rise because of a bias due to a short memory component. In such an ideal situation one would choose m from the middle regime. In practice, however, such an ideal situation will rarely be encountered unless the sample size is very large.

As an alternative to graphical means, data-driven techniques for bandwidth choice have been proposed. They rely on a minimization of the asymptotic mean squared error and have been proposed by Delgado and Robinson [2], Henry and Robinson [11], Henry [10], Giraitis, Robinson and Samarov [5] and Hurvich and Deo [16]. While the latter two contributions concentrate on the log-periodogram regression estimator of d, Henry and Robinson [11] and Henry [10] derive an algorithm to obtain an optimal bandwidth for the LW estimator.

The approximative optimal spectral bandwidth derived heuristically in Henry and Robinson [11] must be iterated until convergence to an optimal bandwidth value is achieved. It is defined as

$$\begin{aligned} \widehat{d}^{(k)} &= \arg\min R(\widehat{m}^{(k)}; d)\,, \\ \widehat{m}^{(k+1)} &= \left(\frac{3T}{4\pi}\right)^{4/5} \left|\theta + \frac{\widehat{d}^{(k)}}{12}\right|^{-2/5}, \end{aligned} \tag{3.2}$$

with initial value $\widehat{m}^{(0)} = T^{0.8}$. Strictly speaking, the optimal rate of $T^{0.8}$ in (3.2) violates the condition in (3.1). In the case of the long memory representation defined in (2.4), if the first and second derivatives of $f_x(\lambda)$ exist, the parameter θ from (3.2) is defined as $\theta = f_x''(0)/2\,f_x(0)$, as shown by Delgado and Robinson [2]. The authors propose a simple feasible approximation for the unknown parameter θ, which is motivated by a Taylor expansion and consists in regressing the periodogram $I_y(\lambda_j)$ on the regressors $Z_{j\ell}(\widehat{d}^{(0)})$, $\ell = 0,1,2$:

$$I_y(\lambda_j) = \sum_{\ell=0}^{2} Z_{j\ell}(\widehat{d}^{(0)})\, \widetilde{\varphi}_\ell + \widetilde{\varepsilon}_j \quad j = 1,\dots,\widehat{m}^{(0)},$$

where $Z_{j\ell}(d) = |1-\exp\{i\lambda_j\}|^{-2d}\, \lambda_j^\ell/\ell!$. The estimates of $f_x(0)$ and $f_x''(0)$ are $\widetilde{\varphi}_0$ and $\widetilde{\varphi}_2$, respectively, so that the estimated parameter is given by

$$\widehat{\theta} = \frac{\widetilde{\varphi}_2}{2\widetilde{\varphi}_0}\,. \tag{3.3}$$

In principle, one could include the determination of $\widehat{\theta}$ as part of the kth iteration step, but Delgado and Robinson [2] advice against doing so, see also Henry [10].

3.2. Approximation of the asymptotic variance

Let now $\widehat{d}$ stand generically for the local Whittle estimator $\widehat{d}_{LW}$or the two-step mean-corrected version $\widehat{d}_{2ELW}$ of the ELW estimator by Shimotsu [27]. If we wish to test a null hypothesis about d_0, the asymptotic version of the test statistic becomes

$$\mathcal{R} = 2\sqrt{m}\,(\widehat{d}-d_0)\,, \tag{3.4}$$

which is compared with critical values from the standard normal distribution because of (2.10). In order to improve the size properties of the tests in finite samples, the bandwidth m in (3.4) can be replaced with an approximation m^* where

$$m^* = \sum_{j=1}^{m} \nu_j^2 \quad \text{with} \quad \nu_j = \log j - \frac{1}{m}\sum_{j=1}^{m}\log j\,, \tag{3.5}$$

because

$$\frac{m^*}{m} = 1 + O\left(\frac{\log^2 m}{m}\right) \to 1 \quad \text{as} \quad m\to\infty\,,$$

see Robinson [22, p.1645] or Hurvich and Beltrao [14, Lemma 1]. The rationale behind m^* stems from the Hessian of $R(m,d)$ from (2.9) evaluated at maximum likelihood,

$$\frac{\partial^2 R(m,\widehat{d})}{\partial d^2} = 4\,\frac{m^*}{m} + o_p(1)\,,$$

see Robinson [22, (4.10)] or Hurvich and Chen [15, p.163], and Shimotsu and Phillips [28, p.1916] for ELW. This motivates the usual maximum-likelihood approximation building on the Fisher information:

$$\sqrt{m}\,(\widehat{d}-d_0) \;\sim\; \mathcal{N}\left(0,\,\frac{m}{4\,m^*}\right).$$

