AISM (2007) 59: 697–725
DOI 10.1007/s10463-006-0074-4

D. S. Poskitt

Autoregressive approximation in nonstandard situations: the fractionally integrated and non-invertible cases

Received: 29 June 2005 / Revised: 5 January 2006 /
Published online: 30 August 2006

Abstract Autoregressive models are commonly employed to analyze empirical time series. In practice, however, any autoregressive model will only be an approximation to reality and in order to achieve a reasonable approximation and allow for full generality the order of the autoregression, h say, must be allowed to go to infinity with T, the sample size. Although results are available on the estimation of autoregressive models when h increases indefinitely with T such results are usually predicated on assumptions that exclude (1) non-invertible processes and (2) fractionally integrated processes. In this paper we will investigate the consequences of fitting long autoregressions under regularity conditions that allow for these two situations and where an infinite autoregressive representation of the process need not exist. Uniform convergence rates for the sample autocovariances are derived and corresponding convergence rates for the estimates of AR(h) approximations are established. A central limit theorem for the coefficient estimates is also obtained. An extension of a result on the predictive optimality of AIC to fractional and non-invertible processes is obtained.

Keywords Autoregression · Autoregressive approximation · Fractional process · Non-invertibility · Order selection · Asymptotic efficiency

1 Introduction

The use of autoregressive (AR) models has a long history that can be traced back to the early papers of Akaike (1969, 1970) and Parzen (1974) and beyond to the

D. S. Poskitt
Department of Econometrics and Business Statistics, Monash University,
Clayton, VIC 3800, Australia
E-mail: Don.Poskitt@Buseco.monash.edu.au

prescient work of Yule (1921). It is not surprising given this long history that there is a substantial literature dealing with such models: using the *Google* web browser with the search word autoregression produced 17,600 sites, the word autoregressive produced 89,700! Nevertheless, there are still gaps in the theory of AR approximation that need to be filled if AR modelling is to be routinely extended to the type of long memory processes currently employed to investigate empirical time series that exhibit long-term persistence. A brief history of the application of long memory processes and a review of various statistical procedures for analyzing such processes is provided in Beran (1992, 1994), see also Baillie (1996).

In order to set the scene, let $y(t)$ for $t \in \mathrm{Z}$ denote a linearly regular, covariance-stationary process,

$$y(t) = \sum_{j=0}^{\infty} \kappa(j)\varepsilon(t-j), \tag{1}$$

where $\varepsilon(t)$, $t \in \mathrm{Z}$, is a zero mean white noise process with variance σ^2 and the impulse response coefficients satisfy the conditions $\kappa(0) = 1$ and $\sum_{j\geq 0} \kappa(j)^2 < \infty$.

Assumption 1 *Let $\mathcal{E}_t$ denote the σ-algebra of events determined by $\varepsilon(s)$, $s \leq t$. It will be supposed throughout the paper that $\varepsilon(t)$ is ergodic and that*

$$E\big[\varepsilon(t) \mid \mathcal{E}_{t-1}\big] = 0 \quad \textit{and} \quad E\big[\varepsilon(t)^2 \mid \mathcal{E}_{t-1}\big] = \sigma^2 . \tag{2}$$

Furthermore, $E\big[\varepsilon(t)^4\big] < \infty$.

Assumption 1 imposes a classical martingale difference structure on the innovations $\varepsilon(t)$. The significance of this assumption here is that it implies that the minimum mean squared error predictor of $y(t)$ given $\mathcal{E}_{t-1}$, $\bar{y}_{\langle t|t-1,\ldots,-\infty\rangle}$ say, is the linear predictor, Hannan and Deistler (1988, Theorem 1.4.2).

Consider now the best linear predictor of $y(t)$ based on the finite past $y(t-j)$, $j = 1, \ldots, h$. Let $\gamma(\tau) = \gamma(-\tau) = E[y(t)y(t+\tau)] = \sigma^2 \sum_{r\geq 0} \kappa(r)\kappa(\tau + r)$, $\tau = 0, 1, \ldots$, denote the autocovariance function of the process $y(t)$. The coefficients of the minimum mean squared error predictor are obtained by solving the Yule–Walker equations

$$\sum_{j=0}^{h} \alpha_h(j)\gamma(j-k) = \delta_{0k}\sigma_h^2 \quad k = 0, 1, \ldots, h, \tag{3}$$

for $\alpha_h(j)$, $j = 1, \ldots, h$, where δ_{0k} is Kronecker's delta, $\alpha_h(0) = 1$ and $\sigma_h^2 = E[\epsilon_h(t)^2]$ is the minimizing value of the prediction error variance of the prediction error

$$\epsilon_h(t) = y(t) - \bar{y}_{\langle t|t-1,\ldots,t-h\rangle} = \sum_{j=0}^{h} \alpha_h(j)y(t-j) \tag{4}$$

associated with the best linear predictor $\bar{y}_{\langle t|t-1,\ldots,t-h\rangle} = -\alpha_h(1)y(t-1) - \cdots - \alpha_h(h)y(t-h)$.

Heuristically speaking it is clear that h must be allowed to go to infinity in order to capture the influence of effects in the remote past and it seems reasonable to suppose that as $h \to \infty$ the best linear predictor $\bar{y}_{\langle t|t-1,\ldots,t-h\rangle}$ determined from the AR(h) model implicit in the Yule–Walker calculations will form a good approximation to the optimal predictor $\bar{y}_{\langle t|t-1,\ldots,-\infty\rangle}$.

Results currently available on the properties of AR(h) models when the autoregressive order h is allowed to increase with the sample size T are usually predicated on the assumption that the process admits an infinite AR representation with coefficients that tend to zero at an appropriate rate. These assumptions are often expressed in terms of particular summability conditions on the AR coefficients themselves, or equivalently the Wold representation. Thus it is commonly assumed that (1) the transfer function associated with Wold's representation,

$$k(z) = \sum_{j=0}^{\infty} \kappa(j) z^j,$$

is invertible, which following common practice is defined to mean $k(z) \neq 0$, $|z| \leq 1$, and, (2) a summability condition such as $\sum_{j\geq 0} |\kappa(j)| < \infty$, or $\sum_{j\geq 0} j|\kappa(j)|^2 < \infty$, or $\sum_{j\geq 0} j^{\frac{1}{2}}|\kappa(j)| < \infty$ holds. See Hannan and Deistler (1988, Sect. 7.4) for example. There are two critical cases that do not meet such conditions (1) non-invertible processes, of course, and (2) fractionally integrated processes. One contribution of this paper is to show that such assumptions can be relaxed and that results on the statistical properties of AR approximations can be extended to allow for fractionally integrated and non-invertible processes.

Fractionally integrated processes were introduced by Granger and Joyeux (1980) and were independently described in Hosking (1980). The class of fractionally integrated processes can be characterized by the specification

$$y(t) = \sum_{j\geq 0} \kappa(j)\varepsilon(t-j) = k(z)\varepsilon(t) = \frac{m(z)}{(1-z)^d}\varepsilon(t)$$

wherein $m(z) = \sum_{j\geq 0} \mu(j) z^j$ and, as will be done henceforth in expressions of this type, the indeterminate z is interpreted as the lag operator, that is $z\varepsilon(t) = \varepsilon(t-1)$. For any $b > -1$ we can expand the operator $(1-z)^b$ via a binomial expansion and rearranging terms in $k(z) = m(z)/(1-z)^d$ we obtain the result that

$$\kappa(j) = \sum_{r=0}^{j} \frac{\mu(j-r)\Gamma(r+d)}{\Gamma(r+1)\Gamma(d)} \quad j = 1, 2, \ldots$$

where the gamma function $\Gamma(x) = \int_0^\infty t^{x-1}\mathrm{e}^{-t}\mathrm{d}t$ for $x \geq 0$ and the relation $\Gamma(x+1) = x\Gamma(x)$ defines $\Gamma(x)$ for $x < 0$. If $m(z)$ is such that $\sum_{j\geq 0} |\mu(j)| < \infty$, then using Sterling's approximation it can be shown that

$$\kappa(j) \sim \frac{m(1)}{\Gamma(d)} j^{d-1} \quad \text{as } j \to \infty\,. \tag{5}$$

From (5) it follows that $\sum_{j\geq 0}|\kappa(j)|^2 < \infty$ if $|d| < 0.5$ and $y(t)$ is well-defined as the limit in mean square of a covariance-stationary process with spectral density

$$f(\omega) = \frac{\sigma^2|k(\mathrm{e}^{\mathrm{i}\omega})|^2}{2\pi} = \frac{\sigma^2|m(\mathrm{e}^{\mathrm{i}\omega})|^2}{2\pi|1-\mathrm{e}^{\mathrm{i}\omega}|^{2d}}\,.$$

Using the result that $|1-\mathrm{e}^{\mathrm{i}\omega}|^{2d} = |2\sin(\omega/2)|^{2d}$ and $\sin(\omega/2) \sim \omega/2$ as $\omega \to 0$ we find that the spectral density obeys the inverse power law $f(\omega) \sim \sigma^2|m(1)|^2/2\pi\omega^{2d}$ as ω approaches zero. Similarly, the autocovariance function declines at a hyperbolic rate, $\gamma(\tau) \sim C\tau^{2d-1}$, $C \neq 0$, as $\tau \to \infty$. Throughout the paper C will stand for a universal, though not the same, constant. Note that the sequence $\gamma(\tau)$ is absolutely summable if $d \in (-0.5, 0)$, but not if $d \in (0, 0.5)$. In the former case $y(t)$ is sometimes said to have intermediate memory, and in the latter case long memory.

This paper extends the theory of AR approximation to both intermediate and long memory processes, and coincidentally non-invertible processes. Beran et al. (1998) discuss the modelling of finite order AR processes driven by fractional Gaussian noise; here we consider the application of long AR approximations to general fractionally integrated processes. We establish uniform convergence rates for the sample autocovariances and derive corresponding convergence rates for the estimates of AR(h) approximations under regularity conditions that allow for fractionally integrated and non-invertible processes. A central limit theorem for the coefficient estimates is also obtained. All these results are, to the authors knowledge, new to the literature. A major contribution of this paper is to provide a verification of a conjecture of Beran (1992, p. 410) concerning the extension of a result on the predictive optimality of AIC due to Shibata (1980) to fractional and non-invertible processes.

The paper proceeds as follows. Section 2 reviews results from the prediction theory of stochastic processes that provide a rationale for a consideration of AR approximations in more general settings than are currently considered. Section 3 outlines the estimation techniques to be discussed. As well as providing basic background, Sects. 2 and 3 establish further notation and present some basic assumptions. Section 4 lists some of the fundamental results that underly the statistical properties of the estimators considered. The properties of AR approximations are discussed in detail in Sects. 5 and 6 the consequences of noninvertibility are investigated. Section 7 of the paper presents a central limit result for the AR estimator. Section 8 closes the paper with a small simulation study illustrating the (finite sample) practical impact of the (asymptotic) theoretical results obtained. Proofs are assembled together in the appendix.

2 Linear prediction and autoregressive approximation

Since by assumption $y(t)$ is a regular process then we know from a famous result due to Szegö (1939) and Kolmorgorov (1941) that $\int_{-\pi}^{\pi}\log\{f(\omega)\}\,\mathrm{d}\omega > -\infty$ and it is not possible to determine $y(t+1)$ precisely from its own history up to time t, i.e.

$$\sigma^2 = 2\pi \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\{f(\omega)\}\,\mathrm{d}\omega\right\} > 0\,. \tag{6}$$

Factorization of $f(\omega)$ implies that $k(z)$ belongs to the Hardy class $\mathcal{H}^2$, has no zeroes inside the unit circle, and $|k(\mathrm{e}^{\mathrm{i}\omega})|^2 > 0$ almost everywhere (a.e.) where $|k(\mathrm{e}^{\mathrm{i}\omega})|^2 = \lim_{\rho\uparrow 1} |k(\rho\mathrm{e}^{\mathrm{i}\omega})|^2$ (see *inter alia* Anderson, 1971, Sect. 7.6). The condition $k(z) \neq 0$, $|z| \leq 1$, need not hold, however, and $k(z)$ does not have to be invertible in order for there to be an autoregression that yields an appropriate approximation to $y(t)$.

