Empirical Economics (1997) 22:103–116

EMPIRICAL ECONOMICS

Fractional Integration with Drift: Estimation in Small Samples

ANTHONY A. SMITH JR.[1], FALLAW SOWELL AND STANLEY E. ZIN

Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA

Abstract: We examine the finite-sample behavior of estimators of the order of integration in a fractionally integrated time-series model. In particular, we compare exact time-domain likelihood estimation to frequency-domain approximate likelihood estimation. We show that over-differencing is of critical importance for time-domain maximum-likelihood estimation in finite samples. Over-differencing moves the differencing parameter (in the over-differenced model) away from the boundary of the parameter space, while at the same time obviating the need to estimate the drift parameter. The two estimators that we compare are asymptotically equivalent. In small samples, however, the time-domain estimator has smaller mean squared error than the frequency-domain estimator. Although the frequency-domain estimator has larger bias than the time-domain estimator for some regions of the parameter bias, it can also have smaller bias. We use a simulation procedure which exploits the approximate linearity of the bias function to reduce the bias in the time-domain estimator.

JEL Classification System-Numbers: C13, C15, C22, C51

1 Introduction

Consider the stationary Gaussian fractionally integrated model with drift:

$$x_t = \mu + (1 - L)^{-d}\varepsilon_t \ , \tag{1}$$

where L is the lag operator, ε_t are independently and identically distributed as $N(0, \sigma^2)$, and $d < \frac{1}{2}$. Variations of this model[2] have gained popularity with empirical researchers as a way of capturing "long-memory" dynamics (see, among others, Diebold and Rudebusch (1989), Lo (1991), Haubrich and Lo (1991), Shea (1991), Cheung and Lai (1993), Sowell (1992a), and Backus and Zin

[1] We would like to thank Frank Diebold, John Geweke, James MacKinnon, and several anonymous referees for helpful comments. We also would like to thank seminar participants at the 1994 Meetings of the Canadian Econometrics Study Group and, in particular, Russell Davidson, Angelo Melino, Peter Robinson, Peter Schmidt, and Tony Wirjanto, for helpful comments. The authors' affiliations are, respectively: GSIA, Carnegie Mellon University; GSIA, Carnegie Mellon University; and GSIA, Carnegie Mellon University and NBER.

[2] The most common generalization of this model is the straightforward addition of stationary autoregressive and moving average dynamics.

0377-7332/97/103-116 $2.50

(1993)). Since a primary goal of this research is to infer long-run economic behavior from observations measured over a relatively short time interval, the finite-sample properties of estimators of the parameters of the fractional model, especially d, have become an important practical consideration.

Recently a number of papers (see, for example, Cheung and Diebold (1994) and Hauser (1992)) have commented on potential problems associated with the practice of centering sample observations around the sample mean and estimating d by maximum likelihood assuming a zero drift (a procedure we refer to as mean-filtered maximum-likelihood estimation). By the very nature of the long memory, the sample mean converges to μ relatively slowly when $d > 0$. This slow convergence may increase the chance that a large sampling error in a small-sample estimate of the mean biases the estimate of d.

We argue that the basic message in these papers is correct, i.e. estimators that avoid estimation of the mean prior to d are preferable, but that mean filtering is not the only source of bias in maximum-likelihood estimators of d when $d > 0$. In particular, an important source of bias is the behavior of the likelihood function near the boundary of the parameter space. Since the likelihood function is not defined for nonstationary models, i.e. for $d \geq \frac{1}{2}$, the sampling distribution of the maximum-likelihood estimator lacks symmetry for values of d near $\frac{1}{2}$. The resulting skewness in the sampling distribution of the maximum likelihood estimator leads this estimator to be biased in small samples. Since the frequency-domain objective function is well-defined for all values of d in finite samples, it is immune to this source of bias. Nonetheless, the frequency-domain estimator is not necessarily preferable to the exact time-domain maximum-likelihood estimator in finite samples. In fact, the contrary is true. In this paper we show how maximum-likelihood estimation *should* be applied in small samples. We find that, after a simple bias correction, exact time-domain estimation typically yields estimators with less bias and smaller mean-squared error than the best frequency-domain estimator.

The paper is organized as follows. Section 2 presents the time-domain and frequency-domain estimators. Section 3 discusses the role of over-differencing. Section 4 compares the small samples properties of the time-domain and frequency-domain estimators. Section 5 discusses ways to correct for the bias of the time-domain estimator. Section 6 presents an empirical application of our proposed methods. Section 7 concludes.