The approximate test statistic $\mathcal{R}^*$ is therefore defined as

$$\mathcal{R}^* = 2\sqrt{m^*}\,(\widehat{d}-d_0)\,, \tag{3.6}$$

to be compared with standard normal percentiles.

Normalizing with m^* instead of m comes in naturally for the log-periodogram regression, too, due to studentizing the estimator with the regressor being $\log\left(4\sin^2(\lambda_j/2)\right) \sim 2\log\lambda_j$. In fact, this corresponds to the original proposal by Geweke and Porter-Hudak [4], which was found experimentally to outperform differing normalizations by Hassler, Marmol and Velasco [8, eqn. (7)]. For LW estimation, an asymptotically equivalent approximation of m building on $\log(2\sin(\lambda_j/2))$ instead of $\log j$ in (3.5) was advocated by Hurvich and Chen [15], while our asymptotically equivalent choice of m^* in (3.5) was used by Shimotsu [26].

In the next section we report experimental size properties of $\mathcal{R}$ and $\mathcal{R}^*$ under the null hypothesis that d_0 is the true value.

4. Experimental evidence

To investigate the finite-sample behaviour of the variants of the LW estimator, we now report the empirical size and RMSE from a simulation study for different bandwidth selection rules. These are deterministic as well as data-driven according to the iterative procedure described in (3.2).

The data generating process [DGP] is an ARFI$(1,d)$ model, where the short memory component $\{x_t\}$ is a stable AR(1) sequence,

$$x_t = a\,x_{t-1} + u_t\,, \tag{4.1}$$

with standard normal innovations $u_t \sim ii\mathcal{N}(0,1)$. We consider the following cases:

1. ARFI$(0,d)$, $d \in \{0, 0.45, 0.7\}$ and
2. ARFI$(1,0)$ with $a = 0.5$.

ARFI$(0,d)$ denotes the case of fractionally integrated noise where $a = 0$. The scheme to generate fractional integration is throughout of type II, see (2.6), where we set $\mu = 0$ without loss of generalization. For each Monte Carlo DGP, 1000 replications with $T \in \{256, 512, 1024\}$ observations were performed. All computations were performed with MATLAB. To estimate d, we minimize over the interval $[-1;3]$ using the routine "fminbnd".

From our experiments we present as empirical size $100\,\widehat{\alpha}$, where $\widehat{\alpha}$ is the relative frequency of rejection under the null at nominal level $\alpha_0 \in \{0.01, 0.05, 0.10\}$. Since $\widehat{\alpha}$ converges to α_0, the approximate 95% confidence interval is given by

$$\left[\widehat{\alpha} \pm 1.96\sqrt{\frac{\alpha_0(1-\alpha_0)}{1000}}\right] \quad \text{or} \quad \left[100\,\widehat{\alpha} \pm 1.96\,\sqrt{10\,\alpha_0(1-\alpha_0)}\right]. \tag{4.2}$$

The following table presents lengths of such intervals that allow to judge whether the percentages of rejections reported in the next two subsections are conformable with the corresponding nominal levels or not.

α_0	1%	5%	10%
CI from (4.2)	$[100\,\widehat{\alpha} \pm 0.617]$	$[100\,\widehat{\alpha} \pm 1.351]$	$[100\,\widehat{\alpha} \pm 1.859]$

In what follows, we highlight in bold face those experimental levels that are not significantly different from the nominal ones at the 95% level according to (4.2).

