DEPARTMENT OF ECONOMICS

Working Paper

Estimation of Fractional Integration in the Presence of Data Noise

Niels Haldrup and Morten Ø. Nielsen

Working Paper No. 2003-10

ISSN 1396-2426

UNIVERSITY OF AARHUS • DENMARK

INSTITUT FOR ØKONOMI

AFDELING FOR NATIONALØKONOMI - AARHUS UNIVERSITET - BYGNING 322
8000 AARHUS C - ☎ 89 42 11 33 - TELEFAX 86 13 63 34

WORKING PAPER

Estimation of Fractional Integration in the Presence of Data Noise

Niels Haldrup and Morten Ø. Nielsen

Working Paper No. 2003-10

DEPARTMENT OF ECONOMICS

SCHOOL OF ECONOMICS AND MANAGEMENT - UNIVERSITY OF AARHUS - BUILDING 350
8000 AARHUS C - DENMARK ☎ +45 89 42 11 33 - TELEFAX +45 86 13 63 34

Estimation of Fractional Integration in the Presence of Data Noise

Niels Haldrup and Morten Ørregaard Nielsen[*]

July 18, 2003

Abstract. The paper presents a comparative study on the performance of commonly used estimators of the fractional order of integration when data is contaminated by noise. In particular, measurement errors, additive outliers, temporary change outliers, and structural change outliers are addressed. It occurs that when the sample size is not too large, as is frequently the case for macroeconomic data, then non-persistent noise will generally bias the estimators of the memory parameter downwards. On the other hand, relatively more persistent noise like temporary change outliers and structural changes can have the opposite effect and thus bias the fractional parameter upwards. Surprisingly, with respect to the relative performance of the various estimators, the parametric conditional maximum likelihood estimator with modelling of the short run dynamics clearly outperforms the semiparametric estimators in the presence of noise that is not too persistent.

Keywords: Fractional integration, long memory, outliers, measurement errors, structural change

JEL Classification: C2, C13, C22

1. Introduction

The past decade or so has witnessed an increasing interest in fractionally integrated processes as a convenient way of describing the long memory properties of many time series, see e.g. Sowell (1992a), and Baillie (1996) for a review. There is now a broad range of applications in finance, see e.g. Andersen *et al.* (2001) and Andersen *et al.* (2003), in macroeconomics, e.g. Diebold and Rudebusch (1989), Crato and Rothman (1994), Hassler and Wolters (1995), and Gil-Alana and Robinson (1997), and in electoral studies, see e.g. Box-Steffensmeir and Smith (1996), Davidson *et al.* (1997), and Dolado *et al.* (2003).

The dominating feature of fractionally integrated processes is that the autocorrelation function dies out very slowly at a hyperbolic rate, thus suggesting distant observations to be highly correlated. When the focus is on the analysis of financial time series loads of high-quality data will typically be available at a high sampling

[*] Department of Economics, University of Aarhus, Building 322, DK-8000 Aarhus C, Denmark. E-mail: nhaldrup@econ.au.dk and monielsen@econ.au.dk

frequency and often also the span of data is relatively long. On the other hand, although macroeconomic time series being analyzed for fractional integration and long memory typically cover a wide span of data sampled annually or quarterly there are good reasons to believe that the data quality is much more questionable compared to financial data.

The purpose of the present paper is to provide a comparative study of the implications of noisy data (measurement errors, outliers, and structural breaks) for sample sizes typically dealt with in the analysis of macroeconomic data, on a number of commonly used estimators of the fractional integration parameter. Our conjecture is that measurement errors and outliers that appear to be temporary will generally tend to bias the parameter of fractional integration downwards whereas outliers that are more permanent like e.g. structural shifts may tend to bias the fractional parameters upwards. Non-linearities can frequently be approximated by multiple level shift models and hence a derived conjecture is that long memory can be caused by the presence of non-linearities of the time series.

There is already some work done in the literature to examine these questions. For instance, Chong and Liu (1999) consider the properties of an estimator based on the partial autocorrelation function when data is measured with noise, but their study is only concerned with biases for a particular estimator which is only little used in the literature. For their estimator a downward bias is found. A similar conclusion is drawn by Maynard and Phillips (2001) with reference to an empirical study of various persistence measures of the forward premium. In Bos *et al.* (1999) an empirical study of G7 inflation rates together with simulations indicate that if the underlying series have level shifts, then the evidence of long memory and fractional integration can be spuriously exaggerated. Work by Granger and Ding (1996) and Diebold and Inoue (2001) also indicate that non-linear models, e.g. regime switching models, can give rise to processes being fractionally integrated.

In the present paper we consider a model setup allowing for a range of different outlier and measurement error components which can be temporary as well as persistent. The biases of fractional d estimators for various parametric and semiparametric estimators which are rather popular in applied work are examined in a Monte Carlo simulation study. The estimators considered are the fully parametric maximum likelihood estimators of Sowell (1992b) and Tanaka (1999) and the semiparametric estimators of Geweke and Porter-Hudak (1993), Kunsch (1987), and Robinson (1995a,b). It occurs that different kinds of noise may affect fractional integration inference differently and also the type of estimation method may have different properties. One general, and perhaps surprising, finding is that the conditional maximum likelihood estimator of Tanaka (1999) clearly ranks as the best, i.e. having the smallest biases, when the short run dynamics is being modelled and the noise and outliers tend not to be too persistent. For more persistent outliers and structural changes there is no clear pattern concerning which estimator to use; overall, the biases are positive and can be rather large in these cases. Amongst the semiparametric estimators the choice of a relatively low bandwidth parameter tends to bias estimators less when noise is not too persistent.

The paper is organized as follows. In section 2 we present the experimental design of the long memory models under scrutiny, estimation methods are described in section 3 followed by a discussion of the simulation results. The final section concludes.

2. The design of a fractional integration with noise process

Consider the univariate fractionally integrated process

$$\begin{aligned} (1-L)^d y_t &= \varepsilon_t, \quad t = 1,2,... & (1)\\ y_t &= 0, \quad t \leq 0, & (2) \end{aligned}$$

where ε_t is distributed as $i.i.d.(0,\sigma_\varepsilon^2)$ and $(1-L)^d$ is the fractional integration filter. y_t is a latent process which cannot be observed due to data contamination. Instead, we observe the series z_t defined as

$$z_t = y_t + v_t \tag{3}$$

with v_t being the error term contaminating y_t. In particular, we consider the noise mechanism

$$v_t = \frac{\theta}{(1-\alpha L)}\delta_t + \eta_t, \tag{4}$$

where $\eta_t \sim i.i.d(0,\sigma_\eta^2)$ is a measurement error and δ_t is a Bernoulli variable which can take either of the values 1 or -1 with a specified probability $p/2$. Otherwise, the value of δ_t equals zero. The first term in (4) is a general outlier component where we assume $|\alpha| \leq 1$, with L being the lag-operator. If $\alpha = 0$, $\theta\delta_t$ is a noise term generated by irregularly observed additive outliers (AO). The parameter θ is the magnitude of the outliers. Hence AO's are characterized by some non-repetitive events which occur irregularly and are unaffected by the dynamics of the y_t process. A different kind of outliers occur when α is non-zero and less than unity. In this situation the outliers also appear irregularly but tend to be persistent although eventually their effect will die out given the assumption $|\alpha| < 1$. We will refer to such outliers as temporary change (TC) outliers following Chen and Liu (1993). Finally, by letting $\alpha = 1$ the outliers have a permanent effect and the v_t component will consist of the sum of all past outlier shocks to the process. In this case the series contaminating y_t will behave as a series with structural level shifts that can be both positive and negative.