2 Maximum-Likelihood and Whittle-Likelihood Estimation

The log of the likelihood function for a sample of size T, denoted by the $T \times 1$ vector $X_T = [x_1 x_2 \ldots x_T]'$, generated by the process in equation (1) is:

$$L(\theta; X_T) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2}\log|\Sigma_T(d)|$$

$$-\frac{1}{2\sigma^2}(X_T - \mu J_T)'\Sigma_T(d)^{-1}(X_T - \mu J_T) \ , \tag{2}$$

where $\theta = [\mu \ \sigma^2 \ d]$, J_T is a $T \times 1$ vector of ones and $\Sigma_T(d)$ is a $T \times T$ matrix defined by $\text{Var}(X_T) = \sigma^2\Sigma_T(d)$. In other words, the vector of observations satisfies

$$X_T \sim N(\mu J_T, \sigma^2\Sigma_T(d)) \ . \tag{3}$$

The functional form of $\text{Var}(X_T)$ for a general fractionally integrated ARMA model is given in Sowell (1992b). Equation (2) is the *exact* likelihood function and values, $\hat{\mu}_T$, $\hat{d}_T$ and $\hat{\sigma}_T$, that maximize this function are the maximum-likelihood estimates (MLE).

Since each evaluation of the likelihood function involves the inversion of the $(T \times T)$ matrix $\Sigma_T(d)$, computing the MLE can be costly. As outlined in Sowell (1989), however, the Toeplitz structure of $\Sigma_T(d)$ can be used to alleviate much of this burden. Sowell (1992b) studies the properties of this estimator for the model in equation (1) and its autoregressive and moving average extensions, assuming $\mu = 0$.

To avoid some of the computation associated with exact MLE, Cheung and Diebold (1994) suggest using the Fox and Taqqu (1986) frequency-domain approximation to the likelihood function. Hauser (1992) reports small sample properties of a frequency-domain estimator similar to the one suggested by Cheung and Diebold (1994). Hauser's estimator, which minimizes the "Whittle-likelihood" function, appears to perform better in small samples than the estimator suggested by Fox and Taqqu (1986). The Whittle-likelihood function is defined as

$$L^W(d; X_T) = \sum_{j=1}^{m} \log(f(\lambda_j; d)) + \sum_{j=1}^{m} \frac{I_T(\lambda_j; X_T)}{f(\lambda_j; d)} \tag{4}$$

where $m = (T-1)/2$, $\lambda_j = 2\pi j/T$, $I_T(\lambda_j; X_T)$ is the periodogram of the sample X_T and $\sigma^2 f(\lambda_j; d)$ is the spectral density function[3] of the model in (1). The Whittle-likelihood estimate (WLE) is the value for d that minimizes the function in (4). This estimator is asymptotically equivalent to the maximum-likelihood estimator. Note that the Whittle-likelihood function does not depend on μ. Moreover, computation of the WLE does not require the inversion of a $T \times T$ matrix.

Another reason that the MLE may appear to be more difficult to compute than the WLE is that it requires numerical optimization over three dimensions, μ, σ, and d, rather than one-dimensional optimization over d. The computa-

[3] For the ARMA extension of the model in equation (1), the spectral density depends on the autoregressive and moving-average parameters as well as on d. To evaluate the Whittle likelihood in this case, simply substitute the appropriate spectral-density function into equation (4).

tional burden associated with the MLE, however, can be reduced substantially by concentrating μ and σ out of the likelihood function (see, for example, Brockwell and Davis (1987)). In particular, the maximum likelihood of d can be computed by maximizing with respect to d the concentrated log-likelihood function given below:

$$l(d; X_T) = \frac{1}{T}\log|\Sigma_T(d)| + \log\left[\frac{1}{T}(X_T - \hat{\mu}_T(d)J_T)'\Sigma_T(d)^{-1}(X_T - \hat{\mu}_T(d)J_T)\right], \tag{5}$$

where

$$\hat{\mu}_T(d) = [J_T'\Sigma_T(d)^{-1}J_T]^{-1}J_T'\Sigma_T(d)^{-1}X_T \ . \tag{6}$$

Either the MLE or the WLE of d can be inserted into equation (6) to obtain an asymptotically efficient estimate of the drift.[4]

3 Over-Differencing

In this section we show the critical role played by over-differencing in maximum-likelihood estimation of the fractional model with drift. By over-differencing we mean differencing of the observed data beyond what is necessary to achieve stationarity. The differencing data transformation accomplishes two tasks that are critical to proper application of MLE in finite samples. First, it moves the differencing parameter (in the over-differenced model) away from the boundary of the parameter space. Second, it eliminates the need to estimate the drift before estimating d.

