ISSN 1440-771X

MONASH University

Australia

Department of Econometrics
and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

The Finite-Sample Properties of Autoregressive Approximations of Fractionally-Integrated and Non-Invertible Processes

S. D. Grose and D. S. Poskitt

July 2006

Working Paper 15/06

The Finite-Sample Properties of Autoregressive Approximations of Fractionally-Integrated and Non-Invertible Processes

S. D. Grose[*] and D. S. Poskitt

Department of Econometrics and Business Statistics, Monash University

Abstract

This paper investigates the empirical properties of autoregressive approximations to two classes of process for which the usual regularity conditions do not apply; namely the non-invertible and fractionally integrated processes considered in Poskitt (2006). In that paper the theoretical consequences of fitting long autoregressions under regularity conditions that allow for these two situations was considered, and convergence rates for the sample autocovariances and autoregressive coefficients established. We now consider the finite-sample properties of alternative estimators of the AR parameters of the approximating $AR(h)$ process and corresponding estimates of the optimal approximating order h. The estimators considered include the Yule-Walker, Least Squares, and Burg estimators.

AMS subject classifications: Primary 62M15: Secondary 62G05, 60G18

Key words and phrases: Autoregression, autoregressive approximation, fractional process, non-invertibility, order selection.

JEL classification: C14, C22, C53

June 23, 2006

[*] Corresponding author: Simone Grose, Department of Econometrics and Business Statistics, Monash University, Victoria 3800, Australia. Email: simone.grose@buseco.monash.edu.au

1 Introduction

The so-called "long memory", or strongly dependent, processes have come to play an important role in time series analysis. Statistical procedures for analyzing such processes have ranged from the likelihood-based methods studied in Fox and Taqqu (1986), Sowell (1992) and Beran (1995), to the non-parametric and semi-parametric techniques advanced by Robinson (1995), among others. These techniques typically focus on obtaining an accurate estimate of the parameter (or parameters) governing the long-term behaviour of the process, and while maximum likelihood is asymptotically efficient in this context, that result as always depends on the correct specification of a parametric model.

An alternate approach looks for an adequate "approximating" model; with a finite-order autoregression being a computationally convenient candidate whose asymptotic properties in the context of certain classes of data generating processes are well-known. However, an exception has, until recently, been the class of processes exhibiting strong dependence, in which case standard asymptotic results no longer apply. Yet it is in these cases that it might be most useful to have recourse to a valid approximating model; either in its own right, or as the basis for subsequent estimation. Indeed, long-order autoregressive models are often used as benchmarks against which the performance of more complex models is measured; see, for instance, Baillie and Chung (2002), Barkoulas and Baum (2003) for a couple of recent examples. Accordingly, Poskitt (2006) considers the statistical consequences of fitting an autoregression to processes exhibiting long memory under regularity conditions that allow for both non-invertible and fractionally integrated processes, providing a theoretical underpinning for the use of finite-order autoregressive approximations in these instances.

We now consider the empirical properties of the AR approximation, particularly the finite-sample properties of alternative estimators of the AR parameters of the approximating $AR(h)$ process and corresponding estimates of the optimal approximating order h.

The paper proceeds as follows. Section 2 summarizes the statistical properties of long memory processes, and their implications for autoregressive approximation. In Section 3 we outline the various autoregressive estimation techniques to be considered. Details of the simulation study are given in Section 4, followed by the results presented in Section 5. Section 6 closes the paper.

2 Autoregressive approximation in non-standard situations

Let $y(t)$ for $t \in \mathcal{Z}$ denote a linearly regular, covariance-stationary process,

$$y(t) = \sum_{j=0}^{\infty} k(j)\varepsilon(t-j) \tag{2.1}$$

where $\varepsilon(t)$, $t \in \mathcal{Z}$, is a zero mean white noise process with variance σ^2 and the impulse response coefficients satisfy the conditions $k(0) = 1$ and $\sum_{j\geq 0} k(j)^2 < \infty$. The innovations $\varepsilon(t)$ are further assumed to conform to a classical martingale difference structure (Assumption 1 of Poskitt, 2006); from which it follows that the minimum mean squared error predictor

(MMSEP) of $y(t)$ is the *linear* predictor

$$\bar{y}(t) = \sum_{j=1}^{\infty} \varphi(j) y(t-j). \tag{2.2}$$

The MMSEP of $y(t)$ based only on the finite past is then

$$\bar{y}_h(t) = \sum_{j=1}^{h} \varphi_h(j) y(t-j) \equiv -\sum_{j=1}^{h} \phi_h(j) y(t-j); \tag{2.3}$$

where the minor reparameterization from φ_h to ϕ_h allows us, on also defining $\phi_h(0) = 1$, to conveniently write the corresponding prediction error as

$$\epsilon_h(t) = \sum_{j=0}^{h} \phi_h(j) y(t-j). \tag{2.4}$$

The finite-order autoregressive coefficients $\phi_h(1), \ldots, \phi_h(h)$ can be deduced from the Yule-Walker equations

$$\sum_{j=0}^{h} \phi_h(j) \gamma(j-k) = \delta_0(k) \sigma_h^2 \,, \quad k = 0, 1, \ldots, h, \tag{2.5}$$

in which $\gamma(\tau) = \gamma(-\tau) = E[y(t)y(t-\tau)]$, $\tau = 0, 1, \ldots$ is the autocovariance function of the process $y(t)$, $\delta_0(k)$ is Kronecker's delta (i.e., $\delta_0(k) = 0 \; \forall \; k \neq 0$; $\delta_0(0) = 1$), and

$$\sigma_h^2 = E\big[\epsilon_h(t)^2\big] \tag{2.6}$$

is the prediction error variance associated with $\bar{y}_h(t)$.

The use of finite-order AR models to approximate an unknown (but suitably regular) process therefore requires that the optimal predictor $\bar{y}_h(t)$ determined from the autoregressive model of order h be a good approximation to the "infinite-order" predictor $\bar{y}(t)$ for sufficiently large h.

However, established results on the estimation of autoregressive models when $h \to \infty$ with the sample size T are generally built on the assumption that the process admits an infinite autoregressive representation with coefficients that tend to zero at an appropriate rate, which is to say (i) the transfer function associated with the Wold representation (2.1) is invertible; and (ii) the coefficients of (2.1), or, equivalently, the autoregressive coefficients in (2.2), satisfy a suitable summability condition. This obviously precludes non-invertible processes, which, failing condition (i), do not even have a infinite-order AR representation, and "persistent", or long-memory, processes, which fail condition (ii). The former would arise if the transfer function $k(z)$ contains a unit root, such as might be induced by over-differencing; the latter, observed in a very wide range of empirical applications, is characterized by an autocovariance structure that decays too slowly to be summable. Specifically, rather than the autocovariance function declining at the exponential rate characteristic of a stable and invertible *ARMA* process, it declines at a hyperbolic rate dependent on a "long memory" parameter $\alpha \in (0,1)$; i.e.,

$$\gamma(\tau) \sim C\tau^{-\alpha}, C \neq 0, \text{ as } \tau \to \infty \,.$$

A detailed description of the properties of such processes can be found in Beran (1994).

Perhaps the most popular model of such a process is the fractionally integrated ($I(d)$) process introduced by Granger and Joyeux (1980) and Hosking (1980). This class of processes can be characterized by the specification

$$y(t) = \frac{\kappa(z)}{(1-z)^d}\,\varepsilon(t)$$

where z is here interpreted as the lag operator ($z^j y(t) = y(t-j)$) and $\kappa(z) = \sum_{j\geq 0} \kappa(j) z^j$. The behaviour of this process naturally depends on the fractional integration parameter d; for instance, if $d \geq 1/2$ the process is no longer stationary, although it may be made so by differencing. More pertinently, the impulse response coefficients of the Wold representation (2.1) characterized by $k(z)$ are now not absolutely summable for any $d > 0$; and the autocovariances decline at the rate $\gamma(\tau) \sim C\tau^{2d-1}$ (i.e., $\alpha = 1 - 2d$). Such processes have been found to exhibit dynamic behaviour very similar to that observed in many empirical time series.

Nonetheless, if the "non-fractional" component $\kappa(z)$ is absolutely summable (i.e., $\kappa(z)$ is the transfer function of a stable, invertible ARMA process) and $|d| < 0.5$, then the coefficients of $k(z)$ are square-summable ($\sum_{j\geq 0} |k(j)|^2 < \infty$), in which case $y(t)$ is well-defined as the limit in mean square of a covariance-stationary process. The model is now essentially a generalization of the classic Box-Jenkins ARIMA model (Box and Jenkins, 1970),

$$(1-z)^d \Phi(z) y(t) = \Theta(z)\varepsilon(t)$$

in that we now allow non-integer values of the integrating parameter d. In this case $y(t)$ satisfies Assumption 2 of Poskitt (2006), the order-h prediction error $\epsilon_h(t)$ converges to $\varepsilon(t)$ in mean-square, the estimated sample-based covariances converge to their population counterparts, though at a slower rate than for a conventionally stationary process, and the Least Squares and Yule-Walker estimators of the coefficients of the approximating autoregression are asymptotically equivalent and consistent. Furthermore, order selection by AIC is asymptotically efficient in the sense of being equivalent to minimizing Shibata's (1980) figure of merit, discussed in more detail in section 3.2.

The non-invertible case is a little different, in that the autoregressive coefficients $\phi(j)$, $j = 1, \ldots$ are determined as the limit of ϕ_h as $h \longrightarrow \infty$. However, $y(t)$ is still linearly regular and covariance stationary, and so the results developed for the ARFIMA model still hold, although the convergence rates given in Poskitt (2006) may be conservative.

3 Model Fitting

We wish to fit an autoregression of order h to a realisation of T observations from an unknown process $y(t)$. For compactness, in this section $y(t)$ will be denoted y_t, $t = 1, \ldots, T$. The model to be estimated is therefore

$$y_t = -\sum_{j=1}^{h} \phi_h(j) y_{t-j} + e_t\,; \tag{3.1}$$

which we may write as $e_t = \Phi_h(z)y_t$, where $\Phi_h(z) = 1+\phi_h(1)z+\cdots+\phi_h(h)z^h$ is the h^{th}-order prediction error filter. We define the "normalized" parameter vector $\boldsymbol{\phi}_h = \begin{pmatrix} 1 & \phi_h \end{pmatrix}$ where $\phi_h = (\phi_h(1), \ldots, \phi_h(h))$.

A variety of techniques have been developed for estimating autoregressions. MATLAB, for instance, offers at least five, including the standards, Yule-Walker and Least Squares, plus two variants of Burg's (1968) algorithm, and a "Forward-Backward" version of least squares. Each of these techniques is reviewed below. Note that in the following $\hat{\phi}_h$ merely indicates an estimator of ϕ_h; in this section it will be clear from the context which estimator is meant.

3.1 Estimation Procedures for Autoregression

3.1.1 Yule-Walker

As already observed, the "true" $AR(h)$ coefficients (i.e., those yielding the minimum mean squared error predictor based on $y_{t-1}, \ldots, y_{t-h}$) correspond to the solution of the Yule-Walker equations (2.5). Rewriting (2.5) in matrix-vector notation yields

$$\boldsymbol{\Gamma}_h \boldsymbol{\phi}_h = \mathbf{v}_h \tag{3.2}$$

where $\boldsymbol{\Gamma}_h$ is the $(h+1)\times(h+1)$ Toeplitz matrix with $(i,j)^{th}$ element equal to $\gamma(i-j)$, $i,j = 0,1,\ldots,h$, which we may for convenience write as $\mathsf{toeplitz}(\gamma(0),\ldots,\gamma(h))$, and $\mathbf{v}_h = (\sigma_h^2, 0, \ldots, 0)$. Removing the "zeroth" case from this system yields

$$\Gamma_h \phi_h = -\gamma_h \tag{3.3}$$

where $\Gamma_h = \mathsf{toeplitz}(\gamma(0),\ldots,\gamma(h-1)))$, and $\gamma_h = (\gamma(1),\ldots,\gamma(h))$.

Yule-Walker estimates of the parameters of (3.1) are obtained by substituting the sample autocorrelation function (ACF) into (3.3) and solving for $\hat{\phi}_h$:

$$\hat{\phi}_h = -R_h^{-1} r_h$$

where $R_h = \mathsf{toeplitz}(r(0),\ldots,r(h-1))$, $r_h = (r(0), r(1), \ldots, r(h))$, $r(k) = \hat{\gamma}(k)/\hat{\gamma}(0)$, and

$$\hat{\gamma}(k) = \frac{1}{T}\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})$$

is the sample autocovariance at lag k. The innovations variance is then estimated as

$$\hat{\sigma}_h^2 = \hat{\gamma}(0) + \sum_{j=1}^{h} \hat{\phi}_h(j)\hat{\gamma}(j).$$

This estimator has the advantage that it can be readily calculated without requiring matrix inversion via Levinson's (1947) recursion[1], and being based on Toeplitz calculations the corresponding filter $\hat{\Phi}_h(z)$ will be stable. However, while the Yule-Walker equations give the minimum mean squared error predictor given the actual ACF of the underlying process, this is not the case when based on sample autocorrelations. Hence the Yule-Walker variance estimate $\hat{\sigma}_h^2$ does not in general minimize the empirical mean squared error.

