ISSN 1440-771X

MONASH University

Australia

Department of Econometrics
and Business Statistics

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

Properties of the Sieve Bootstrap for Fractionally Integrated and Non-Invertible Processes

D.S. Poskitt

July 2006

Working Paper 12/06

Properties of the Sieve Bootstrap for Fractionally Integrated and Non-Invertible Processes

D. S. Poskitt*

Department of Econometrics and Business Statistics, Monash University

Abstract

In this paper we will investigate the consequences of applying the sieve bootstrap under regularity conditions that are sufficiently general to encompass both fractionally integrated and non-invertible processes. The sieve bootstrap is obtained by approximating the data generating process by an autoregression whose order h increases with the sample size T. The sieve bootstrap may be particularly useful in the analysis of fractionally integrated processes since the statistics of interest can often be non-pivotal with distributions that depend on the fractional index d. The validity of the sieve bootstrap is established and it is shown that when the sieve bootstrap is used to approximate the distribution of a general class of statistics admitting an Edgeworth expansion then the error rate achieved is of order $O(T^{\beta+d-1})$, for any $\beta > 0$. Practical implementation of the sieve bootstrap is considered and the results are illustrated using a canonical example.

Key words and phrases : autoregressive approximation, fractional process, non-invertibility, rate of convergence, sieve bootstrap. (JEL CLASSIFICATION: C15, C22)

ACKNOWLEDGEMENTS: This research was partially supported by the Australian Research Council under Grant DP0452717. I am grateful to Simone Grose and Benedikt Pötscher for helpful and constructive comments on a previous version of this paper. The computations leading to the results presented in Section 5.2 were carried out by Simone Grose using MATLAB.

*Corresponding address: Don Poskitt, Department of Econometrics and Business Statistics, Monash University, Victoria 3800, Australia Tel.:+61-3-9905-9378; fax:+61-3-9905-5474.

E-mail address: Don.Poskitt@Buseco.monash.edu.au

1 Introduction

It is well known that, under a variety of conditions that hold in many econometric applications, improvements in the accuracy of first order large sample approximations can be obtained using bootstrap techniques. Such improvements require that the bootstrap re-sampling be conducted in such a way as to capture the essential features of the data generating process and in the context of time series analysis there are two basic methods that can be employed, the block bootstrap (Künsch, 1989) and the sieve bootstrap (Bühlmann, 1997). Both techniques are second order accurate, but the errors made by the bootstrap converge to zero more slowly than those of the bootstrap based on data drawn from a simple random sample. The error in the coverage probability of a one sided confidence interval is $O(T^{-3/4})$ for the block bootstrap, for example, compared to the $O(T^{-1})$ rate achieved with simple random samples, where here, as in what follows, T is used to denote sample size. The relatively poor performance of the block bootstrap has lead to the search for other ways to implement the bootstrap with dependent data and to the development of adaptations designed to increase the asymptotic refinement of the block bootstrap, see the recent contributions of Horowitz (2003) and Andrews (2004) and the references contained therein, for example. Choi and Hall (2000) have shown, however, that when the sieve bootstrap is applied to a linear process then the error in the coverage probability of a one sided confidence interval is $O(T^{\beta-1})$, for any $\beta > 0$, which is only slightly larger than $O(T^{-1})$. Choi and Hall concur with the conclusion of Bühlmann and they argue that for linear time series the sieve bootstrap has substantial advantages and superior performance over blocking methods.

The sieve bootstrap is obtained by approximating the data generating process by an autoregression of order h where h increases with the sample size. The bootstrap samples are then drawn from the autoregressive approximation. Details are presented below. Heuristically speaking, it is clear that the order of the autoregression must be allowed to go to infinity in order to achieve full generality and results on the properties of autoregressive models when $h \to \infty$ as $T \to \infty$, such that $h/T \to 0$, have been available for some time, see Hannan and Deistler (1988, Section 7.4) for example. However, such results are usually predicated on the presumption that the process admits an infinite autoregressive representation with coefficients that tend to zero at an appropriate rate, conditions that are not met by (i) fractionally integrated and (ii) non-invertible processes. One of the contributions of this paper is to show that, subject to appropriate adaptation, results on the properties of the sieve bootstrap can be extended to allow for both fractionally integrated and non-invertible processes.

Fractional 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

$$y(t) = \sum_{j\geq 0} k(j)\varepsilon(t-j) = k(z)\varepsilon(t) = \frac{\kappa(z)}{(1-z)^d}\varepsilon(t) \qquad (1.1)$$

wherein $\varepsilon(t)$ denotes a white noise process and, as will be done henceforth in expressions of this type, the indeterminate z in $k(z) = \sum_{j\geq 0} k(j)z^j$ is interpreted as the lag operator, that is $z\varepsilon(t) = \varepsilon(t-1)$. For any $b > -1$ the operator $(1-z)^b$ is defined via the binomial expansion

$$(1-z)^b = 1 - bz + \frac{b(b-1)z}{2!} - \frac{b(b-1)(b-2)z^3}{3!} + \cdots ,$$

which yields the result that

$$\frac{1}{(1-z)^d} = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)z^j}{\Gamma(j+1)\Gamma(d)} ,$$

where the gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ for $x \geq 0$ and the relation $\Gamma(x+1) = x\Gamma(x)$ defines $\Gamma(x)$ for $x < 0$. Hence

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

where $\kappa(z) = \sum_{j\geq 0} \kappa(j)z^j$. If $\kappa(z)$ is such that $\sum_{j\geq 0} |\kappa(j)| < \infty$, $\kappa(z)$ might be the transfer function of a stable and invertible autoregressive moving-average ($ARMA$) process for example, then using Sterling's approximation it can be shown that

$$k(j) \sim \frac{\kappa(1)}{\Gamma(d)} j^{d-1} \quad \text{as} \quad j \to \infty . \tag{1.2}$$

From (1.2) it follows that $\sum_{j\geq 0} |k(j)|^2 < \infty$ if $|d| < 0.5$ and $y(t)$ is well-defined as the limit in mean square of a covariance-stationary process with spectral density

$$f(\omega) = \frac{\sigma^2|k(e^{\imath\omega})|^2}{2\pi} = \frac{\sigma^2|\kappa(e^{\imath\omega})|^2}{2\pi|1-e^{\imath\omega}|^{2d}} .$$

Using the result that $|1-e^{\imath\omega}|^{2d} = |2\sin(\omega/2)|^{2d}$ and $\sin(\omega/2) \sim \omega/2$ as $\omega \to 0$ it can be shown that the spectral density obeys the inverse power law $f(\omega) \sim \sigma^2|\kappa(1)|^2/2\pi\omega^{2d}$ as ω approaches zero. Similarly, the autocovariance function declines at a hyperbolic rate, $\gamma(\tau) \sim C\tau^{2d-1}$, $C \neq 0$, as $\tau \to \infty$, and not at an exponential rate as it would for a stable and invertible $ARMA$ process. Throughout the paper C will stand for a universal, though not the same, constant. For a more detailed examination of the properties outlined above see Beran (1994).