In addition to Whittle-likelihood estimation, over-differencing provides another simple estimator for d that does not require knowledge of μ. Operating on both sides of equation (1) with $(1 - L)$, i.e. first-differencing equation (1), yields a fractionally integrated series, $(1 - L)x_t$, that has no drift. The over-differenced model can be written:

$$\Delta x_t \equiv (1 - L)x_t = (1 - L)^{-\delta}\varepsilon_t \ , \tag{7}$$

where $\delta \equiv d - 1$. Note that, by construction, the drift parameter μ equals 0 in the over-differenced model. Therefore, by working with the first differences $\{\Delta x_t\}$ of the observed series $\{x_t\}$, we can always estimate the fractional parameter δ without regard to the drift parameter: simply add 1 to the estimate of δ in the over-differenced model given by equation (7). In finite samples, differencing

[4] Note that equation (5), which is the generalized least squares (GLS) estimator from a regression of X_T on J_T, can be derived in the usual way from the first-order conditions associated with maximizing the likelihood function (2). A similar equation can be derived to obtain an asymptotically efficient GLS estimate of σ^2.

reduces the number of observations, thereby entailing a loss of information. This loss of information is analogous to dropping the zero-frequency observation in the Whittle likelihood. We explore the implications of this loss of information below.

As we have discussed above, the estimation of the drift in long-memory models has become an area of some controversy. Cheung and Diebold (1994) and Hauser (1992) show by means of Monte Carlo experiments that the bias in the mean-filtered maximum likelihood estimator increases substantially as d approaches $\frac{1}{2}$. They attribute this bias to the slow convergence of the sample mean $\bar{x}_T$ to the drift parameter μ when $0 < d < \frac{1}{2}$.[5] Hauser (1992) concludes that this problem is sufficiently serious that the Whittle-likelihood estimator is always preferable to the MLE.

Although the slow convergence of estimators of μ may explain part of the poor finite-sample performance of MLE, we argue here that there is an alternative reason for the finite-sample bias of MLE when $d > 0$: in particular, the proximity of the true value of d to the boundary of the parameter space (i.e. $\frac{1}{2}$). To assess the impact of the boundary of the parameter space on the sampling distribution of the MLE, we conduct a Monte Carlo study of the MLE of d in model (1) with $d = 0.45$ and $\mu = 0$.[6] We generate 10,000 simulated samples, each consisting of $T = 100$ observations, from the joint distribution given in equation (3). For each of these samples, we compute the MLE of d (with σ^2 concentrated out) and the WLE of d. Since the likelihood function given by equation (5) is not defined for $d \geq \frac{1}{2}$ (and tends to $-\infty$ as d tends to $\frac{1}{2}$), all the MLE's of d are less than $\frac{1}{2}$. Figure 1 plots histograms using the 10,000 MLE and WLE estimates.

Note the high degree of skewness in the MLE's sampling distribution. The downward bias reported in Cheung and Diebold (1994) and Hauser (1992) reflects this skewness. The WLE's sampling distribution is much more symmetric around 0.45: this symmetry helps to explain the WLE's relative lack of bias when $d = 0.45$.

In this Monte Carlo study, it is clear that proximity to the boundary of the parameter space, rather than slow convergence in estimating the drift, underlies the bias in the MLE: by construction, the drift in the over-differenced model is zero and we therefore do not need to estimate it. This finding suggests that before MLE is undertaken in small samples, and especially when d is near $\frac{1}{2}$, the observed series should be differenced beyond what would be necessary to yield stationarity. This is the maximum-likelihood estimation procedure that we compare to the Whittle-likelihood estimator in the next section.

---

[5] Note that the concentrated log-likelihood function given in equation (5) depends on the GLS estimate $\hat{\mu}_T(d)$ of μ. For $d \in (-\frac{1}{2}, \frac{1}{2})$ it can be shown that $\hat{\mu}_T(d)$ and $\bar{x}_T$ converge at the same rate to μ. Thus concentrated maximum likelihood does not circumvent any potential problems introduced by the need to estimate μ.

[6] Without loss of generality, σ^2 can be normalized to 1.