We also note that the Yule-Walker estimator of ϕ_h is well known to suffer from substantial bias in finite samples, even relative to the Least Squares approach discussed below. Tjøstheim and Paulsen (1983) present theoretical and empirical evidence of this phenomenon and show that when y_t is a finite autoregression then the first term in an asymptotic expansion of the bias of $\hat{\phi}_h$ has order of magnitude $O(T^{-1})$ but the size of the constant varies inversely with the distance of the zeroes of the true autoregressive operator from the unit circle. Hence, when the data generating mechanism shows strong autocorrelation it is possible for the bias in the Yule-Walker coefficient estimates to be substantial. Given that fractional processes can display long-range dependence with autocovariances that decay much slower than exponentially, similar effects are likely to be manifest when using the Yule-Walker method under the current scenario.

3.1.2 Least-Squares

Least Squares is perhaps the most commonly-used estimation technique, with implementations on offer in just about every numerical package and application. In this case (3.1) is fitted by minimizing the sum of squared errors $\sum_{t=h+1}^{T} \hat{e}_t^2$, where $\hat{e}_t = y_t - \hat{y}_t$, and

$$\hat{y}_t = -\sum_{j=1}^{h} \hat{\phi}_h(j) y_{t-j}$$

is the h^{th}-order linear predictor. In other words, the forward prediction error is minimized in the least squares sense. This corresponds to solving the normal equations

$$\mathbf{M}_h \hat{\boldsymbol{\phi}}_h = -\mathbf{m}_h$$

where

$$\mathbf{M}_h = \sum_{t=h+1}^{T} \left[\begin{pmatrix} y_{t-1} \\ \vdots \\ y_{t-h} \end{pmatrix} \begin{pmatrix} y_{t-1} & \ldots & y_{t-h} \end{pmatrix} \right]$$

and

$$\mathbf{m}_h = \sum_{t=h+1}^{T} y_t \begin{pmatrix} y_{t-1} \\ \vdots \\ y_{t-h} \end{pmatrix} .$$

Note that, following standard practice, the LS estimator presented here is based on the

[1] Generally referred to as Durbin-Levinson recursion (see Durbin, 1960). For a summary of the algorithm see Brockwell & Davis, §5.2.

last $T-h$ values of y; i.e., on $y_t,\ t=h+1,\dots,T$, making the effective[2] sample size $T-h$. The least squares estimate of the variance is then

$$\hat{\sigma}_h^2 = (T-h)^{-1}\sum_{t=h+1}^{T}(y_t-\hat{y}_t)^2.$$

By way of contrast with the Yule-Walker estimator, Least Squares minimizes the observed mean squared error but there is no guarantee that the corresponding AR filter $\hat{\Phi}_h(z)$ will be stable.

3.1.3 Least-Squares (Forward-backward)

The conventional least squares approach discussed above obtains $\hat{\phi}_h$ such that the sum of squared forward prediction errors

$$SSE_1 = \sum_{t=h+1}^{T}\left(y_t + \sum_{j=1}^{h}\phi(j)y_{t-j}\right)^2$$

is minimized. However, we can also define a LSE based on the equivalent time-reversed formulation; i.e., we now minimize the sum of squared backward prediction errors,

$$SSE_2 = \sum_{t=1}^{T-h}\left(y_t + \sum_{j=1}^{h}\phi(j)y_{t+j}\right)^2 .$$

The combination of the two yields "forward-backward" least squares (FBLS), sometimes called the modified covariance method, in which $\hat{\phi}_h$ is obtained such that $SSE_1 + SSE_2$ is minimized. The normal equations are now $\mathbf{M}_h\hat{\phi}_h = -\mathbf{m}_h$ with

$$\mathbf{M}_h = \sum_{t=h+1}^{T}\left[\begin{pmatrix} y_{t-1} \\ \vdots \\ y_{t-h}\end{pmatrix}\begin{pmatrix} y_{t-1} & \dots & y_{t-h}\end{pmatrix}\right] + \sum_{t=1}^{T-h}\left[\begin{pmatrix} y_{t+1} \\ \vdots \\ y_{t+h}\end{pmatrix}\begin{pmatrix} y_{t+1} & \dots & y_{t+h}\end{pmatrix}\right]$$

and

$$\mathbf{m}_h = \sum_{t=h+1}^{T}\begin{pmatrix} y_ty_{t-1} \\ \vdots \\ y_ty_{t-h}\end{pmatrix} + \sum_{t=1}^{T-h}\begin{pmatrix} y_ty_{t+1} \\ \vdots \\ y_ty_{t+h}\end{pmatrix} .$$

This may be thought of as "stacking" a time-reversed version of y_t; i.e., y_t for $t = T-h,\dots,1$, on top of y_t, $t=h+1,\dots,T$, and regressing the resulting $2(T-h)$-vector on its first h lags. See Kay (1988, Chpt.7) or Marple (1987, Chpt.8) for further details.

[2] An obvious alternative is to take the range of summation for the least squares estimator as $t=1,\dots,T$, and assume the pre-sample values $y_{1-h},\dots,y_0$ are zero. The effect of the elimination of the initial terms is, for given h, asymptotically negligible, but may well have significant impact in small samples.

3.1.4 Burg's method

The "Burg" estimator for the coefficients of an autoregressive process (Burg, 1967, 1968), while a standard in spectral analysis, is not well known in the econometrics literature. It does however, have several nice features, chief among which is that parameter stability is imposed without the sometimes large biases involved in Yule-Walker estimation. As we shall see, its properties in that regard tend to mimic those of Least Squares; making it something of a "best of both worlds" estimator. The estimator essentially performs a Least Squares optimization with respect to the partial autocorrelation coefficient alone (called the 'reflection coefficient' in the related literature), with the remaining coefficients determined by Levinson recursion (Durbin, 1960; Levinson, 1947). The result is a set of prediction error filter coefficients which solve

$$\mathbf{\Gamma}_h \boldsymbol{\phi}_h = \mathbf{v}_h \tag{3.4}$$

(*cf.* equation (3.2)) where in this case $\mathbf{v}_h = (v_h, 0, \ldots, 0)$ in which v_h is the output 'power' of prediction error filter Φ_h; that is, the mean-squared error of the order-h autoregression.

Burg (1968) outlined a recursive scheme for solving (3.4); later formalized by Andersen (1974)[3]. Essentially, (3.4) is solved via Levinson recursion as per the Yule-Walker procedure, except that the partial autocorrelation coefficient at each stage (m, say) is now obtained by minimizing the average of the forward and backward mean squared prediction errors described in 3.1.3. This is equivalent to obtaining the reflection coefficient as the harmonic mean of the forward and backward partial correlation coefficients, for which reason the Burg algorithm is sometimes referred to as the "harmonic" method.

The so-called "geometric" procedure (or 'geometric Burg' procedure) differs in implementation only in that the m^{th}-order partial autocorrelation coefficient ϕ_{mm} is calculated as

$$\phi_{mm} = \sum_{t=1}^{T-m} b_{mt} b'_{mt} \Bigg/ \sum_{t=1}^{T-m} \sqrt{b^2_{mt} b'^2_{mt}}$$

rather than as

$$\phi_{mm} = \sum_{t=1}^{T-m} b_{mt} b'_{mt} \Bigg/ \sum_{t=1}^{T-m} \frac{1}{2} \left(b^2_{mt} + b'^2_{mt} \right)$$

(notation as per Andersen). This corresponds to obtaining the m^{th}-order PAC by minimizing the geometric rather than harmonic mean of the forward and backward partial correlation coefficients. In either case the PAC produced at each stage is by construction less than unity in absolute magnitude, ensuring a stable AR filter.

3.2 Selecting the optimal AR order

Undoubtedly more important in determining the accuracy or otherwise of any autoregressive approximation than a particular choice of model *fitting* technique, is the choice of h, the order of the approximating model. If we suppose, for a moment, that the order-h prediction error variance, σ^2_h, is known to us (i.e., we know the theoretical ACF), then we might also suppose

[3] Fortran code implementing Andersen's procedure is given in Ulrych and Bishop (1975).

the existence of an "optimal" order for the AR approximation, h^*; where h^* corresponds to that value of h which minimizes a suitably-penalized function[4] of σ_h^2. This value may then be taken as the basis for comparison of the estimation techniques under consideration.

The problem is analogous to model selection in an empirical setting, except that we are choosing between approximating models based on their theoretical properties, rather than between models according to their "fit" to a set of observed data. We therefore consider the "figure of merit" function

$$L_T(h) = (\sigma_h^2 - \sigma^2) + h\sigma^2/T$$

proposed by Shibata (1980) in the context of fitting autoregressive models to a truly infinite-order process. Shibata showed that if an $AR(h)$ model is fitted to a stationary Gaussian process that has an $AR(\infty)$ representation and this model is then used to predict an independent realization of the same process then the difference between the mean squared prediction error of the fitted model ($\hat{\sigma}_h^2$) and the innovation variance (σ^2) converges in probability to $L_T(h)$. Poskitt (2006) showed that this is also true for the non-standard processes considered here.

Accordingly, if we define h_T^*, for a given process and sample size, as the value of h that minimizes $L_T(h)$ over the range $h = 0, 1, \ldots, H_T$, h_T^* is then asymptotically "efficient" in the sense of minimizing the difference between the mean squared prediction error of the fitted model and the innovation variance; and a sequence of selected model orders, say h_T', is likewise asymptotically efficient if $L_T(h_T') \longrightarrow L_T(h_T^*)$ as $T \longrightarrow \infty$.

Returning to the empirical setting, there are of course any number of candidate criteria for model selection, of which perhaps the best known is that due to Akaike (1970), namely

$$AIC(h) = \ln(\hat{\sigma}_h^2) + 2h/T\,,$$

where $\hat{\sigma}_h^2$ is the finite-order (mean squared error) estimator of the innovations variance as produced by the estimation technique under consideration.

AIC is a member of the class of so-called "Information Criteria", based on the maximized log-likelihood, plus a penalty of the form hC_T/T, where $C_T > 0$ is chosen such that $C_T/T \rightarrow 0$ as $T \rightarrow \infty$. Further criteria in this style were subsequently proposed by numerous authors, in particular Schwarz (1978) ($C_T = \log T$) and Hannan and Quin (1979) ($C_T = \log\log T$), whose criteria are known to be consistent in the sense that they will asymptotically correctly identify the true model if it is included in the selection set. In our case, of course, we are looking for an optimal means of choosing the order of an *approximating* model, the true process being infinite-order, so consistency arguments along these lines cannot apply.

AIC, on the other hand, corresponds to setting $C_T = 2$, which in more conventional (i.e., finite order) situations tends to result in over-parameterized models (i.e., AIC is not "consistent" in the sense of Schwarz's BIC). However, AIC *is* asymptotically efficient, in the sense of Shibata (1980) outlined above, under Shibata's original regularity conditions. Poskitt (2006) showed that this is still the case for the long memory and non-invertible

[4] We cannot, of course, minimize σ_h^2 itself, as this is monotonic decreasing in h, and in fact equals σ^2 in the limit as $h \rightarrow \infty$.

processes that are the focus of this paper, subject to a suitable rate of increase in the maximum order H_T.

3.3 Other asymptotically efficient selection criteria

Other methods of autoregressive order determination that do not share the same structure as the information criteria mentioned above have been proposed; these include the criterion autoregressive transfer function suggested by Parzen (1974), the mean squared prediction error criterion of Mallows (1973), and the final prediction error criterion of Akaike (1970). The simplest of these is undoubtedly Akaike's FPE:

$$FPE_T(h) = \left(\frac{T+h}{T-h}\right)\hat{\sigma}_h^2 \; ;$$

being, like the various IC, based only on the finite-order (mean squared error) estimator of the innovations variance as produced by the estimation technique under consideration.

The Parzen and Mallows criteria, on the other hand, essentially compare the magnitude of the MSE corresponding to an autoregression of order h to an "infinite"-order estimator of σ^2; i.e., one that does rely on a (necessarily truncated) approximating process. The usual candidate is the nonparametric estimator for σ^2 constructed by analogy with Kolmogorov's Formula[5] for the one-step ahead mean square prediction error,

$$\sigma^2 = 2\pi \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\{f(\omega)\}\, d\omega\right\} \; ;$$

namely

$$\tilde{\sigma}_\infty^2 = 2\pi \exp\left\{\gamma' + N^{-1}\sum_{j=1}^{N} \ln I_T(\omega_j)\right\} \tag{3.5}$$

where $N = [(T-1)/2]$, $\omega_j = 2\pi j/T$, $\gamma' = 0.57721\ldots$ is Euler's constant, required for consistency, and

$$I_T(\omega) = (2\pi T)^{-1}\left|\sum_{t=1}^{T} e^{-\omega i t} y(t)\right|^2$$

is the periodogram of the T-vector y.

Mallow's (1973) statistic is then calculated as

$$MC_T(h) = \left(\frac{\hat{\sigma}_h^2}{\tilde{\sigma}_\infty^2} - 1\right) + \frac{2h}{T} \; ;$$

while Parzen's (1974) criterion is

$$CAT_T(h) = 1 - \frac{\tilde{\sigma}_\infty^2}{\tilde{\sigma}_h^2} + \frac{h}{T} \, ,$$

where $\tilde{\sigma}_h^2 = \dfrac{T}{T-h}\,\hat{\sigma}_h^2$ is the "unbiased" estimator of the innovation variance σ^2. Further

[5] Szegö (1939), and Kolmorgorov (1941).

criteria in this style have been suggested; for instance, the CAT_2 criterion of Bhansali (1985)

$$CAT_{2,T}(h) = 1 - \frac{\tilde{\sigma}^2_\infty}{\hat{\sigma}^2_h} + \frac{2h}{T}\,,$$

which is based on a penalized comparison of $\hat{\sigma}^2_h$ with $\tilde{\sigma}^2_\infty$, and so very similar in appearance to Mallow's statistic.