Many empirical time series exhibit dynamic behaviour typical of a fractional process, and Beran (1992, 1994) and Baillie (1996) provide a brief history of the application of fractional models and a review of various statistical procedures for analyzing such processes. The use of fractional models depends, of course, on the practitioner being able to conduct appropriate inference and the inferential procedures currently available are, for the most part, based on first-order asymptotic theory. A natural alternative to using large-sample asymptotics to analyse the properties of different statistical procedures is application the bootstrap, and the bootstrap may be particularly useful in the analysis of fractionally integrated processes since the statistics of interest can often be non-pivotal with distributions that depend on d.

Examination of non-invertible processes is motivated by the observation that, although it might be argued that processes observed in the real world are unlikely to exhibit spectral zeroes, lack of invertibility might be induced by the actions of the practitioner, by over-differencing for example. The consequences of such over-differencing for the subsequent analysis of any techniques applied to the observed time series would then be of interest.

The paper proceeds as follows. The sieve bootstrap is described in the following section. In Section 3 results from the theory of stochastic processes that provide a rationale for a consideration of the sieve bootstrap in more general settings than are currently considered are reviewed. This section also outlines the estimation techniques to be used. These two sections provide the background, establish notation and present the basic assumptions. Section 4 lists some of the fundamental results that justify the sieve bootstrap and establishes a convergence rate of $O(T^{\beta+d-1})$ for any $\beta > 0$ for a general class of statistics that admit an Edgeworth expansion. Some additional practical issues are discussed in Section 5, where an illustration of the performance of the sieve bootstrap is also presented. Proofs and technical lemmas are assembled together in Section 6.

2 The Sieve Bootstrap

Consider a statistic $\mathbf{S}_T = (s_{1T}, \ldots, s_{mT})'$ where $s_{iT} = s_i(y(1), \ldots, y(T))$ and each $s_i(\cdot)$ for $i = 1, \ldots, m$ is a suitably smooth function of the time series values $y(1), \ldots, y(T)$. Let $F_{\mathbf{S}_T}(\mathbf{s})$ be the distribution function of $\mathbf{S}_T$ under the probability law $\mathcal{P}_{\{y(1),\ldots,y(T)\}}$ of the data generating mechanism. Bootstrap procedures are designed to construct an approximation to $F_{\mathbf{S}_T}(\mathbf{s})$ by approximating $\mathcal{P}_{\{y(1),\ldots,y(T)\}}$ and for the sieve bootstrap the approximation is constructed in the following manner.

Let $\mathcal{Y}_T = \{y(1), \ldots, y(T)\}$ denote a realization of a stochastic process. From $\mathcal{Y}_T$ estimate the parameters of the hth order autoregressive approximation using the Levinson (1947)-Durbin (1960) algorithm, denoted by $\bar{\boldsymbol{\phi}}_h = (\bar{\phi}_h(1) \cdots \bar{\phi}_h(h))$ and $\bar{\sigma}_h^2$, and evaluate the residuals

$$\bar{\epsilon}_h(t) = \sum_{j=0}^{h} \bar{\phi}_h(j) y(t-j) \, , \; t = 1, \ldots, T \, .$$

From $\bar{\epsilon}_h(t)$, $t = 1, \ldots, T$, construct the standardized residuals $\tilde{\epsilon}_h(t) = (\bar{\epsilon}_h(t) - \bar{\epsilon}_h)/s_{\bar{\epsilon}_h}$ where $\bar{\epsilon}_h = T^{-1}\sum_{t=1}^{T} \bar{\epsilon}_h(t)$ and $s_{\bar{\epsilon}_h}^2 = T^{-1}\sum_{t=1}^{T}(\bar{\epsilon}_h(t) - \bar{\epsilon}_h)^2$.

Denote by $U_{\tilde{\epsilon}_h,T}(e)$ the distribution function of the probability distribution that puts probability mass $1/T$ at each $\tilde{\epsilon}_h(t)$, $t = 1, \ldots, T$, and let $\epsilon_h^+(t)$, $t = 1, \ldots, T$, denote a simple random sample of *i.i.d.* values drawn from

$$U_{\tilde{\epsilon}_h,T}(e) = T^{-1} \sum_{t=1}^{T} \mathbf{1}\{\tilde{\epsilon}_h(t) \le e\} \, .$$

Define the bootstrap realization $\mathcal{Y}_T^* = \{y^*(1), \ldots, y^*(T)\}$ where $y^*(t)$ is the autoregressive process defined by

$$\sum_{j=0}^{h} \bar{\phi}_h(j) y^*(t-j) = \epsilon_h^*(t)$$

where $\epsilon_h^*(t) = \bar{\sigma}_h \epsilon_h^+(t)$. Now define $\mathbf{S}_T^*$ as for $\mathbf{S}_T$ but with the observed realization $\mathcal{Y}_T$

replaced by $\mathcal{Y}_T^*$, so that $\mathbf{S}_T^* = (s_{1T}^*, \ldots, s_{mT}^*)'$ where $s_{iT}^* = s_i(y^*(1), \ldots, y^*(T))$.