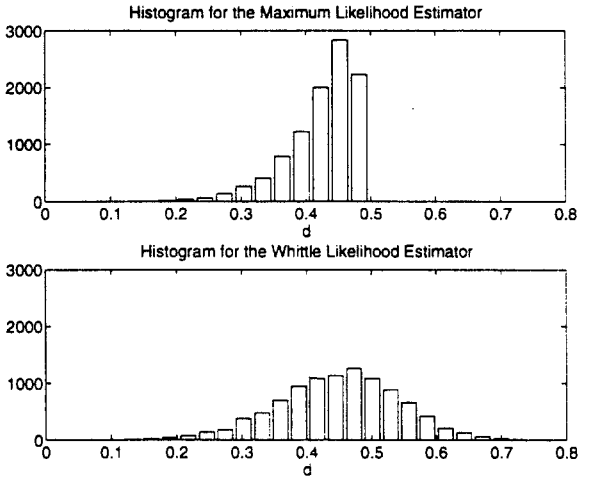

Fig. 1. Sampling distributions for MLE and WLE: $T = 100$, $d = 0.45$

4 Finite-Sample Properties of MLE and WLE

Sowell (1992b) studies the small-sample properties of the MLE of d when $\mu = 0$. Since, as we showed in Section 3, over-differencing can be used to eliminate the drift parameter, these results are also applicable in models with unknown drift. Since we advocate over-differencing as a way of eliminating finite-sample bias in MLE, we need to explore the properties of MLE for smaller values of d in equation (1) than are covered in Sowell (1992b). We likewise extend the Hauser (1992) Monte Carlo results for WLE to a larger range of values for the fractional-differencing parameter.

Figures 2–5 summarize Monte Carlo results for three different estimators of the fractional differencing parameter d in the model given by equation (1), with $\mu = 0$. The three estimators are (concentrated) MLE, WLE, and bias-corrected MLE, which we discuss in Section 5.

Figures 2 and 4 graph estimates of bias and mean squared error (MSE), respectively, of the three estimators of d as a function of the true value of d when the sample size $T = 50$. Figures 3 and 5 display corresponding graphs for the sample size $T = 100$.[7] Estimates of bias and MSE are based on 1,000 independent replications (with different seeds for the random number generator for each

[7] Existing results in Cheung and Diebold (1994) and Hauser (1992) suggest that MLE and WLE perform equally well in samples of size 200 or larger.

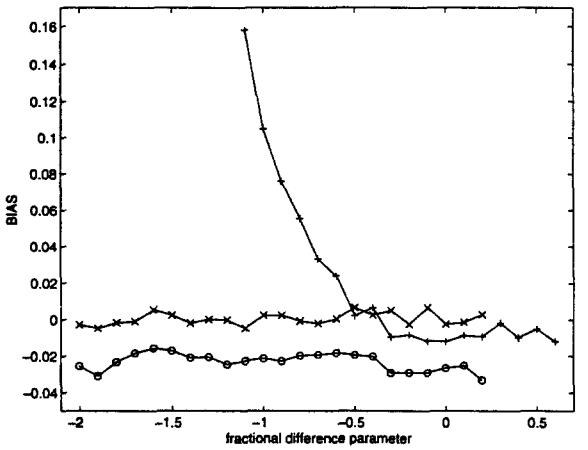

Fig. 2. Bias for $T = 50$ (MLE are o's, WLE are +'s, bias-corrected MLE are ×'s)

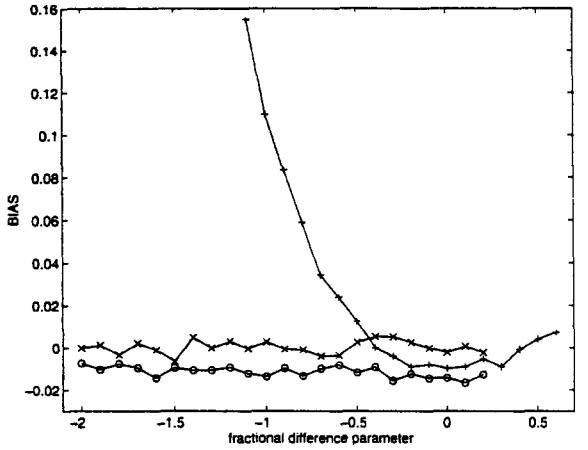

Fig. 3. Bias for $T = 100$ (MLE are o's, WLE are +'s, bias-corrected MLE are ×'s)

value of d). To take into account the loss of information entailed by first differencing, the Monte Carlo results for WLE are based on samples of size $T + 1$ while the Monte Carlo results for MLE are based on samples of size T. For MLE and bias-corrected MLE, we compute bias and MSE for 23 equally spaced values of d in the interval $[-2, 0.2]$. For WLE, we compute bias and MSE for 18 equally spaced values of d in the range $[-1.1, 0.6]$.[8]

[8] To generate simulated samples for values of $d \geq \frac{1}{2}$, we use the last T partial sums of 400 observations from a model with fractional differencing parameter equal to $d - 1$.