Parzen (1977) subsequently suggested an alternative "autoregressive transfer" criterion, not involving $\tilde{\sigma}^2_\infty$,

$$CAT^*_T(h) = \left(T^{-1} \sum_{j=1}^{h} \tilde{\sigma}_j^{-2} \right) - \tilde{\sigma}_h^{-2}\,.$$

All these criteria are asymptotically efficient in the sense of Shibata (1980) (see Bhansali, 1986) and so asymptotically equivalent. Accordingly, in large samples we anticipate that these criteria will move together and will be minimized at the same value of h.

In finite samples, of course, there are likely to be significant differences between the autoregressive order selected by the various criteria, with the final selected order also depending on the estimation technique employed, as we shall see.

4 Simulation Experiment

We will initially focus our attention on the simplest of non-invertible and fractionally integrated processes: in the first instance the first-order moving average process

$$y(t) = \varepsilon(t) - \varepsilon(t-1)\,; \tag{4.1}$$

and in the second, the fractional noise process

$$y(t) = \varepsilon(t)/(1-L)^d\,,\ 0 < d < 0.5. \tag{4.2}$$

In both cases $\varepsilon(t)$ will be taken to be Gaussian white noise with unit variance.

The theoretical ACF's of these processes are well known: for (4.1) we have $\gamma(0) = 2$, $\gamma(1) = -1$, and zero otherwise. For (4.2) the ACF is as given in (for instance) Brockwell and Davis (1991, §13.2), and accordingly very simply computed, for $k > 1$, via the recursion

$$\gamma(k) = \gamma(k-1)\,\frac{k-1+d}{k-d},$$

initialized at $\gamma(0) = \Gamma(1-2d)/\Gamma^2(1-d)$.

Knowledge of the ACF allows both simulation of the process itself and computation of the coefficients of the h-step ahead linear filter via Levinson recursion. As we might expect, for the simple models considered here the coefficient solutions simplify very nicely: for model (4.1) we have:

$$\phi_h(j) = \frac{h+1-j}{h+1}\,,$$

while for the fractional noise process (4.2) the coefficients are given by the recursion

$$\phi_h(j+1) = \phi_h(j)\frac{(j-d)(h-j)}{(j+1)(h-d-j)}\ ,\ j = 0, 1, 2, \ldots$$

We also note that for the moving average model (4.1) the prediction error variance falls out of the recursion as

$$\sigma_h^2 = \sigma^2 \left\{1 + \frac{1}{h+1}\right\}$$

from which we can quickly deduce

$$h_T^* = \sqrt{T} - 1.$$

For the fractional noise models we have

$$\sigma_h^2 = \sigma^2 \frac{\Gamma(h+1)\Gamma(h+1-2d)}{\Gamma^2(h+1-d)},$$

in which case h_T^* is obtained most simply by calculating $L_T(h)$ for $h = 1, 2, \ldots$, stopping when the criterion starts to increase.

The first stage of the simulation experiment is based around comparing the properties of alternative estimators of the parameters of the optimal autoregressive approximation, where "optimal" is here defined in terms of minimizing Shibata's figure of merit, $L_T(h)$.

In the second stage we consider the problem of empirically choosing the order of the approximation, viewed both as a problem in model selection, and in terms of estimating the theoretically optimal order h_T^*. Restricting the selection criteria to be considered to those known to be asymptotically efficient in the infinite order setting, we shall begin with the most obvious choice, the Akaike information criterion, or AIC. Accordingly, having obtained h^* for each model and sample size, we then "estimate" it by finding the value $\hat{h}$ that minimizes

$$AIC(h) = \ln(\hat{\sigma}_h^2) + 2h/T,$$

where $\hat{\sigma}_h^2$ is the mean squared error delivered by each of the five estimation techniques. We will denote this empirically optimal value by $\hat{h}_T^{AIC}$.

4.1 Monte Carlo Design

The simulation experiments presented here are based on a total of five data generating mechanisms: the non–invertible moving average process (4.1), and the fractional noise process (4.2) with $d = 0.125,\ 0.3,\ 0.375$ and 0.45, labelled as follows:

Model	Description
MA1	Non–invertible MA(1) as per (4.1)
FN125	Fractional Noise as per (4.2), with $d = 0.125$
FN30	Fractional Noise as per (4.2), with $d = 0.3$
FN375	Fractional Noise as per (4.2), with $d = 0.375$
FN45	Fractional Noise as per (4.2), with $d = 0.45$

The fractional noise processes listed here are all stationary with increasing degrees of long-range dependence; however, for $d < 0.25$ the distribution of $T^{1/2}(\hat{\gamma}_T(\tau) - \gamma(\tau))$ is asymptotically normal, while for $d \geq 0.25$ the autocovariances are no longer even $\sqrt{T}$-consistent (see Hosking, 1996, for details). Results for the $d = 0.125$ case are therefore expected to differ qualitatively from those for which $d > 0.25$.

For all processes $\varepsilon(t)$ is standard Gaussian white noise ($\sigma^2 = 1.0$). For each process we considered sample sizes T = 50, 100, 200, 500 and 1000. The maximum AR order for the model search phase was set at $H_T = 2\sqrt{T}$, and all results based on $R = 1000$ replications.

The "optimal" autoregressive approximation for each DGP and sample size is obtained by calculating $h_T^* = \text{argmin}_{1,\ldots,H_T} L_T(h)$ as per §3.2, with the parameters (the coefficient vector $\phi_{h^*} = (\phi_{h^*}(1), \ldots, \phi_{h^*}(h^*))$ and the corresponding mean squared prediction error $\sigma^2_{h^*}$ following as outlined above[6].

The empirical distribution of the various statistics of interest (see below) is obtained by using the N realized values for each statistic as the basis for a kernel density estimate of the associated distribution. We use the Gaussian kernel, with bandwidth equal to 75% of the Wand and Jones (1995) over-smoothed bandwidth; i.e.,

$$h = 0.75 \sqrt[5]{\frac{243}{35R}}\, s(X)$$

where $s(X)$ is the empirical standard deviation of the R-element series X.

The experiment is conducted as follows: for each replication $r = 1, 2, \ldots, R$

1. A data vector of length T is generated according to the selected design.
2. The optimal AR order h_T^* is obtained as outlined above, and the parameters of the corresponding AR approximation estimated by each of the five methods described in Section 3.
3. Summary statistics are computed and saved for subsequent computation of their empirical distributions. These include (with h taken to be h_T^* in all cases)
   - the estimated coefficients $\hat{\phi}_h = (\hat{\phi}_h(1), \ldots, \hat{\phi}_h(h))$
   - the estimation error $\hat{\phi}_h(j) - \phi_h(j),\ j = 1, \ldots, h$
   - the squared and absolute estimation error:

     $$\left(\hat{\phi}_h(j) - \phi_h(j)\right)^2 \text{ and } \left|\hat{\phi}_h(j) - \phi_h(j)\right|,\ j = 1, \ldots, h$$

   - and the sum and average of the error in estimating the vector ϕ_h:

     $$\sum_{j=1}^{h}\left(\hat{\phi}_h(j) - \phi_h(j)\right) \text{ and } \frac{1}{h}\sum_{j=1}^{h}\left(\hat{\phi}_h(j) - \phi_h(j)\right).$$

4. The best AR order is estimated empirically for each of the five methods by estimating autoregressions of all orders $h = 1, 2, \ldots, H_T$ and computing the corresponding AIC. $\hat{h}_T^{AIC}$ is taken to be the value of h that yields the smallest AIC in each case.

[6] We omit the subscript T when using h^* in this context.

5. Finally, the behaviour of the various selection criteria discussed in subsection 3.3 is assessed by using the Burg algorithm to estimate autoregressions of orders $h = 1, 2, \ldots, H_T$, computing the corresponding values of the several criteria, and so the set of minimizing orders $\hat{h}_T$.

5 Empirical Distributions

This section discusses the results presented in Appendices A and B. Note that in the tables we use the following shorthand notation: 'MA' indicates the non-invertible moving average (4.1); 'FN' the fractional noise process (4.2). The five estimation techniques (Yule-Walker, Least-Squares, "Harmonic" Burg, "Geometric" Burg, and Forward-Backward least squares) are designated YW, LS, HB, GB, and FB respectively.

5.1 The optimal AR order

The relative frequency of occurrence of the empirical order selected by minimizing the AIC is presented in Table 1, and in Figures 1 and 2.

Table 1 displays the AR order selected by minimizing AIC, $\hat{h}_T^{AIC}$, averaged over $N = 1000$ Monte-Carlo realizations, by estimation method, model, and sample size. Shibata's h^*, and the "theoretical" h_{AIC} are included for comparative purposes.

Figure 1 presents the relative frequency of $\hat{h}_T^{AIC}$ for the Least Squares, Forward-Backward, Yule-Walker, and Burg estimators when $T = 100$; Figure 2 plots the same quantities for $T = 500$. The maximum order is $H_T = 2\sqrt{T}$ in each case. The results for Geometric-Burg are indistinguishable from those for "harmonic" Burg on this scale, and so are omitted for clarity.

It is notable that that the average AIC-selected order is generally quite close to h_T^*, and in all cases much closer to h_T^* than to h_{AIC}. In fact for the moving average model the AIC estimates based on Least Squares are pretty much spot on, with FB being next closest, followed by the Burg estimators, and finally, Yule-Walker.

However, the distribution of $\hat{h}_T^{AIC}$ is highly skewed to the right, with the degree of skewness being greatest for smaller d and least for the non-invertible moving average. The dispersion of $\hat{h}_T^{AIC}$ about h_T^* is correspondingly large, increasing with d, and being greatest for the non-invertible MA. The figures also show that the higher average $\hat{h}_T^{AIC}$ for Least Squares is caused by a greater proportion of large orders being selected, with, for $T = 100$, the distribution of $\hat{h}_T^{AIC}$ for LS not quite falling away to zero by $h = H_T$.

For the fractional noise models $\hat{h}_T^{AIC}$ exceeds h_T^* for all values of d, T, and estimators; and since $\hat{h}_T^{AIC}$ is invariably largest for LS, and smallest for YW, Least Squares is now generally the "worst"-performing estimator in this sense. However, in accordance with the predictions of Poskitt (2006, Section 5), $\hat{h}_T^{AIC}$ for all five estimators approaches h_T^* as T increases, with the differences between the estimators diminishing accordingly. This is reflected in Figures 1 and 2, where we see that as T increases the difference between the distributions of $\hat{h}_T^{AIC}$ for each of the five estimators becomes negligible, and the distributions become more concentrated around h_T^*.

Repeating the experiment for the set of asymptotically efficient criteria discussed in 3.3 with $\hat{\sigma}_h^2$ as produced by the Burg algorithm, we find that there is indeed little to choose between them, even for quite small sample sizes. The behaviour of AIC and FPE is essentially identical, for instance, with a minimum "rate of agreement" of 97% for the fractional noise models, and 95.2% for the non-invertible Moving Average (Table 3). Disagreement between the six criteria was greatest for the non-invertible MA in all cases; and least for fractional noise with small d.

The empirical distributions of the autoregressive order as selected using Akaike's IC, Parzen's CAT, Bhansali's CAT_2 and Mallows' criterion, for fractional noise with $d = 0.3$, are displayed in Figure 3 for sample sizes from 50 to 1000. Akaike's FPE and Parzen's CAT^* are omitted for clarity, there being little visible difference between these and AIC. The distributions are highly skewed, though becoming less so as T increases, and not notably different from each other; although CAT tends to select smaller h than the others, particularly for $T = 50$ and 100, and so has less weight in the long right-hand tail of the distribution. This is borne out by the average order as selected by each of the six criteria presented in Table 2; for the fractional noise models CAT invariably produces the smallest order on average.

With respect to the accuracy with which the six criteria estimate h^*, we find that, at least for small and "moderate" fractional integration, Parzen's CAT results in the smallest average error (measured as the average difference between $\hat{h}$ and h^* in $R = 1000$ Monte-Carlo replications). The picture was much more mixed for the non-invertible MA and fractional noise with $d > 0.3$, with the smallest error shared between CAT, MC, CAT^*, and CAT_2, in that order. AIC trumped the others just once, and FPE not at all. In general, for the fractional noise processes the six criteria tended to select $h > h^*$, particularly for small sample sizes, and $d \leq 0.3$. For the non-invertible moving average all criteria exceeded h^* on average, with the exception of Mallows' criterion, which tended to underfit.

5.2 Autoregressive coefficients

Turning to the empirical distributions of the coefficient estimators themselves, we focus on the estimation error (bias) in the first and last coefficients; i.e., $(\hat{\phi}_h(1) - \phi(1))$ and $(\hat{\phi}_h(h) - \phi(h))$ respectively, the sum of the estimation errors in all h coefficients, $\sum_{j=1}^{h}(\hat{\phi}_h(j) - \phi(j))$ and the corresponding average, $h^{-1}\sum_{j=1}^{h}(\hat{\phi}_h(j) - \phi(j))$. h equals h_T^* in all cases. The density estimates are constructed from the simulated values as outlined in the preceding section; i.e., using a Gaussian kernel and bandwidth $0.75\,\xi\,\sqrt[5]{(243/35R)} \simeq 0.278\xi$ where ξ is the empirical standard deviation of the $R = 1000$ Monte Carlo realizations of the relevant quantity.