Table 1. White noise: ARFI(0,0)

Test statistic	n.s.	$T=256$			$T=512$			$T=1024$		
		$T^{0.55}$	$T^{0.65}$	$T^{0.75}$	$T^{0.55}$	$T^{0.65}$	$T^{0.75}$	$T^{0.55}$	$T^{0.65}$	$T^{0.75}$
	1%	6.7	4.5	3.4	3.9	3.2	2.3	3.8	2.4	**1.5**
$\mathcal{R}_{LW}$	5%	16.9	11.8	8.8	12.3	10.0	7.5	10.7	8.8	**6.0**
	10%	24.3	17.2	16.9	19.1	15.9	13.3	15.6	15.1	12.2
	1%	2.4	2.1	**1.5**	**1.6**	1.7	**1.1**	2.2	**1.2**	**1.1**
$\mathcal{R}^*_{LW}$	5%	7.8	6.7	**6.1**	**5.9**	**5.8**	**6.1**	6.6	**6.1**	**4.6**
	10%	14.4	12.1	**11.3**	**11.5**	**11.3**	**10.7**	**11.6**	**11.8**	**10.0**
RMSE		0.145	0.103	0.072	0.115	0.079	0.055	0.089	0.062	0.038
	1%	5.9	4.6	3.6	4.5	3.4	2.5	3.8	2.6	2.5
$\mathcal{R}_{2ELW}$	5%	15.6	11.9	9.1	12.6	10.7	7.7	10.6	8.6	9.3
	10%	24.5	18.7	18.5	19.5	16.0	14.5	16.7	15.9	14.4
	1%	1.8	2.3	1.9	1.9	1.9	**1.4**	2.3	**1.6**	1.7
$\mathcal{R}^*_{2ELW}$	5%	6.6	6.9	6.6	6.6	**6.2**	6.4	6.8	6.5	7.2
	10%	14.1	12.5	11.9	**11.5**	**11.7**	**11.8**	12.2	**11.8**	12.6
RMSE		0.148	0.108	0.076	0.118	0.085	0.058	0.092	0.067	0.042

Root mean squared errors and frequencies of rejections at nominal significance level n.s. when testing for the true value $d_0 = 0$ based on 1000 replications. The test statistics $\mathcal{R}$ and $\mathcal{R}^*$ are from (3.4) and (3.6), respectively. The bandwidths are $m = T^{0.55}, T^{0.65}, T^{0.75}$. In bold face experimental levels not significantly different from the nominal ones at 95% level according to (4.2).

4.1. Deterministic choice of m

As deterministic rules for bandwidth selection we include $m = T^{0.55}, T^{0.65}$ and $T^{0.75}$. Tables 1-4 report the rejection frequencies of the test statistics (3.4) and (3.6) for the LW estimator and the two-step mean-corrected ELW estimator, respectively. All empirical sizes are computed for the bandwidth values m and for the sample sizes T mentioned above.

Table 2. Stationary fractionally integrated noise: ARFI(0,0.45)

		$T=256$			$T=512$			$T=1024$		
Test statistic	n.s.	$T^{0.55}$	$T^{0.65}$	$T^{0.75}$	$T^{0.55}$	$T^{0.65}$	$T^{0.75}$	$T^{0.55}$	$T^{0.65}$	$T^{0.75}$
	1%	5.3	4.1	3.1	4.5	2.6	**1.2**	3.1	1.7	2.4
$\mathcal{R}_{LW}$	5%	14.2	13.4	10.3	12.2	8.6	8.4	10.5	7.8	7.2
	10%	23.2	20.1	17.1	20.3	15.2	14.6	17.3	13.6	12.1
	1%	2.0	**1.4**	**1.1**	**1.6**	**1.1**	**0.7**	1.7	**1.0**	1.7
$\mathcal{R}^*_{LW}$	5%	**5.9**	7.8	6.4	6.5	**5.5**	**5.8**	**5.8**	**5.2**	**5.2**
	10%	12.0	13.7	12.6	**11.7**	**11.0**	11.9	**11.7**	**10.9**	**10.7**
RMSE		0.149	0.103	0.073	0.115	0.083	0.055	0.090	0.060	0.040
	1%	5.7	4.2	3.4	4.6	2.6	1.7	3.5	2.2	2.5
$\mathcal{R}_{2ELW}$	5%	14.5	12.8	10.5	12.6	8.7	8.7	11.5	7.8	8.4
	10%	22.6	19.3	19.7	20.4	16.5	14.5	18.5	14.2	14.7
	1%	2.8	2.1	1.9	1.7	**1.5**	1.8	1.8	**1.6**	1.9
$\mathcal{R}^*_{2ELW}$	5%	8.1	7.5	6.8	**6.3**	**5.8**	**5.9**	**5.9**	**5.8**	7.3
	10%	13.9	13.7	13.9	11.9	**11.2**	11.9	**11.6**	**11.2**	12.9
RMSE		0.148	0.110	0.078	0.117	0.082	0.056	0.092	0.062	0.045

Results when testing for the true value $d_0 = 0.45$; for further comments see Table 1.