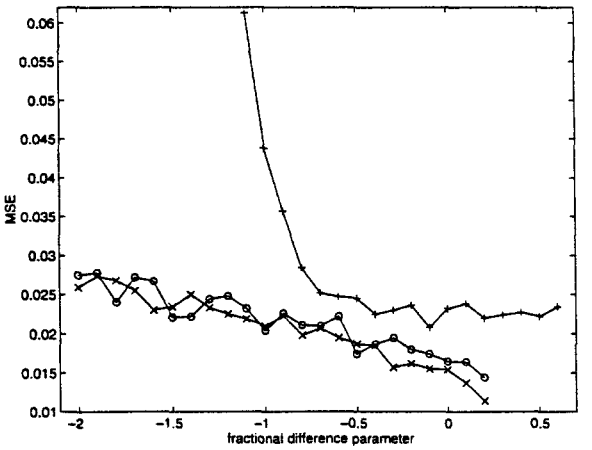

Fig. 4. MSE for $T = 50$ (MLE are o's, WLE are +'s, bias-corrected MLE are ×'s)

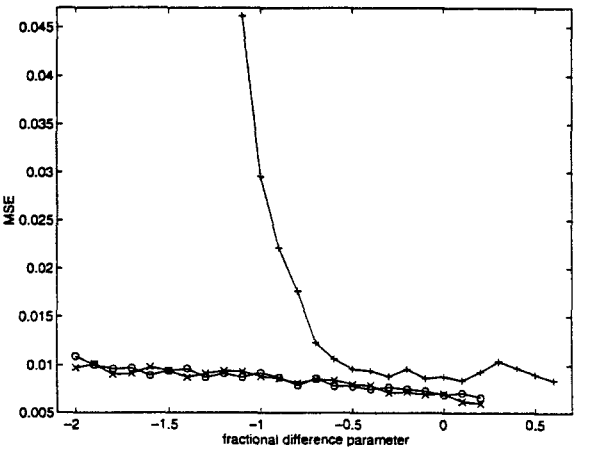

Fig. 5. MSE for $T = 100$ (MLE are o's, WLE are +'s, bias-corrected MLE are ×'s)

In the Monte Carlo experiments for MLE, d ranges from -2 to 0.2. As discussed in Section 3, for values of $d \in (0.2, 0.5)$ the sampling distribution of the MLE becomes skewed, leading to large negative biases of the MLE and therefore to poor small-sample performance. The range $[-2, 0.2]$ is sufficiently large to encompass estimation (after differencing) of the time-trend model in Sowell (1992a) and Hauser (1992) (see the discussion at the end of this section). The WLE results range from -1.1 to 0.6 to cover a comparable range for the fractional differencing parameter in an *undifferenced model.*

The most notable feature of the Figures 2–5 is the poor small-sample performance of the WLE when $d < -\frac{1}{2}$. Bias and MSE increase substantially as d falls.

By contrast, the bias of the MLE is nearly constant for $d \in [-2, 0.2]$ while the MSE of the MLE increases moderately as d falls. For $d \geq 0.2$, both MLE and WLE are downward biased, but the bias for WLE is less than the bias for MLE. These results are consistent with those in Cheung and Diebold (1994) and Hauser (1992). Nonetheless, for all values of d in the range $[-2, 0.2]$, MSE for MLE is less than MSE for WLE, although the differences for some values of d are not significantly different from zero (recall that we use only 1,000 replications to generate the Monte Carlo results).

As we argue in Section 3. for $d \in (0.2, \frac{1}{2})$ we recommend differencing prior to estimation by MLE. To compare WLE in an undifferenced model to MLE in a differenced model, one must compare bias and MSE at two different values of d in Figures 2–5. For example, if the true value of d is 0.4, one must compare the performance of WLE at d to the performance of MLE at $d - 1 = -0.6$. It is clear from the figures that MLE (in an over-differenced model) continues to perform well relative to WLE (in an undifferenced model) when $d \in (0.2, \frac{1}{2})$. For example, when $T = 50$ and $d = 0.4$, MSE for WLE is 0.0227, whereas MSE for MLE at $d = -0.6$ is 0.0222. These results show that the loss of information entailed by first differencing is no more severe than the loss of information entailed by dropping the zero frequency when using WLE to estimate d.

The dramatic increase in the bias of WLE when $d < -\frac{1}{2}$ has important implications for the use of fractionally-intergrated models to test between trend-stationarity and difference-stationarity. Consider the deterministic-trend model: $y_t = \mu t + \varepsilon_t$, where $\varepsilon_t \sim iidN(0, \sigma^2)$. Lagging this equation one period and subtracting the result from the original equation, one obtains the model: $\Delta y_t = \mu + (1 - L)\varepsilon_t$. Next consider the difference-stationary model: $y_t = \mu + y_{t-1} + \varepsilon_t$, which can be rewritten as $\Delta y_t = \mu + \varepsilon_t$. Both of these alternative models can be nested within the following fractionally-integrated model:

$$\Delta y_t = \mu + (1 - L)^{-d}\varepsilon_t \,. \tag{8}$$

Note that the right hand side of this equation is identical to the right hand side of equation (1). When $d = -1$, equation (8) reduces to the trend-stationary model with a deterministic linear time trend; when $d = 0$, equation (8) reduces to the difference stationary model. To test for trend-stationarity, one must therefore test the null hypothesis that $d = -1$. As Figures 2 and 4 show, WLE produces severely biased estimates of d when $d < -\frac{1}{2}$. Since this bias is positive, WLE tends to favor the difference-stationary model over the trend-stationary model. Since MLE displays much smaller bias than WLE when $d < -\frac{1}{2}$, MLE is clearly preferred to WLE when using the nesting model (8) to distinguish between trend-stationarity and difference-stationarity.[9]