Rewriting the Yule–Walker equations in matrix–vector notation yields $\boldsymbol{\Gamma}_h\boldsymbol{\alpha}_h = -\boldsymbol{\gamma}_h$ where $\boldsymbol{\Gamma}_h = [\gamma(i-j)]_{i,j=1,\ldots,h}$, $\boldsymbol{\alpha}_h = (\alpha_h(1),\ldots,\alpha_h(h))'$ and $\boldsymbol{\gamma}_h = (\gamma(1),\ldots,\gamma(h))'$. Regularity of $y(t)$ implies that $\boldsymbol{\Gamma}_h$ is nonsingular for all h and it follows that $\boldsymbol{\alpha}_h$ is unique. Using the Levinson (1947) and Durbin (1960) algorithm to solve (3) it can be shown that $a_h(z) = \sum_{j=0}^{h} \alpha_h(j) z^j \neq 0$, $|z| \leq 1$, and that σ_h^2 is monotonically decreasing in h. Moreover, basic Hilbert space arguments imply that $\lim_{h\to\infty} \sigma_h^2 = \sigma^2$. The later is an immediate consequence of the following result, see Anderson (1971, Theorem 7.6.6) for example.

Lemma 1 *If $y(t)$ is a linearly regular, covariance-stationary process then the limit of $E[(\epsilon_h(t) - \varepsilon(t))^2]$ as $h \to \infty$ is zero.*

We can therefore think of an infinite autoregressive, AR(∞), representation of $y(t)$ as arising, not by inverting $k(z)$, but from the limit of the AR operators $a_h(z)$ as $h \to \infty$. Indeed, Wold (1938) first derived (1) by fitting autoregressions of ever increasing order.

Example 1 Suppose that $y(t) = \varepsilon(t) - \varepsilon(t-1)$. Then $y(t)$ is regular and the predictor $\bar{y}_{\langle t|t-1,\ldots,t-h\rangle}$ is well defined for all $h \geq 1$. Solving the Yule–Walker equations with

$$\boldsymbol{\Gamma}_h = \sigma^2 \begin{bmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -1 \\ 0 & \ldots & 0 & -1 & 2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\gamma}_h = \sigma^2 \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

it is easily verified that the solution is

$$\boldsymbol{\alpha}_h = \left(\frac{h}{h+1}, \frac{h-1}{h+1}, \ldots, \frac{1}{h+1}\right)' \quad \text{and} \quad \sigma_h^2 = \sigma^2\left\{1 + \frac{1}{h+1}\right\}.$$

Clearly $y(t)$ is not invertible, and although substituting $k(z) = 1 - z$ into the recursions, $\alpha(0) = \kappa(0) = 1$, $\sum_{i=0}^{j} \kappa(i)\alpha(j-i) = 0$, $j = 1, 2, \ldots$, that define the reciprocal $a(z) = \sum_{j\geq 0} \alpha(j) z^j = 1/k(z)$, yields the algebraic result $\alpha(j) = 1$ for all j, the series $\sum_{j\geq 0} y(t-j)$ is not convergent in mean square since

$E[(\sum_{j\geq n} y(t-j))^2] = \sigma^2$ for all $n \geq 1$. Nevertheless, expanding the prediction error process

$$\epsilon_h(t) = \sum_{j=0}^{h} \left\{ 1 - \frac{j}{h+1} \right\} y(t-j)$$

as a function of the innovations and rearranging terms gives

$$\epsilon_h(t) = \varepsilon(t) - \frac{1}{h+1} \sum_{j=1}^{h+1} \varepsilon(t-j)\,.$$

Since $\varepsilon(t)$ are martingale differences the second term in this last expression obeys the law of the iterated logarithm in h and $\epsilon_h(t)$ converges to $\varepsilon(t)$ almost surely (a.s.) as $h \to \infty$, not just in mean square. □

Example 2 Now suppose that $y(t)$ is a fractional noise process, $y(t) = (1-z)^{-d}\varepsilon(t)$, $|d| < 0.5$. Set $\psi(j) = \Gamma(j-d)/[\Gamma(j+1)\Gamma(-d)]$, $j = 1, 2, \ldots$, the coefficients in the binomial expansion of $(1-z)^d$. Then it can be shown that $y(t)$ is the solution to the stochastic difference equation $\sum_{j\geq 0} \psi(j) y(t-j) = \varepsilon(t)$. Thus $y(t)$ admits an infinite AR representation for all $d \in (-0.5, 0.5)$ even though $k(z) = (1-z)^{-d}$ is not invertible in the conventional sense if $-0.5 < d < 0$. Inserting the recursion $\gamma(h) = \gamma(h-1)(h+d-1)/(h+d)$, $h = 1, 2, \ldots$, $\gamma(0) = \sigma^2\Gamma(1-2d)/\Gamma^2(1-d)$, into the Levinson–Durbin algorithm we find that

$$\alpha_h(j) = \psi(j) \left\{ \frac{\Gamma(h+1)\Gamma(h+1-d-j)}{\Gamma(h+1-j)\Gamma(h+1-d)} \right\}$$

for $j = 1, \ldots, h$ and

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

Now, from Sterling's approximation it follows that $\Gamma(x+1+a)/\Gamma(x+1) = x^a\{1+o(1)\}$ for $|a| < 1$ as $x \to \infty$ and from this it is straightforward to show that

$$\frac{\Gamma(h+1)\Gamma(h+1-2d)}{\Gamma^2(h+1-d)} = \left\{ 1 + \frac{(1-d)}{h} \right\}^d \{1+o(1)\}$$

and $\sigma_h^2 \to \sigma^2$ as $h \to \infty$, illustrating directly the consequence of Lemma 1 in this case. It also follows that $|\alpha_h(j) - \psi(j)| \to 0$ for all $j = 1, \ldots, h$ as $h \to \infty$ for, as might have been anticipated, the sequence of autoregressions characterized by $a_h(z)$, $h = 1, 2, \ldots$, converge in mean square to the infinite AR representation $(1-z)^d y(t) = \varepsilon(t)$. □

From the preceding discussion it is apparent that it is the regularity of $y(t)$ that is important in the context of AR modelling rather than invertibility. This observation motivates the following assumption:

Assumption 2 *The series $y(t)$ is a linearly regular, covariance-stationary process with Wold representation $y(t) = \sum_{j\geq 0} \kappa(j)\varepsilon(t-j)$ where $k(z) = m(z)/(1-z)^d$ for $|d| < 0.5$ and $m(z)$ is a causal transfer function with impulse response coefficients satisfying $\sum_{j\geq 0} |\mu(j)| < \infty$.*

3 Model fitting

Let $y(t),\ t = 1, \ldots, T$ denote a realisation of T observations on an observed process and set

$$c_T(r) = c_T(-r) = T^{-1} \sum_{t=r+1}^{T} y(t-r)y(t) \quad r = 0, 1, \ldots, T-1\,, \tag{7}$$

the sample autocovariance function. Substituting $c_T(r)$ for $\gamma(r)$ in the Yule–Walker equations and solving for $\alpha_h(j),\ j = 1, \ldots, h$, and σ_h yields estimates of the parameters in the AR(h) model. We will denote the Yule–Walker estimator and its associated estimates by the use of an over-bar. This estimator can be readily calculated via the Levinson–Durbin recursions, and being based on Toeplitz calculations $\bar{a}_h(z)$ will be stable. The variance estimate $\bar{\sigma}_h^2 = c_T(0) + \sum_{j=1}^{h} \bar{\alpha}_h(j)c_T(j)$ need not minimize the empirical mean squared error however.

Estimating the parameters by directly minimizing the observed mean squared error leads to the least squares estimates of course, which we shall denote by use of a carét. The least squares estimator is obtained by solving the normal equations $\mathbf{M}_h\hat{\boldsymbol{\alpha}}_h = -\mathbf{m}_h$ where

$$\mathbf{M}_h = T^{-1} \sum_{t=1}^{T} \begin{bmatrix} y(t-1) \\ \vdots \\ y(t-h) \end{bmatrix} \big(y(t-1, \ldots, y(t-h)\big)$$

and

$$\mathbf{m}_h = T^{-1} \sum_{t=1}^{T} y(t) \begin{bmatrix} y(t-1) \\ \vdots \\ y(t-h) \end{bmatrix}.$$

By way of contrast with the Yule–Walker estimator, the prediction error variance estimate $\hat{\sigma}_h^2 = T^{-1} \sum_{t=1}^{T} (y(t) + \hat{\alpha}_h(1)y(t-1) + \cdots + \hat{\alpha}_h(h)y(t-h))^2$ minimizes the observed mean squared error, but there is no guarantee that $\hat{a}_h(z)$ will be stable.

In the above expressions the pre-sample values $y(1-h), \ldots, y(0)$ are assumed to be equal to zero. For ease of exposition and notational simplicity summations will continue to be expressed in this manner in what follows. In practice the range of summation for the least squares estimator is often taken as $t = h+1, \ldots, T$. The effects of the elimination of the initial terms will, for given h, be asymptotically negligible. Efficient numerical methods for solving least squares problems of this type are readily available of course.

As will be shown below, $(\hat{\boldsymbol{\alpha}}_h', \hat{\sigma}_h^2)$ and $(\bar{\boldsymbol{\alpha}}_h', \bar{\sigma}_h^2)$ are asymptotically equivalent under the regularity conditions employed here, but they can exhibit quite different finite sample behaviour.

4 Some asymptotic theory

We begin with some asymptotic properties of the basic statistics that form the building blocks of the Yule–Walker and least squares estimators.

Theorem 1 *Suppose that $y(t)$ is a covariance-stationary process that satisfies Assumption* 1 *and Assumption* 2 *and that $H_T = o\{(T/\log T)^{\frac{1}{2}-d'}\}$ where $d' = \max\{0, d\}$. Then with probability one*

$$\max_{0\le\tau\le H_T} |c_T(\tau) - \gamma(\tau)| = O\left\{\left(\frac{\log T}{T}\right)^{\frac{1}{2}-d'}\right\}.$$

This result is of interest in its own right for it indicates that the convergence rate of the autocovariance estimates of a fractional process equals that that obtains in the standard stationary case if $d < 0$ and $y(t)$ has intermediate memory, but if $d > 0$ and $y(t)$ exhibits long memory then the convergence can be much slower.

Let

$$\begin{aligned} c_T(j,k) &= T^{-1}\sum_{t=1}^{T} y(t-j)y(t-k) \\ &= T^{-1}\sum_{t=\max\{j,k\}+1}^{T} y(t-j)y(t-k) \quad j,k = 0,1,\ldots,H_T\,. \end{aligned}$$

Whereas the autocovariance estimates $c_T(\tau)$ are used to calculate $\bar{\boldsymbol{\alpha}}_h$, it is the lag covariances $c_T(j,k)$ that determine the normal equations that define $\hat{\boldsymbol{\alpha}}_h$.

Theorem 2 *Under the same conditions as for Theorem* 1

$$\max_{0\le\tau\le H_T}\max_{|j-k|=\tau} |c_T(j,k) - \gamma(\tau)| = O\{(\log T/T)^{\frac{1}{2}-d'}\} \quad a.s.$$

uniformly in $j,k = 0,1,\ldots,H_T$.

Combining Theorem 1 with Theorem 2 gives rise to the following corollary.

Corollary 1 *If $y(t)$ satisfies assumptions* 1 *and* 2 *then the Yule–Walker and least squares AR estimators $\bar{\boldsymbol{\alpha}}_h$ and $\hat{\boldsymbol{\alpha}}_h$ are asymptotically equivalent and*

$$\begin{aligned} \|\hat{\boldsymbol{\alpha}}_h - \bar{\boldsymbol{\alpha}}_h\|^2 &= O\left\{\left(\frac{h^{1+4d}}{\lambda_{\min}(\boldsymbol{\Gamma}_h^4)}\right)\left(\frac{\log T}{T}\right)^{1-2d'}\right\} \\ &\quad + O\left\{\left(\frac{h}{\lambda_{\min}(\boldsymbol{\Gamma}_h^2)}\right)\left(\frac{\log T}{T}\right)^{1-2d'}\right\} \end{aligned}$$

with probability one.

In the light of Corollary 1 the results that follow will be expressed and proven in terms of the least squares or the Yule–Walker estimates, whichever is most convenient, it being understood that equivalent asymptotic properties will hold for both estimators.

In what follows consideration will be given to the properties of the estimates obtained by fitting an AR(h) model where the order h is allowed to increase with T. In the conventional case where an AR(∞) representation exists it is common

practice to analyse the truncation effects due to employing an AR(h) approximation using Baxter's inequality, Baxter (1962). Since under present assumptions an infinite AR representation is not guaranteed to exist this technique is not available to us. We can, nevertheless, handle the consequences of using an AR(h) approximation if we know something of the relationship between the statistical properties of realizations of the innovations $\varepsilon(t)$ and realizations of the prediction errors $\epsilon_h(t)$.