Construct B independent bootstrap realizations $\mathcal{Y}_{T,b}^*$ and calculate $\mathbf{S}_{T,b}^*$ for $b = 1, \ldots, B$. Approximate $F_{\mathbf{S}_T}(\mathbf{s})$ by the empirical distribution function

$$\bar{F}_{\mathbf{S}_T^*,B}(\mathbf{s}) = B^{-1} \sum_{b=1}^{B} \mathbf{1}\{\mathbf{S}_{T,b}^* \leq \mathbf{s}\}\,.$$

The idea behind the sieve bootstrap is that the distribution of $\mathbf{S}_T^*$ under $\mathcal{P}_{\{y^*(1),\ldots,y^*(T)\}}$ should mimic that of $\mathbf{S}_T$ under $\mathcal{P}_{\{y(1),\ldots,y(T)\}}$ and therefore we can expect $F_{\mathbf{S}_T^*}(\mathbf{s})$ to approximate $F_{\mathbf{S}_T}(\mathbf{s})$ reasonably well provided $\mathcal{P}_{\{y^*(1),\ldots,y^*(T)\}}$ is in some sense close to $\mathcal{P}_{\{y(1),\ldots,y(T)\}}$. The analytical determination of $F_{\mathbf{S}_T^*}(\mathbf{s})$ is generally intractable, but by simulating a large number of independent bootstrap realizations we can approximate $F_{\mathbf{S}_T^*}(\mathbf{s})$ by $\bar{F}_{\mathbf{S}_T^*,B}(\mathbf{s})$. By the Glivenko-Cantelli Theorem $\bar{F}_{\mathbf{S}_T^*,B}(\mathbf{s})$ converges to $F_{\mathbf{S}_T^*}(\mathbf{s})$ a.s. uniformly in $\mathbf{s}$ as $B \to \infty$. Thus, we can approximate $F_{\mathbf{S}_T^*}(\mathbf{s})$ arbitrarily closely by taking the number of bootstrap realizations sufficiently large and we can anticipate that $\bar{F}_{\mathbf{S}_T^*,B}(\mathbf{s})$ will also approximate $F_{\mathbf{S}_T}(\mathbf{s})$ closely provided $F_{\mathbf{S}_T^*}(\mathbf{s})$ is sufficiently near to $F_{\mathbf{S}_T}(\mathbf{s})$.

3 Rationale

Let $y(t)$ for $t \in Z$ denote a linearly regular, covariance-stationary process with Wold representation,

$$y(t) = \sum_{j=0}^{\infty} k(j)\varepsilon(t-j) \tag{3.1}$$

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

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*

$$E\big[\varepsilon(t) \mid \mathcal{E}_{t-1}\big] = 0 \quad \textit{and} \quad E\big[\varepsilon(t)^2 \mid \mathcal{E}_{t-1}\big] = \sigma^2\,. \tag{3.2}$$

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).

Since by assumption $y(t)$ is a regular process then we know from a famous result due to Szegö (1939) and Kolmorgorov (1941) that it is not possible to determine $y(t+1)$ precisely from its own history up to time t and

$$\sigma^2 = 2\pi \exp\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left\{\frac{\sigma^2|k(e^{\imath\omega})|^2}{2\pi}\right\} d\omega\} > 0\,. \tag{3.3}$$

where $\sigma^2 = E[(y(t) - \bar{y}_{\langle t|t-1,\ldots\infty\rangle})^2]$. The transfer function $k(z)$ has no zeroes inside the unit circle and $|k(e^{\imath\omega})|^2 > 0$ almost everywhere (a.e.) where $|k(e^{\imath\omega})|^2 = \lim_{\rho\uparrow 1} |k(\rho e^{\imath\omega})|^2$, the radial limit of $k(z)$ on the boundary of the unit circle $|z| = 1$. In the context of autoregressive modelling and the sieve bootstrap it is common practice to strengthen the condition, $k(z) \neq 0$, $|z| < 1$, by adding the restriction that $k(z)$ has no zeroes on the unit circle, and to assume that a condition such as $\sum_{j\geq 0} |k(j)| < \infty$, or $\sum_{j\geq 0} j|k(j)|^2 < \infty$, holds, see *inter alia* Bühlmann (1997, Section 3.1). It is not necessary for $k(z)$ to be invertible, however, in order for there to be an autoregression that yields an appropriate approximation to the process, and note that the impulse response coefficients of $k(z)$ in (1.1) will not satisfy the above summability conditions if $d > 0$ and the process exhibits long memory, a case commonly encountered.

3.1 Autoregressive Approximation

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

$$\sum_{j=0}^{h} \phi_h(j)\gamma(j-k) = \delta_{0k}\sigma_h^2\,, \quad k = 0, 1, \ldots, h, \tag{3.4}$$

where δ_{0k} is Kronecker's delta, $\phi_h(0) = 1$ and

$$\sigma_h^2 = E\big[\epsilon_h(t)^2\big] \tag{3.5}$$

is the minimising value of the prediction error variance associated with the prediction error

$$\epsilon_h(t) = \sum_{j=0}^{h} \phi_h(j)y(t-j)\,. \tag{3.6}$$

Rewriting the Yule-Walker equations in matrix-vector notation yields $\mathbf{\Gamma}_h\boldsymbol{\phi}_h = -\boldsymbol{\gamma}_h$ where $\mathbf{\Gamma}_h = [\gamma(i-j)]_{i,j=1,\ldots,h}$, $\boldsymbol{\phi}_h = (\phi_h(1), \ldots, \phi_h(h))'$ and $\boldsymbol{\gamma}_h = (\gamma(1), \ldots, \gamma(h))'$. Note that regularity of $y(t)$ implies that $\mathbf{\Gamma}_h$ is nonsingular for all h and it follows that $\boldsymbol{\phi}_h$ is unique and $\phi_h(z) = \sum_{j=0}^{h} \phi_h(j)z^j \neq 0$, $|z| \leq 1$. Solving (3.4) using the Levinson (1947)-Durbin (1960) algorithm

$$\begin{aligned}
\phi_h(j) &= \phi_{h-1}(j) + \phi_h(h)\phi_{h-1}(h-j), \quad \phi_h(0) = 1, \quad j = 1, \ldots, h-1 \\
\phi_h(h) &= \sum_{j=0}^{h-1} \phi_{h-1}(j)\gamma(h-j)/\sigma_{h-1}^2 \\
\sigma_h^2 &= \sigma_{h-1}^2(1 - \phi_h(h)^2)
\end{aligned} \tag{3.7}$$

initiated at $\phi_0(0) = 1$ and $\sigma_0^2 = \gamma(0)$, and using the relationship $\sigma_h^2 = \det(\mathbf{\Gamma}_{h+1})/\det(\mathbf{\Gamma}_h)$, which leads to the conclusion that $|\phi_h(h)| < 1$ for all h, we can see that σ_h^2 is monotonically decreasing in h. Basic Hilbert space arguments can also be used to show that $\lim_{h\to\infty} \sigma_h^2 = \sigma^2$.