[9] For a variety of macroeconomic time series, d lies between 0 and -1 in the model given by equation (8). In this case, therefore, it is not necessary to over-difference in order to move d away from the boundary of the parameter space.

5 Bias Correction

Although we find in Section 4 that MLE tends to have smaller MSE than WLE for $d \in [-2, 0.2]$, we also find that MLE tends to have larger bias than WLE when $d \geq -\frac{1}{2}$. In this section. we propose methods to correct the bias of the MLE.

Figures 2 and 4 show that the bias of the maximum-likelihood estimator of d is nearly constant over the range $[-2, 0.2]$. In particular, the average bias of MLE over the range $[-2, 0.2]$ is -0.0226 for $T = 50$ and -0.0115 for $T = 100$. This finding suggests a simple procedure for correcting the bias of the MLE when $d \in [-2, 0.2]$: add a constant to the MLE. For samples of size $T = 100$ one would add 0.0115 and for samples of size $T = 50$ one would add 0.0226. This adjustment does not affect the standard deviation of the estimator but does generally lower the MSE by reducing the bias. For other sample sizes, one would have to compute these quantities through comparable simulations.

An obvious extension of this procedure allows for bias correction to vary not only with sample size but also with the value of d. Let $\hat{\theta}_T$ be a consistent estimate of a vector of parameters in a general finite-dimensional model. Define $b_T(\theta) \equiv E(\hat{\theta}_T|\theta) - \theta$, where θ is the population "true" parameter vector. $b_T(\theta)$ is the *bias function*: it maps the population parameters θ into the bias of the estimator of θ in a sample of size T. In general, $b_T(\theta) \neq 0$, so that $\hat{\theta}_T$ is a biased estimator of θ. If $b_T(\theta)$ is linear in θ, however, this bias can be eliminated. Define $h_T(\theta) \equiv \theta + b_T(\theta)$. If b_T is linear, then the inverse of h_T evaluated at $\hat{\theta}_T$ is an unbiased estimator of θ.[10] To see this, note that linearity of b_T implies linearity of h_T. Thus $E(h_T^{-1}(\hat{\theta}_T)|\theta) = h_T^{-1}(E(\hat{\theta}_T|\theta)) = h_T^{-1}(h_T(\theta)) = \theta$. This fact motivates the following bias-corrected estimator:

$$\tilde{\theta}_T \equiv h_T^{-1}(\hat{\theta}_T) \ . \tag{9}$$

When b_T is nonlinear, $\tilde{\theta}_T$ is, in general, biased. In many circumstances, however, bias correction can still do a very good job of reducing bias (see, for example, MacKinnon and Smith (1995)).

In general, h_T is unknown. To make this estimator operational, we use simulation to estimate this function and a simple iterative scheme to compute its inverse. To estimate $h_T(\theta)$, we proceed as follows. First, given the value of θ, generate n i.i.d. simulated samples from the given model, each with T observations. Next, for each simulated sample, estimate θ; let $\hat{\theta}_T^{(i)}$ be the ith such estimate. Finally, define $\hat{h}_{T,n}(\theta) \equiv n^{-1} \sum_{i=1}^{n} \hat{\theta}_T^{(i)}(\theta)$. $\hat{h}_{T,n}$ is a consistent (in n) estimate of h_T.

---

[10] Andrews (1993) proposes a similar procedure to obtain median-unbiased estimates for autoregressive/unit root models. However, we are concerned with obtaining mean-unbiased estimates. MacKinnon and Smith (1995) consider this mean-unbiased estimator in other models.

Using $\hat{h}_{T,n}$ in place of h_T, $\tilde{\theta}_T$ can then be computed iteratively as follows:

$$\tilde{\theta}_T^{(j+1)} = s\tilde{\theta}_T^{(j)} + (1-s)(\hat{\theta}_T + \tilde{\theta}_T^{(j)} - \hat{h}_{T,n}(\tilde{\theta}_T^{(j)})) \ , \tag{10}$$

where $s \in [0, 1)$, $\tilde{\theta}_T^{(j)}$ is the bias-corrected estimate at the beginning of the jth iteration, and $\tilde{\theta}_T^{(1)} = \hat{\theta}_T$. Note that if $\tilde{\theta}_T^{(j+1)} = \tilde{\theta}_T^{(j)} = \tilde{\theta}_T^{(\infty)}$, then equation (10) reduces to: $\hat{\theta}_T = h_{T,n}(\tilde{\theta}_T^{(\infty)})$. Thus, up to the approximation error in $h_{T,n}$ (which we can make arbitrarily small by increasing n), $\tilde{\theta}_T^{(\infty)}$ is the bias-corrected estimate defined by equation (9). In practice, we stop the iterations when the difference between successive estimates is smaller than a given tolerance.