Theorem 3 *Let $\varepsilon(t)$ and $\epsilon_h(t)$ denote the innovations and prediction errors associated with the minimum mean squared error predictors $\bar{y}_{\langle t|t-1,\ldots,-\infty\rangle}$ and $\bar{y}_{\langle t|t-1,\ldots,t-h\rangle}$ of $y(t)$ where $y(t)$ satisfies Assumptions* 1 *and* 2. *Then*

$$T^{-1}\sum_{t=1}^{T}\varepsilon(t)\{\epsilon_h(t)-\varepsilon(t)\}=O\left\{\left(\frac{\log\log T}{T}\right)^{\frac{1}{2}}\right\}$$

with probability one, uniformly in h.

Theorem 3 implies that

$$T^{-1}\sum_{t=1}^{T}\epsilon_h(t)^2-T^{-1}\sum_{t=1}^{T}\varepsilon(t)^2=T^{-1}\sum_{t=1}^{T}\{\epsilon_h(t)-\varepsilon(t)\}^2 +O\{(\log\log T/T)^{\frac{1}{2}}\},\tag{8}$$

which provides an empirical counterpart to the result that $\sigma_h^2\geq\sigma^2$ in that the first term on the right hand side of (8) will converge to $E[(\epsilon_h(t)-\varepsilon(t))^2]\geq 0$, by ergodicity, and thus for T sufficiently large $T^{-1}\sum_{t=1}^{T}\epsilon_h(t)^2$ will be bounded below by $T^{-1}\sum_{t=1}^{T}\varepsilon(t)^2$, with the difference converging to zero as h increases, see Lemma 1. It will be seen that (8) plays an important role in determining the behaviour of model selection devices for large T, as does the following result.

Theorem 4 *Let $y(t)$ and $\epsilon_h(t)$ be as in Theorem* 3. *Then uniformly in $h\leq H_T$*

$$\max_{1\leq j\leq h}T^{-1}\sum_{t=1}^{T}\epsilon_h(t)y(t-j)=O\left\{\left(\frac{\log T}{T}\right)^{\frac{1}{2}-d'}\right\}\quad a.s..$$

Theorem 4 is the empirical counterpart of the result that the prediction error $\epsilon_h(t)$ is, by construction, orthogonal to $y(t-1),\ldots,y(t-h)$, that is $E[\epsilon_h(t)y(t-j)]=0$, $j=1,\ldots,h$.

5 Autoregressive modelling

In practice, of course, neither $\varepsilon(t)$ nor $\epsilon_h(t)$ can be observed and their properties will have to be deduced by fitting AR models to the data. We begin, therefore, by first establishing the consistency of the coefficient estimates of the AR(h) model to those of the AR(h) approximation to the process.

Theorem 5 *If $y(t)$ is a stationary process that satisfies Assumption* 1 *and Assumption* 2 *then uniformly in $h \leq H_T$*

$$\sum_{j=1}^{h} |\hat{\alpha}_h(j) - \alpha_h(j)|^2 = O\left\{\left(\frac{h}{\lambda_{\min}(\mathbf{\Gamma}_h^2)}\right)\left(\frac{\log T}{T}\right)^{1-2d'}\right\} \quad a.s..$$

The following theorem relates to the residuals

$$\hat{\epsilon}_h(t) = \sum_{j=0}^{h} \hat{\alpha}_h(j) y(t-j)$$

as estimates of the prediction errors $\epsilon_h(t)$.

Theorem 6 *Under the same assumptions as for Theorem* 5

$$T^{-1}\sum_{t=1}^{T} \epsilon_h(t)\{\hat{\epsilon}_h(t) - \epsilon_h(t)\} = O\left\{\left(\frac{h}{\lambda_{\min}(\mathbf{\Gamma}_h)}\right)\left(\frac{\log T}{T}\right)^{1-2d'}\right\}$$

with probability one, uniformly in $h \leq H_T$.

Comparison of Theorem 6 with Theorem 3 indicates that whereas the deviation of $\epsilon_h(t)$ from $\varepsilon(t)$ relative to the magnitude of $\varepsilon(t)$, as measured by their covariation, converges to zero at a rate that is independent of d the same is not true of the corresponding relationship between $\hat{\epsilon}_h(t)$ and $\epsilon_h(t)$. The relevance of this observation stems from the fact that it is common practice to determine the order of the model to be employed by minimizing a model selection criterion of the form

$$\mathrm{SC}_T(h) = \log(\hat{\sigma}_h^2) + \frac{hC_T}{T}$$

over the range $h = 0, 1, \ldots, M_T$ where $\hat{\sigma}_h^2 = T^{-1}\sum_{t=1}^{T} \hat{\epsilon}_h(t)^2$ and $C_T > 0$ is chosen by the practitioner such that $C_T/T \to 0$ as $T \to \infty$, as is $M_T < H_T$. If $C_T = 2$ we have AIC, if $C_T = \log T$ we have BIC, Schwarz (1978), and setting $C_T = \log\log T$ we obtain the criterion advanced in Hannan and Quinn (1979).

Consider the function

$$L_T(h) = (\sigma_h^2 - \sigma^2) + \frac{h\sigma^2}{T}.$$

Shibata (1980) introduced $L_T(h)$ as a figure of merit in the context of fitting AR models to a truly infinite-order process. Shibata shows that if an AR(h) model is fitted to a stationary Gaussian process that has an AR(∞) representation and it is used to predict an independent realization of the same process then the difference between the mean squared prediction error of the fitted model and the innovation variance converges in probability to $L_T(h)$. Thus, if $y(t)'$ denotes an independent realization of the process $y(t)$ then

$$E\left[\left(y(t)' + \sum_{i=1}^{h} \hat{\alpha}_h(i) y(t-i)'\right)^2\right]$$

$$= \sigma_h^2 + E\left[\sum_{j=1}^{h}\sum_{i=1}^{h} (\hat{\alpha}_h(i) - \alpha_h(i))(\hat{\alpha}_h(j) - \alpha_h(j))\gamma(j-i)\right]$$

and given that $T^{\frac{1}{2}}(\hat{\boldsymbol{\alpha}}_h - \boldsymbol{\alpha}_h) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \sigma^2 \boldsymbol{\Gamma}_h^{-1})$ then the asymptotic expectation and probability limit of the second term is $h\sigma^2/T$. Noting that $E[(y(t)' - \bar{y}'_{\langle t|t-1,\ldots,t-h\rangle})^2] = \sigma_h^2 \geq \lim_{h\to\infty} E[\epsilon_h(t)'^2] = \sigma^2$ we can see that the first term of $L_T(h)$ measures the fit of the model and the second reflects the inaccuracy or uncertainty in the determination of the parameters of $\bar{y}'_{\langle t|t-1,\ldots,t-h\rangle} = -\sum_{j=1}^{h} \alpha_h(j) y(t-j)'$. Now, $L_T(h)$ is bounded below by $L_T(h_T^*)$ in the range $h = 0, 1, \ldots, M_T$ where $L_T(h_T^*) = \min_{h=1,\ldots,M_T} L_T(h)$ and Shibata defines a sequence of selected orders h'_T as being efficient if $\lim_{T\to\infty} L_T(h'_T)/L_T(h_T^*) = 1$.

Although the regularity conditions imposed by Shibata (1980) are too restrictive to be applicable here a similar rationale for consideration of $L_T(h)$ can be given. Observe also that by Theorem 6 the empirical difference $T^{-1}\sum_{t=1}^{T} \hat{\epsilon}_h(t)^2 - T^{-1}\sum_{t=1}^{T} \varepsilon(t)^2$ equals

$$T^{-1}\sum_{t=1}^{T} \epsilon_h(t)^2 - T^{-1}\sum_{t=1}^{T} \varepsilon(t)^2 + T^{-1}\sum_{t=1}^{T} (\hat{\epsilon}_h(t) - \epsilon_h(t))^2 + O\left\{ h\left(\frac{\log T}{T}\right)^{1-2d'} \right\}.$$

The limit of $T^{-1}\sum_{t=1}^{T} \epsilon_h(t)^2 - T^{-1}\sum_{t=1}^{T} \varepsilon(t)^2$ is $\sigma_h^2 - \sigma^2$ and the third term

$$T^{-1}\sum_{t=1}^{T} (\hat{\epsilon}_h(t) - \epsilon_h(t))^2 = T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{h}\sum_{i=1}^{h} (\hat{\alpha}_h(i) - \alpha_h(i)) \times (\hat{\alpha}_h(j) - \alpha_h(j)) y(t-i) y(t-j).$$

is a consistent estimate of $\sum_{j=1}^{h}\sum_{i=1}^{h} (\hat{\alpha}_h(i) - \alpha_h(i))(\hat{\alpha}_h(j) - \alpha_h(j))\gamma(j-i)$ by Theorem 2. Thus $L_T(h)$ can be viewed as providing a limiting bound to the empirical difference in the mean squared prediction error and the innovation variance.

Set

$$\bar{L}_T(h) = \log\left(1 + \frac{\sum_{t=1}^{T} \epsilon_h(t)^2 - \sum_{t=1}^{T} \varepsilon(t)^2}{\sum_{t=1}^{T} \varepsilon(t)^2}\right) + \frac{h}{T}$$

and let $\bar{h}_T^*$ denote a sequence of non-negative integers at each of which the minimum of $\bar{L}_T(h)$ with respect to h is attained, that is

$$\bar{L}_T(\bar{h}_T^*) = \min_{0\leq h\leq M_T} \bar{L}_T(h)$$

or equivalently $\bar{h}_T^* = \operatorname{argmin}_{0,1,\ldots,M_T} \bar{L}_T(h)$.

Theorem 7 *If $y(t)$ is a covariance-stationary process that satisfies Assumptions* 1 *and* 2 *then*

$$\lim_{T\to\infty} \left| \frac{\sigma^2 \bar{L}_T(\bar{h}_T^*)}{L_T(h_T^*)} - 1 \right| = 0$$

almost surely where $h_T^* = \operatorname{argmin}_{0,1,\ldots,M_T} L_T(h)$.

The criterion $\bar{L}_T(h)$ is unfeasible, but letting $\mathrm{AIC}_T(h)$ denote the criterion $\mathrm{SC}_T(h)$ when $C_T = 2$ we can deduce from Theorem 8 presented immediately below that

$$\max_{1\le h\le M_T}\left|\mathrm{AIC}_T(h)-\bar{L}_T(h)\right| \le \left|\log\left(T^{-1}\sum_{t=1}^{T}\varepsilon(t)^2\right)\right| + \frac{M_T}{T} + O\left\{\left(\frac{M_T}{\lambda_{\min}(\mathbf{\Gamma}_{M_T})}\right)\left(\frac{\log T}{T}\right)^{1-2d'}\right\}. \quad (9)$$

Theorem 8 *Under the same assumptions as for Theorem 7*

$$\mathrm{SC}_T(h) = \log\left(T^{-1}\sum_{t=1}^{T}\varepsilon(t)^2\right) + \log\left(1+\frac{\sum_{t=1}^{T}\epsilon_h(t)^2-\sum_{t=1}^{T}\varepsilon(t)^2}{\sum_{t=1}^{T}\varepsilon(t)^2}\right) + \frac{hC_T}{T} + O\left\{\left(\frac{h}{\lambda_{\min}(\mathbf{\Gamma}_h)}\right)\left(\frac{\log T}{T}\right)^{1-2d'}\right\}$$

with probability one, uniformly in $h = 0, 1, \ldots, H_T$.

Note that if $\left(M_T/\lambda_{\min}(\mathbf{\Gamma}_{M_T})\right)(\log T/T)^{1-2d'}$ converges to zero as $T \to \infty$ then the only non-vanishing term on the right hand side of (9) (the first term) is independent of both d and h. We can therefore conclude that $h_T^{\mathrm{AIC}}/\bar{h}_T^{*} \to 1$ as $T \to \infty$ where h_T^{AIC} is the autoregressive order determined by $\mathrm{AIC}_T(h)$ provided that $M_T/\lambda_{\min}(\mathbf{\Gamma}_{M_T}) = o\left\{(T/\log T)^{1-2d'}\right\}$.

Theorem 9 *Suppose that* $y(t)$ *is a covariance-stationary process that satisfies Assumptions* 1 *and* 2*, and let*

$$h_T^{\mathrm{AIC}} = \mathrm{argmin}_{0,1,\ldots,M_T}\mathrm{AIC}_T(h)$$

where $\lim_{T\to\infty}\left(M_T/\lambda_{\min}(\mathbf{\Gamma}_{M_T})\right)(\log T/T)^{1-2d'} = 0$. *Then the* $\mathrm{AR}(h_T^{\mathrm{AIC}})$ *model is asymptotically efficient in the sense that*

$$L_T(h_T^{\mathrm{AIC}}) = L_T(h_T^{*})\{1+o(1)\}$$

almost surely as $T \to \infty$.