Although results are obtained for all five estimators, only the Yule-Walker results are distinguishable from Least Squares in the plots, so only these are presented graphically. Each figure also includes a plot of the Normal distribution with zero mean and variance equal to the observed variance of the quantity being plotted. The estimation error in the first and last coefficients, and averaged over the coefficient vector, for each model, sample size, and estimation technique is presented in Tables 4 – 7.

Beginning with the behaviour of the various estimators of the first and last coefficients (Figures 4 – 7, and Tables 6, 7), we observe that, as we might expect, departures from normality worsen as d increases, with the worst case represented by the non-invertible MA. More notable is that the "degree" of non-normality increases with sample size; the distributions become noticeably less symmetric, with, for the distribution $\hat{\phi}_h(1)$, an interesting "bump" appearing on the left-hand side. Only for the FN45 and MA1 models is there much difference between the estimators, with the Yule-Walker results typically appearing less "normal" than the Least Squares.

The Yule-Walker departures from normality are reflected in the summaries of estimation error and mean squared error presented in the tables. For the fractional noise processes the error in Yule-Walker estimates of the autoregressive coefficients is generally larger than for the other four estimation techniques, particularly for the smaller sample sizes. However, despite the Yule-Walker estimator having a distinctly more "off-center" empirical distribution than the other estimators, its relative performance with respect to average estimation error (bias) tended to *improve* with the degree of fractional integration, being best for $d = 0.45$. Nonetheless, the bias in Yule-Walker estimates of the partial autocorrelation coefficient was almost invariably greater than for the other estimators; with the single exception being for the non-invertible moving average and $T = 50$. In all other instances the average error in the Yule-Walker estimates of the PAC was greater than for all the other estimators, sometimes by an order of magnitude.

For the first coefficient the outcome is not quite so one-sided, with Yule-Walker in fact being *more* accurate than its competitors for $d = 0.375$ and 0.45. The worst case for bias, mean squared error, and general non-normality was undoubtedly the non-invertible moving average, with the worst affected estimator being, unsurprisingly, the Yule-Walker; partly because the accuracy of the Yule-Walker estimator improves only very slowly as T increases from 100 to 1000.

When estimation error and squared error was averaged over the h-vector of coefficients, we find that in every case Yule-Walker is most biased, while its mean squared error is often least. Similarly, Least Squares was the *best* performer with respect to $h^{-1}\sum_{j=1}^{h}(\hat{\phi}_h(j)-\phi(j))$, but the *worst* with respect to $h^{-1}\sum_{j=1}^{h}(\hat{\phi}_h(j)-\phi(j))^2$. Nonetheless, there is very little difference in the relative accuracy of the five estimators; and while Yule-Walker stands out somewhat, the Least Squares, Forward-Backward, Burg and Geometric-Burg are essentially indistinguishable from each other.

Turning to the total coefficient error $\sum_{j=1}^{h}(\hat{\phi}_h(j)-\phi(j))$, comparison of the estimated distributions of this quantity with a normal curve of error with zero mean and variance ξ^2 (figures 8 – 10) indicates that when $d = 0.125$ the distribution is reasonably close to normal for all estimators. When $d > 0.25$, however, the presence of the Rosenblatt process in the limiting behaviour of the underlying statistics (see Hosking, 1996, §3; also Rosenblatt, 1961) is manifest in a marked distortion in the distribution relative to the shape anticipated of a normal random variable, particularly in the right hand tail of the distribution. This distortion is still present when $T = 1000$ and does not disappear asymptotically.

The situation with the moving average process is a little different (Figure 11), firstly in that the marked skew to the right is not evident, and secondly in the degree of difference between Yule-Walker and the other estimators. The Yule-Walker estimator seems to result in a considerable negative bias in the coefficient estimates when summed over the coefficient vector, and this bias becomes worse as T increases.

5.3 Central Limit Theorem

Finally, we consider the "standardized weighted sum of coefficients":

$$\hat{\varphi}_{\lambda,T} = T^{1/2}\frac{\boldsymbol{\lambda}_h'\boldsymbol{\Gamma}_h(\hat{\boldsymbol{\phi}}_h - \boldsymbol{\phi}_h)}{\sqrt{\boldsymbol{\lambda}_h'(\boldsymbol{\Phi}_h\boldsymbol{\Delta}_h\boldsymbol{\Phi}_h')\boldsymbol{\lambda}_h}},$$

where $\boldsymbol{\lambda}_h$ is a "differencing" vector, $\boldsymbol{\Delta}_h$ is the $h\times h$ limiting covariance matrix of $T^{1/2}\{\hat{\gamma}_T(\tau)-\hat{\gamma}_T(0)-(\gamma(\tau)-\gamma(0))\}$, $\tau = 1,\ldots,h$ and $\boldsymbol{\Phi}_h$ is the sum of the lower triangular Toeplitz matrix based on $(1,\phi_h(1),\ldots,\phi_h(h-1))$ and a Hankel matrix based on $(\phi_h(2),\ldots,\phi_h(h),0)$ (see Poskitt, 2006, section 6, for details).

$\hat{\varphi}_{\lambda,T}$ was shown by Poskitt to have a standard normal limiting distribution; a result that follows from the observation by Hosking (1996) that the "non-normal" component of the limiting distribution of the autocovariances of a long-memory process with $d \in [0.25, 0.5)$ can be removed by some form of differencing; for instance, by computing $\hat{\gamma}_T(\tau)-\hat{\gamma}_T(0)$. Applied to the autoregressive coefficients this means that while the limiting distribution of the Least-Squares (or, equivalently, Yule-Walker) estimators is non-normal, this is not the case for a suitably weighted function of the coefficient vector. In accordance with Hosking's findings, the weights are based on centering (or differencing) the toeplitz matrix of autocovariances; for instance, if we define $\boldsymbol{\lambda}_h' = (1,0,\ldots,0,-1)$, then $\boldsymbol{\lambda}_h'\boldsymbol{\Gamma}_h(\hat{\boldsymbol{\phi}}_h - \boldsymbol{\phi}_h) = \sum_{j=1}^{h}(\gamma(j-1) - \gamma(h-j))(\hat{\phi}_h(j)-\phi_h(j))$. The main proviso here is that the elements of $\boldsymbol{\lambda}_h$ must sum to zero, though this need not be true if the centering matrix is included explicitly.

Figure 12 plots the observed distribution of $\hat{\varphi}_{\lambda,T}$ with $\boldsymbol{\lambda}_h' = (1,0,\ldots,0,-1)$, for $T = 1000$, based on $\hat{\boldsymbol{\phi}}_h$ obtained from N realizations of the fractionally-integrated process $y(t) = \varepsilon(t)/(1-z)^d$, with $d = 0.3$ and 0.45, and $h = h_T^*$ (i.e., $h = 9$ and 14 respectively). The estimated empirical distributions are, as before, obtained using the Gaussian kernel and the Wand and Jones bandwidth, and overlayed with a standard normal density. Although some bias is still apparent even at this sample size, more so for the Yule-Walker estimator than for Least Squares, the skewness and kurtosis of the type observed previously with this process has now gone. Figure 13 plots the same quantities for $d = 0.375$ and varying T; comparing this with panel (c) of figures 8 to 10 shows that the sample need not be especially large for the operation of Poskitt (2006, Theorem 6.1) to become apparent.

6 Conclusion

While the Least Squares and Yule-Walker estimators of ϕ_h and σ_h^2 are shown to be asymptotically equivalent under the regularity conditions employed in Poskitt (2006), that paper, and the more extensive work presented here, shows their finite-sample behaviour to be quite different, particularly as regards the "normality" or otherwise of the empirical distributions of the estimated coefficients. The error in Yule-Walker estimation of the autoregressive coefficients is generally larger than for the other four techniques, although this varies according to which coefficient is under examination, and, of course, with the degree of fractional integration. For the non-invertible moving average the relative bias in the Yule-Walker estimator is notable, particularly when averaged over the entire h-vector of coefficients.

For the fractional noise processes the differences were not nearly so marked, although Yule-Walker was still generally less accurate than the other estimators. The bias in Yule-Walker estimates of the partial autocorrelation coefficient, for instance, was almost invariably greater than for the other estimators; with the single exception being for the non-invertible moving average and $T = 50$. There was very little to choose between Least Squares, Forward-Backward Least Squares, and the two Burg estimators; Least Squares did best with respect to estimation of the PAC, while the Burg estimators tended to be slightly more accurate otherwise.

The effect of fractional integration on the finite sample distributions of the coefficients can be quite startling, particularly for d close to 0.5, and particularly if we consider the sum of the coefficients. The distributions are quite heavily skewed in that case, and generally of somewhat irregular appearance; more importantly, these "irregularities" do not disappear as T increases. Weighting the coefficients so as to take advantage of Hosking's result regarding the differences of autocovariances substantially removes the skewness, resulting in a statistic with a standard Normal limiting distribution; unfortunately, the weights are fairly specific functions of the process autocovariances, so it is difficult to see an immediate practical application for this result.

The asymptotic efficiency of AIC as an order-selection tool in the infinite-order setting is borne out by the results presented in Table 1; however, we also found that the selected order is extremely variable, with a highly skewed distribution. Thus while the average AIC-selected order approaches the "optimal" order h^* reasonably quickly, the actual order selected in any given instance can lie anywhere between one and the upper limit of the search ($2\sqrt{T}$, in this case). Repeating the experiment with other asymptotically efficient selection criteria did not make a great deal of difference to the distribution of selected orders; all displayed a similar degree of variability and skewness, although the alternate criteria tended to better approximate the "optimal" order on average, the best performer in this regard being the 'criterion autoregressive transfer' function of Parzen (1974).

In summary, with the possible exception of the Yule-Walker approach, neither the choice of estimation method nor model selection technique would seem unduly critical as regards the average estimation outcome over a large number of realizations, at least for larger sample sizes. It must be noted, however, that our examination of the properties of these estimated autoregressive approximations has been largely in terms of the theoretical finite order approximation of a known infinite order process. The implications of the combination of autoregressive estimator and order selection criteria for data fitting and forecasting performance are yet to be examined.

Appendix A: Figures

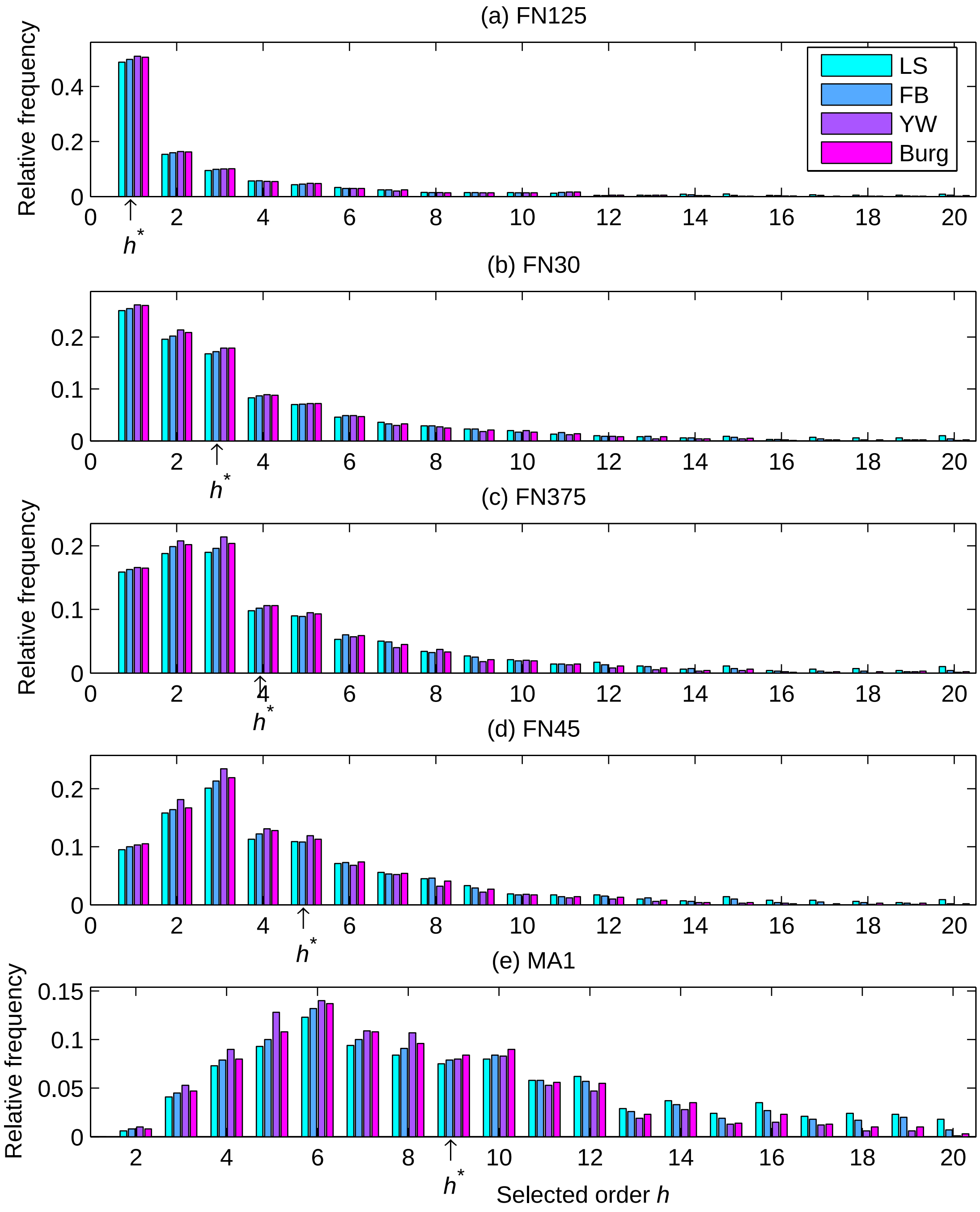

Figure 1: Relative frequency of occurrence of h_T^{AIC}, $T = 100$, for the fractional noise process $y(t) = \varepsilon(t)/(1-z)^d$ with (a) $d = 0.125$ (b) $d = 0.3$ (c) $d = 0.375$ and (d) $d = 0.45$ and (e) the moving average process $y(t) = \varepsilon(t) - \varepsilon(t-1)$.