Generally, it can be observed that the size distortion of the test tends to decrease as T and m increase. In the case of white noise (Table 1), $\mathcal{R}$ is noticeably oversized, and the performance of all estimators is very similar. The rejection probabilities seem indeed to be markedly sensitive to the bandwidth choice, though this problem is reduced for large sample sizes. The variance approximation used in $\mathcal{R}^*$ from (3.6) clearly reduces the size distortion in all cases and for both estimators. When $d = 0.45$, the size distortion is at least as large as in the white noise case. The performance of the LW estimator in terms of size and RMSE is a bit better than that of 2ELW in almost all cases, also when the sample size is small, but the overall performance of the two approaches is quite similar. Again, the variance approximation used in (3.6) reduces the size distortion of all tests remarkably. In the case d takes on the nonstationary value 0.7, the two-stage ELW estimator outperforms the LW estimator, above all in terms of size distortion, see Table 3.

Table 3. Nonstationary fractionally integrated noise: ARFI(0,0.7)

		$T=256$			$T=512$			$T=1024$		
Test statistic	n.s.	$T^{0.55}$	$T^{0.65}$	$T^{0.75}$	$T^{0.55}$	$T^{0.65}$	$T^{0.75}$	$T^{0.55}$	$T^{0.65}$	$T^{0.75}$
	1%	5.8	3.7	3.4	4.0	3.4	4.2	3.6	2.9	4.4
$\mathcal{R}_{LW}$	5%	15.1	11.7	10.1	12.3	11.8	11.6	13.1	10.9	12.0
	10%	22.5	19.5	16.0	19.1	18.2	17.8	18.4	17.0	19.8
	1%	**1.6**	**1.3**	**1.5**	**1.5**	1.9	2.6	1.9	1.8	3.6
$\mathcal{R}_{LW^*}$	5%	7.2	6.6	7.0	**5.8**	7.3	8.5	7.4	7.3	10.6
	10%	13.1	12.4	12.5	**11.8**	13.4	15.0	13.8	13.6	17.1
RMSE		0.148	0.103	0.075	0.114	0.083	0.058	0.092	0.065	0.047
	1%	4.4	3.1	2.1	3.8	2.4	3.1	2.7	2.3	2.0
$\mathcal{R}_{2ELW}$	5%	12.5	9.9	9.6	11.8	10.8	9.2	9.7	7.3	8.0
	10%	21.2	17.2	14.6	18.9	17.5	15.3	17.2	13.8	14.4
	1%	1.7	**1.6**	**1.4**	**1.3**	**1.5**	**1.6**	**1.1**	**1.0**	**1.5**
$\mathcal{R}^*_{2ELW}$	5%	**5.6**	**5.0**	**5.2**	**5.8**	7.1	7.5	**5.6**	**5.0**	**5.7**
	10%	**11.3**	**11.1**	**11.7**	**10.9**	**11.7**	12.5	**11.4**	**9.6**	11.9
RMSE		0.144	0.102	0.072	0.112	0.081	0.056	0.088	0.058	0.042

Results when testing for the true value $d_0 = 0.7$; for further comments see Table 1.

When short-run dynamics are added to the model, the selection of an adequate bandwidth becomes even more decisive. Table 4 shows the simulation results for an ARFI(1,0) process with a moderate autoregressive coefficient $a = 0.5$. The influence of the short-term component dominates the behaviour of all estimators, and the contamination of the periodogram leads to a large bias in the estimation of d. The size distortion caused by the autoregressive term increases excessively as m gets large. If we allowed for a moving-average component instead of the AR(1) dynamics, similarly detrimental effects are obtained. The theoretical charm of semiparametric estimators consists in their (asymptotic) robustness against the presence of short memory. The simulations show, however, that the finite sample performance may be far from the asymptotic promise. We conclude that even for samples as large as $T = 1000$ the bandwidth must be chosen very conservatively ($m < T^{0.65}$) to get somewhere close to the nominal size. Once more, the variance approximation used in (3.6) clearly outperforms the test from (3.4).