Thus, for h sufficiently large it seems reasonable to suppose that the optimal predictor $\bar{y}_{\langle t|t-1,\dots,t-h\rangle} = \phi_h(1)y(t-1)+\cdots+\phi_h(h)y(t-h)$ determined from the autoregressive model in (3.6) will form a good approximation to the best predictor $\bar{y}_{\langle t|t-1,\dots\infty\rangle}$ and hence that $\epsilon_h(t)$ will be close to $\varepsilon(t)$. We can therefore think of an infinite autoregression as arising, not by inverting $k(z)$, but as the limit of the autoregressive approximations obtained as $h \to \infty$. Indeed, Wold (1938) first derived (3.1) by fitting autoregressions of ever increasing order.

From the preceding discussion it is apparent that it is the regularity of $y(t)$ that is important in the context of autoregressive modelling rather than invertibility. This observation gives rise to the following:

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

3.2 Data Modelling

The sieve bootstrap is obtained by approximating the data generating process by an autoregression of order h and then resampling from the autoregressive approximation where the parameters of the $AR(h)$ approximation are determined by fitting autoregressive models to the data. More explicitly, given a realisation of T observations $y(t),\ \ t = 1,\dots,T$, set

$$c_T(r) = c_T(-r) = T^{-1}\sum_{t=r+1}^{T} y(t-r)y(t)\,, \quad r = 0,1,\dots,T-1\,, \tag{3.8}$$

the sample autocovariance function. Substituting $c_T(r)$ for $\gamma(r)$ in the Yule-Walker equations 3.4 and solving for $\phi_h(j),\ j = 1,\dots,h$ and σ_h yields estimates of the parameters in the $AR(h)$ model. Noting the correspondence with the method of moments we denote the Yule-Walker estimator and its associated estimates by the use of an over-bar. This estimator has the advantage that it can be readily calculated via the Levinson-Durbin recursions, and being based on Toeplitz calculations the operator $\bar{\phi}_h(z)$, like $\phi_h(z)$, will be stable.

In order to implement the sieve bootstrap the order of the autoregressive approximation must be prescribed. Following Bühlmann (1997) we suppose that h is chosen using Akaike's information criterion, Akaike (1969), that is, the order of the model to be employed is obtained by minimizing the model selection criterion

$$AIC_T(h) = \log(\bar{\sigma}_h^2) + 2h/T$$

over the range $h = 0,1,\dots,M_T$ where $M_T = [c(\log T)^a]$, the integer part of $c(\log T)^a$ for some $a \geq 1$ and $c > 0$. Bühlmann (1997) justifies the use of AIC by reference to the predictive optimality property of AIC due to Shibata (1980). The regularity conditions imposed by Shibata *op. cit.* are too restrictive to be applicable here. Nevertheless, a similar justification for consideration of AIC can be given and Poskitt (2004) shows that if $y(t)$ is a covariance-stationary process that satisfies Assumptions 1 and 2, and $h_T^{AIC} = \operatorname{argmin}_{0,1,\dots,M_T} AIC_T(h)$ where $\lim_{T\to\infty}(M_T/\lambda_{\min}(\mathbf{\Gamma}_{M_T}))(\log T/T)^{1-2d'} = 0$, then the $AR(h_T^{AIC})$ model is asymptotically efficient in the sense that if it is used to predict a future value of the same process then the mean squared prediction error achieves an asymptotic lower bound.

Poskitt (2004)'s result extends the predictive optimality property of AIC to fractional and non-invertible processes and thereby provides the theoretical background to the fundamental step of selecting h in application of the sieve bootstrap to such processes. It is based in part on the following theorem.

Theorem 3.1 *If $y(t)$ is a stationary process that satisfies Assumption 1 and Assumption 2 then uniformly in $h \leq H_T$*

$$\sum_{j=1}^{h} |\bar{\phi}_h(j) - \phi_h(j)|^2 = O\left\{ \left(\frac{h}{\lambda_{\min}(\mathbf{\Gamma}_h^2)} \right) \left(\frac{\log T}{T} \right)^{1-2d'} \right\}.$$

where $H_T = o\{(T/\log T)^{\frac{1}{2}-d'}\}$ where $d' = \max\{0, d\}$

Theorem 3.1 establishes the consistency of the coefficient estimates of the $AR(h)$ model to those of the $AR(h)$ approximation to the process and indicates that the parameter estimation errors converge to zero at a rate that is dependent on d. The relevance of this observation stems from the fact that the convergence rate of the sieve bootstrap itself depends upon the convergence rate of these estimates. It is also apparent that the presence of spectral zeroes has an important impact via it's influence on the proximity of $\lambda_{\min}(\mathbf{\Gamma}_h)$ to zero. To investigate this impact in further detail it is necessary to give explicit structure to the spectral zeroes of the process. This is done by extending Assumption 2.

Assumption 3 *There exists a set of frequencies $\theta_j \in (0, \pi)$ and numbers $\nu_j > 0$, $j = 1, \dots, n$, such that $|k(\omega)|^2 \sim |2\sin(\omega/2)|^{-2d} |\mu_j(\omega)|^2 |\omega - \theta_j|^{2\nu_j}$ as $\omega \to \theta_j$, where $\mu_j(\omega)$ is of bounded variation on $(0, \pi)$ and slowly varying at θ_j, for each $j = 1, \dots, n$.*

By appropriate choice of n and the θ_j and ν_j the factors $|\mu_j(\omega)|^2 |\omega - \theta_j|^{2\nu_j}$ can be thought of as modeling spectral zeroes or troughs.

4 Some Asymptotic Theory

The following Lipschitz-type condition determines the degree of smoothness that the statistic $\mathbf{S}_T$ must satisfy in order for the results presented here to hold.