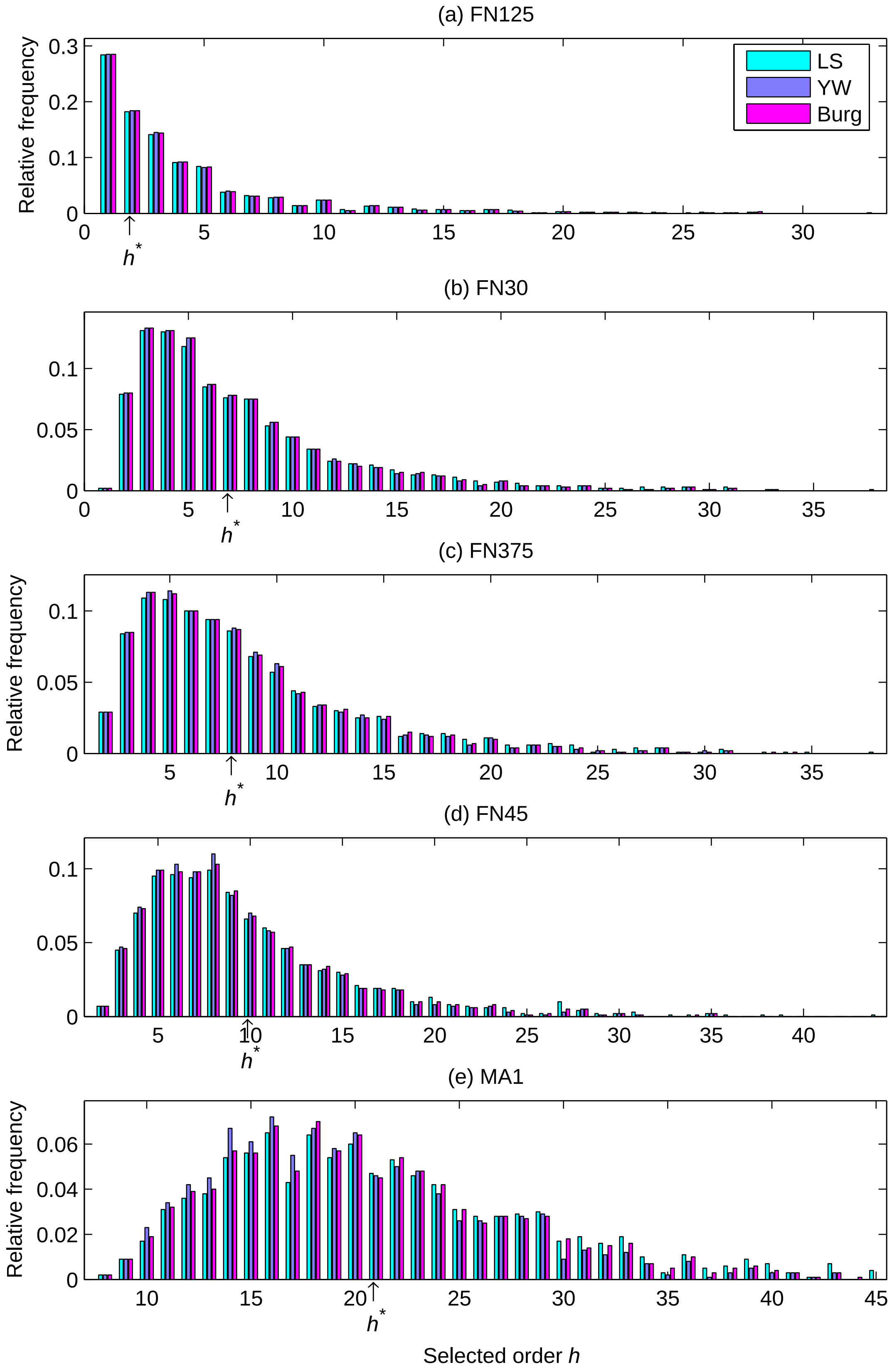

Figure 2: Relative frequency of occurrence of h_T^{AIC}, $T = 500$, for the fractional noise process $y(t) = \varepsilon(t)/(1-z)^d$ with (a) $d = 0.125$ (b) $d = 0.3$ (c) $d = 0.375$ and (d) $d = 0.45$ and (e) the moving average process $y(t) = \varepsilon(t) - \varepsilon(t-1)$.

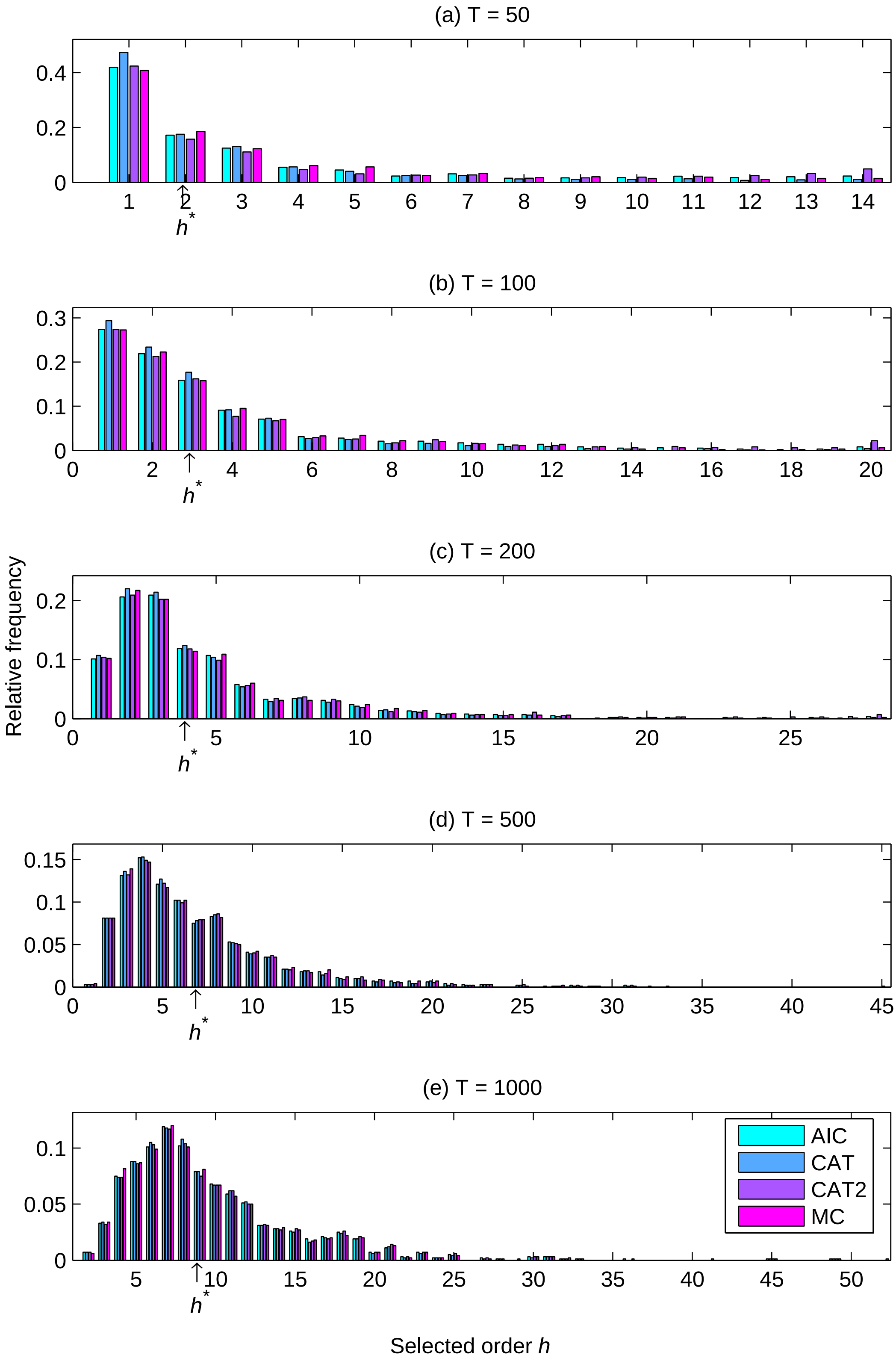

Figure 3: Relative frequency of occurrence of h_T as determined using Akaike's IC, Parzen's CAT, Bhansali's CAT_2 and Mallows' criterion *(MC)*, with $T = 50$, 100, 200, 500 and 1000, for the fractional noise process $y(t) = \varepsilon(t)/(1-z)^{0.3}$.

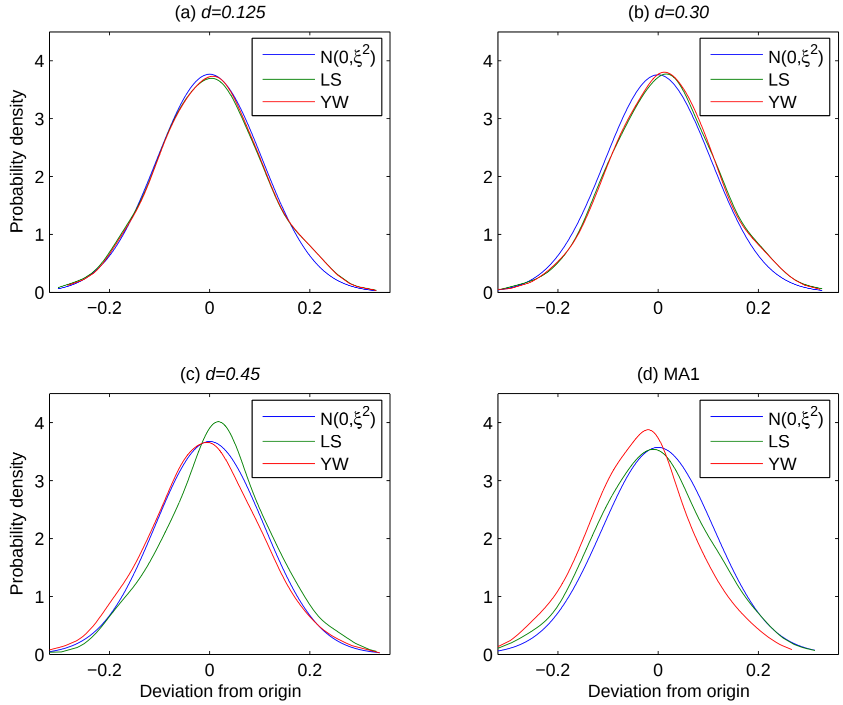

Figure 4: Empirical distribution of $(\hat{\phi}_h(1) - \phi_h(1))$ for the fractional noise process $y(t) = \varepsilon(t)/(1-z)^d$ with (a) $d = 0.125$ (b) $d = 0.3$ (c) $d = 0.45$, and (d) the moving average process $y(t) = \varepsilon(t) - \varepsilon(t-1)$, $h = h_T^*$ and $T = 100$.

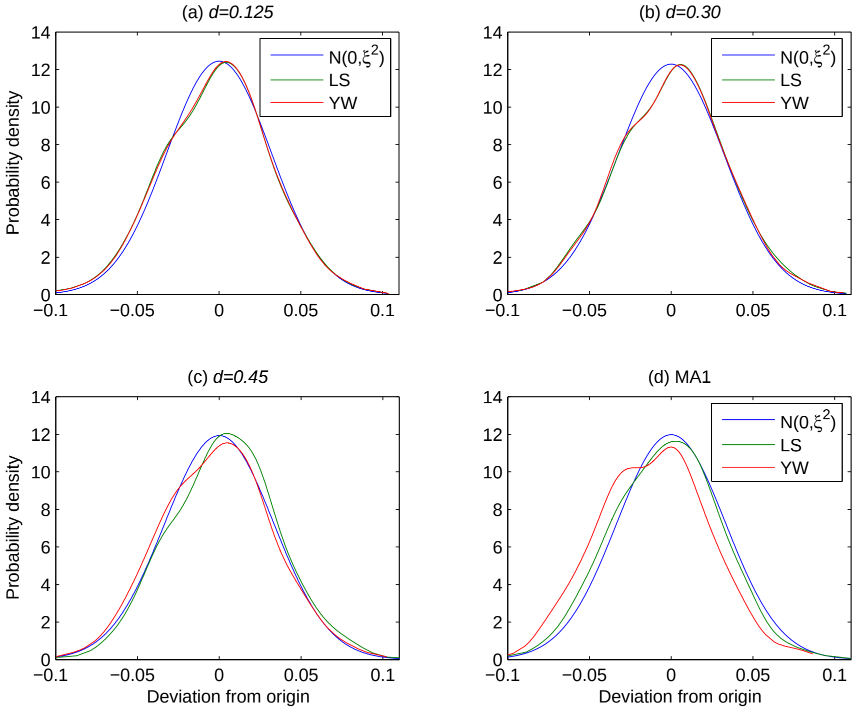

Figure 5: Empirical distribution of $(\hat{\phi}_h(1) - \phi_h(1))$ for the fractional noise process $y(t) = \varepsilon(t)/(1-z)^d$ with (a) $d = 0.125$ (b) $d = 0.3$ (c) $d = 0.45$, and (d) the moving average process $y(t) = \varepsilon(t) - \varepsilon(t-1)$, $h = h_T^*$ and $T = 1000$.