Table 4. AR(1)

		$T=256$			$T=512$			$T=1024$		
Test statistic	n.s.	$T^{0.55}$	$T^{0.65}$	$T^{0.75}$	$T^{0.55}$	$T^{0.65}$	$T^{0.75}$	$T^{0.55}$	$T^{0.65}$	$T^{0.75}$
	1%	8.1	23.1	91.8	4.3	16.6	94.6	3.6	11.2	96.9
$\mathcal{R}_{LW}$	5%	18.5	41.8	97.3	13.1	31.2	98.4	13.0	24.3	99.2
	10%	25.7	52.3	97.9	20.1	43.0	99.4	19.5	34.7	99.5
	1%	3.1	13.4	88.2	**1.6**	9.8	92.0	1.9	7.2	95.7
$\mathcal{R}^*_{LW}$	5%	9.6	30.6	95.6	**6.1**	24.2	97.9	7.2	19.4	99.2
	10%	15.9	42.2	97.5	12.5	35.2	99.2	14.6	30.0	99.5
RMSE		0.158	0.172	0.276	0.120	0.120	0.220	0.091	0.087	0.174
	1%	7.1	28.6	95.6	4.8	20.5	98.5	3.4	10.9	98.4
$\mathcal{R}_{2ELW}$	5%	17.8	45.1	98.6	12.6	38.5	99.6	10.7	25.9	99.6
	10%	26.2	55.0	99.4	19.5	47.7	99.8	19.3	35.3	99.7
	1%	2.6	16.2	94.0	**1.6**	13.0	96.4	1.7	7.6	97.1
$\mathcal{R}^*_{2ELW}$	5%	8.4	35.5	97.6	6.7	31.0	99.6	7.1	20.7	99.6
	10%	15.5	46.1	99.3	12.3	42.6	99.8	13.0	31.2	99.7
RMSE		0.159	0.183	0.316	0.116	0.130	0.244	0.095	0.087	0.192

The model is ARFI(1,0) with $a=0.5$; for further comments see Table 1.

4.2. Data-driven choice of m

In this subsection we study the iterative procedure defined in (3.2), concentrating on the LW estimator in order to save space. In the white noise case we observe in Table 5 similar empirical sizes as under deterministic bandwidth selection in Table 1. For the ARFI(0,d) cases (Tables 5 and 6) the size distortion is larger than under the deterministic rules reported in Tables 2 and 3. Finally, in the ARFI(1,0) case (Table 6) the data-driven selection is superior only to large bandwidths chosen deterministically ($T^{0.65}$ or $T^{0.75}$). This reinforces the above warning to select m rather conservatively in practice to circumvent bias and size distortion due to short-memory. All in all, data-driven bandwidth determination has no benevolent effect on the size distortion. What is more, there were many cases where no convergence according to (3.2) was achieved (see Tables 5 and 6).

Table 5. LW and data-driven bandwidth selection

		White noise			ARFI(0,0.45)		
Test statistic	n.s.	$T = 256$	$T = 512$	$T = 1024$	$T = 256$	$T = 512$	$T = 1024$
	1%	4.9	3.5	2.7	8.2	6.5	4.1
$\mathcal{R}_{LW}$	5%	12.3	8.5	7.1	17.6	15.7	12.3
	10%	19.9	15.2	12.5	25.0	23.2	18.2
	1%	2.5	2.2	2.1	3.3	2.3	2.0
$\mathcal{R}^*_{LW}$	5%	7.5	**5.5**	**5.4**	9.7	9.8	8.3
	10%	13.7	11.3	**9.8**	15.9	15.9	14.3
Non convergence		70	50	26	29	8	8

Based on 1000 replications of the ARFI$(0,0)$ and ARFI(0,0.45) models. The number of non convergence cases is also reported. For further comments see Table 1.