Figures 2 and 3 plot estimates of the bias functions for the MLE of d in samples of size 50 and 100. These figures show that the bias functions are very close to linear over the range $[-2, 0.2]$: in fact, as discussed above, they are very nearly constant. Thus one would expect bias correction to reduce the bias of the MLE almost to zero.

Figures 2–5 also report estimates of the bias and MSE of the bias-corrected[11] maximum likelihood estimates.[12] As expected, bias correction is very successful at reducing the bias, typically by a factor of 10 or more. Mean squared error also tends to fall as a result of the reduction in bias. In principle, however, MSE can be larger for the bias-corrected estimator estimator than for the original MLE. To see this, consider the case where $h_T(d)$ is linear. In this case, the bias-corrected estimator is related to the MLE by $\tilde{d}_T = \frac{1}{m}(\hat{d}_T - b)$, where m is the slope of $h_T(d)$ and b is the intercept. The relationship between the variance of $\tilde{d}_T$ and the variance of $\hat{d}_T$, therefore, depends on the value of m:

$$\mathrm{Var}(\tilde{d}_T) = \frac{\mathrm{Var}(\hat{d}_T)}{m^2} \ .$$

When $m > 1$, the bias-corrected estimator has smaller variance than the MLE, so that MSE unambiguously falls. When $m < 1$, the bias-corrected estimator has larger variance than the MLE: this increase in variance can offset the reduction in bias, leading possibly to an increase in MSE. For the present model with $d \in [-2, 0.2]$, however, m is only slightly smaller than 1. Thus the variances of the MLE and of the bias-corrected MLE are nearly identical, implying that the bias-corrected MLE tends to have smaller MSE than the MLE.

---

[11] Given the approximate linearity of the bias function for the MLE, we adopt a method for finding the approximate inverse of h_T that is faster than the general iterative algorithm described above. In particular, we compute the bias of the MLE using 10,000 simulated samples on a fine grid of values for d in the interval $[-2, 0.1]$. We then fit a line through these points using ordinary least squares and use the inverse of the fitted line to calculate the bias-corrected MLE's. For the empirical application in Section 6, we use the iterative algorithm summarized in equation (10).

[12] Note that in principle we could also apply a similar bias correction procedure to the WLE. However, the nonlinearity of the bias function for WLE suggests that bias correction would not perform as well for the WLE as it does for the MLE.

6 An Empirical Application

This section applies the bias-correction method described in Section 5 to the estimation of a fractionally integrated model of the natural log of quarterly U.S. real GNP for the time period 1947:1 to 1989:4.[13] The fractionally integrated model which we estimate takes the form of equation (8) with the addition of autoregressive and moving average dynamics for the error term ε_t. Specifically, it is assumed that:

$$(1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p)\varepsilon_t = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\eta_t \qquad (11)$$

where $\eta_t \sim iidN(0, \sigma^2)$ and L is the lag operator. Equations (8) and (11) define a fractional ARIMA(p, d, q) model.

Sowell (1992a) argues that a fractional ARIMA(3, d, 2) model provides a good fit to the behavior of log real GNP. This model has eight unknown parameters to be estimated: the fractional differencing parameter d, three autoregressive parameters (ϕ_1, ϕ_2, and ϕ_3), two moving average parameters (θ_1 and θ_2), the drift μ, and the innovation variance σ^2. The model is estimated using exact time-domain maximum likelihood, with μ and σ^2 concentrated out of the likelihood function as described in Section 2. Sowell (1992b) shows how to compute Var(Y_T), where $Y_T \equiv [\Delta y_1 \Delta y_2 \ldots \Delta y_T]'$, for a fractional ARIMA(p, d, q) model. For scaling purposes, each observation Δy_t is divided by the sample standard deviation of Δy_t (i.e. 0.010728).

The first row of Table 1 reports the maximum likelihood point estimates and the second row reports estimated asymptotic standard errors. These estimates differ slightly from the estimates reported in Sowell (1992a) because we use concentrated rather than mean-filtered maximum likelihood to obtain the estimates.