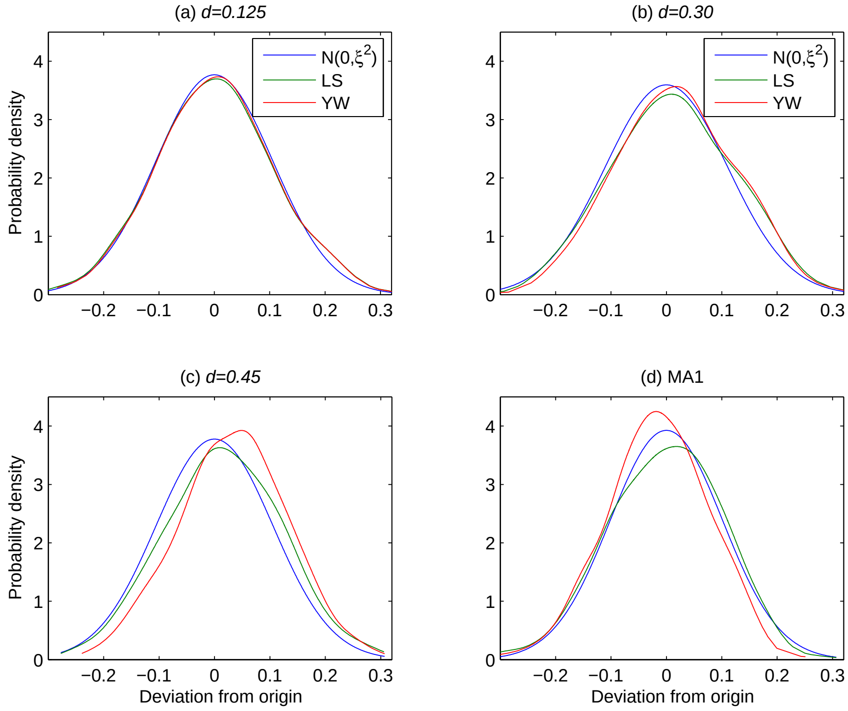

Figure 6: Empirical distribution of $(\hat{\phi}_h(h) - \phi_h(h))$ for the fractional noise process $y(t) = \varepsilon(t)/(1-z)^d$ with (a) $d = 0.125$ (b) $d = 0.3$ (c) $d = 0.45$, and (d) the moving average process $y(t) = \varepsilon(t) - \varepsilon(t-1)$, $h = h_T^*$ and $T = 100$.

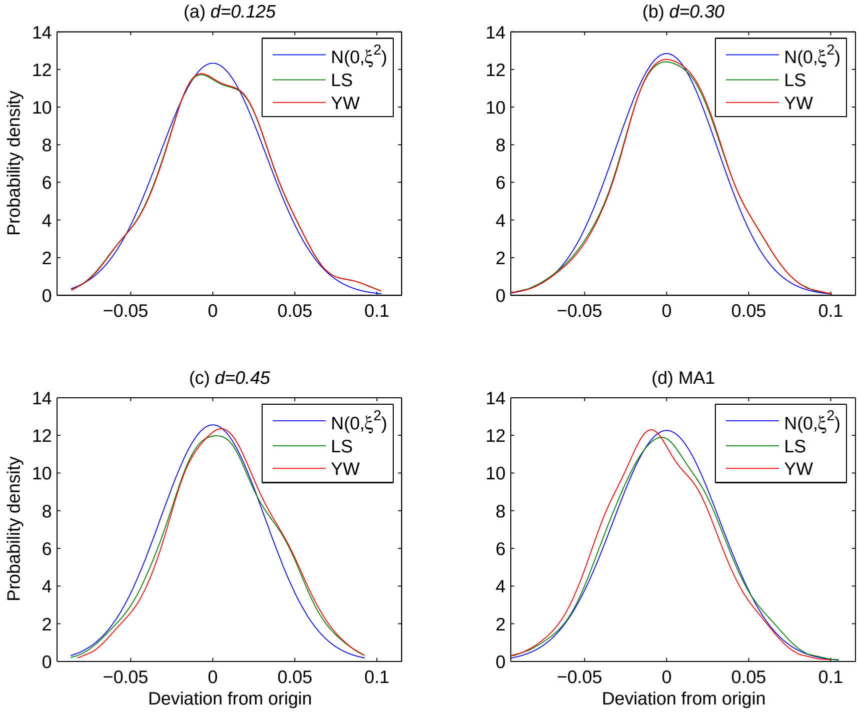

Figure 7: Empirical distribution of $(\hat{\phi}_h(h) - \phi_h(h))$ for the fractional noise process $y(t) = \varepsilon(t)/(1-z)^d$ with (a) $d = 0.125$ (b) $d = 0.3$ (c) $d = 0.45$, and (d) the moving average process $y(t) = \varepsilon(t) - \varepsilon(t-1)$, $h = h_T^*$ and $T = 1000$.

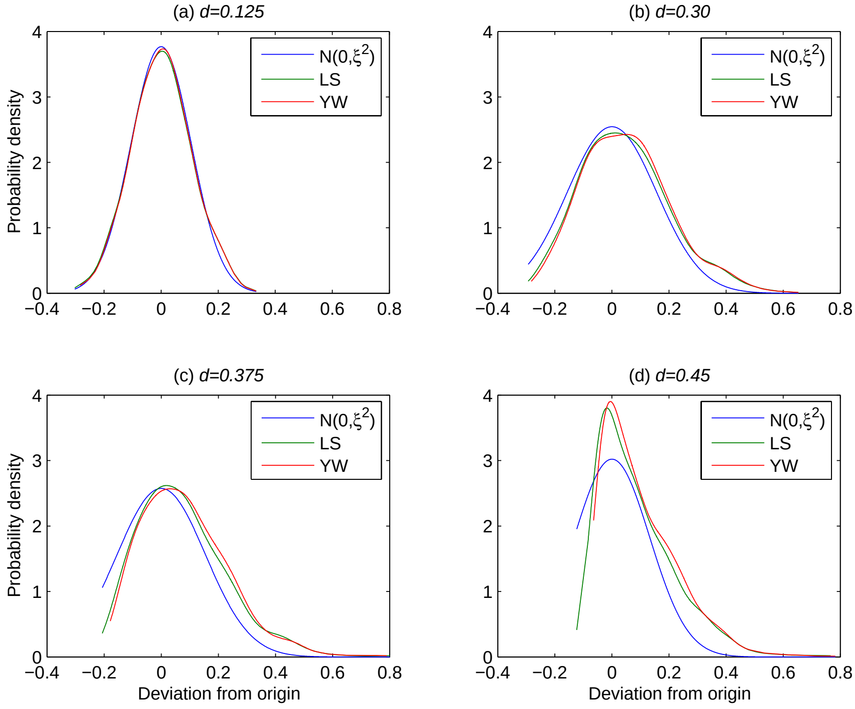

Figure 8: Empirical distribution of $\sum_{j=1}^{h}(\hat{\phi}_h(j) - \phi(j))$ for the fractional noise process $y(t) = \varepsilon(t)/(1-z)^d$ with (a) $d = 0.125$ (b) $d = 0.3$ (c) $d = 0.375$ and (d) $d = 0.45$, $h = h_T^* = 1, 3, 4$ and 5 respectively, and $T = 100$.

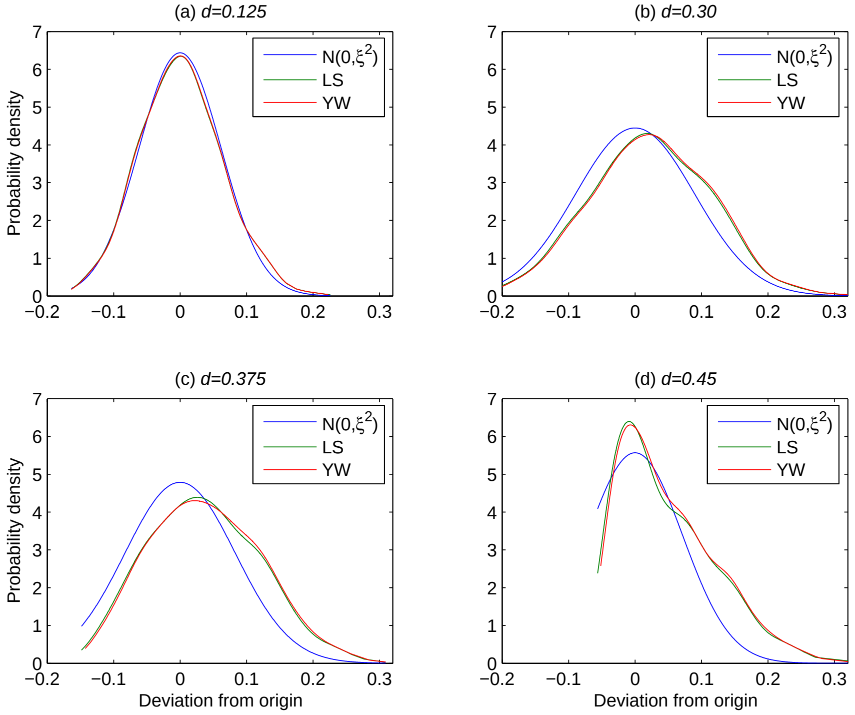

Figure 9: Empirical distribution of $\sum_{j=1}^{h}(\hat{\phi}_h(j) - \phi_h(j))$ for the fractional noise process $y(t) = \varepsilon(t)/(1-z)^d$ with (a) $d = 0.125$ (b) $d = 0.3$ (c) $d = 0.375$ and (d) $d = 0.45$, $h = h_T^* = 2, 7, 8$ and 10 respectively, and $T = 500$.

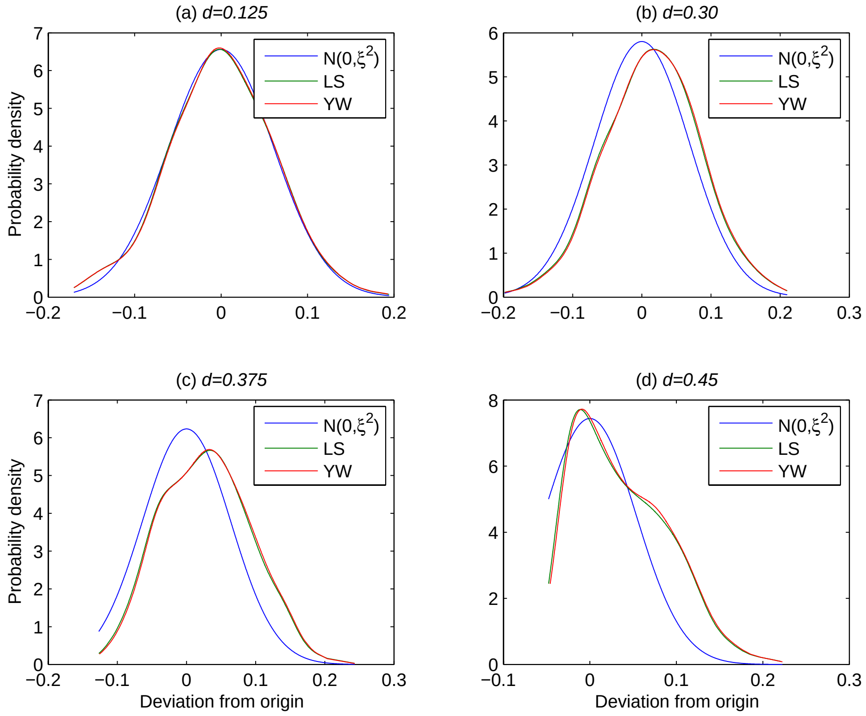

Figure 10: Empirical distribution of $\sum_{j=1}^{h}(\hat{\phi}_h(j) - \phi_h(j))$ for the fractional noise process $y(t) = \varepsilon(t)/(1-z)^d$ with (a) $d = 0.125$ (b) $d = 0.3$ (c) $d = 0.375$ and (d) $d = 0.45$, $h = h_T^* = 4, 9, 12$ and 14 respectively, and $T = 1000$.