Alternative methods of autoregressive order determination that do not share the same structure as $\mathrm{SC}_T(h)$ above have been proposed in the literature. The criterion autoregressive transfer function suggested by Parzen (1974) and the mean squared prediction error criterion of Mallows (1973), for example. Parzen's criterion can be expressed as

$$\mathrm{CAT}_T(h) = 1 - \frac{(T-h)\tilde{\sigma}^2}{T\hat{\sigma}_h^2} + \frac{h}{T}$$

and Mallow's statistic

$$\mathrm{MC}_T(h) = T\left(\frac{\hat{\sigma}_h^2}{\tilde{\sigma}^2}-1\right)+2h$$

where

$$\tilde{\sigma}^2 = 2\pi \exp\left\{(2\pi N)^{-1}\sum_{j=1}^{N}\log\left\{(2\pi)^{-1}\sum_{\tau=1-T}^{T-1}c_T(\tau)\cos(2\pi j\tau/T)\right\}+\gamma'\right\},$$

$\gamma' = 0.57721$ (Euler's constant) and $N = [(T-1)/2]$, a nonparametric estimate of the innovation variance constructed from the periodogram by analogy with (6). Simple algebra shows that

$$\mathrm{CAT}_T(h) - \mathrm{CAT}_T(h-1) = \left\{\frac{(T-h+1)\hat{\sigma}_h^2 - (T-h)\hat{\sigma}_{h-1}^2}{T\hat{\sigma}_h^2\hat{\sigma}_{h-1}^2}\right\}\tilde{\sigma}^2 + \frac{1}{T}$$

and

$$\mathrm{MC}_T(h) - \mathrm{MC}_T(h-1) = T\left(\frac{\hat{\sigma}_h^2 - \hat{\sigma}_{h-1}^2}{\tilde{\sigma}^2}\right) + 2\,,$$

while from Theorem 6 and expression (8) it follows that

$$\mathrm{AIC}_T(h) - \mathrm{AIC}_T(h-1) = \frac{\hat{\sigma}_h^2 - \hat{\sigma}_{h-1}^2}{T^{-1}\sum_{t=1}^{T}\varepsilon(t)^2} + \frac{2}{T} + o(\hat{\sigma}_h^2 - \hat{\sigma}_{h-1}^2)\,.$$

Similarly, it is straightforward to show that the final prediction error criterion

$$\mathrm{FPE}_T(h) = \left(\frac{T+h}{T-h}\right)\hat{\sigma}_h^2$$

introduced by Akaike (1970) satisfies $\log \mathrm{FPE}_T(h) = \mathrm{AIC}_T(h) + O(T^{-2})$. Thus, bare remainder terms, we can anticipate that these criteria will move together and will be minimized at the same value of h. This suggests, and it can be shown, that $\mathrm{CAT}_T(h)$, $\mathrm{MC}_T(h)$ and $\mathrm{FPE}_T(h)$ will also be asymptotically efficient selection criteria.

6 The non-invertible case

Heretofore, specific reference has not been made to the non-invertible case. This is because the existence of unit roots does not invalidate our basic assumptions and the results presented thus far will hold regardless. It is apparent from Corollary 1 and Theorems 5, 6, 8 and 9, however, that the behaviour of $\lambda_{\min}(\mathbf{\Gamma}_h)$ will play an important role via its influence on the various orders of magnitude presented in these results.

To investigate the impact of $\lambda_{\min}(\mathbf{\Gamma}_h)$ in further detail it is necessary to impose additional structure upon the process. This is done in the following lemma.

Lemma 2 *Suppose Assumption* 2 *holds. If* $m(z) \neq 0$, $|z| \leq 1$, *then as* $h \to \infty$

$$\lambda_{\min}(\mathbf{\Gamma}_h) \begin{cases} \geq \inf_\omega \sigma^2 |m(\mathrm{e}^{i\omega})|^2 / |1 - \mathrm{e}^{i\omega}|^{2d} > 0 & \text{when } d \geq 0, \\ = O(h^{2d}) & \text{when } d < 0. \end{cases}$$

If, however, there exists a set of frequencies $\theta_j \in (0, \pi)$ *and numbers* $\nu_j > 0$, *such that for each* $j = 1, \ldots, n$, $|m(\mathrm{e}^{i\omega})| \sim |\phi_j(\omega)||\omega - \theta_j|^{\nu_j}$ *as* $\omega \to \theta_j$, *where* $\phi_j(\omega)$ *is slowly varying at* θ_j, *then*

$$\lambda_{\min}(\mathbf{\Gamma}_h) \begin{cases} = O(h^{-2\max\{\nu_1, \nu_2, \ldots, \nu_n\}}) & \text{when } d \geq 0, \\ = O(h^{-2\max\{-d, \nu_1, \nu_2, \ldots, \nu_n\}}) & \text{when } d < 0. \end{cases}$$

The factors $|\phi_j(\omega)||\omega - \theta_j|^{\nu_j}$ introduced in Lemma 2 mimic the effect of an operator $m(z)$ with roots on the unit circle and the lemma indicates how zeroes in $|k(\mathrm{e}^{i\omega})|^2$ can translate into a measure of the proximity of $\lambda_{\min}(\mathbf{\Gamma}_h)$ to zero as $h \to \infty$. When taken in conjunction with the results presented in Sects. 4 and 5, Lemma 2 leads to a consideration of terms of order $O\{h^{1+4q}(\log T/T)^{1-2d'}\}$, or smaller, where $q \geq 0$. In particular, to operationalize the estimation procedures examined above a value for M_T must be chosen that satisfies the requirements of Theorem 9, namely, M_T must be such that $M_T^{1+4q}(\log T/T)^{1-2d'} \to 0$ as $T \to \infty$ for all possible values of q and d, both of which are unknown to the practitioner of course. One such choice is $M_T = [c(\log T)^a]$, the integer part of $c(\log T)^a$ for some $a \geq 1$ and $c > 0$.

It can be argued that processes observed in the real world are unlikely to exhibit spectral zeroes, and hence that the supplementary conditions of Lemma 2, although technically convenient, are unrealistic. As pointed out by a referee, a more meaningful extension might be to consider situations where $k(z) = m(z)/[(1 - z)^d (1 + z)^s \prod_{j=1}^n (1 - 2\cos(\theta_j)z + z^2)^{\nu_j}]$ and $y(t)$ is a member of the class of Gegenbauer processes, Grey et al. (1989). An analysis of AR approximations to Gegenbauer processes would take us too far afield in this paper.

7 A central limit theorem

We now wish to establish the asymptotic distribution of the AR estimator $\hat{\boldsymbol{\alpha}}_h$, or equivalently $\bar{\boldsymbol{\alpha}}_h$, under the regularity conditions considered in this paper. The difficulty is that the convergence rate of the autocovariance estimates upon which the coefficient estimators are based depends on the value of d. If $-0.5 < d < 0.25$ then the asymptotic distribution of $T^{\frac{1}{2}}(c_T(\tau) - \gamma(\tau))$ is normal, but when $d \geq 0.25$ the autocovariances are no longer $\sqrt{T}$ consistent. See Hosking (1996) for details. Given that in practice d will not be known, we seek a transformation that will lead to a conventional $\sqrt{T}$ consistent, asymptotic normal approximation in which the parameters of the approximating distribution can be determined without explicit knowledge of d.

An interesting feature of the autocovariances noted by Hosking (1996, p.268) is that when $d \in [0.25, 0.5)$ they contain a common slowly varying component that can be removed by differencing. Indeed, in this case $T^{\frac{1}{2}}(c_T(\tau) - \gamma(\tau)) =$

$T^{2d-\frac{1}{2}}\varrho_T - \zeta_T(\tau)$ where ϱ_T and $\zeta_T(\tau)$ have non-degenerate limiting distributions, a Rosenblatt process and a Normal distribution respectively. Thus, if $\upsilon(\tau) = \gamma(\tau) - \gamma(0)$ and $u_T(\tau) = c_T(\tau) - c_T(0)$ then from Hosking (1996, Theorem 5) it follows that, whatever is the value of d, $T^{\frac{1}{2}}\{u_T(\tau) - \upsilon(\tau)\}$, for $\tau = 1, \ldots, h$, will have a non-degenerate multivariate Normal limiting distribution with mean zero and covariance matrix

$$\boldsymbol{\Delta}_h = \left[\frac{1}{2}\sum_{s=-\infty}^{\infty}(\gamma(s) - \gamma(s-k) - \gamma(s+l) + \gamma(s-k+l))^2 + K_4\upsilon(k)\upsilon(l)\right]_{k,l=1,\ldots,h}$$

where K_4 is the fourth cumulant of $\varepsilon(t)$. This suggests that some form of differencing, or centering, may be necessary to achieve our desired outcome and ultimately gives rise to the following result.

Theorem 10 *Let* $\mathbf{C}_h = \mathbf{I}_h - h^{-1}\mathbf{1}\mathbf{1}'$ *denote the hth order centering matrix where* $\mathbf{1} = (1, 1, \ldots, 1)'$ *is the h element sum vector. Set* $\boldsymbol{\Phi}_h = \mathbf{I}_h + \mathbf{P}_h$ *where* $\mathbf{P}_h$ *equals*

$$\begin{bmatrix} \alpha_h(2) & \alpha_h(3) & \cdots & \cdots & \alpha_h(h) & 0 \\ \alpha_h(3) & \alpha_h(4) & \cdots & \alpha_h(h) & 0 & 0 \\ \vdots & \vdots & & 0 & 0 & 0 \\ \alpha_h(h-1) & \alpha_h(h) & & & & \\ \alpha_h(h) & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \cdots & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 & 0 \\ \alpha_h(1) & 0 & \cdots & \cdots & 0 & 0 \\ \alpha_h(2) & \alpha_h(1) & 0 & \cdots & \cdots & 0 \\ \vdots & & & & & \vdots \\ \alpha_h(h-2) & \alpha_h(h-3) & \cdots & \alpha_h(1) & 0 & 0 \\ \alpha_h(h-1) & \alpha_h(h-2) & \cdots & \cdots & \alpha_h(1) & 0 \end{bmatrix}.$$

Then for any h component vectors $\boldsymbol{\lambda}_h$, *where* $1 \leq h \leq M_T$, $M_T = [c(\log T)^a]$, $a \geq 1, c > 0$, *such that* $0 < \|\boldsymbol{\lambda}_h\| < \infty$ *the scalars* $T^{\frac{1}{2}}\boldsymbol{\lambda}_h'\mathbf{C}_h\boldsymbol{\Gamma}_h(\bar{\boldsymbol{\alpha}}_h - \boldsymbol{\alpha}_h)$ *form a triangular array equal to* $\boldsymbol{\beta}_{h,T} + \boldsymbol{\rho}_{h,T}$ *where* $\boldsymbol{\rho}_{h,T} = o_p(1)$ *and* $\boldsymbol{\beta}_{h,T}/\eta_h \xrightarrow{\mathcal{L}} N(0,1)$ *where* $\eta_h^2 = \boldsymbol{\lambda}_h'(\mathbf{C}_h\boldsymbol{\Phi}_h\boldsymbol{\Delta}_h\boldsymbol{\Phi}_h'\mathbf{C}_h)\boldsymbol{\lambda}_h$.

A corollary of Theorem 10, that follows from Bernstein's Lemma, is that if $\boldsymbol{\lambda}_h = \mathbf{C}_h\boldsymbol{\lambda}_h$ then a zero mean normal distribution with variance $\boldsymbol{\lambda}_h'(\boldsymbol{\Phi}_h\boldsymbol{\Delta}_h\boldsymbol{\Phi}_h')\boldsymbol{\lambda}_h$ can be used as an asymptotic approximation to the large sample distribution of $T^{\frac{1}{2}}\boldsymbol{\lambda}_h'\boldsymbol{\Gamma}_h(\bar{\boldsymbol{\alpha}}_h - \boldsymbol{\alpha}_h)$. The condition that $\boldsymbol{\lambda}_h = \mathbf{C}_h\boldsymbol{\lambda}_h$ implies, of course, that the elements of $\boldsymbol{\lambda}_h$ must sum to zero.

8 Empirical illustrations

This section of the paper reports the outcome of some simulation experiments designed to illustrate the theoretical results and properties discussed above. The experiments are based on three data generating mechanisms, the non-invertible moving average process $y(t) = \varepsilon(t) - \varepsilon(t-1)$ of Example (1) and two cases of the fractional noise process $y(t) = \varepsilon(t)/(1-z)^d$ of Example (2) with $d = 0.125$ and 0.375. For all three processes $\varepsilon(t)$ is standardized, Gaussian white noise. For each process the sample sizes $T = 100, 200, 500, 1,000$ were considered and the values and figures presented here are all based on $R = 1,000$ replications. In light of the discussion in Sect. 3, the behaviour of both the Yule–Walker and least squares estimates is reported here. The properties of the estimation procedure proposed by Burg (1968) were also investigated. Burg's estimator produces an estimate of $a_h(z)$ that is, like the Yule–Walker estimate, stable, but the finite sample properties of Burg's estimator prove to be almost indistinguishable from those of the least squares estimate. Detailed particulars of Burg's algorithm and other features of the simulations reported here can be found in Grose and Poskitt (2005), where a more extensive range of simulation experiments are documented.