The third row of Table 1 reports the bias-corrected maximum likelihood estimates.[14] These estimates are computed using the iterative algorithm described in Section 5, with $n = 200$.[15] Forty iterations starting from the (uncorrected) maximum likelihood estimates are sufficient to obtain convergence to three decimal places. Our success in computing bias-corrected estimates for a richly parameterized model with eight parameters suggests that the bias correction procedures advocated in Section 5 can be applied in a wide variety of circumstances. The reported standard errors are asymptotically correct for both estimates.

Some of the bias-corrected estimates differ substantially from the original estimates. The bias-corrected estimate of d, for example, moves 25% closer to

---

[13] Since we work with the first differences of log real GNP, the data set consists of 171 quarterly observations.

[14] Note that the standard errors reported in the second row of Table 1 are asymptotically valid both for the uncorrected and for the corrected ML estimates.

[15] We find that when $n = 200$ simulation error is very small relative to the uncertainty in the observed data.

Table 1. Parameter estimates for the fractional ARIMA(3, d, 2) model standard errors reported in parenthesis

	d	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	σ^2
Maximum Likelihood (ML)	−0.61	−1.20	0.94	−0.52	−0.29	0.81	0.78
	(.29)	(.30)	(.25)	(.16)	(.10)	(.12)	(.08)
Bias-Corrected ML	−0.46	−1.03	0.73	−0.46	−0.25	0.76	0.79

zero. More importantly, the bias-corrected estimate of d is closer to $-\frac{1}{2}$ than the original estimate. Recall from the end of Section 4 that $d = -1$ corresponds to a trend-stationary model and $d = 0$ corresponds to a difference-stationary model. The proximity of the bias-corrected estimate of d to $-\frac{1}{2}$ reinforces the finding in Sowell (1992a) that the postwar U.S. time series for real GNP are not informative enough to distinguish between trend-stationarity and difference-stationarity.

7 Final Remarks

We advocate the estimation of fractionally integrated models by first differencing the observed time series and then using time-domain maximum likelihood. This procedure eliminates the boundary problems associated with positive values of d, eliminates the nuisance drift parameter, and avoids the large bias of WLE of $d < -\frac{1}{2}$. The conclusions that we have drawn for the simple fractionally integrated model may not generalize to more highly parameterized models (i.e. models with stationary autoregressive and moving average dynamics). Further Monte Carlo analysis is needed before we can draw any more general conclusions. Nonetheless, in a more complicated model, it is doubtful that the frequency-domain estimator will exhibit less bias than it does in the simple model. The bias correction procedure that we use in this paper works quite well for the maximum-likelihood estimator in the fractional model. This approach appears quite promising and warrants further study in more general settings.

References

Andrews DWK (1993) Exactly median-unbiased estimation of first order autoregressive/unit-root Models. Econometrica 61:139–165

Backus DK, Zin SE (1993) Long-memory inflation uncertainty: Evidence from the term structure of interest rates. Journal of Money, Credit, and Banking 25:681–700
Brockwell PJ, Davis RA (1987) Time series: Theory and methods. Springer-Verlag, New York
Diebold FX, Rudebusch GD (1989) Long memory and persistence in aggregate output. Journal of Monetary Economics 24:189–209
Fox R, Taqqu MS (1986) Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. The Annals of Statistics 14:517–532
Haubrich JG, Lo AW (1989) The sources and nature of long-term memory in the business cycle. manuscript, University of Pennsylvania
Hauser M (1992) Long range dependence in international output series: A reexamination. manuscript, University of Economics and Business Administration, Vienna
Lo A (1989) Long-term memory in stock market prices. Econometrica 59:1279–1313
MacKinnon JG, Smith AA Jr (1995) Approximate bias correction in econometrics. manuscript, Queen's Institute for Economic Research Discussion Paper No. 919
Shea GS (1990) Uncertainty and implied variance bounds in long-memory models of the interest rate term structure. Empirical Economics 16:387–412
Sowell F (1989) A decomposition of block Toeplitz matrices with applications to vector time series. manuscript, Carnegie Mellon University
Sowell F (1990) The fractional unit root distribution. Econometrica 58:495–505
Sowell F (1992a) Modeling long-run behavior with the fractional ARIMA model. Journal of Monetary Economics 29:277–302
Sowell F (1992b) Maximum Likelihood estimation of stationary univariate fractionally integrated time series models. Journal of Econometrics 53:165–188
Yin-Wong C, Diebold F (1994) On Maximum-Likelihood estimation of the differencing parameter of fractionally-integrated noise with unknown mean. Journal of Econometrics 62:301–316
Yin-Wong C, Lai K (1993) A fractional cointegration analysis of purchasing power parity. Journal of Business and Economic Statistics 11:103–112

First version received: March 1995
Final version received: July 1996