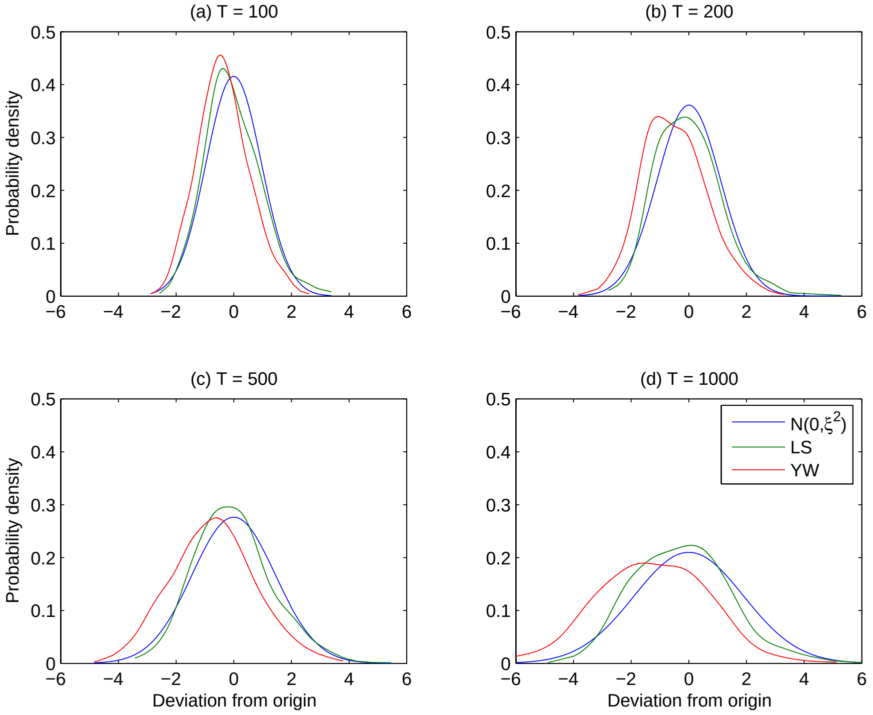

Figure 11: Empirical distribution of $\sum_{j=1}^{h}(\hat{\phi}_h(j) - \phi_h(j))$ for the moving average process $y(t) = \varepsilon(t) - \varepsilon(t-1)$ for $h = h_T^*$ and (a) $T = 100$ (b) $T = 200$ (c) $T = 500$ and (d) $T = 1000$.

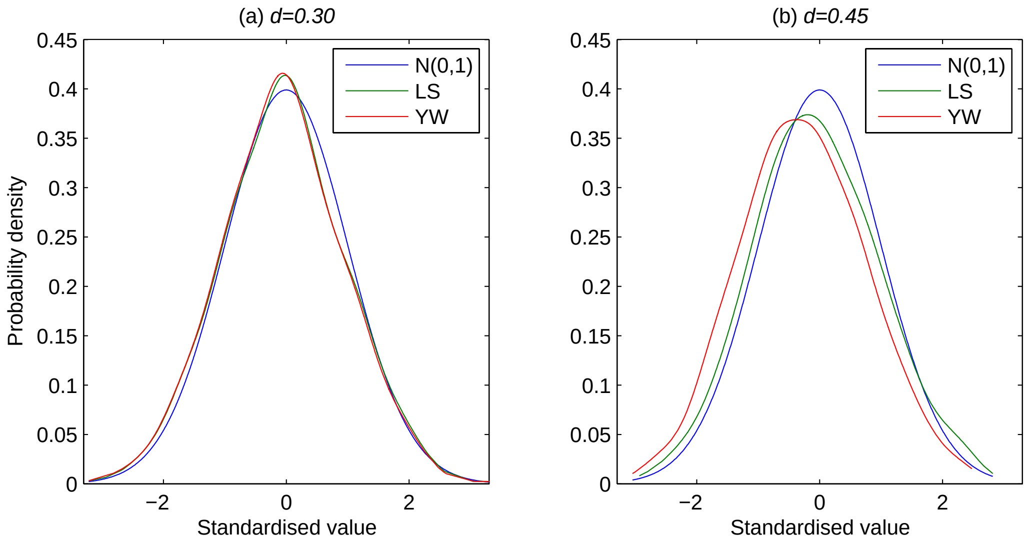

Figure 12: Observed distribution of $\varphi_{\lambda,T}$ for (a) $y(t) = \varepsilon(t)/(1-z)^{0.3}$, and (b) $y(t) = \varepsilon(t)/(1-z)^{0.45}$, when $\boldsymbol{\lambda}_h' = (1, 0, \ldots, 0, -1)$, $T = 1000$.

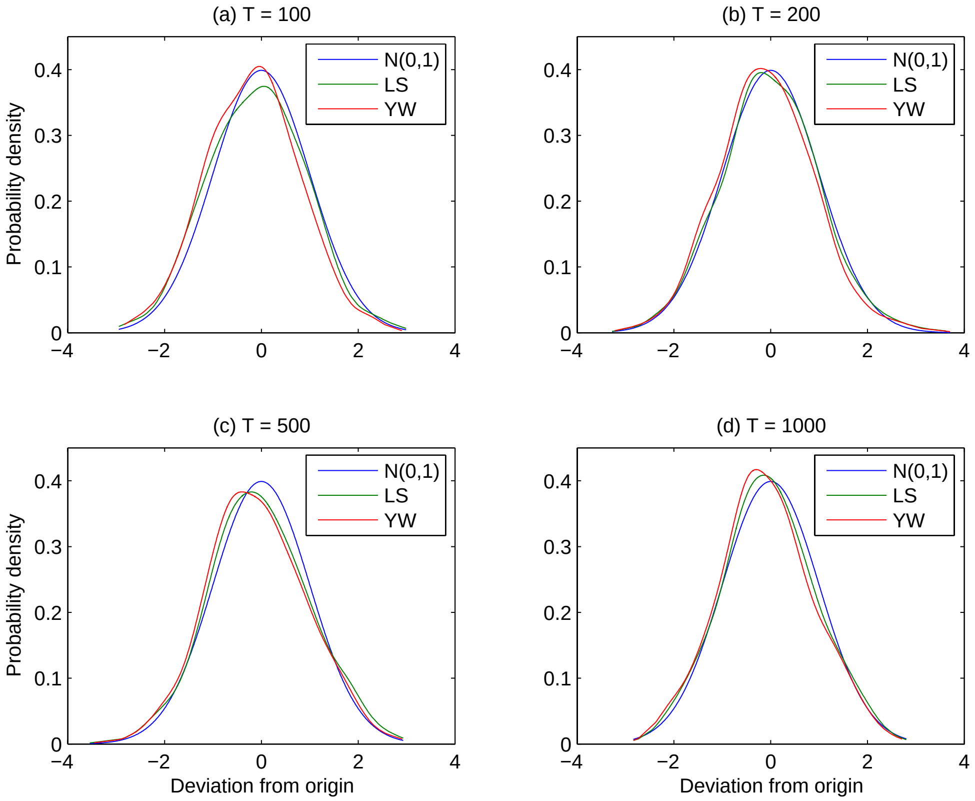

Figure 13: Observed distribution of $\varphi_{\lambda,T}$ for $y(t) = \varepsilon(t)/(1-z)^{0.375}$, when $\boldsymbol{\lambda}_h' = (1, 0, \ldots, 0, -1)$, for (a) $T = 100$ (b) $T = 200$ (c) $T = 500$ and (d) $T = 1000$.

Appendix B: Tables

Table 1

AIC-based estimates of h. Average in $R = 1000$ replications, by estimation method, model, and sample size.

				Estimation method				
Model: $y(t) =$	T	h^*	h_{AIC}	LS	YW	FB	Burg	GB
$(1-L)\varepsilon(t)$	50	6	4	6.53	5.14	5.85	5.52	5.51
	100	9	6	9.22	7.86	8.71	8.31	8.31
	200	13	9	13.21	11.94	12.88	12.42	12.42
	500	21	14	21.21	19.9	20.87	20.57	20.57
	1000	31	21	30.99	29.29	30.62	30.19	30.19
$\dfrac{\varepsilon(t)}{(1-L)^{0.125}}$	50	1	1	3.34	2.37	2.7	2.5	2.5
	100	1	1	3.27	2.72	2.96	2.78	2.78
	200	1	1	3.18	2.98	3.1	2.99	2.99
	500	2	2	4.22	4.09	4.15	4.12	4.12
	1000	4	2	5.42	5.33	5.39	5.33	5.33
$\dfrac{\varepsilon(t)}{(1-L)^{0.3}}$	50	2	1	3.71	2.74	3.15	2.9	2.9
	100	3	2	4.17	3.56	3.87	3.65	3.65
	200	4	3	5.19	4.74	4.96	4.77	4.77
	500	7	5	7.59	7.29	7.46	7.3	7.3
	1000	9	7	10.19	10.04	10.11	10.05	10.05
$\dfrac{\varepsilon(t)}{(1-L)^{0.375}}$	50	3	2	4	2.88	3.32	3.09	3.08
	100	4	3	4.66	3.92	4.29	4.1	4.1
	200	5	4	6.08	5.48	5.74	5.56	5.56
	500	8	6	8.69	8.33	8.62	8.4	8.4
	1000	12	8	11.99	11.72	11.87	11.74	11.74
$\dfrac{\varepsilon(t)}{(1-L)^{0.45}}$	50	3	2	4.35	3.01	3.71	3.38	3.38
	100	5	3	5.13	4.21	4.75	4.48	4.48
	200	6	4	6.87	5.99	6.59	6.39	6.39
	500	10	7	9.94	9.32	9.8	9.52	9.52
	1000	14	10	13.76	13.26	13.63	13.43	13.43

Table 2

Estimates of h, as produced by minimizing Akaike's AIC and FPE, Parzen's CAT and CAT*, Mallows' statistic, and Bhansali's CAT$_2$. Average in $R = 1000$ replications, by sample size T and model.

				Criterion					
Model	d	T	h^*	AIC	FPE	CAT	CAT*	MC	CAT2
FN	0.125	50	1	2.805	2.677	2.31	2.538	2.578	3.425
		100	1	2.656	2.602	2.39	2.494	2.495	3.242
		200	1	2.534	2.528	2.456	2.464	2.519	2.878
		500	2	3.319	3.319	3.238	3.263	3.326	3.365
		1000	4	4.676	4.676	4.555	4.67	4.66	4.662
	0.3	50	2	3.4	3.204	2.769	3.005	3.227	3.853
		100	3	3.787	3.718	3.237	3.56	3.665	4.194
		200	4	4.719	4.701	4.405	4.624	4.666	4.988
		500	7	6.848	6.834	6.634	6.701	6.785	6.899
		1000	9	9.648	9.648	9.487	9.61	9.626	9.72
	0.375	50	3	3.652	3.466	2.929	3.301	3.485	4.103
		100	4	4.341	4.3	3.852	4.095	4.242	4.741
		200	5	5.744	5.723	5.343	5.585	5.621	6.075
		500	8	8.062	8.05	7.84	7.922	8.041	8.223
		1000	12	11.25	11.25	11.11	11.17	11.30	11.32
	0.45	50	3	4.092	3.916	3.307	3.739	4.023	4.478
		100	5	4.837	4.722	4.373	4.554	4.793	5.271
		200	6	6.503	6.481	6.206	6.38	6.405	6.902
		500	10	9.249	9.24	8.936	9.104	9.324	9.283
		1000	14	12.94	12.94	12.71	12.87	12.9	13.0
MA		50	6	7.179	6.917	6.874	6.63	5.43	9.206
		100	9	10.02	9.901	9.685	9.504	8.255	12.33
		200	13	14.09	14.02	13.8	13.61	12.71	16.21
		500	21	21.79	21.74	21.43	21.3	20.4	24.17
		1000	31	31.7	31.67	31.12	31.15	30.18	33.45

Table 3

Minimum "rate of agreement" between Akaike's AIC and: Akaike's FPE, Parzen's CAT* and CAT, Mallows' statistic, and Bhansali's CAT$_2$, by sample size T and model class.

		Criterion				
Model	T	FPE	CAT*	CAT	MC	CAT2
FN	50	0.97	0.93	0.847	0.845	0.775
	100	0.987	0.955	0.905	0.879	0.842
	200	0.995	0.967	0.931	0.887	0.879
	500	0.997	0.977	0.947	0.914	0.914
	1000	0.999	0.988	0.963	0.95	0.942
MA	50	0.952	0.878	0.898	0.648	0.589
	100	0.984	0.914	0.926	0.692	0.646
	200	0.991	0.923	0.932	0.789	0.726
	500	0.997	0.946	0.944	0.848	0.799
	1000	0.997	0.952	0.945	0.87	0.837

Table 4

Estimation error in the coefficients, averaged over the coefficient vector $(h^{-1}\sum_{j=1}^{h}(\hat{\phi}_h(j) - \phi_h(j)),\ h = h_T^*)$. Average in $R = 1000$ replications, by estimation method, model, and sample size.