As can be seen, the design of the model is such that we can control the impact on the various estimators when the frequency, the (relative) magnitude, and the persistence of the outliers changes. Also, the noise resulting from measurement errors can be controlled via the variance (inverse) signal-to-noise ratio $(\sigma_\eta/\sigma_\varepsilon)^2$.

In fact, our model[1] can cover a vast range of different contamination problems which potentially will have different effects on the degree of long memory in the *observed* time series z_t. Looking at some extreme situations will illuminate this. For instance, assuming that the measurement error component is very large, i.e. the (inverse) signal-to-noise ratio $(\sigma_\eta/\sigma_\varepsilon)^2$ is large, then obviously the noise component will have a huge impact on the observed series even though the long memory component y_t will still dominate asymptotically as long as $d > 0$. For a finite stretch of data the series may thus appear to have a lower memory parameter than indicated by d. The question is of course how a potential bias will depend upon the signal-to-noise ratio for a given sample size and for a range of different estimation methods. A similar situation is expected when e.g. $\alpha = 0$, the parameter θ is relatively large, and/or the probability of outlier occurences is large. If instead, we allow the parameter α tending to one in the limit a rather different situation will occur. In this case there will be frequent level shifts in the noise component and hence in the observed series. If the parameter d is relatively small we will thus expect an upward bias in the estimated d due to the persistence of the jump process. Bos *et al.* (1999) already report some results in support of spuriously finding fractional integration in the presence of time series with level shifts. It is also worth mentioning that in many cases a non-linear component of a process can be arbitrarily well approximated by level shift and jump processes and hence the question we want to ask relates to the topic of how structural level breaks and non-linearities may potentially affect the behaviour of estimators of fractionally integrated processes.

In Figures 1 and 2 some simulated series are displayed to give an idea of the processes we have in mind and to support the above intuition. Figure 1 compares y_t and z_t for $d = 0.45$, $\theta = 0$, and a signal-to-noise ratio $(\sigma_\eta/\sigma_\varepsilon)^2 = 3$ for a sample of 200 obsevations. Obviously, the levels of the latent and the observed processes move together but clearly the high frequency element of the observed process plays a larger role than in the latent variable. In Figure 2 the adverse result seems apparent. In this case y_t and z_t are graphed for $\theta = 5, p = .05, \alpha = 0.99$, $d = 0.45$, and $(\sigma_\eta/\sigma_\varepsilon)^2 = 0$. As seen, when a (large) discrete level shift appears in the series it is likely to be generated by an outlier which subsequently will die out very slowly. The decay of the outlier will naturally depend upon the persistence parameter α : the larger α the stronger low frequency dominance of the series.

One way of clarifying the signifcance of data noise is by interpretation of the power spectrum. It is easily shown that the power spectrum of the latent process can be

[1] In a different context, the present model setup for outliers and noise is similar to Franses and Haldrup (1994) and Haldrup, Montanes, and Sanso (2003).

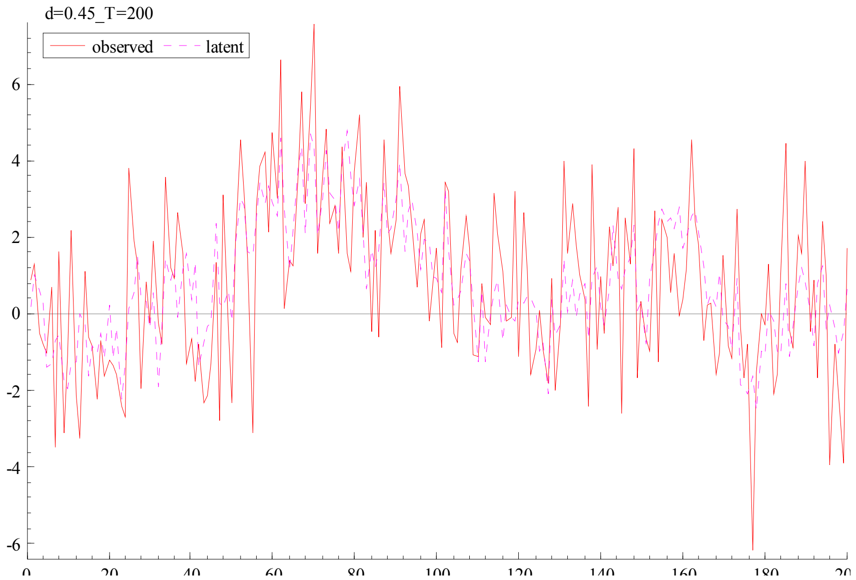

Figure 1: Fractional integration with measurement error: Observed and latent process simulated for $d = 0.45$, $\theta = 0$, and a signal to noise ratio $(\sigma_\eta/\sigma_\varepsilon)^2 = 3$.

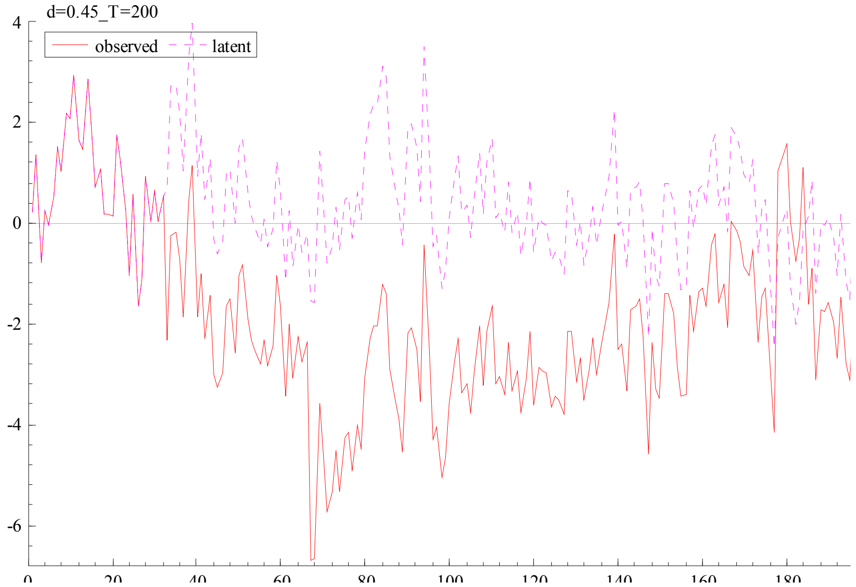

Figure 2: Fractional integration and persistent tempory change outlier: Observed and latent process simulated for $\theta = 5, p = .05, \alpha = 0.99$, $d = 0.45$, and $(\sigma_\eta/\sigma_\varepsilon)^2 = 0$.