Assumption 4 *Let $\mathfrak{Y}$ be a Borel set in $\mathbb{R}^T$. Then for all $\mathcal{Y}_T, \mathcal{Y}_T^* \in \mathfrak{Y}$ there exists a family of Borel measurable functions $B_t : \mathbb{R}^2 \to [0, \infty)$, satisfying*

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} E[B_t(y_t, y_t^*)^2] < \infty \,,$$

for which

$$\|\mathbf{S}_T - \mathbf{S}_T^*\|^2 \leq \frac{1}{T} \sum_{t=1}^{T} B_t(y_t, y_t^*) |y_t - y_t^*| \,.$$

Assumption 4 can be verified directly in some cases. For the standard deviation s_y, where $s_y^2 = T^{-1} \sum_{t=1}^{T} (y_t - \bar{y})^2$, $\bar{y} = T^{-1} \sum_{t=1}^{T} y_t$, the bound $|s_y - s_{y^*}|^2 \leq (1/T) \sum_{t=1}^{T} |y_t - y_t^*|^2$

follows from the triangular inequality, for example. More generally, Assumption 4 can be ensured by imposing suitable sufficient conditions on the elements of $\mathbf{S}_T$. Thus, if

$$s_{iT} = g_i\left((T-m)^{-1}\sum_{t=1}^{T} f_{i,t}(y_t, y_{t-1}, \ldots, y_{t-m})\right) \tag{4.1}$$

where, for all $i = 1, \ldots, m$, the function $g_i : \mathbb{R} \to \mathbb{R}$ is differentiable on $\mathbb{R}$ and $f_{i,t} : \mathbb{R}^{m+1} \to \mathbb{R}$ has continuous partial derivatives for each t, then $\mathbf{S}_T$ will be differentiable on $\mathfrak{Y}$ and Lipschitzian. Functions of linear statistics of the type given in 4.1 include the sample autocovariances, autocorrelations and partial-autocorrelations.

Theorem 4.1 *Let $\eta(F_X, F_Y)$ denote Mallow's measure of the distance between two probability distributions F_X and F_Y, defined as $\inf\{E\|X-Y\|^2\}^{\frac{1}{2}}$ where the infimum is taken over all square integrable random variables X and Y in $\mathbb{R}^m$ with marginal distributions F_X and F_Y. Then*

$$\eta(F_{\mathbf{S}_T^*}, F_{\mathbf{S}_T}) = O\left\{\left(\frac{h^5}{\lambda_{\min}(\mathbf{\Gamma}_h^2)}\right)^{\frac{1}{2}}\left(\frac{\log T}{T}\right)^{\frac{1}{2}-d'}\right\}$$

under Assumptions 1, 2 and 4, and

$$\eta(F_{\mathbf{S}_T^*}, F_{\mathbf{S}_T}) = o\left\{\left(\frac{T^{\beta}}{T^{1-2d'}}\right)^{\frac{1}{2}}\right\}$$

for any $\beta > 0$ under Assumptions 1 through 4 inclusive.

See Bickel and Freedman (1981, Section 8) for a discussion of the properties of $\eta(F_X, F_Y)$. Since $\eta(\bar{F}_{\mathbf{S}_T^*,B}, F_{\mathbf{S}_T^*}) = o(1)$ (Bickel and Freedman, 1981, Lemma 8.4) it follows from the triangular inequality, $\eta(\bar{F}_{\mathbf{S}_T^*,B}, F_{\mathbf{S}_T}) \leq \eta(\bar{F}_{\mathbf{S}_T^*,B}, F_{\mathbf{S}_T^*}) + \eta(F_{\mathbf{S}_T^*}, F_{\mathbf{S}_T})$, and Theorem 4.1 that $\eta(\bar{F}_{\mathbf{S}_T^*,B}, F_{\mathbf{S}_T}) = o(1)$. This implies that $\bar{F}_{\mathbf{S}_T^*,B}$ converges in probability to $F_{\mathbf{S}_T}$ and validates the sieve bootstrap under the scenarios being considered here.

The topology induced by Mallows metric is relatively weak, however, and the convergence rate given in Theorem 4.1 is no better than that achieved using known central limit properties of fractional processes, as described in Hosking (1996) for example. In order to obtain better convergence rates let us suppose that $F_{\mathbf{S}_T}(\mathbf{s})$ is absolutely continuous with respect to Lebesgue measure, differentiable for all $\mathbf{s}$, and that the following assumption is satisfied.

Assumption 5 *Let $\psi_T(\boldsymbol{\theta}) = E[\exp(\imath\boldsymbol{\theta}'\mathbf{S}_T)]$ denote the characteristic function of $\mathbf{S}_T$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ and let $\partial^j \log\psi_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^j$ denote the vector of jth order partial derivatives corresponding to $\partial^j \log\psi_T(\boldsymbol{\theta})/\partial\theta_1^{j_1}\cdots\partial\theta_m^{j_m}$ for all non-negative integers $j_1, \ldots, j_m$ satisfying $\sum_{l=1}^m j_l = j$. Then firstly, for any $\delta > 0$ the conditions*

$$\begin{aligned}\int_{\|\boldsymbol{\theta}\|>\delta\sqrt{T}} |\psi_T(\boldsymbol{\theta})|^2 d\boldsymbol{\theta} &= o(T^{2-r}) \text{ and}\\ \int_{\|\boldsymbol{\theta}\|>\delta\sqrt{T}} \left|\frac{\partial^s\psi_T(\boldsymbol{\theta})}{\partial\theta_l^s}\right|^2 d\boldsymbol{\theta} &= O(T^{1-r}),\ l = 1, \ldots, m,\end{aligned}$$

hold where $s = [m/2]+1$ and $r \geq 3$, and secondly, $\partial^q \log\psi_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^q$ exists for all $\boldsymbol{\theta}$ in a neighbourhood of the origin and $\lim_{\|\boldsymbol{\theta}\|\to 0} T^{-1}\partial^q \log\psi_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^q$ exists as $T \to \infty$ for all

$q = 1, \ldots, q' = \max\{s, r+1\}$.

Here E denotes the expectation taken with respect P, the probability measure induced by $\mathcal{P}_{\{y(1),\ldots,y(T)\}}$. Assumption 5 summarizes Assumptions 1 and 2 of Taniguchi (1984), which in turn are related to the conditions imposed by Durbin (1980) in order to validate the Edgeworth expansion in dependent data settings.