						Estimation method		
Model	d	T	h	LS	YW	FB	Burg	GB
MA		50	6	0.00017	-0.04026	-0.00039	0.00467	0.00517
		100	9	-0.00351	-0.0393	-0.00387	0.00025	0.00043
		200	13	-0.00289	-0.038	-0.00312	-0.00155	-0.00148
		500	21	-0.00131	-0.03346	-0.00126	-0.00039	-0.00037
		1000	31	-0.0076	-0.04102	-0.00753	-0.00698	-0.00697
FN	0.125	50	1	0.00444	0.00732	0.00466	0.00466	0.00463
		100	1	0.00173	0.00312	0.00177	0.00177	0.00176
		200	1	0.00026	0.00097	0.00026	0.00026	0.00026
		500	2	0.00064	0.00083	0.00061	0.00058	0.00058
		1000	4	0.00004	0.00013	0.00003	0.00002	0.00002
	0.3	50	2	0.03275	0.0376	0.0328	0.03283	0.03278
		100	3	0.01505	0.01712	0.01499	0.01482	0.01481
		200	4	0.00856	0.00936	0.0084	0.00829	0.00829
		500	7	0.00316	0.00348	0.00315	0.00316	0.00316
		1000	9	0.00162	0.00177	0.00162	0.00162	0.00162
	0.375	50	3	0.03307	0.03768	0.03292	0.03264	0.0326
		100	4	0.01848	0.02067	0.01849	0.01826	0.01825
		200	5	0.01039	0.01126	0.01025	0.01016	0.01016
		500	8	0.00447	0.00479	0.00447	0.00445	0.00445
		1000	12	0.00224	0.00237	0.00224	0.00223	0.00223
	0.45	50	3	0.03685	0.04242	0.03655	0.0364	0.03637
		100	5	0.01702	0.01929	0.01704	0.01686	0.01685
		200	6	0.01064	0.01153	0.01054	0.01044	0.01044
		500	10	0.00463	0.00492	0.00463	0.00461	0.00461
		1000	14	0.00255	0.00267	0.00255	0.00254	0.00254

Table 5

Mean-squared estimation error in the coefficients, averaged over the coefficient vector $(h^{-1}\sum_{j=1}^{h}(\hat{\phi}_h(j)-\phi_h(j))^2$, $h = h_T^*)$. Average in $R = 1000$ replications, by estimation method, model, and sample size.

						Estimation method		
Model	d	T	h	LS	YW	FB	Burg	GB
MA		50	6	0.0349	0.0309	0.0337	0.0333	0.0334
		100	9	0.0217	0.0204	0.0213	0.021	0.021
		200	13	0.0141	0.0152	0.0139	0.0138	0.0138
		500	21	0.0081	0.0099	0.008	0.008	0.008
		1000	31	0.0058	0.0086	0.0058	0.0058	0.0058
FN	0.125	50	1	0.0218	0.0209	0.0217	0.0217	0.0217
		100	1	0.0112	0.011	0.0112	0.0112	0.0112
		200	1	0.0056	0.0056	0.0056	0.0056	0.0056
		500	2	0.002	0.002	0.002	0.002	0.002
		1000	4	0.0011	0.001	0.0011	0.0011	0.0011
	0.3	50	2	0.0243	0.0232	0.0241	0.024	0.024
		100	3	0.0127	0.0121	0.0127	0.0126	0.0126
		200	4	0.0057	0.0055	0.0057	0.0056	0.0056
		500	7	0.0022	0.0022	0.0022	0.0022	0.0022
		1000	9	0.0011	0.0011	0.0011	0.0011	0.0011
	0.375	50	3	0.0273	0.0248	0.0269	0.0267	0.0268
		100	4	0.013	0.0121	0.0128	0.0127	0.0127
		200	5	0.0061	0.0059	0.0061	0.0061	0.0061
		500	8	0.0023	0.0023	0.0023	0.0023	0.0023
		1000	12	0.0012	0.0012	0.0012	0.0012	0.0012
	0.45	50	3	0.0279	0.0254	0.0275	0.0275	0.0275
		100	5	0.0139	0.0129	0.0137	0.0136	0.0136
		200	6	0.0062	0.006	0.0061	0.0061	0.0061
		500	10	0.0025	0.0024	0.0025	0.0025	0.0025
		1000	14	0.0012	0.0012	0.0012	0.0012	0.0012

Table 6

Estimation error in the first coefficient ($\hat{\phi}_h(1) - \phi_h(1)$, $h = h_T^*$). Average in $R = 1000$ replications, by estimation method, model, and sample size.

					Estimation method				
Model	d	T	h	$\phi_h(1)$	LS	YW	FB	Burg	GB
MA		50	6	0.8571	-0.0233	-0.0534	-0.0232	-0.015	-0.0146
		100	9	0.9	-0.0117	-0.0353	-0.0132	-0.0077	-0.0076
		200	13	0.9286	-0.0061	-0.0241	-0.0068	-0.0044	-0.0044
		500	21	0.9545	-0.0035	-0.0145	-0.0033	-0.0024	-0.0024
		1000	31	0.9688	-0.003	-0.0112	-0.003	-0.0024	-0.0024
FN	0.125	50	1	-0.1429	0.0044	0.0073	0.0047	0.0047	0.0046
		100	1	-0.1429	0.0017	0.0031	0.0018	0.0018	0.0018
		200	1	-0.1429	0.0003	0.001	0.0003	0.0003	0.0003
		500	2	-0.1333	-0.0009	-0.0008	-0.001	-0.001	-0.001
		1000	4	-0.129	-0.0013	-0.0012	-0.0013	-0.0012	-0.0012
	0.3	50	2	-0.3529	0.02	0.0233	0.0196	0.0199	0.0198
		100	3	-0.3333	0.0103	0.0104	0.0099	0.0097	0.0097
		200	4	-0.3243	0.006	0.0054	0.0055	0.0055	0.0055
		500	7	-0.3134	0.002	0.0017	0.002	0.0021	0.0021
		1000	9	-0.3103	0.0008	0.0005	0.0008	0.0008	0.0008
	0.375	50	3	-0.4286	0.0256	0.0203	0.0239	0.0232	0.0232
		100	4	-0.4138	0.0135	0.0091	0.0128	0.0126	0.0126
		200	5	-0.4054	0.0092	0.006	0.0086	0.0085	0.0085
		500	8	-0.3934	0.0044	0.0027	0.0043	0.0043	0.0043
		1000	12	-0.3871	0.0021	0.0009	0.0021	0.002	0.002
	0.45	50	3	-0.5294	0.0307	0.0053	0.0285	0.0276	0.0276
		100	5	-0.4945	0.0147	-0.0078	0.0134	0.0135	0.0135
		200	6	-0.4865	0.0111	-0.0035	0.0103	0.0104	0.0104
		500	10	-0.4712	0.0054	-0.0023	0.0054	0.0055	0.0055
		1000	14	-0.4649	0.0032	-0.0015	0.0032	0.0031	0.0031

Table 7

Estimation error in the partial autocorrelation coefficient ($\hat{\phi}_h(h) - \phi_h(h)$, $h = h_T^*$). Average in $R = 1000$ replications, by estimation method, model, and sample size.

					Estimation method				
Model	d	T	h	$\phi_h(h)$	LS	YW	FB	Burg	GB
MA		50	6	0.1429	0.0224	-0.0017	0.0212	0.0205	0.0207
		100	9	0.1	-0.0028	-0.0164	-0.0035	-0.0035	-0.0034
		200	13	0.0714	-0.0015	-0.0111	-0.0018	-0.0019	-0.0018
		500	21	0.0455	-0.0003	-0.0057	-0.0003	-0.0003	-0.0003
		1000	31	0.0313	-0.0009	-0.0047	-0.0009	-0.0009	-0.0009
FN	0.125	50	1	-0.1429	0.0044	0.0073	0.0047	0.0047	0.0046
		100	1	-0.1429	0.0017	0.0031	0.0018	0.0018	0.0018
		200	1	-0.1429	0.0003	0.001	0.0003	0.0003	0.0003
		500	2	-0.0667	0.0022	0.0024	0.0022	0.0022	0.0022
		1000	4	-0.0323	0.002	0.0022	0.002	0.002	0.002
	0.3	50	2	-0.1765	0.0455	0.0519	0.046	0.0458	0.0457
		100	3	-0.1111	0.0115	0.0159	0.0116	0.0116	0.0116
		200	4	-0.0811	0.0132	0.0154	0.0132	0.0132	0.0132
		500	7	-0.0448	0.0019	0.0026	0.0019	0.0019	0.0019
		1000	9	-0.0345	0.0032	0.0036	0.0032	0.0032	0.0032
	0.375	50	3	-0.1429	0.0243	0.0391	0.0251	0.0251	0.025
		100	4	-0.1034	0.0284	0.0357	0.0287	0.0286	0.0286
		200	5	-0.0811	0.0099	0.0138	0.0099	0.0099	0.0099
		500	8	-0.0492	0.007	0.0087	0.0071	0.0071	0.0071
		1000	12	-0.0323	0.0038	0.0046	0.0039	0.0039	0.0039
	0.45	50	3	-0.1765	0.0277	0.06	0.0288	0.0288	0.0287
		100	5	-0.0989	0.0158	0.0331	0.0164	0.0164	0.0164
		200	6	-0.0811	0.0164	0.0257	0.0165	0.0165	0.0165
		500	10	-0.0471	0.0085	0.0124	0.0086	0.0086	0.0086
		1000	14	-0.0332	0.005	0.0069	0.0051	0.0051	0.0051

References

Akaike, H. (1970). Statistical predictor identification. *Annals of Institute of Statistical Mathematics* **22** 203–217.

Andersen, N. (1974). On the calculation of filter coefficients for maximum entropy spectral analysis. *Geophysics* **39** 69–72.

Baillie, R. T. and Chung, S.-K. (2002). Modeling and forecasting from trend-stationary long memory models with applications to climatology. *International Journal of Forecasting* **18** 215–226.

Barkoulas, J. and Baum, C. F. (2003). Long-memory forecasting of U.S. monetary indices. Working Paper 558, Department of Economics, Boston College, Chestnut Hill, MA 02467 U.S.A.

Beran, J. (1994). *Statistics for Long Memory Processes.* Chapman and Hall, New York.

Beran, J. (1995). Maximum likelihood estimation of the differencing parameter for invertible short and long memory autoregressive integrated moving average models. *Journal of the Royal Statistical Society* **B 57** 654–672.

Bhansali, R. J. (1985). The criterion autoregressive transfer function of Parzen. *Journal of Time Series Analysis* .

Bhansali, R. J. (1986). Asymptotically efficient selection of the order by the criterion autoregressive transfer function. *Annals of Statistics* **14** 315–325.

Box, G. and Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control.* Holden Day, San Francisco.

Brockwell, P. L. and Davis, R. A. (1991). *Time Series: Theory and Methods.* 2nd ed. Springer Series in Statistics, Springer-Verlag, New York.

Burg, J. (1967). Maximum entropy spectral analysis. Paper presented at the 37th International Meeting, Society of Exploratory Geophysics, Oklahoma City, Oklahoma.

Burg, J. (1968). A new analysis technique for time series data. Paper presented at the Advanced Study Institute on Signal Processing, N.A.T.O., Enschede, Netherlands.

Durbin, J. (1960). The fitting of time series models. *Review of International Statistical Institute* **28** 233–244.

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

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.

Hannan, E. J. and Quin, B. G. (1979). The determination of the order of an autoregression. *Journal of Royal Statistical Society* **B 41** 190–195.

Hosking, J. R. M. (1980). Fractional differencing. *Biometrika* **68** 165–176.

Hosking, J. R. M. (1996). Asymptotic distributions of the sample mean, autocovariances, and autocorrelations of long memory time series. *Journal of Econometrics* **73** 261–284.

Kay, S. M. (1988). *Modern Spectral Estimation.* Prentice-Hall.

Kolmorgorov, A. N. (1941). Interpolation und extrapolation von stationaren zufalligen folgen. *Bulletin Academy Science U. S. S. R., Mathematics Series* **5** 3–14.

Levinson, N. (1947). The Wiener RMS (root mean square) error criterion in filter design and prediction. *Journal of Mathematical Physics* **25** 261–278.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

Marple, S. L. (1987). *Digital Spectral Analysis with Applications.* Prentice-Hall.

Parzen, E. (1974). Some recent advances in time series modelling. *IEEE Transactions on Automatic Control* **AC-19** 723–730.

Parzen, E. (1977). Multiple times series: Determining the order of approximating autoregressive schemes. In *Multivariate Analysis* (P. R. Krishnaiah, ed.), vol. IV. North-Holland, Amsterdam, 283–295.

Poskitt, D. S. (2006). Autoregressive approximation in nonstandard situations: The fractionally integrated and non-invertible cases. *Annals of Institute of Statistical Mathematics* forthcoming.

Robinson, P. M. (1995). Log periodogram regression of time series with long memory. *Annals of Statistics* **23** 1048–1072.

Rosenblatt, M. (1961). Independence and dependence. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 2. University of California, Berkeley, University of California Press, Berkeley, CA.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6** 461–464.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* **8** 147–164.

Sowell, F. (1992). Maximum likelihood estmation of stationary univariate fractionally integrated time series models. *Journal of Econometrics* **53** 165–188.

Szegö, G. (1939). *Orthogonal Polynomials.* American Mathematical Society Colloquium Publication.

Tjøstheim, D. and Paulsen, J. (1983). Bias of some commonly-used time series estimates. *Biometrika* **70** 389–399.

Ulrych, T. J. and Bishop, T. N. (1975). Maximum entropy spectral analysis and autoregressive decomposition. *Reviews of Geophysics and Space Physics* **13** 183–200.

Wand, M. and Jones, M. (1995). *Kernel Smoothing.* Chapman and Hall.