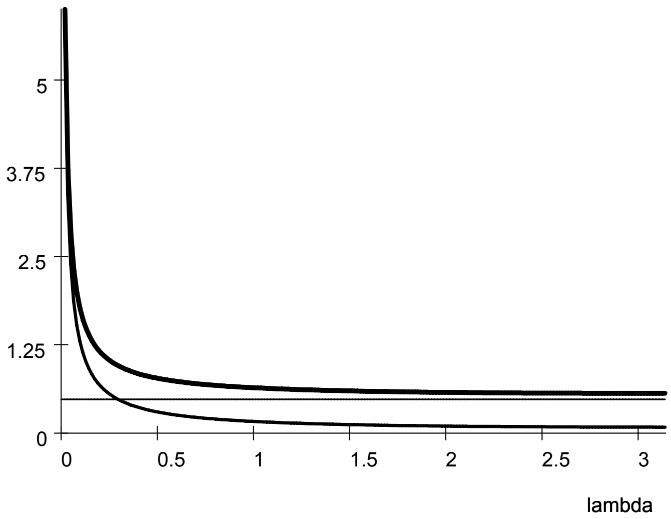

Figure 3: The spectrum $f_z(\lambda) = f_y(\lambda) + f_v(\lambda)$ and its components $f_v(\lambda)$, (thin line), and $f_y(\lambda)$, (medium line), of the process used to generate Figure 1.

written as

$$f_y(\lambda) = \mid 1 - \exp(i\lambda) \mid^{-2d} f_\varepsilon(\lambda),$$

where $f_\varepsilon(\lambda) = \sigma_\varepsilon^2 / 2\pi$ is the spectrum of ε_t whereas the spectrum of the contamination part reads

$$f_v(\lambda) = \frac{\sigma_\eta^2}{2\pi} + \frac{\theta^2 p}{2\pi} \left(1 + \alpha^2 - 2\alpha \cos(\lambda)\right)^{-1}.$$

Thus, for the observed series

$$f_z(\lambda) = \mid 1 - \exp(i\lambda) \mid^{-2d} \frac{\sigma_\varepsilon^2}{2\pi} + \frac{\sigma_\eta^2}{2\pi} + \frac{\theta^2 p}{2\pi} \left(1 + \alpha^2 - 2\alpha \cos(\lambda)\right)^{-1}.$$

Around the origin the spectrum can be approximated as

$$f_z(\lambda) = \lambda^{-2d} \frac{\sigma_\varepsilon^2}{2\pi} + \frac{\sigma_\eta^2}{2\pi} + \frac{\theta^2 p}{2\pi(1-\alpha)^2} \text{ for } |\alpha| < 1 \text{ and for } \lambda \to 0 \tag{5}$$

$$f_z(\lambda) = \lambda^{-2d} \frac{\sigma_\varepsilon^2}{2\pi} + \frac{\sigma_\eta^2}{2\pi} + \lambda^{-2} \frac{\theta^2 p}{2\pi} \text{ for } \alpha = 1 \text{ and for } \lambda \to 0. \tag{6}$$

As seen from (5), the high frequency variability will become relatively more dominant as the (inverse) signal-to-noise ratio $\sigma_\eta^2/\sigma_\varepsilon^2$, the magnitude of outliers, θ, and/or

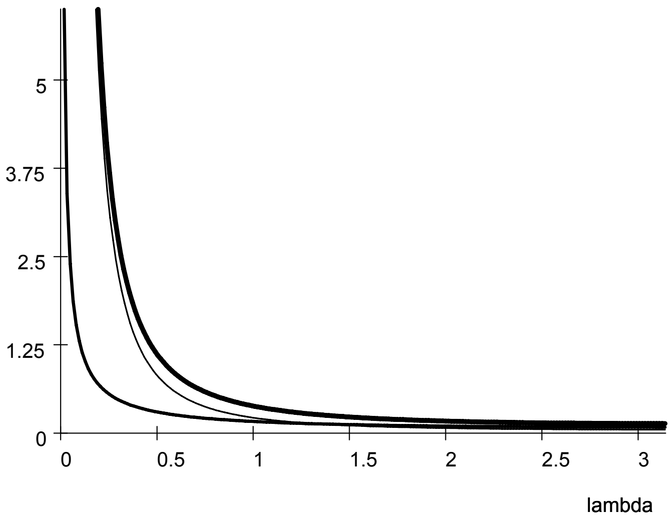

Figure 4: The spectrum $f_z(\lambda) = f_y(\lambda) + f_v(\lambda)$ and its components $f_v(\lambda)$, (thin line), and $f_y(\lambda)$, (medium line), of the process used to generate Figure 2.

the probability of outliers, p, increases. On the other hand, as α tends to unity and thus gives rise to persistent outliers (or structural shifts) the lower end of the frequency band of the power spectrum of the observed series tends to be dominated by the contamination term for $d < 1$ as can be seen from (6). Our intuition is thus, that as long as noise is moderate and temporary the low frequency component is affected little by noise unless, of course, the variance σ_ε^2 is relatively small. If σ_ε^2 is relatively small the low frequency element will tend to reduce the spectral density around the origin relative to the higher frequencies. On the other hand the peakedness of the power spectrum around the origin will tend to increase as α tends to unity and thus tending to bias the integration parameter upwards relative to the non-contaminated series.

In Figures 3 and 4 these features are displayed by plots of the relevant spectra underlying the processes simulated in Figures 1 and 2.

3. Estimators of fractionally integrated processes

We consider four commonly employed estimators for fractionally integrated processes. Two fully parametric maximum likelihood estimators due to Sowell (1992b) and Tanaka (1999) and two semiparametric estimators by Geweke and Porter-Hudak (1983), Kunsch (1987), and Robinson (1995a,b).

The data generating process (DGP) is the univariate ARFIMA(p, d, q) model which is given by (1)-(2) where ε_t is a stationary and invertible ARMA (p, q) process, $\phi(L)\varepsilon_t = \theta(L)\tilde{\varepsilon}_t$, where $\tilde{\varepsilon}_t$ is $iid(0, \sigma^2)$. The parametric estimators are based on the likelihood function of the ARFIMA(p, d, q) model with or without the initialization (2).