Table 6. LW and data-driven bandwidth selection

		ARFI(0,0.7)			ARFI(1,0)		
Test statistic	n.s.	$T = 256$	$T = 512$	$T = 1024$	$T = 256$	$T = 512$	$T = 1024$
	1%	14.2	9.8	9.7	14.2	8.3	5.7
$\mathcal{R}_{LW}$	5%	23.6	19.0	18.4	24.0	16.4	15.0
	10%	31.2	26.6	26.1	31.8	23.4	21.6
	1%	8.1	4.1	3.8	7.1	4.0	3.0
$\mathcal{R}^*_{LW}$	5%	15.4	10.4	10.6	15.1	10.6	10.5
	10%	21.5	18.9	18.5	21.6	18.3	16.1
Non convergence		48	17	8	15	7	8

Based on 1000 replications of ARFI(0,0.7), and ARFI(1,0) with $a = 0.5$; for further comments see Table 5.

5. Concluding remarks

The LW estimator is asymptotically more efficient than other popular semiparametric long memory estimators. Moreover, the procedure has been proven to be robust to heteroskedasticity of a certain degree, see for instance Robinson and Henry [24]. At the same time it has at least three limitations. First, consistency and limiting normality only hold for a restricted parameter range excluding relevant cases of nonstationarity (see Robinson [22] and Velasco [30]). To overcome this shortcoming, Shimotsu and Phillips [28] proposed a computationally more involved variant called the exact LW estimator. It has the same limiting properties as LW but covers also the region of nonstationarity. In most practical situations, a mean-corrected version of the exact LW has to be worked with, see Shimotsu [27]. Second, the normal distribution of the normalized estimator holds of course only asymptotically. In finite samples it may be hard to control the probability of a type I error. It is a priori not clear how the estimator should be normalized to get a test statistic with satisfactory size properties in finite samples. Third, the semiparametric nature of LW hinges on an appropriate choice of a bandwidth m that balances the trade-off between bias and variance. The optimal rate of divergence for m was determined such as to minimize the asymptotic mean squared error [mse]. The resulting rate, however, violates the rate given in (3.1) ensuring limiting normality. Hence, the question arises how the normalized estimator will behave as a test statistic when the bandwidth is chosen according to a data-driven criterion that is mse optimal.

All three issues just raised with the (exact) LW estimator are addressed in our paper by means of a Monte Carlo study. Our contributions can be summarized as follows. First, the 2ELW by Shimotsu [27] is superior to the LW estimator when the process is indeed nonstationary. In the region of stationarity, however, there are many situations where LW dominates the exact variant in terms of size distortion and mse. Not knowing in practice, whether the true d is less than 0.5 or not, both ELW and LW can equally be advised. Second, with respect to the normal approximation we compared the usual asymptotic version of the test statistic $\mathcal{R}$ from (3.4) with the finite sample modification $\mathcal{R}^*$ from (3.6). Our experimental results on size distortions are very clear-cut: the finite sample approximation $\mathcal{R}^*$ constitutes a uniform improvement in that it always outperforms the classical variant. This holds for LW just as well as for 2ELW. Finally, we ran a horserace between deterministic bandwidth selection and data-driven selection rules relying on the mse optimal bandwidth rate. It turns out, generally speaking, that a data-driven bandwidth choice affects subsequent inference even in large samples. Resulting size distortions are large compared to the case of careful, moderate choices of deterministic values of m. Our experiments show that even a small bandwidth choice of $m = T^{0.55}$ may result in a too liberal size performance under short memory for samples with length between $T = 250$ and $T = 1000$.

References

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

[2] Delgado, M. and Robinson, P. M. (1996). Optimal spectral bandwidth for long memory. *Statistica Sinica*, **6**, 97–112.

[3] Faÿ, G., Moulines, E., Roueff, F. and Taqqu, M. S. (2009). Estimators of long memory: Fourier versus wavelets. *Journal of Econometrics*, **151**, 159–177.

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

[5] Giraitis, L., Robinson, P. M. and Samarov, A. (2000). Adaptive semiparametric estimation of the memory parameter. *Journal of Multivariate Analysis*, **72**, 183–207.

[6] Granger, C. W. J. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, **1**, 15–29.