Let $\mathbf{V}_T = E\left[(\mathbf{S}_T - E[\mathbf{S}_T])(\mathbf{S}_T - E[\mathbf{S}_T])'\right] = O(T)$ and set $\mathbf{Z}_T = \mathbf{V}_T^{-\frac{1}{2}}(\mathbf{S}_T - E[\mathbf{S}_T])$. Assumption 5 ensures the validity of the formal Edgeworth expansion

$$\mathrm{P}(\mathbf{Z}_T \le \mathbf{z}) = G(\mathbf{z}) + \sum_{j=3}^{r} T^{1-j/2}\pi_j(\mathbf{z}, \mathbf{K}_r)g(\mathbf{z}) + o(T^{1-r/2}) \tag{4.2}$$

uniformly in $\mathbf{z}$, where $G(\mathbf{z})$ denotes the distribution function of a Gaussian $N(\mathbf{0}, \mathbf{I}_m)$ random vector, $g(\mathbf{z})$ the corresponding density, and $\pi_j(\mathbf{z}, \mathbf{K}_r)$ is a polynomial function of degree j in $\mathbf{z}$ whose coefficients are polynomials in the elements of the cumulants $\mathbf{K}_r = (\mathbf{k}_1', \ldots, \mathbf{k}_r')'$, $\mathbf{k}_r = \imath^{-r}\partial^r \log \psi_T(\mathbf{0})/\partial\boldsymbol{\theta}^r$. See Theorem 1 of Taniguchi (1984).

Similarly, if E^* is used to denote the expectation taken with respect to the probability measure P^* induced by $\mathcal{P}_{\{y^*(1),\ldots,y^*(T)\}}$ and $\mathbf{Z}_T^* = \mathbf{V}_T^{*-\frac{1}{2}}(\mathbf{S}_T^* - E^*[\mathbf{S}_T^*])$ where $\mathbf{V}_T^* = E^*\left[(\mathbf{S}_T^* - E^*[\mathbf{S}_T^*])(\mathbf{S}_T^* - E^*[\mathbf{S}_T^*])'\right]$, then

$$\mathrm{P}^*(\mathbf{Z}_T^* \le \mathbf{z}) = G(\mathbf{z}) + \sum_{j=3}^{r} T^{1-j/2}\pi_j(\mathbf{z}, \mathbf{K}_r^*)g(\mathbf{z}) + o(T^{1-r/2}) \tag{4.3}$$

where $\mathbf{K}_r^* = (\mathbf{k}_1^{*\prime}, \ldots, \mathbf{k}_r^{*\prime})'$, $\mathbf{k}_r^* = \imath^{-r}\{\partial^r \log \psi_T^*(\mathbf{0})/\partial\boldsymbol{\theta}^r\}$, $\psi_T^*(\boldsymbol{\theta}) = E^*[\exp(\imath\boldsymbol{\theta}'\mathbf{S}_T^*)]$.

Note that P^* depends on $\mathcal{Y}_T$ and the elements of $\mathbf{K}_r^*$, which are constants relative to P^*, are random variables relative to P. A comparison of (4.2) and (4.3) for $r \ge 3$ now indicates that

$$\sup_{\mathbf{z}} |\mathrm{P}^*(\mathbf{Z}_T^* \le \mathbf{z}) - \mathrm{P}(\mathbf{Z}_T \le \mathbf{z})| = T^{-\frac{1}{2}}O(\|\mathbf{K}_r^* - \mathbf{K}_r\|) + o(T^{-\frac{1}{2}})\,. \tag{4.4}$$

Expression (4.4) forms the background to the following theorem since, as is shown below, under the regularity conditions of the theorem $\|\mathbf{K}_r^* - \mathbf{K}_r\| = O(T^{-\frac{1}{2}(1-2d')+\beta})$.

Theorem 4.2 *Suppose that the statistic $\mathbf{S}_T$ satisfies Assumption 4 and Assumption 5 with $r \ge 3$ when calculated from any process $y(t)$ that satisfies Assumptions 1, 2 and 3. Then*

$$\sup_{\mathbf{z}} |P^*(\mathbf{Z}_T^* \le \mathbf{z}) - P(\mathbf{Z}_T \le \mathbf{z})| = O(T^{-(1-d')+\beta})$$

for all $\beta > 0$.

Theorem 4.2 indicates the refinements that are possible using the sieve bootstrap. Assumption 5 is a relatively high level condition, however, that will need to be verified on a case by case basis. If an Edgeworth expansion of the form implicit in Assumption 5 can be established independently, or if (4.2) and (4.3) are known to obtain *a priori*, then Assumption 5 can be dispensed with and the result in Theorem 4.2 will continue to hold. More importantly, Theorem 4.2 is expressed in terms of standardized statistics and in practice it is unlikely that the mean vectors and covariance matrices required to construct such quantities will be known. Standardization can be circumvented, however, and the sieve bootstrap can be implemented without prior knowledge of the moments of $\mathbf{S}_T$. Formally we have the following result.

Theorem 4.3 *Suppose that for any process $y(t)$ that satisfies Assumptions 1, 2 and 3 the statistic $\mathbf{S}_T$ admits the formal Edgeworth expansion*

$$P(\mathbf{V}_T^{-\frac{1}{2}}(\mathbf{S}_T - E[\mathbf{S}_T]) \leq \mathbf{z}) = G(\mathbf{z}) + \sum_{j=3}^{r} T^{1-j/2}\pi_j(\mathbf{z}, \mathbf{K}_r)g(\mathbf{z}) + o(T^{1-r/2})$$

uniformly in $\mathbf{z}$ for some $r \geq 3$. Suppose also that $\mathbf{S}_T$ satisfies Asumption 4. Then

$$|P^*(T^{-\frac{1}{2}}(\mathbf{S}_T^* - \mathbf{S}_T) \leq \mathbf{s}) - P(T^{-\frac{1}{2}}(\mathbf{S}_T - E[\mathbf{S}_T]) \leq \mathbf{s})| = O_p(T^{-(1-d')+\beta})$$

for all $\beta > 0$ uniformly in $\mathbf{s}$.

Statistics for which Edgeworth expansions have been established in the context of fractional processes, and to which the results given here can be applied, include quadratic forms in Gaussian long memory processes and Gaussian maximum likelihood estimates, Lieberman, Rousseau and Zucker (2001, 2003), and semiparametric Whittle estimates of long memory, Giraitis and Robinson (2003).