First, ignoring the initialization in (2), the definition (1) of fractional integration is valid for $d < 1/2$ and is denoted a type I fractional process by Marinucci and Robinson (1999). This leads to the maximum likelihood objective function

$$L_E\left(d, \phi, \theta, \sigma^2\right) = -\frac{T}{2}\ln|\Omega| - \frac{1}{2}Y'\Omega^{-1}Y,$$

where $Y = (y_1, ..., y_T)'$, ϕ and θ are the parameters of $\phi(L)$ and $\theta(L)$, and Ω is the variance matrix of Y, which is a complicated function of d and the remaining parameters of the model, see Sowell (1992b). We call the estimator $\hat{d}_{EML} = \arg\max L_E$ the exact maximum likelihood (EML) estimator. Even though we generate the data according to the DGP (1)-(2) which is defined for any d, the EML estimator is valid only when $d < 1/2$. Thus, in the Monte Carlo study we employ the EML estimator only when this restriction is satisfied by the DGP under consideration.

Second, imposing the initialization (2), the definition (1)-(2) is valid for any value of d and is a type II fractional process in the terminology of Marinucci and Robinson (1999). The objective function corresponding to this DGP considered by Tanaka (1999) is

$$L_C(d, \phi, \theta) = -\frac{T}{2}\ln\left\{\sum_{t=1}^{T}\left(\frac{\phi(L)}{\theta(L)}(1-L)^d y_t\right)^2\right\}$$

and we call the estimator $\hat{d}_{CML} = \arg\max L_C$ the conditional maximum likelihood (CML) estimator. Maximizing L_C is equivalent to minimizing the usual (conditional) sum of squares and hence this estimator is also referred to as the CSS estimator by some authors.

Both the EML and the CML estimators are $\sqrt{T}$-consistent and asymptotically normal. We shall not give the asymptotic normal distributions here, nor the conditions under which they are derived, but instead refer the reader to Sowell (1992b) and Tanaka (1999). Note also that both estimators are asymptotically efficient in the classical sense when the model is correctly specified.

The semiparametric estimators are based on the power spectrum around the origin, i.e.

$$f(\lambda) = \lambda^{-2d}\frac{\sigma_\varepsilon^2}{2\pi} \text{ for } \lambda \to 0. \tag{7}$$

One of the two commonly used semiparametric estimators is the log-periodogram or Geweke and Porter-Hudak (GPH) estimator introduced by Geweke and Porter-Hudak (1983) and analyzed in detail by Robinson (1995a). Taking logs in (7) and inserting sample quantities we get the approximate regression relationship

$$\ln\left(I\left(\lambda_j\right)\right) = c + -2d\ln\left(\lambda_j\right) + \text{error}, \tag{8}$$

where c is a constant term, $\lambda_j = 2\pi j/T$ are the Fourier frequencies, and the quantity $I\left(\lambda\right) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T}(y_t - \bar{y})e^{it\lambda}\right|^2$ is the periodogram of y_t. The estimator $\hat{d}_{GPH}$ is defined as the OLS estimator in the regression (8) using $j = 1, ..., m$, where $m = m\left(T\right)$ is a bandwidth number which tends to infinity as $T \to \infty$. Note that the estimator is invariant to non-zero means since $j = 0$ is left out of the regression. Under suitable regularity conditions, including y_t being Gaussian and a restriction on the bandwidth, Robinson (1995a) derived the asymptotically normal limit distribution for $\hat{d}_{GPH}$ when $d \in (-1/2, 1/2)$ is in the stationary and invertible range. Recently, Kim and Phillips (1999) demonstrated that, for the model in (1)-(2), the range of consistency is $d \in (-1/2, 1]$ and the range of asymptotic normality is $d \in (-1/2, 3/4)$.

The other semiparametric estimator we consider is the Gaussian semiparametric (GSP) estimator (or local Whittle estimator) which is attractive because of its nice asymptotic properties, the very mild assumptions underlying it, and the likelihood interpretation. The estimator $\hat{d}_{GSP}$ is defined as the maximizer of the (local Whittle likelihood) function

$$Q\left(g, d\right) = -\frac{1}{m}\sum_{j=1}^{m}\left\{\ln\left(g\lambda_j^{-2d}\right) + \frac{\lambda_j^{2d}}{g}I\left(\lambda_j\right)\right\}. \tag{9}$$

Like the GPH estimator, this estimator is invariant to non-zero means since $j = 0$ is absent from the summation. One drawback compared to log-periodogram estimation is that numerical optimization is needed. However, this estimator does not require the Gaussianity condition and Robinson (1995b) showed that $\sqrt{m}(\hat{d}_{GSP} - d) \xrightarrow{d} N(0, 1/4)$. This is an extremely simple asymptotic distribution facilitating easy asymptotic inference. The ranges of consistency and asymptotic normality for the model (1)-(2) have been shown by Phillips and Shimotsu (2003) to be the same as those of the GPH estimator.

Many variants of the GPH and the GSP estimators have appeared in the literature, extending the range of consistency and asymptotic normality or reducing the order of the asymptotic bias. However, we shall not consider those here and generally expect them to behave in a similar manner as the original GPH and GSP estimators.

The drawback for the semiparametric approach is that only $\sqrt{m}$-consistency is achieved in comparison to $\sqrt{T}$-consistency (and efficiency) in the parametric case. Thus, the semiparametric approach is much less efficient than the parametric one since it requires at least $m/T \to 0$. However, the semiparametric estimators are robust to short run dynamics since they use only information from the periodogram ordinates in the vicinity of the origin.

Hence, we expect that in our setup with noisy data, the semiparametric estimators will outperform the parametric ones whenever the noise contaminates the higher frequencies only. Indeed, the semiparametric estimators should be asymptotically unaffected by such noise.

4. Monte Carlo findings

For our simulation study we apply the four common estimators described above. Following the semiparametric approach we apply the GPH and GSP estimators with bandwidths equal to 10 and 20, which are given in parenthesis, e.g. GPH(10) denotes the GPH estimator with bandwidth equal to 10. When the bandwidth is small we expect more robustness to the presence of (temporary) noise at the cost of a higher variability of the estimate. The fully parametric EML and CML estimators are both applied with no short run dynamics in the estimation procedure, denoted by (0,d,0), and with one AR and one MA term estimated, denoted (1,d,1). Thus, the (0,d,0) estimators are correctly specified according to the latent unobserved process whereas the (1,d,1) estimators overfit the latent process. The additional AR and MA terms are expected to pick up some of the contamination from the noise term, v_t, and therefore the (1,d,1) estimators are expected to be less affected by the presence of the noise in the observed series. The EML estimator is only reported for situations with the true d being less than one half because this estimator by construction cannot exceed this value.