Figure 1 presents the relative frequency of occurrence of the different orders given by h_T^{AIC} when $T = 100$ and the value $M_T = 2\sqrt{T} = [(\log T)^{1.962}] = 20$ is

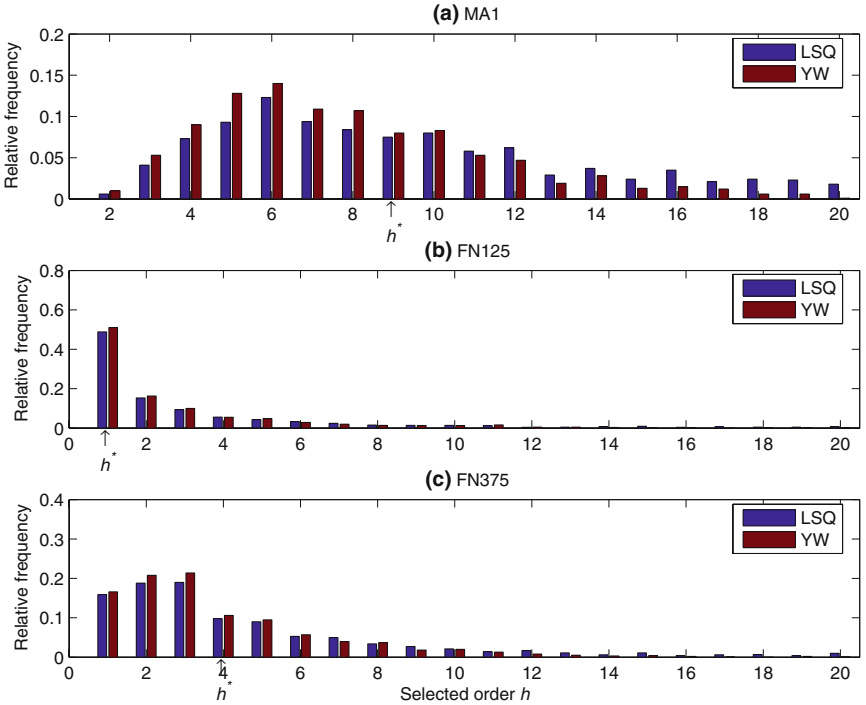

Fig. 1 Relative frequency of occurrence of h_T^{AIC}, $T = 100$, for **a** $y(t) = \varepsilon(t) - \varepsilon(t-1)$, **b** $y(t) = \varepsilon(t)/(1-z)^{0.125}$ and **c** $y(t) = \varepsilon(t)/(1-z)^{0.375}$

employed. At this sample size the dispersion of h_T^{AIC} about h_T^* is quite large, for all three processes, and there are no obvious differences in the observed performance of the two estimators. As T increases, however, first, the orders chosen by h_T^{AIC} become more concentrated around h_T^*, in accord with the predictions of Sect. 5, and secondly, the values of h_T^{AIC} produced by the Yule–Walker estimator are generally smaller than those given by least squares. The latter feature is shown in Table 1, which gives the average value of h_T^{AIC} compared to h_T^* for each model and sample size.

Some indication of why the Yule–Walker procedure produces smaller values of h_T^{AIC} than least squares can be found in Table 2, which presents the empirical variance and the empirical bias of the estimates of the partial autocorrelation $\alpha_h(h)$ for $h = h_T^*$. For all three processes and at all sample sizes the Yule–Walker estimate exhibits a larger bias than does least squares. The bias pushes the estimate of $\alpha_h(h)$ towards the origin, leading to smaller values of h being selected. This behaviour is most noticeable in the case of the non-invertible moving average process, where the bias of the Yule–Walker estimate exceeds that of least squares by an order of magnitude even when $T = 1,000$.

Table 1 Average value of h_T^{AIC} compared to h_T^*

Process	T	h_T^*	h_T^{AIC}(LS)	h_T^{AIC}(YW)
$y(t) = \varepsilon(t) - \varepsilon(t-1)$	100	9	9.218	7.855
	200	13	13.209	11.936
	500	21	21.208	19.904
	1,000	31	30.993	29.29
$(1-z)^{0.125}y(t) = \varepsilon(t)$	100	1	3.27	2.716
	200	1	3.181	2.982
	500	2	4.224	4.091
	1,000	4	5.422	5.331
$(1-z)^{0.375}y(t) = \varepsilon(t)$	100	4	4.658	3.916
	200	5	6.078	5.484
	500	8	8.694	8.326
	1,000	12	11.992	11.72

Table 2 Partial autocorrelation estimates for $h = h_T^*$

Process	T	$\alpha_h(h)$	LS		YW	
			Variance	Bias	Variance	Bias
$y(t) = \varepsilon(t)$	100	0.1	0.010335	−0.00278	0.008645	−0.016423
$-\varepsilon(t-1)$	200	0.071429	0.004958	−0.00151	0.004601	−0.011105
	500	0.045455	0.001991	−0.00028	0.001871	−0.005715
	1,000	0.031250	0.001059	−0.00087	0.001041	−0.004680
$(1-z)^{0.125}y(t)$	100	−0.142857	0.011214	0.001727	0.010982	0.003122
$= \varepsilon(t)$	200	−0.142857	0.005617	0.000256	0.005558	0.000974
	500	−0.066667	0.002027	0.002196	0.002014	0.002440
	1,000	−0.032258	0.001051	0.002046	0.001043	0.002185
$(1-z)^{0.375}y(t)$	100	−0.103448	0.012071	0.028433	0.011374	0.035746
$= \varepsilon(t)$	200	−0.081081	0.005793	0.009894	0.005540	0.013757
	500	−0.049180	0.002156	0.007040	0.002104	0.008661
	1,000	−0.032258	0.001024	0.003841	0.001007	0.004590

These results are consistent with findings reported elsewhere. Indeed, it is known that in finite samples the stability of $\bar{a}_h(z)$ can give rise to significant biases that are not present with $\hat{a}_h(z)$. Tjøstheim and Paulsen (1983) present empirical evidence of this phenomenon and show that for finite autoregressions the first term in an asymptotic expansion of the bias of $\bar{\boldsymbol{\alpha}}_h$ has order of magnitude $O(T^{-1})$, with the size of the constant varying inversely with the distance of the zeroes of the true AR operator from the unit circle. Thus, when the data generating mechanism shows strong autocorrelation the bias in the Yule–Walker coefficient estimates can be substantial. This bias is known to feed through to other quantities of interest such as the prediction error variance, Paulsen and Tjøstheim (1985), and estimates of power spectra, Lysne and Tjøstheim (1987). Given that fractional processes can display long-range dependence with autocovariances that decay much slower than exponentially, similar effects can be anticipated when employing the Yule–Walker estimates under the current scenario *a-fortiori*.

Figures 2 and 3 illustrate the impact of the distributional properties discussed in Sect. 7 for the two fractional noise processes.

Figure 2 plots the empirical distribution of $h^{-1}\sum_{j=1}^{h}(\bar{\alpha}_h(j) - \alpha(j))$ and $h^{-1}\sum_{j=1}^{h}(\hat{\alpha}_h(j) - \alpha(j))$, the average deviation or coefficient error of the Yule–Walker and least squares estimators, when $h = h_T^*$ and $T = 500$. The density estimates are constructed from the simulated values using a Gaussian kernel with bandwidth equal to 75% of the over-smoothed bandwidth i.e., $0.75\,\xi\,\sqrt[5]{(243/35R)}$ where ξ is the standard deviation observed over the replications, see Wand and Jones (1995). Comparison of the estimated distributions to a normal curve of error with zero mean and variance ξ^2 indicates that when $d = 0.125$ the distribution of the average coefficient error is reasonably close to normal for both estimators. When $d = 0.375$, however, the presence of the Rosenblatt process in the

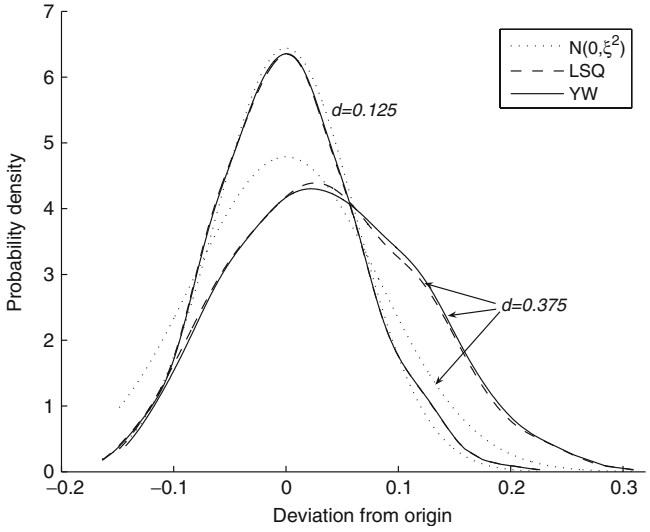

Fig. 2 Empirical distribution of $h^{-1}\sum_{j=1}^{h}(\bar{\alpha}_h(j) - \alpha(j))$ for fractional noise processes $y(t) = \varepsilon(t)/(1-z)^d$ with $d = 0.125$ and $d = 0.375$, $h = h_T^*$ and $T = 500$

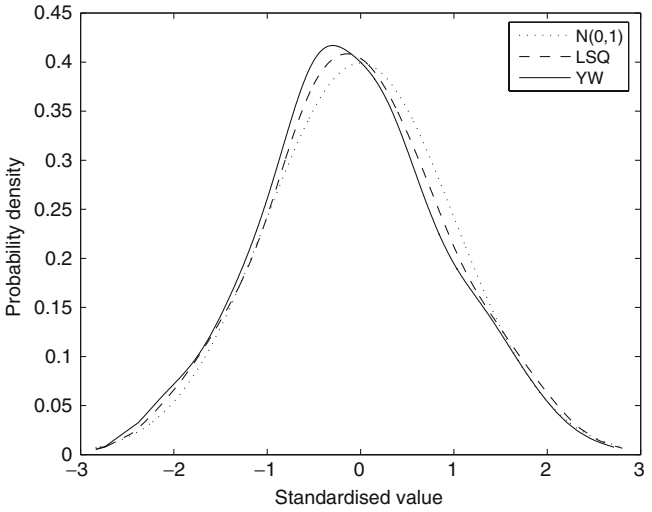

Fig. 3 Observed distribution of $\varphi_{\lambda,T}$ for $y(t) = \varepsilon(t)/(1-z)^{0.375}$, when $\boldsymbol{\lambda}_h' = (1, 0, \ldots, 0, -1)$, $T = 1,000$

limiting behaviour of the underlying statistics 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 $T = 1,000$ and does not disappear asymptotically.By way of contrast, Fig. 3 plots the observed distributions of $\bar{\varphi}_{\lambda,T} = T^{\frac{1}{2}}\boldsymbol{\lambda}_h'\boldsymbol{\Gamma}_h(\bar{\boldsymbol{\alpha}}_h - \boldsymbol{\alpha}_h)/(\boldsymbol{\lambda}_h'(\boldsymbol{\Phi}_h\boldsymbol{\Delta}_h\boldsymbol{\Phi}_h')\boldsymbol{\lambda}_h)^{\frac{1}{2}}$, $h = h_T^*$, and, to use an obvious notation, $\hat{\varphi}_{\lambda,T}$, obtained from realizations of the process $y(t) = \varepsilon(t)/(1-z)^{0.375}$ when $\boldsymbol{\lambda}_h' = (1, 0, \ldots, 0, -1)$ and $T = 1,000$. The empirical distributions are overlayed with a standard normal density. Although some bias is still apparent even at this sample size, more so for $\bar{\boldsymbol{\alpha}}_h$ than $\hat{\boldsymbol{\alpha}}_h$, kurtosis and skewness of the type observed previously with this process has now gone and the operation of Theorem 10 is apparent.