[7] Hassler, U. and Kokoszka, P. (2010). Impulse responses of fractionally integrated processes with long memory. *Econometric Theory*, **26**, 1855–1861.

[8] —, Marmol, F. and Velasco, C. (2006). Residual log-periodogram inference for long-run relationships. *Journal of Econometrics*, **130**, 165–207.

[9] Hauser, M. A. (1997). Semiparametric and nonparametric testing for long memory: A monte carlo study. *Empirical Economics*, **22**, 247–271.

[10] Henry, M. (2001). Robust automatic bandwidth for long memory. *Journal of Time Series Analysis*, **22**, 293–316.

[11] — and Robinson, P. M. (1996). Bandwidth choice in gaussian semiparametric estimation of long range dependence. In P. M. Robinson and M. Rosenblatt (eds.), *Athens Conference on Applied Probability and Time Series, Vol. II*, New York: Springer, pp. 220–232.

[12] — and Zaffaroni, P. (2003). The long-range dependence paradigm for macroeconomics and finance. In P. Doukhan, G. Oppenheim and M. S. Taqqu (eds.), *Theory and Applications of Long-Range dependence*, New York: Birkhäuser Boston, pp. 417–438.

[13] Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, **68**, 165–176.

[14] Hurvich, C. M. and Beltrao, K. I. (1994). Automatic semiparametric estimation of the memory parameter of a long memory time series. *Journal of Time Series Analysis*, **15**, 285–302.

[15] — and Chen, W. W. (2000). An efficient taper for overdifferenced series. *Journal of Time Series Analysis*, **21**, 155–180.

[16] — and Deo, R. S. (1999). Plug-in selection of the number of frequencies in regression estimates of the memory parameter of a long-memory time series. *Journal of Time Series Analysis*, **20**, 331–341.

[17] — and Ray, B. K. (1995). Estimation of the memory parameter for nonstationary or noninvertible fractionally integrated processes. *Journal of Time Series Analysis*, **16**, 17–41.

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

[19] Marinucci, D. and Robinson, P. M. (1999). Alternative forms of fractional Brownian motion. *Journal of Statistical Planning and Inference*, **80**, 111–122.

[20] Nielsen, M. Ø. and Frederiksen, P. H. (2005). Finite sample comparison of parametric, semiparametric, and wavelet estimators of fractional integration. *Econometric Reviews*, **24**, 405–443.

[21] Robinson, P. M. (1994). Rates of convergence and optimal bandwidth in spectral analysis of processes with long range dependence. *Probability Theory and Related Fields*, **99**, 443–473.

[22] — (1995). Gaussian semiparametric estimation of long range dependence. *Annals of Statistics*, **23**, 1630–1661.

[23] — (2005). The distance between rival nonstationary fractional processes. *Journal of Econometrics*, **128**, 283–399.

[24] — and Henry, M. (1999). Long and short memory conditional heteroskedasticity in estimating the memory parameter of levels. *Econometric Theory*, **15**, 299–336.

[25] Shao, X. and Wu, W. B. (2007). Local whittle estimation of fractional integration for nonlinear processes. *Econometric Theory*, **23**, 899–951.

[26] Shimotsu, K. (2006). Simple (but effective) tests of long memory versus structural breaks. *Working Paper, Queen's University.*

[27] — (2010). Exact local Whittle estimation of fractional integration with unknown mean and trend. *Econometric Theory*, **26**, 501–540.

[28] — and Phillips, P. C. B. (2005). Exact local Whittle estimation of fractional integration. *The Annals of Statistics*, **33**, 1890–1933.

[29] Taqqu, M. S. and Teverovsky, V. (1996). Semi-parametric graphical estimation techniques for long-memory data. In P. M. Robinson and M. Rosenblatt (eds.), *Athens Conference on Applied Probability and Time Series, Vol. II*, New York: Springer, pp. 420–432.

[30] Velasco, C. (1999). Gaussian semiparametric estimation of non-stationary time series. *Journal of Time Series Analysis*, **20**, 87–127.

[31] Whittle, P. (1951). *Hypothesis testing in time series analysis.* Almqvist and Wiksell.

[32] — (1953). Estimation and information in stationary time series. *Arkiv för Matematik*, **2**, 423–434.

1 306010 270419 >