The series generated in the simulation study follows the scheme:

$$\begin{aligned} z_t &= y_t + v_t \\ (1-L)^d y_t &= \varepsilon_t \text{ for } t = 1, 2, ...T \\ y_t &= 0 \ \text{ for } t \leq 0 \\ \varepsilon_t &\sim N(0, \sigma_\varepsilon^2 = 1) \end{aligned} \tag{10}$$

for a range of values of d. The design of v_t is described below in the discussion of the separate experiments. For all cases a sample of $T = 100$ observations was used. This is considered to be a typical sample size in many macroeconomic studies, e.g. 25 years of quarterly observations. The Ox programing language, see Doornik (2001), was used in the simulation study with 1,000 replications in each experiment

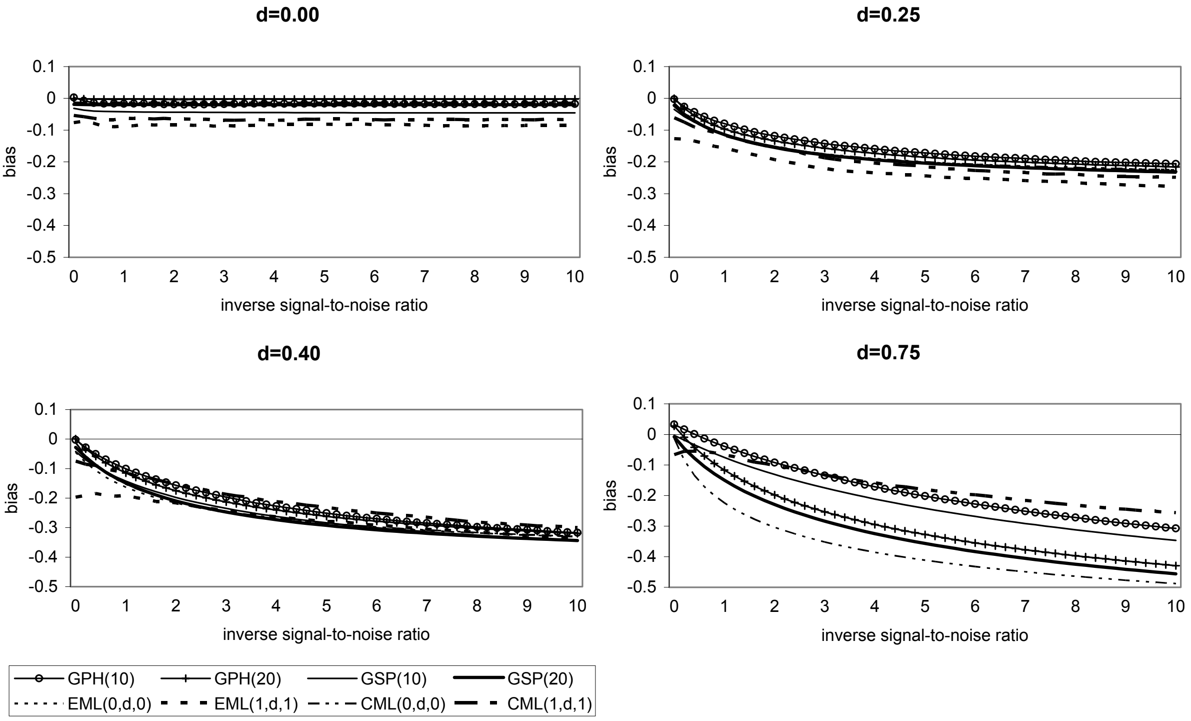

Figure 5: Simple Measurement Errors

4.1. Simple Measurement Errors. For the simple measurement error model we have

$$\begin{aligned} v_t &= \eta_t \\ \eta_t &\sim N(0, \sigma_\eta^2). \end{aligned}$$

Figure 5 displays the biases of the estimators for the (inverse) signal-to-noise ratio $(\sigma_\eta^2/\sigma_\varepsilon^2)$ (*noise ratio* in the sequel) for the latent process y_t with fractional integration orders $d = \{0, 0.25, 0.40, 0.75\}$.

For $d = 0$ the observed process is clearly a sum of two white noise processes. In this case it is seen that biases are relatively minor. For the remaining values of d, biases are generally found to be negative and to increase with the order of integration, and obviously the biases tend to increase for all estimators when the noise ratio increases. For $d = 0.75$ a pattern concerning the relative performance of the single estimators is revealed. The CML(0,d,0) is seen to have the largest biases. For instance, when the noise ratio equals 5 the bias is approximately -0.4 which is

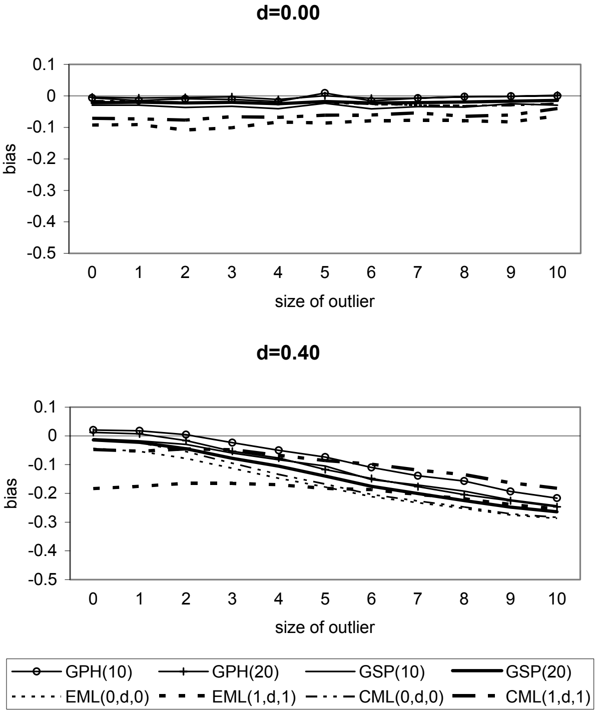

Figure 6: Additive Outliers

huge given the true value of d. However, by permitting the CML estimator to allow for short run dynamics, i.e. by considering the CML(1,d,1) estimator, biases are reduced significantly (around -0.2 in the above example). In fact, this parametric estimator turns out to perform the best amongst the range of estimators considered. Although the semiparametric estimators are generally performing worse than the CML(1,d,1) there is a clear indication that the GPH and GSP estimators with a low bandwidth, e.g. $m = 10$, have relatively lower biases compared to the higher bandwidth estimators.