5 Practical Considerations and An Illustration

5.1 Practical Considerations

Thus far we have couched our discussion of the sieve bootstrap in terms of the Yule-Walker estimates. Estimating the parameters of the autoregressive approximation by directly minimizing the observed mean squared error $T^{-1}\sum_{t=1}^{T}(y(t) - \phi_h(1)y(t-1) + \cdots + \phi_h(h)y(t-h))^2$ leads to the least squares estimates of course. By way of contrast, whereas the least squares estimator minimizes the observed mean squared error, the Yule-Walker estimator need not, but there is no guarantee that the least squares estimate of $\phi_h(z)$ will, like $\bar{\phi}_h(z)$, be stable. The difference in the two estimators is due to edge effects and, as is shown in Poskitt (2004), although these effects are asymptotically negligible the two estimators can have quite different finite sample behaviour. In particular, when applied to noninvertible and fractional processes the Yule-Walker coefficient estimates exhibit a substantial finite sample bias which feeds through to the prediction error variance and order estimates, *c.f.* Tjøstheim and Paulsen (1983) and Paulsen and Tjøstheim (1985). Such biases are not present with the least squares estimator, suggesting that the sieve bootstrap be constructed from statistics based on least squares calculations. In the context of the sieve bootstrap, however, we require an estimate of $\phi_h(z)$ that is stable. A suitable compromise is given by the algorithm due to Burg (1968). Burg's algorithm generates a stable estimator of $\phi_h(z)$ that shares the superior finite sample properties of least squares, see Poskitt (2004). It is therefore recommended that for practical purposes Burg's algorithm be used at the first step of the sieve bootstrap rather than the Levinson-Durbin algorithm.

It is useful to note that alternative methods of autoregressive order determination that generate asymptotically efficient selection criteria have been proposed in the literature. The criterion autoregressive transfer function suggested by Parzen (1974),

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

and the mean squared prediction error criterion of Mallows (1973),

$$MC_T(h) = T\left(\frac{\bar{\sigma}_h^2}{\tilde{\sigma}^2} - 1\right) + 2h\,,$$

where

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

$\gamma' = 0.57721$ (Eulers constant) and $N = [(T-1)/2]$, a nonparametric estimate of the innovation variance constructed from the periodogram by analogy with (3.3), for example. Any of these criteria could be used to determine h in place of AIC.

5.2 An Illustration

This section illustrates the main results of the paper by means of a small simulation experiment. The experiments follow Lieberman et al. (2003) and examine the distribution of the maximum likelihood estimator $\tilde{d}_T$ of d for the fractional noise process $(1-z)^d y(t) = \varepsilon(t)$ where $\varepsilon(t)$ is standard Gaussian white noise. The true values of d considered where 0.1, 0.2, 0.3 and 0.4, and the sample sizes examined were $T = 20, 40, 80, 160$. To obtain $\tilde{d}_T$ the exact Gaussian likelihood was maximized over the interval $[-0.49, 0.49]$.

The performance of the sieve bootstrap is summarized graphically in the following figures, which present the exact distribution and the bootstrap distribution. The exact, or Monte–Carlo, distribution of $\zeta_T = \pi\sqrt{T/6}(\tilde{d}_T - d)$ was calculated empirically, as in Lieberman et al. (2003), using 1000 simulation replications. For each realization of the process the sieve bootstrap distribution of $\zeta_T^* = \pi\sqrt{T/6}(\tilde{d}_T^* - d)$ was constructed from $B = 500$ bootstrap re-samples, with the order h_T chosen via AIC, using Burg's algorithm to estimate the models. Note that the value of B used here ensures that $\mathrm{P}(\sup_{\mathbf{s}} |\bar{F}_{\mathbf{S}_T^*,B}(\mathbf{s}) - F_{\mathbf{S}_T^*}(\mathbf{s})| < \delta) > 1 - 2\exp(-\delta^2(1000))$. Both distributions were evaluated using a kernel density estimate based on a Gaussian kernel with bandwidth equal to $0.75\, s\sqrt[5]{(243/35N)}$ where s is the standard deviation calculated from the N data values being smoothed, see Wand and Jones (1995). To provide a basis for comparison the exact and an approximate Edgeworth expansion that provide asymptotic expansions of the distribution of ζ_T up to terms $o(T^{-\frac{1}{2}})$, as described in Lieberman and Phillips (2004), are also plotted, as is the asymptotic normal approximation.

Figure 1 presents the distributions obtained when $d = 0.2$ and $T = 40, 160$. The sieve bootstrap does not give as accurate an approximation to the exact (Monte–Carlo) distribution as the Edgeworth expansions, nor the normal approximation, and the ability of the sieve bootstrap to provide a reasonable representation of the distribution of ζ_T appears to be called into question.

[Figure 1 about here.]

The effects of increasing d are seen in Figure 2, which presents the distributions obtained when $d = 0.4$ and $T = 40, 160$.

[Figure 2 about here.]

The slower convergence rate for ζ_T^* inherent in having a larger value of d does not seem to have harmed the performance of the sieve bootstrap. Rather, the larger value of d has clearly produced a deterioration in the performance of the normal approximation, particularly when $T = 40$. The relativities seen in Figure 2 are not too unexpected, of course, since the convergence rate of both Edgeworth expansions is $o(T^{-\frac{1}{2}})$, compared to the $O(T^{-(1-d')+\beta})$, $\beta > 0$, rate for the sieve bootstrap given in Theorem 4.2, compared to $O_p(1)$ for the asymptotic approximation.

In both Figure 1 and Figure 2 the distribution of ζ_T^* appears to have fatter tails than the true distribution, to be rather more skewed, and more platykurtic in the case of Figure 1. A detailed examination of individual replications reveals the cause; the likelihood surface is often very flat in an asymmetric interval around d,

[Figure 3 about here.]

as illustrated in Figure 3. This figure plots the log-likelihood for ten randomly chosen realizations when $d = 0.4$ and $T = 160$. Roughly speaking, for each of these realizations any value of d in the interval $(0.375, 0.45)$ seems equally likely and $\tilde{d}_T$ can fall anywhere in the interval. For such realizations the sieve bootstrap values $\tilde{d}_T^*$ are also concentrated in the same interval, and hence the overall skewness. This suggests that previous distortions to the sieve bootstrap distribution can be removed by re-centering, the justification for which lies in Theorem 4.3. The consequences of re-centering are illustrated in Figures 4 and 5,

[Figure 4 about here.]

which present the counterparts to Figures 1 and 2 for $\tilde{\zeta}_T^* = \pi\sqrt{T/6}(\tilde{d}_T^* - \tilde{d}_T) = \zeta_T^* - \zeta_T$.