Appendix: Proofs

Proof of Theorem 1 Assume that $\frac{1}{4} < d < \frac{1}{2}$. By Theorem 3 of Hosking (1996) $E[(c_T(\tau) - \gamma(\tau))^2] = O(T^{-2(1-2d)})$ and from Chebychev's inequality

$$\Pr\left(|c_T(\tau) - \gamma(\tau)| > \delta\left(\frac{\log T}{T}\right)^{\frac{1}{2}-d}\right) \leq \frac{C}{\delta^2}\frac{1}{(T\log T)^{1-2d}}\,.$$

Now set $\Delta_\tau(T) = (c_T(\tau) - \gamma(\tau))(\log T/T)^{d-\frac{1}{2}}$. Then for $T' = N'^{4/(1-2d)}$

$$\sum_{N'=1}^{\infty}\Pr\left(\max_{|\tau|\leq H_{T'}}|\Delta_\tau(T')| > \delta\right) \leq \frac{C}{\delta^2}\sum_{N'=1}^{\infty}\left(\frac{1-2d}{4\log N'}\right)^{3(1-2d)/2}\frac{1}{N'^2} < \infty$$

and by the Borel–Cantelli lemma $\Delta_\tau(T') \to 0$ a.s. uniformly in τ, $|\tau| \leq H_{T'}$.

Let $N^2 = T'$. Then for all T such that $N^2 < T < (N+1)^2$

$$\Delta_\tau(T) = \left(\frac{T}{\log T}\frac{\log(N+1)^2}{(N+1)^2}\right)^{\frac{1}{2}(1-2d)} \frac{(N+1)^2}{T}\Delta_\tau((N+1)^2) - \left(\frac{T}{\log T}\right)^{\frac{1}{2}(1-2d)} \frac{1}{T}\sum_{t=T+1}^{(N+1)^2}(y(t)y(t-\tau)-\gamma(\tau))$$

and

$$\max_{|\tau|\leq H_T}|\Delta_\tau(T)| \leq \max_{|\tau|\leq H_T}\left(\frac{T}{\log T}\frac{\log(N+1)^2}{(N+1)^2}\right)^{\frac{1}{2}(1-2d)}\left(1+\frac{1}{N}\right)^2|\Delta_\tau((N+1)^2)| + \max_{|\tau|\leq H_T}\left(\frac{T}{\log T}\right)^{\frac{1}{2}(1-2d)}\frac{1}{T}\left|\sum_{t=T+1}^{(N+1)^2}(y(t)y(t-\tau)-\gamma(\tau))\right|. \tag{10}$$

But

$$\frac{T}{\log T}\frac{\log(N+1)^2}{(N+1)^2} \leq \frac{\log(N+1)}{\log N} \to 1 \text{ as } N \to \infty$$

and by what has already been shown it follows that the first term on the right hand side of (10) converges to zero a.s..

Moreover, using Chebychev's inequality once more we can bound

$$\Pr\left(\max_{|\tau|\leq H_T}\left|\sum_{t=T+1}^{(N+1)^2}(y(t)y(t-\tau)-\gamma(\tau))\right| \geq \delta(\log T)^{\frac{1}{2}(1-2d)}T^{\frac{1}{2}(1+2d)}\right) \tag{11}$$

by

$$\left(\frac{(N+1)^2}{\log(N+1)^2}\right)^{\frac{1}{2}(1-2d)} \cdot \frac{C}{\delta^2(\log T)^{(1-2d)}T^{(1+2d)}} \cdot (2N+1)^{4d}$$

where the first factor accounts for the maximum being taken over $H_T < H_{(N+1)^2}$ terms and the last factor arises because the sum contains $(N+1)^2 - T < (2N+1)$ summands. Thus we can deduce that the probability in (11) is less than

$$\frac{(2N+1)^{4d}(N+1)^{(1-2d)}}{(2\log N)^{3(1-2d)/2}N^{2+4d}} \leq \frac{18}{N^{1+2d}}$$

and hence, via the Borel–Cantelli lemma, that the second term of (10) converges to zero with probability one since the series $\{N^{-(1+2d)}\}$ is convergent.

A similar proof using the method of subsequences can be employed to establish the result for the remaining cases, $d = \frac{1}{4}$, when $E[(c_T(\tau)-\gamma(\tau))^2] = O(\log T/T)$, and $d \in (-\frac{1}{2}, \frac{1}{4})$, when $E[(c_T(\tau)-\gamma(\tau))^2] = O(T^{-1})$. □

Proof of Theorem 2 The following relationship exists between the elements of the sequence $c_T(\tau)$ and the $c_T(j,k)$ for $j-k=\tau=0,\pm1,\ldots,\pm H_T$:

$$T\{c_T(\tau)-c_T(j,k)\}=\sum_{s=B_T(j,k)}^{T} y(s-|\tau|)y(s)=D_T(j,k,\tau) \tag{12}$$

where $B_T(j,k)=T+1-\min\{j\,,k\}$. Note that $D_T(j,k,\tau)$ contains $\min\{j\,,k\}$, or at most H_T, summands. Now, since $\varepsilon(t)$ has finite fourth moment

$$E[y(t)^4]=\sigma^4\left(\sum_{j=0}^{\infty}\kappa(j)^2\right)^2+K_4\sum_{j=0}^{\infty}\kappa(j)^4<\infty,$$

where K_4 is the fourth cumulant of $\varepsilon(t)$, and the variance of $D_T(j,k,\tau)$ is dominated by CH_T^2 uniformly in $j\,,k=0,1,\ldots,H_T$. Thus

$$\Pr\left(\max_{|\tau|\le H_T}|D_T(j,k,\tau)|\ge\delta T\right)<H_T\frac{CH_T^2}{\delta^2T^2}\le\frac{C}{(\log T)^{3/2}T^{\frac12}} \tag{13}$$

where the final inequality follows since for $0<d<\frac12$ $H_T=o\{(T/\log T)^{\frac12}(\log T/T)^d\}$ and $(\log T/T)^d<1$ and $H_T=o\{(T/\log T)^{\frac12}\}$ for $-\frac12<d\le0$.

Along the subsequence $T'=N'^4$ it follows that $\lim_{N'\to\infty}\max_{|\tau|\le H_{T'}}T'^{-1}|D_{T'}(j,k,\tau)|=0$ a.s. because $\{N'^{-2}\}$ is a convergent series. Furthermore, letting $N^2=T'$, then for all T between N^2 and $(N+1)^2$ we can bound $\left|N^{-2}D_{N^2}(j,k,\tau)-T^{-1}D_T(j,k,\tau)\right|$ by

$$\left|\frac{(T-N^2)D_{N^2}(j,k,\tau)}{TN^2}\right|+\left|\frac{\sum_{s=B_{N^2}(j,k)}^{B_T(j,k)}y(s-|\tau|)y(s)-\sum_{s=N^2+1}^{T}y(s-|\tau|)y(s)}{T}\right|. \tag{14}$$

The first term in (14) converges to zero uniformly in j, k and τ by what has already been proved since $(T-N^2)/TN^2\le(2N+1)/N^4$ and the second term converges similarly via an application of Chebychev's inequality and the Borel–Cantelli lemma.

To show the latter, consider the case $d \in (-\frac{1}{2}, \frac{1}{4})$ for example. By Theorems 1 and 2 and Theorem 3 of Hosking (1996) the variance of the numerator can be bounded by $C(T - N^2) \leq C(2N + 1)$ uniformly in $j, k = 0, 1, \ldots, H_T$ so

$$\Pr\left(\max_{|\tau|\leq H_T}\left|\sum_{s=B_{N^2}(j,k)}^{B_T(j,k)} y(s-|\tau|)y(s) - \sum_{s=N^2+1}^{T} y(s-|\tau|)y(s)\right| \geq \delta T\right) < H_T \cdot \frac{C(2N+1)}{\delta^2 T^2}$$

and $H_T(2N + 1)/T^2 \leq 2(N + 1)(2N + 1)/N^4 \log(N + 1) < 6/N^2$. □

For convenience and completeness we now state a result taken from Poskitt (2000).

Lemma 3 *Let* $\mathbf{A}_T$ *and* $\mathbf{B}_T$ *denote two* $h \times h$ (*stochastic*) *matrices such that* $\|\mathbf{A}_T - \mathbf{B}_T\|$ *equals* $O(C_T)$ *where* $C_T \to 0$ *as* $T \to \infty$ *and suppose that* $\liminf_{T\to\infty} \lambda_{min}[\mathbf{B}_T] \geq \delta_h > 0$. *Then* $\mathbf{A}_T$ *is nonsingular for all* T *sufficiently large and* $\|\mathbf{A}_T^{-1} - \mathbf{B}_T^{-1}\| = (\delta_h)^{-2} O(C_T)$.

Proof of Corollary 1 Let $\bar{\mathbf{\Gamma}}_h = [c_T(i - j)]_{i,j=1,\ldots,h}$ and $\bar{\boldsymbol{\gamma}}_h = (c_T(1), \ldots, c_T(h))'$. Then

$$\begin{aligned}\bar{\boldsymbol{\alpha}}_h - \hat{\boldsymbol{\alpha}}_h &= \bar{\mathbf{\Gamma}}_h^{-1}\bar{\boldsymbol{\gamma}}_h - \mathbf{M}_h^{-1}\mathbf{m}_h \\ &= (\bar{\mathbf{\Gamma}}_h^{-1} - \mathbf{M}_h^{-1})\mathbf{m}_h + \bar{\mathbf{\Gamma}}_h^{-1}(\bar{\boldsymbol{\gamma}}_h - \mathbf{m}_h)\,. \end{aligned} \tag{15}$$

From Theorem 1 it follows that

$$\limsup_{T\to\infty} \|\mathbf{\Gamma}_h - \bar{\mathbf{\Gamma}}_h\|^2 = O\left\{h^2\left(\frac{\log T}{T}\right)^{1-2d'}\right\} = o(1)$$

and hence that

$$\begin{aligned}\liminf_{T\to\infty} \lambda_{\min}(\bar{\mathbf{\Gamma}}_h) &\geq \lambda_{\min}(\mathbf{\Gamma}_h) - \limsup_{T\to\infty} \|\mathbf{\Gamma}_h - \bar{\mathbf{\Gamma}}_h\| \\ &= \lambda_{\min}(\mathbf{\Gamma}_h) > 0\,.\end{aligned}$$

It can also be shown that $\liminf_{T\to\infty} \lambda_{\min}(\mathbf{M}_h) \geq \lambda_{\min}(\mathbf{\Gamma}_h)$, using the previous argument in conjunction with Theorem 2. From Lemma 3 it follows that the norm $\|\bar{\mathbf{\Gamma}}_h^{-1} - \mathbf{M}_h^{-1}\|$ is $O\{(h/\lambda_{\min}(\mathbf{\Gamma}_h^2))(\log T/T)^{\frac{1}{2}(1-2d')}\}$. Now the first term on the right hand side of (15) can be bounded in norm by $\|(\bar{\mathbf{\Gamma}}_h^{-1} - \mathbf{M}_h^{-1})\| \cdot (\|\boldsymbol{\gamma}_h\| + \|\mathbf{m}_h - \boldsymbol{\gamma}_h\|)$, which equals

$$O\left\{\left(\frac{h}{\lambda_{\min}}(\mathbf{\Gamma}_h^2)\right)\left(\frac{\log T}{T}\right)^{\frac{1}{2}(1-2d')}\left(h^{\frac{1}{2}(4d-1)} + h^{\frac{1}{2}}\left(\frac{\log T}{T}\right)^{\frac{1}{2}(1-2d')}\right)\right\},$$

and the norm of the second term of (15) is $O\{(h^{\frac{1}{2}}/\lambda_{\min}(\mathbf{\Gamma}_h))(\log T/T)^{\frac{1}{2}-d'}\}$. □

Proof of Theorem 3 Let $r(z) = \sum_{j\geq 1} \rho(j)z^j = a_h(z)k(z) - 1$. Then $\epsilon_h(t) - \varepsilon(t) = r(z)\varepsilon(t) = \sum_{j\geq 1} \rho(j)\varepsilon(t-j)$. From Parseval's relation

$$\sum_{j\geq 1} \rho(j)^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |a_h(\mathrm{e}^{\mathrm{i}\omega})k(\mathrm{e}^{\mathrm{i}\omega}) - 1|^2 \mathrm{d}\omega = \sigma^{-2} E[(\epsilon_h(t) - \varepsilon(t))^2] < \infty$$

and therefore we can conclude that $T^{-1} \sum_{t=1}^{T} \varepsilon(t)\{\epsilon_h(t) - \varepsilon(t)\} = O\{(\log\log T/T)^{\frac{1}{2}}\}$ by Theorem 5.3.5. of Hannan and Deistler (1988). □