4.2. Additive Outliers. The noise component of the additive outlier model reads

$$v_t = \theta\delta_t,$$

where the Bernoulli variable δ_t takes the values of plus or minus one with probability 2.5% and a zero value with probability 95%. In Figure 6 biases are displayed as a function of the size of the outliers, θ. Overall, the conclusions are similar to the case of measurement errors: There are hardly any biases for $d = 0$, and for other values

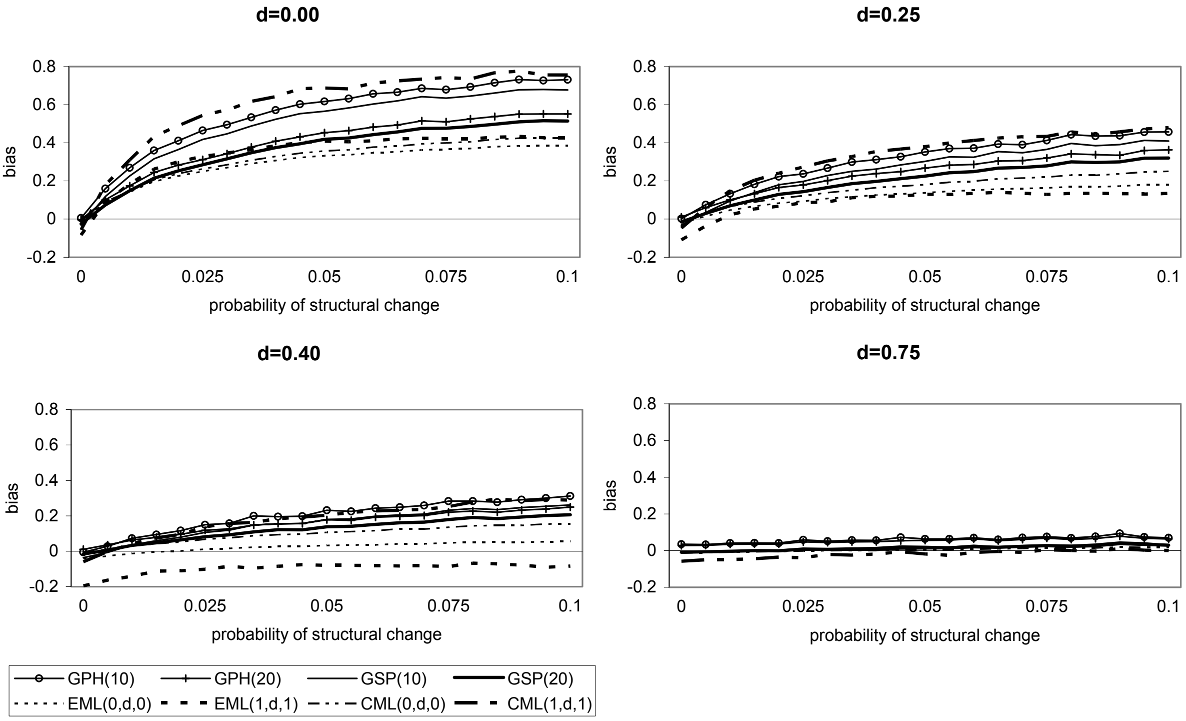

Figure 7: Structural Changes

of d biases tend to be negative and to increase with θ. For a large value of d, e.g. $d = 0.75$, the CML(1,d,1) performs the best and CML(0,d,0) the worst and with the semiparametric estimators with a low bandwidth having relatively lower biases compared to the higher bandwidth ones. Interestingly, for intermediate values of d and for relatively small values of θ, the EML(1,d,1) appears to be rather strongly biased compared to the remaining estimators. For instance, for $d = 0.4$ the EML(1,d,1) estimator has an approximate bias of -0.2 when θ is in the range of 0 to 4. In this range the remaining estimators have hardly any bias.

4.3. Structural Changes. The structural change model of data noise is given by

$$\begin{aligned} v_t &= \frac{\theta}{(1-\alpha L)}\delta_t \\ \alpha &= 1, \end{aligned}$$

where the y_t series is perturbed with a value of $\theta\delta_t = \pm 1$, each with a probability $p/2$, which is displayed for increasing values of p along the horizontal axis of Figure 7. Since $\alpha = 1$ the contamination part has a persistent effect, i.e. $v_t = \sum_{j=1}^{t} \theta\delta_j$ as opposed to the previous cases being scrutinized. In this case the v_t component behaves as a level shift process. When d is large the persistence of the latent process appears to dominate the persistence of the noise component whereby biases of all estimators appear to be negligible. On the other hand, when the degree of long memory is smaller the strong zero frequency contamination of the y_t process will, as expected, induce heavy positive biases of the estimators.

The relative ranking of the estimators with respect to their biases appear to be much similar for various values of d. The largest biases are found for the CML(1,d,1) estimator followed by the bandwidth 10 and 20 semiparametric estimators, respectively. The fact that the bandwidth 20 estimators have smaller biases than the bandwidth 10 estimators follows our intuition given that the bandwidth 10 estimators are more heavily concentrated around the origin where the contamination component has its major dominance. At first sight, the EML estimator is found to be the superior estimator in this case regardless of its short run specification. It is important to notice, however, that because the EML estimator cannot allow estimates of d exceeding 0.5, there will be a limit concerning the observed biases of this estimator for increasing values of d in comparison with other estimators. This feature can give rise to an unfair advantage of the EML estimator.

4.4. Temporary Change Outliers. The contamination component in a model with temporary change outliers is described as

$$v_t = \frac{\theta}{(1-\alpha L)}\delta_t$$

with $\theta\delta_t = \pm 5$, each with probability 2.5%. In Figure 8 biases are displayed for increasing values of the persistence parameter α along the horizontal axis. Note that the extreme value of $\alpha = 0$ corresponds to the case of additive outliers, whereas $\alpha = 1$ corresponds to a level shift (or structural change). For the intermediate values of α the shocks have a temporary effect. The observed biases are thus expected to be described as a convex combination of the situations previously described, i.e. when α is low (high) biases are expected to be similar to those of Figure 6 (Figure 7). Note that the biases for the two extreme cases tend to be of opposite direction. It is thus of interest to analyze which values of the persistence parameter α that tend to create a zero frequency concentration dominating that of the unobserved process.

As visualized from Figure 8, biases generally tend to be negative for small values of α but eventually become positive as α increases. For a large value of the long memory parameter d biases are seen to be rather similar for all estimators and overall the bias