Proof of Theorem 4 By definition $\epsilon_h(t) = \sum_{j=0}^{h} \alpha_h(j)y(t-j)$. Simple substitution now gives us

$$T^{-1} \sum_{t=1}^{T} \epsilon_h(t)y(t-r) = \sum_{j=0}^{h} \alpha_h(j) T^{-1} \sum_{t=1}^{T} y(t-j)y(t-r) = \sum_{j=0}^{h} \alpha_h(j) c_T(j,r),$$

which by Theorem 2 equals

$$\sum_{j=0}^{h} \alpha_h(j)[\gamma(j-r) + O\{(\log T/T)^{\frac{1}{2}-d'}\}].$$

Since $\alpha_h(j)$, $j = 1, \ldots, h$, solve the Yule–Walker equations $\sum_{j=0}^{h} \alpha_h(j)\gamma(j-r) = 0$ for $r = 1, \ldots, h$. Moreover, $a_h(z) \neq 0$, $|z| \leq 1$, and there exists constants $C < \infty$ and $\zeta < 1$ such that $|\alpha_h(j)| < C\zeta$ and $\sum_{j=0}^{h} |\alpha_h(j)| < C(1-\zeta^{h+1})/(1-\zeta) < C(1-\zeta)^{-1}$ so that $\sum_{j=0}^{h} \alpha_h(j) O\{(\log T/T)^{\frac{1}{2}-d'}\} = O\{(\log T/T)^{\frac{1}{2}-d'}\}$. Hence the desired result. □

Proof of Theorem 5 Substituting $\epsilon_h(t) = \sum_{j=0}^{h} \alpha_h(j)y(t-j)$ into the normal equations yields the expression

$$\mathbf{M}_h(\hat{\boldsymbol{\alpha}}_h - \boldsymbol{\alpha}_h) = T^{-1} \sum_{t=1}^{T} \epsilon_h(t) \begin{bmatrix} y(t-1) \\ \vdots \\ y(t-h) \end{bmatrix}.$$

It follows that

$$\|\hat{\boldsymbol{\alpha}}_h - \boldsymbol{\alpha}_h\|^2 \leq \frac{1}{\lambda_{\min}(\boldsymbol{\Gamma}_h^2)} \sum_{j=1}^{h} \left(T^{-1} \sum_{t=1}^{T} \epsilon_h(t)y(t-j) \right)^2$$

and hence that $\|\hat{\boldsymbol{\alpha}}_h - \boldsymbol{\alpha}_h\|^2 = (1/\lambda_{\min}(\boldsymbol{\Gamma}_h^2)) O\{h(\log T/T)^{1-2d'}\}$ by Theorem 4. □

Proof of Theorem 6 From the definition of $\hat{\epsilon}_h(t)$ and $\epsilon_h(t)$ we get

$$\hat{\epsilon}_h(t) - \epsilon_h(t) = \sum_{j=1}^{h} \{\hat{\alpha}_h(j) - \alpha_h(j)\} y(t-j)$$

and from the Cauchy-Schwartz inequality, Theorem 4 and Theorem 5 we have

$$\begin{aligned} |T^{-1} \sum_{t=1}^{T} \epsilon_h(t)\{\hat{\epsilon}_h(t) - \epsilon_h(t)\}| &= |T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{h} \{\hat{\alpha}_h(j) - \alpha_h(j)\} \epsilon_h(t) y(t-j)| \\ &\leq \left[\|\hat{\boldsymbol{\alpha}}_h - \boldsymbol{\alpha}_h\|^2 \cdot \sum_{j=1}^{h} \left(T^{-1} \sum_{t=1}^{T} \epsilon_h(t) y(t-j) \right)^2 \right]^{\frac{1}{2}} \\ &= O\left\{ \frac{h}{\lambda_{\min}(\boldsymbol{\Gamma}_h)} \left(\frac{\log T}{T} \right)^{1-2d'} \right\}, \end{aligned} \tag{16}$$

giving the result of the theorem. □

Proof of Theorem 7 Since $h/T \to 0$ as $T \to \infty$ and $\sigma_h^2 - \sigma^2$ is monotonically decreasing in h it follows that $h_T^* \to \infty$ as $T \to \infty$. Similarly, for T sufficiently large the behaviour of

$$\log\left(1 + \frac{\sum_{t=1}^{T} \epsilon_h(t)^2 - \sum_{t=1}^{T} \varepsilon(t)^2}{\sum_{t=1}^{T} \varepsilon(t)^2} \right)$$

will be determined by that of $\log(1 + (\sigma_h^2 - \sigma^2)/\sigma^2)$. The latter is decreasing in h and it follows that $\bar{h}_T^* \to \infty$ as $T \to \infty$. Indeed, expanding $\bar{L}_T(h)$ using $\log(1+x) = \sum_{r \geq 1} (-)^{r-1} x^r / r$ and recognizing from Lemma 1 and Eq. (8) that $T^{-1} \sum_{t=1}^{T} \epsilon_h(t)^2 - T^{-1} \sum_{t=1}^{T} \varepsilon(t)^2 = E[(\epsilon_h(t) - \varepsilon(t))^2] + o(1)$ will converge to zero as h increases we find that

$$\bar{L}_T(h) = \frac{\sum_{t=1}^{T} \epsilon_h(t)^2 - \sum_{t=1}^{T} \varepsilon(t)^2}{\sum_{t=1}^{T} \varepsilon(t)^2} + \frac{h}{T} + o\left\{ \frac{\sum_{t=1}^{T} \epsilon_h(t)^2 - \sum_{t=1}^{T} \varepsilon(t)^2}{\sum_{t=1}^{T} \varepsilon(t)^2} \right\}$$

and

$$\begin{aligned} \left| \bar{L}_T(h) - \frac{L_T(h)}{\sigma^2} \right| &\leq \left| \left(\frac{\sigma^2}{\sum_{t=1}^{T} \varepsilon(t)^2} \right) \left(\frac{\sum_{t=1}^{T} \epsilon_h(t)^2 - \sum_{t=1}^{T} \varepsilon(t)^2}{\sigma_h^2 - \sigma^2} \right) - 1 \right| \\ &\quad \times \left(\frac{\sigma_h^2 - \sigma^2}{\sigma^2} \right) + o\left\{ \frac{\sum_{t=1}^{T} \epsilon_h(t)^2 - \sum_{t=1}^{T} \varepsilon(t)^2}{\sum_{t=1}^{T} \varepsilon(t)^2} \right\} \\ &= o\left\{ (\sigma_h^2 - \sigma^2)/\sigma^2 \right\}. \end{aligned} \tag{17}$$

From (17) we conclude that

$$\left| \frac{\sigma^2 \bar{L}_T(h)}{L_T(h)} - 1 \right| = \left(\frac{\sigma^2}{(\sigma_h^2 - \sigma^2) + h/T} \right) o\left\{ (\sigma_h^2 - \sigma^2)/\sigma^2 \right\} = o(1).$$

By definition of $\bar{h}_T^*$ and h_T^* as the minimizing values of, respectively, $\bar{L}_T(h)$ and $L_T(h)$ over the common range $h = 0, 1, \ldots, M_T$ it now follows that

$$\frac{\sigma^2 \bar{L}_T(\bar{h}_T^*)}{L_T(h_T^*)} = \frac{\bar{L}_T(\bar{h}_T^*)}{\bar{L}_T(h_T^*)\{1 + o(1)\}} \leq 1 + o(1)$$

and

$$\frac{\sigma^2 \bar{L}_T(\bar{h}_T^*)}{L_T(h_T^*)} = \frac{L_T(\bar{h}_T^*)\{1 + o(1)\}}{L_T(h_T^*)} \geq 1 + o(1)\,,$$

which implies that

$$\left| \frac{\sigma^2 \bar{L}_T(\bar{h}_T^*)}{L_T(h_T^*)} - 1 \right| = o(1)\,,$$

as required. □

Proof of Theorem 8 The least squares residual $\hat{\epsilon}_h(t)$ is by construction orthogonal to $y(t-1), \ldots, y(t-h)$ for $t = 1, \ldots, T$ and thus

$$T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_h(t)\{\hat{\epsilon}_h(t) - \epsilon_h(t)\} = \sum_{j=1}^{h} \{\hat{\alpha}_h(j) - \alpha_h(j)\} T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_h(t) y(t-j) = 0\,.$$

The residual mean square can therefore be re-expressed as

$$T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_h(t)^2 = T^{-1} \sum_{t=1}^{T} \epsilon_h(t)^2 + T^{-1} \sum_{t=1}^{T} \epsilon_h(t)\{\hat{\epsilon}_h(t) - \epsilon_h(t)\}$$

and the right hand side equals

$$T^{-1} \sum_{t=1}^{T} \epsilon_h(t)^2 + O\left\{ \frac{h}{\lambda_{\min}(\mathbf{\Gamma}_h)} \left(\frac{\log T}{T} \right)^{1-2d'} \right\}$$

by Theorem 6. A trivial re-expression of $T^{-1} \sum_{t=1}^{T} \epsilon_h(t)^2$ as the sum of $T^{-1} \sum_{t=1}^{T} \varepsilon(t)^2$ and $T^{-1} \sum_{t=1}^{T} \epsilon_h(t)^2 - T^{-1} \sum_{t=1}^{T} \varepsilon(t)^2$, used in conjunction with the usual McLaurin expansion of $\log(1+x)$ as given above, now yields the result that

$$\log T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_h(t)^2 = \log T^{-1} \sum_{t=1}^{T} \varepsilon(t)^2 + \log\left(1 + \frac{\sum_{t=1}^{T} \epsilon_h(t)^2 - \sum_{t=1}^{T} \varepsilon(t)^2}{\sum_{t=1}^{T} \varepsilon(t)^2} \right) + O\left\{ \left(\frac{\sum_{t=1}^{T} \varepsilon(t)^2}{T} \right)^{-1} \frac{h}{\lambda_{\min}(\mathbf{\Gamma}_h)} \left(\frac{\log T}{T} \right)^{1-2d'} \right\}.$$

But $T^{-1} \sum_{t=1}^{T} \varepsilon(t)^2$ converges to σ^2 a.s., giving the result as stated in the theorem. □

Proof of Lemma 2 Let $\mathbf{x}$ denote a unit eigenvector associated with the eigenvalue $\lambda_{\min}(\mathbf{\Gamma}_h)$. Then $\lambda_{\min}(\mathbf{\Gamma}_h) = \mathbf{x}'\mathbf{\Gamma}_h\mathbf{x}$. Now consider the following cases:

First, $m(z) \neq 0$, $|z| \leq 1$, and $d \geq 0$. Then

$$\lambda_{\min}(\mathbf{\Gamma}_h) = \int_{-\pi}^{\pi} \left| \sum_{s=1}^{h} x_s \exp(-\mathrm{i}\omega s) \right|^2 f(\omega)\mathrm{d}\omega$$

$$\geq \inf_{\omega} f(\omega) \int_{-\pi}^{\pi} \left| \sum_{s=1}^{h} x_s \exp(-\mathrm{i}\omega s) \right|^2 \mathrm{d}\omega = 2\pi \inf_{\omega} f(\omega)$$

and $2\pi f(\omega) = \sigma^2 |m(\mathrm{e}^{\mathrm{i}\omega})|^2 / |1 - \mathrm{e}^{\mathrm{i}\omega}|^{2d} > 0$ for all ω.

Second, $m(z) \neq 0$, $|z| \leq 1$, and $d < 0$. An adaptation of the circulant imbedding argument underlying the simulation technique of Davies and Harte (1987) yields the result that $\mathbf{\Gamma}_h = \mathbf{U}^*\mathbf{\Lambda}\mathbf{U}$ where the $(2h+1) \times h$ matrix

$$\mathbf{U} = [(2h+1)^{-\frac{1}{2}} \exp(-\mathrm{i}2\pi(j-1)(k-1)/(2h+1))]_{j=1,\ldots,2h+1,\, k=1,\ldots,h},$$

and $\mathbf{\Lambda} = 2\pi \mathrm{diag}\{f_h(\omega_0), \ldots, f_h(\omega_{2h})\}$,

$$f_h(\omega) = \frac{1}{2\pi} \sum_{\tau=-h}^{h} \gamma(\tau) \exp(-\mathrm{i}\omega\tau), \quad \omega_j = 2\pi j/(2h+1), \quad j = 0, \ldots, 2h\,,$$

as can be readily verified via straightforward, if somewhat tedious, algebra. Hence $\lambda_{\min}(\mathbf{\Gamma}_h) = \mathbf{x}'\mathbf{\Gamma}_h\mathbf{x} = \mathbf{w}^*\mathbf{\Lambda}\mathbf{w}$ where $\mathbf{w} = \mathbf{U}\mathbf{x}$ and $\|\mathbf{w}\| = 1$ since $\|\mathbf{x}\| = 1$ and $\mathbf{U}^*\mathbf{U} = \mathbf{I}$.