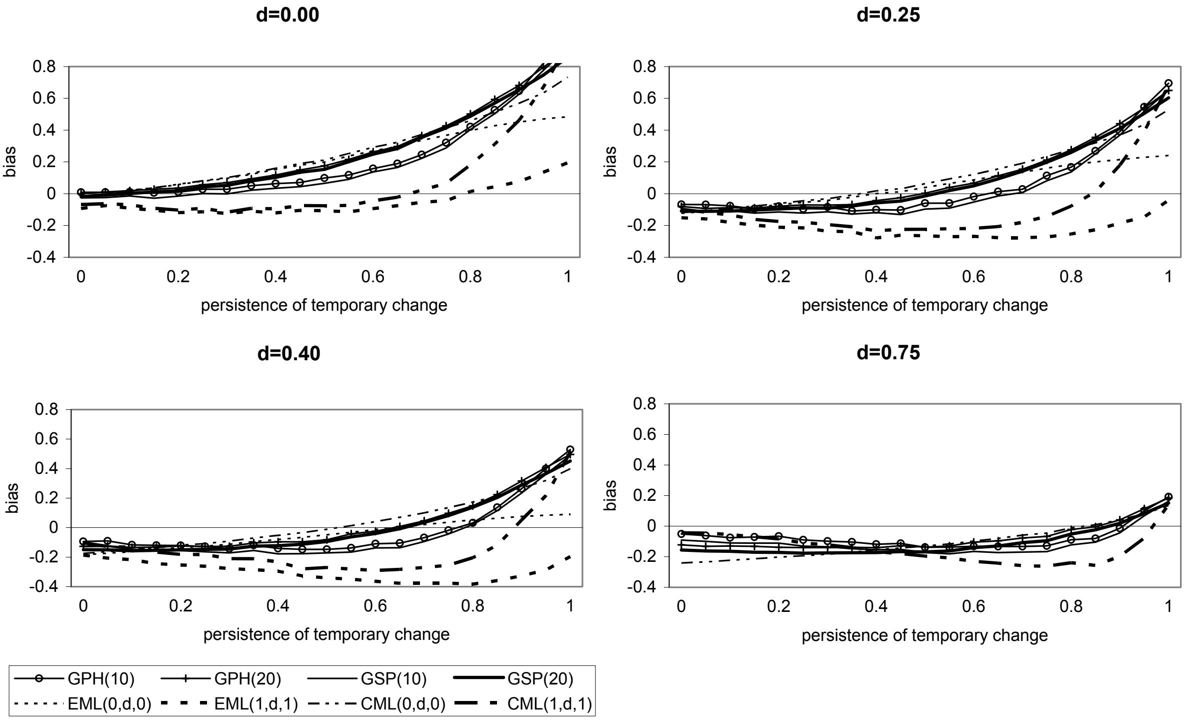

Figure 8: Temporary Change Outliers

tends to be negative, thus indicating that the noise component is dominated by the unobserved series. For small values of d there is a stronger degree of dispersion as the the α parameter increases. The EML(1,d,1) will generally be negatively biased for all values of α which is in contrast to the remaining estimators where a strong positive bias is observed for larger values of α. This reflects the results from the "structural change" case where the EML estimators appeared to perform the best, even though the limitations of these estimators in terms of their area of definition should be recognized.

From Figure 8 it can also be seen that there is no unique ranking of the estimators. The range of biases is very broad, however, and depends a lot on the parameter values of the design.

4.5. Unit Root Case. In Figure 9 we display the biases of the estimators for the case where $d = 1$ in the latent process (10), that is, the y_t series contains a unit root. Each of the previous cases are examined for this special situation. Because the EML estimator does not allow for a value of d exceeding 0.5, we do not consider

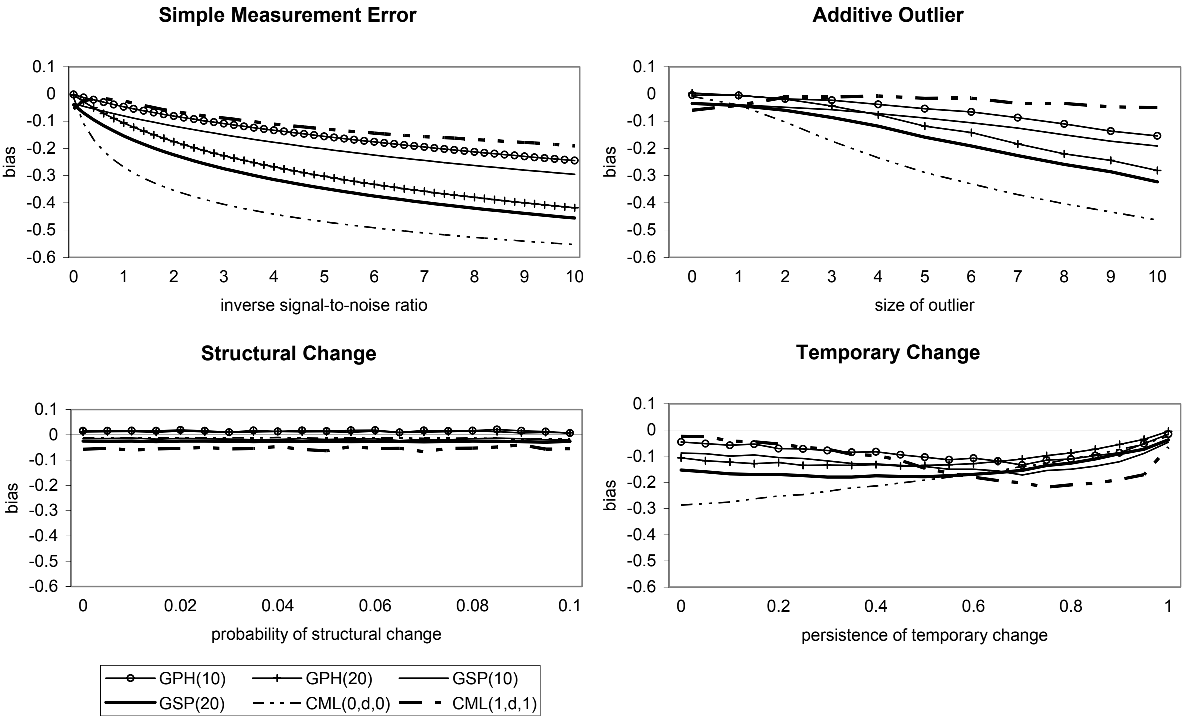

Figure 9: Unit Root Case

this estimator in the present case. The design of each graph corresponds to the models previously described and hence we allow the noise ratio, the size of outliers, the probability of structural shifts, and the persistence of temporary changes to vary according to the model considered.

Both for the case of measurement errors and additive outliers the CML$(1,d,1)$ is the least biased estimator and the CML$(0,d,0)$ is the most biased. In between, the semiparametric estimators can be ranked with the bandwidth 10 estimators having smaller biases than the bandwidth 20 estimators. This is in accordance with the previous findings.

When a unit root is present and the series is contaminated by persistent level shifts, i.e. the lower left graph, there are hardly any biases to be found. This is not surprising given that the contamination component itself resemples a random walk, though not a Gaussian random walk.

For temporary change outliers, i.e. the last graph of Figure 9, no clear ranking of the estimators can be given. Generally, the biases are negative since the contamination component is less persistent than the latent unit root process in this case.

These findings complement the analysis of Bos *et al.* (1999) who argue that G7 inflation rates are estimated as fractionally integrated processes even though the series may be short memory with level shifts. Our results suggest that fractional integration may also be observed if the inflation rate series are unit root processes contaminated by measurement erorrs, additive outliers, or temporary change outliers.

5. Concluding remarks

In this paper we have examined the finite sample performance of a range of common estimators of fractional integration when the potentially fractionally integrated processes are contaminated by data noise. This analysis is important given the wide application of fractional integration estimators to macroeconomic data where significant data contamination is likely to be present. In general, huge biases may occur but the direction of the biases depends upon the type of the noise component and the degree of persistence of both the latent and the noise processes. It appears that the parametric conditional maximum likelihood estimator allowing for short run dynamics performs the best in many situations when noise is not too persistent. For both semiparametric estimators, i.e. the Geweke-Porter-Hudak and the Gaussian semiparametric estimators, the choice of a low bandwidth appeared to be more robust to non-persistent noise. For noise and outliers being relatively more persistent no clear ranking of the estimators is revealed.

Our results suggest that more work needs to be done in this field in the development of estimation methods that are robust to noise and outliers in time series. Some initial work along these lines has already been developed by Sun and Phillips (2003) in a recent paper. Also, our findings call for more research on the development of tests and estimators to discriminate fractionally integrated processes from time series with structural changes and/or non-linearities.

References

[1] Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens, 2001, The distribution of realized stock return volatility, *Journal of Financial Economics* **61**, 43-76.

[2] Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys, 2003, Modelling and forecasting realized volatility, *Econometrica* **71**, 579-625.

[3] Baillie, R. T., 1996, Long memory processes and fractional integration in econometrics, *Journal of Econometrics* **73**, 6-59.

[4] Bos, C., P. H. Franses, and M. Ooms, 1999, Long memory and level shifts: Re-analyzing inflation rates, *Empirical Economics* **24,** 427-449.

[5] Box-Steffensmeir, J. M., R. M. Smith, 1996, The dynamics of aggregate partisanship, *American Political Science Review* **90**, 567-580.

[6] Byers, D., J. Davidson, and D. Peel, 1997, Modelling political popularity: An analysis of long range dependence in opinion poll series, *Journal of the Royal Statistial Society Series A* **160**, 471-490.

[7] Chen, C., and L. Liu, 1993, Joint estimation of model parameters and outlier effects in time series, *Journal of the American Statistical Association* **88,** 284-297.

[8] Chong, T. T., and G. C. Lui, 1999, Estimating the fractionally integrated process in the presence of measurement error, *Economics Letters* **63**, 285-294.

[9] Crato, N., and P. Rothman, 1994, Fractional integration analysis of long-run behavior for US macroeconomic time series, *Economics Letters* **45**, 287-291.

[10] Diebold, F. X., and A. Inoue, 2001, Long memory and regime switching, *Journal of Econometrics* **105**, 131-159.

[11] Diebold, F. X., and G. Rudebusch, 1989, Long memory and persistence in aggregate output, *Journal of Monetary Economics* **24**, 189-209.

[12] Dolado, J. J., J. Gonzalo, and L. Mayoral, 2003, Long-range dependence in Spanish political opinion poll series, *Journal of Applied Econometrics* **18**, 137-155.

[13] Doornik, J. A., 2001, *Ox: An object-oriented matrix language* (4th edition), London: Timberlake Consultants Press.

[14] Franses, P. H., and N. Haldrup, 1994, The effects of additive outliers on tests for unit roots and cointegration, *Journal of Business and Economic Statistics* **12**, 471-478.

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

[16] Gil-Alana, L. A., and P. M. Robinson, 1997, Testing of unit root and other nonstationary hypotheses in macroeconomic time series, *Journal of Econometrics* **80**, 241-268.

[17] Granger, C. W. J., and Z. Ding, 1996, Varieties of long memory models, *Journal of Econometrics* **73**, 61-77.

[18] Haldrup, N., A. Montanés, and A. Sansó, 2003, Measurement errors and outliers in seasonal unit root testing, Working Paper, Aarhus University.

[19] Hassler, U., and J. Wolters, 1995, Long memory in inflation rates: International evidence, *Journal of Business and Economic Statistics* **13**, 37-45.

[20] Künsch, H. R., 1987, Statistical aspects of self-similar processes, pp. 67-74 of: Prokhorov, Y., and V. V. Sazanov (eds), *Proceedings of the first world congress of the bernoulli society*, Utrecht: VNU Science Press.

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

[22] Maynard, A., and P. C. B. Phillips, 2001, Rethinking an old empirical puzzle: Econometric evidence on the forward discount anomaly, *Journal of Applied Econometrics* **16,** 671-708.

[23] Phillips, P. C. B., and K. Shimotsu, 2003, Local Whittle estimation in nonstationary and unit root cases, Cowles foundation discussion paper 1266.

[24] Robinson, P. M., 1995a, Gaussian semiparametric estimation of long range dependence, *Annals of Statistics* **23**, 1630-1661.

[25] Robinson, P. M., 1995b, Log-periodogram regression of time series with long range dependence, *Annals of Statistics* **23**, 1048-1072.

[26] Sowell, F. B., 1992a, Modeling long run behavior with the fractional ARIMA model, *Journal of Monetary Economics* **29**, 277-302.

[27] Sowell, F. B., 1992b, Maximum likelihood estimation of stationary univariate fractionally integrated time series models, *Journal of Econometrics* **53**, 165-188.

[28] Sun, Y., and P. C. B. Phillips, 2003, Nonlinear log-periodogram regression for perturbed fractional processes, *Journal of Econometrics* **115**, 355-389.

[29] Tanaka, K., 1999, The nonstationary fractional unit root, *Econometric Theory* **15**, 549-582.

Working Paper

2002-16: Morten Ørregaard Nielsen, Local Empirical Spectral Measure of Multivariate Processes with Long Range Dependence.

2002-17: Morten Ørregaard Nielsen, Semiparametric Estimation in Time Series Regression with Long Range Dependence

2002-18: Morten Ørregaard Nielsen, Multivariate Lagrange Multiplier Tests for Fractional Integration.

2002-19: Michael Svarer, Determinants of Divorce in Denmark.

2003-01: Helena Skyt Nielsen, Marianne Simonsen and Metter Verner, Does the Gap in Family-Friendly Policies Drive the Family Gap?

2003-02: Torben M. Andersen, The Macroeconomic Policy Mix in a Monetary Union with Flexible Inflation Targeting.

2003-03: Michael Svarer and Mette Verner, Do Children Stabilize Marriages?

2003-04: René Kirkegaard and Per Baltzer Overgaard, Buy-Out Prices in Online Auctions: Multi-Unit Demand.

2003-05: Peter Skott, Distributional consequences of neutral shocks to economic activity in a model with efficiency wages and over-education.

2003-06: Peter Skott, Fairness as a source of hysteresis in employment and relative wages.

2003-07: Roberto Dell'Anno, Estimating the Shadow Economy in Italy: a Structural Equation approach.

2003-08: Manfred J. Holler and Peter Skott: The Importance of setting the agenda.

2003-09: Niels Haldrup: Empirical analysis of price data in the delineation of the relevant geographical market in competition analysis.

2003-10: Niels Haldrup and Morten Ø. Nielsen: Estimation of Fractional Integration in the Presence of Data Noise.