From the Rayleigh–Ritz theorem it follows that

$$\lambda_{\min}(\mathbf{\Gamma}_h) \geq \min\{2\pi f_h(\omega_0), \ldots, 2\pi f_h(\omega_{2h})\}. \tag{18}$$

But $f_h(\omega) = f(\omega) - \int_{-\pi}^{\pi} \{f(\omega) - f(\theta)\} D_h(\omega - \theta)\mathrm{d}\theta$ where $D_h(\theta) = \sin((h + \frac{1}{2})\theta)/2\pi \sin(\theta/2)$, Dirichlet's kernel, and since $f(\cdot)$ is absolutely integrable and continuous a.e. it follows from the Riemann–Lebesgue lemma that

$$\lim_{h\to\infty} \sup_{0\leq\omega\leq\pi} \int_{-\pi}^{\pi} \{f(\omega) - f(\theta)\} D_h(\omega - \theta)\mathrm{d}\theta = 0.$$

We can therefore conclude that for every $\delta > 0$

$$|\mathrm{argmin}_{\omega\in\{\omega_0,\ldots,\omega_{2h}\}} f_h(\omega) - \mathrm{argmin}_{\omega\in\{\omega_0,\ldots,\omega_{2h}\}} f(\omega)| < \delta \tag{19}$$

for h sufficiently large. Now, $f(\omega_0) = 0$ and $f(\omega_i) > 0$, $i = 1, \ldots, 2h$, and therefore

$$\min\{2\pi f_h(\omega_0), \ldots, 2\pi f_h(\omega_{2h})\} = 2\pi f_h(\omega_0) = \sum_{\tau=-h}^{h} \gamma(\tau) = O(h^{2d})$$

for all h sufficiently large.

Third, $|m(e^{i\omega})| \sim |\phi_j(\omega)||\omega-\theta_j|^{\nu_j}$ as $\omega \to \theta_j$, $j = 1, \ldots, n$. Then by assumption $2\pi f(\omega) \sim \sigma^2|m(1)|^2/\omega^{2d}$ as $\omega \to 0$ and $2\pi f(\omega) \sim \sigma^2 |2\sin(\omega/2)|^{-2d}|\phi_j(\omega)|^2|\omega-\theta_j|^{2\nu_j}$ as $\omega \to \theta_j$. Arguing as above we also know that (18) holds and that (19) obtains for h sufficiently large.

Set $j_i(h) = [(2h+1)\theta_i/2\pi]$ for $i = 1, \ldots, n$. Both $|\omega_{j_i(h)}-\theta_i|$ and $|\omega_{(j_i(h)+1)}-\theta_i|$ are less than $2\pi/(2h+1)$. Thus, as $h \to \infty$, $2\pi f(\omega_{j_i(h)}) \sim \sigma^2|2\sin(\omega_{j_i(h)}/2)|^{-2d}|\phi_j(\omega_{j_i(h)})|^2|\omega_{j_i(h)}-\theta_j|^{2\nu_j}$, which is $O(\{2\pi/(2h+1)\}^{2\nu_j})$, and similarly $f(\omega_{(j_i(h)+1)}) \sim O(\{2\pi/(2h+1)\}^{2\nu_j})$. Now let $\bar{\omega}_m = \operatorname{argmin}_{\omega\in\{\omega_0,\ldots,\omega_{2h}\}} f_h(\omega)$. Either $\bar{\omega}_m = \omega_0$ and

$$\min\{2\pi f_h(\omega_0), \ldots, 2\pi f_h(\omega_{2h})\} = \sum_{\tau=-h}^{h} \gamma(\tau) = O(h^{2d}),$$

or $\bar{\omega}_m$ equals $\omega_{j_i(h)}$ or $\omega_{(j_i(h)+1)}$ for some $i \in \{1, \ldots, n\}$ and

$$\min\{2\pi f_h(\omega_0), \ldots, 2\pi f_h(\omega_{2h})\} = \sum_{\tau=-h}^{h} \int_{-\pi}^{\pi} g_i(\omega) e^{i\omega\tau} d\omega = O(h^{-2\nu_i})$$

where $g_i(\omega) = f(\bar{\omega}_m + \omega)$. It follows that $f_h(\bar{\omega}_m) = O(h^{-2m})$ where $m = \max\{\nu_1\, \nu_2, \ldots, \nu_n\}$ if $d \geq 0$ and $m = \max\{-d, \nu_1\, \nu_2, \ldots, \nu_n\}$ if $d < 0$. □

Proof of Theorem 10 By definition of $\bar{\boldsymbol{\alpha}}_h$

$$\begin{aligned} T^{\frac{1}{2}}\mathbf{C}_h\boldsymbol{\Gamma}_h(\bar{\boldsymbol{\alpha}}_h - \boldsymbol{\alpha}_h) = {} & T^{\frac{1}{2}}\mathbf{C}_h(\bar{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h) - T^{\frac{1}{2}}\mathbf{C}_h(\bar{\boldsymbol{\Gamma}}_h - \boldsymbol{\Gamma}_h)\boldsymbol{\alpha}_h \\ & + T^{\frac{1}{2}}\mathbf{C}_h(\boldsymbol{\Gamma}_h - \bar{\boldsymbol{\Gamma}}_h)(\bar{\boldsymbol{\alpha}_h} - \boldsymbol{\alpha}_h). \end{aligned}$$

Multiplying through by $\boldsymbol{\lambda}_h'$ and rearranging terms on the right hand side yields the result that $T^{\frac{1}{2}}\boldsymbol{\lambda}_h'\mathbf{C}_h\boldsymbol{\Gamma}_h(\bar{\boldsymbol{\alpha}}_h - \boldsymbol{\alpha}_h)$ equals $\boldsymbol{\beta}_{hT} + \boldsymbol{\rho}_{hT}$ where $\boldsymbol{\rho}_{hT} = T^{\frac{1}{2}}\boldsymbol{\lambda}_h'\mathbf{C}_h(\boldsymbol{\Gamma}_h - \bar{\boldsymbol{\Gamma}}_h)(\bar{\boldsymbol{\alpha}_h} - \boldsymbol{\alpha}_h)$ and $\boldsymbol{\beta}_{hT} = T^{\frac{1}{2}}\boldsymbol{\lambda}_h'\mathbf{C}_h[(\bar{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h) - (\bar{\boldsymbol{\Gamma}}_h - \boldsymbol{\Gamma}_h)\boldsymbol{\alpha}_h]$. Hence we are lead to consider the limiting behaviour of $\boldsymbol{\beta}_{hT}$ and $\boldsymbol{\rho}_{hT}$.

Let $\mathbf{D}_h = [u_T(i-j) - \upsilon(i-j)]_{i,j=1,\ldots,h}$ and $\mathbf{d}_h = (u_T(1) - \upsilon(1), \ldots, u_T(h) - \upsilon(h))'$. Then it is a simple exercise to show that $\mathbf{C}_h(\bar{\boldsymbol{\Gamma}}_h - \boldsymbol{\Gamma}_h) = \mathbf{C}_h\mathbf{D}_h$ and $\mathbf{C}_h(\bar{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h) = \mathbf{C}_h\mathbf{d}_h$ and it follows that $\boldsymbol{\beta}_{hT} = T^{\frac{1}{2}}\boldsymbol{\lambda}_h'\mathbf{C}_h[\mathbf{d}_h - \mathbf{D}_h\boldsymbol{\alpha}_h]$. Writing $\mathbf{D}_h$ as $\mathbf{T}_h + \mathbf{T}_h'$ where $\mathbf{T}_h$ is the lower triangular Toeplitz matrix with first column $(0, u_T(1) - \upsilon(1), \ldots, u_T(h-1) - \upsilon(h-1))'$ we find, after some straightforward rearrangement, that $\mathbf{D}_h\boldsymbol{\alpha}_h = \mathbf{P}_h\mathbf{d}_h$ and hence that $\mathbf{d}_h - \mathbf{D}_h\boldsymbol{\alpha}_h = \boldsymbol{\Phi}_h\mathbf{d}_h$. Thus $\boldsymbol{\beta}_{h,T} = T^{\frac{1}{2}}\boldsymbol{\lambda}'\boldsymbol{\Phi}_h\mathbf{d}_h$. By Theorem 5 of Hosking (1996), however, $T^{\frac{1}{2}}\mathbf{d}_h$ converges in distribution to $N(\mathbf{0}, \boldsymbol{\Delta}_h)$. We can therefore conclude that $\boldsymbol{\beta}_{h,T}/\eta_h \xrightarrow{\mathcal{L}} N(\mathbf{0}, 1)$ where $\eta_h^2 = \boldsymbol{\lambda}_h'(\mathbf{C}_h\boldsymbol{\Phi}_h\boldsymbol{\Delta}_h\boldsymbol{\Phi}_h'\mathbf{C}_h)\boldsymbol{\lambda}_h$, as stated.

Similarly, $\boldsymbol{\rho}_{hT} = -T^{\frac{1}{2}}\boldsymbol{\lambda}_h'(\mathbf{C}_h\mathbf{D}_h)(\bar{\boldsymbol{\alpha}_h} - \boldsymbol{\alpha}_h)$. Corollary 1 and Theorem 5 imply that $\|(\bar{\boldsymbol{\alpha}_h} - \boldsymbol{\alpha}_h)\| = O(M_T^{1+4q}(\log T/T)^{1-2d'})$ where $q \geq 0$ and from Theorem 5 of Hosking (1996), once again, we have that $T^{\frac{1}{2}}\mathbf{D}_h = O_p(1)$. This leads to the conclusion that $\boldsymbol{\rho}_{hT} = o_p(1)$ and completes the proof. □

Acknowledgements This research was partially supported by the Australian Research Council under Grant DP0452717. The computations leading to the results presented in Sect. 8 where carried out by Simone Grose using MATLAB. I am grateful to Simone Grose, Andy Tremayne, Jon Wellner and an anonymous referee for helpful and constructive comments on previous versions of this paper.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of Institute of Statistical Mathematics*, *21*, 243–247.
Akaike, H. (1970). Statistical predictor identification. *Annals of Institute of Statistical Mathematics*, *22*, 203–217.
Anderson, T. W. (1971). *The statistical analysis of time series*. New York: Wiley.
Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics*, *73*, 5–59.
Baxter, G. (1962). An asymptotic result for the finite predictor. *Mathematica Scandinavica*, *10*, 137–144.
Beran, J. (1992). Statistical methods for data with long-range dependence. *Statistical Science*, *7*, 404–427.
Beran, J. (1994). *Statistics for long memory processes*. New York: Chapman and Hall.
Beran, J., Bhansali, R. J., Ocker, D. (1998). On unified model selection for stationary and nonstationary short- and long-memory autoregressive processes. *Biometrika*, *85*, 921–934.
Burg, J. (1968). A new analysis technique for time series data. Technical report, Advanced Study Institute on Signal Processing, N.A.T.O., Enschede, Netherlands.
Davies, R. B., Harte, D. S. (1987). Tests for hurst effect. *Biometrika*, *74*, 95–101.
Durbin, J. (1960). The fitting of time series models. *Review of International Statistical Institute*, *28*, 233–244.
Granger, C. W. J., Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, *1*, 15–29.
Grey, H. L., Zhang, N.-F., Woodward, W. A. (1989). On generalized fractional processes. *Journal of Time Series Analysis*, *10*, 233–257.
Grose, S., Poskitt, D. S. (2005). Empirical evidence on nonstandard autogregressive approximations. Working Paper, Monash University.
Hannan, E. J., Deistler, M. (1988). *The statistical theory of linear systems*. New York: Wiley.
Hannan, E. J., Quinn, 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.
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.
Lysne, D., Tjøstheim, D. (1987). Loss of spectral peaks in autoregressive spectral estimation. *Biometrika*, *74*, 200–206.
Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, *15*, 661–675.
Parzen, E. (1974). Some recent advances in time series modelling. *IEEE Transactions on Automatic Control*, *AC-19*, 723–730.
Paulsen, J., Tjøstheim, D. (1985). On the estimation of residual variance and order in autoregressive time series. *Journal of the Royal Statistical Society*, *B-47*, 216–228.
Poskitt, D. S. (2000). Strongly consistent determination of cointegrating rank via canonical correlations. *Journal of Business and Economic Statistics*, *18*, 71–90.
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.
Szegö, G. (1939). *Orthogonal polynomials*. American Mathematical Society Colloquium Publication.

Tjøstheim, D., Paulsen, J. (1983). Bias of some commonly-used time series estimators. *Biometrika*, *70*, 389–400.
Wand, M. P., Jones, M. C. (1995). *Kernel smoothing*. London: Chapman and Hall.
Wold, H. (1938). *The analysis of stationary time series* (2nd ed.). Uppsala: Almquist and Wicksell.
Yule, G. U. (1921). On the time correlation problem. *Journal of the Royal Statistical Society*, *84*, 497–510.