[Figure 5 about here.]

The improvement in performance brought about by removing the fluctuations in ζ_T is apparent. The sieve bootstrap now yields a more accurate representation of the true distribution than does the asymptotic normal approximation, even for moderately large T, and it is capturing the second order properties of the estimator quite well, on a par with the analytically derived, but unfeasible, Edgeworth expansions.

Finally, it is of interest to note that the performance of the sieve bootstrap appears to be at least as good as that of the model based bootstrap. The latter is derived in the obvious way, using the residuals from the known model, rather than the autoregressive approximation, as a basis for constructing the bootstrap re–samples.

6 Proofs and Technical Lemmas

Before proceeding let us collect together some properties of the $AR(h)$ approximation and the associated estimates. We begin with a lemma that relates the residuals

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

to the prediction errors $\epsilon_h(t)$. The lemma depends on Theorem 3.1, which we prove first.

Proof of Theorem 3.1: The result follows directly from Corollary 4.1 and Theorem 5.1 of Poskitt (2004). ∎

Lemma 6.1 *Under the same assumptions as for Theorem 3.1*

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

uniformly in $h \le H_T$, $H_T = o\{(T/\log T)^{\frac{1}{2}-d'}\}$ *where* $d' = \max\{0, d\}$.

Proof: From the definition of $\bar{\epsilon}_h(t)$ and $\epsilon_h(t)$ we get

$$\bar{\epsilon}_h(t) - \epsilon_h(t) = \sum_{j=1}^{h}\{\bar{\phi}_h(j) - \phi_h(j)\} y(t-j)$$

and

$$T^{-1}\sum_{t=1}^{T}\{\bar{\epsilon}_h(t)-\epsilon_h(t)\}^2 \le |T^{-1}\sum_{t=1}^{T}\epsilon_h(t)\{\bar{\epsilon}_h(t)-\epsilon_h(t)\}| + |T^{-1}\sum_{t=1}^{T}\bar{\epsilon}_h(t)\{\bar{\epsilon}_h(t)-\epsilon_h(t)\}|. \quad (6.1)$$

From the Cauchy-Schwartz inequality we now have

$$\begin{aligned} |T^{-1}\sum_{t=1}^{T}\epsilon_h(t)\{\bar{\epsilon}_h(t) - \epsilon_h(t)\}| &= |T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{h}\{\bar{\phi}_h(j) - \phi_h(j)\}\epsilon_h(t) y(t-j)| \\ &\le \left[\|\bar{\boldsymbol{\phi}}_h - \boldsymbol{\phi}_h\|^2 \cdot \sum_{j=1}^{h}\left(T^{-1}\sum_{t=1}^{T}\epsilon_h(t) y(t-j)\right)^2\right]^{\frac{1}{2}} \end{aligned}$$

and simple substitution gives us

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

which by Poskitt (2004, Theorem 4.1) equals

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

Since $\phi_h(j)$, $j = 1, \ldots, h$, solve the Yule-Walker equations $\sum_{j=0}^{h}\phi_h(j)\gamma(j-r) = 0$ for $r = 1, \ldots, h$. Moreover, $\phi_h(z) \ne 0$, $|z| \le 1$, and there exists constants $C < \infty$ and $\zeta < 1$

such that $|\phi_h(j)| < C\zeta^j$ and $\sum_{j=0}^h |\phi_h(j)| < C(1-\zeta^{h+1})/(1-\zeta) < C(1-\zeta)^{-1}$ so that $\sum_{j=0}^h \phi_h(j) O\{(\log T/T)^{\frac{1}{2}-d'}\} = O\{(\log T/T)^{\frac{1}{2}-d'}\}$. This result, when combined with Theorem 3.1 gives us the bound $O\left\{(h/\lambda_{\min}(\mathbf{\Gamma}_h))\,(\log T/T)^{1-2d'}\right\}$ for the first term on the right hand side in (6.1).

Similarly,

$$T^{-1}\sum_{t=1}^{T}\bar{\epsilon}_h(t)\{\bar{\epsilon}_h(t)-\epsilon_h(t)\} = \sum_{j=1}^{h}\{\bar{\phi}_h(j)-\phi_h(j)\}T^{-1}\sum_{t=1}^{T}\bar{\epsilon}_h(t)y(t-j)$$

and

$$\begin{aligned} T^{-1}\sum_{t=1}^{T}\bar{\epsilon}_h(t)y(t-r) &= \sum_{j=0}^{h}\bar{\phi}_h(j)T^{-1}\sum_{t=1}^{T}y(t-j)y(t-r) \\ &= \sum_{j=0}^{h}\bar{\phi}_h(j)[c_T(j-r)+O\{(\log T/T)^{\frac{1}{2}-d'}\}] \end{aligned}$$

where the last equality follows as a consequence of Theorems 4.1 and 4.2 of Poskitt (2004). Since $\bar{\phi}_h(j)$, $j=1,\ldots,h$, solve the empirical Yule-Walker equations $\sum_{j=0}^h \bar{\phi}_h(j)c_T(j-r)=0$, $r=1,\ldots,h$, and $\bar{\phi}_h(z)$ is by construction stable, a repetition of the arguments just applied to the first term on the right hand side of (6.1) yields the same order of magnitude for the second term, and the lemma is proved. ∎

Set $\varphi_h(z)=\sum_{j=0}^{\infty}\varphi_h(j)z^j$ where the $\varphi_h(j)$ and $\phi_h(j)$ are related by the recursions

$$\phi_h(0)=\varphi_h(0)=1\,,\quad \sum_{i=0}^{j}\varphi_h(i)\phi_h(j-i)=0,\ j=1,2,\ldots\,. \tag{6.2}$$

Then $\varphi_h(z)=\{\phi_h(z)\}^{-1}$ for $|z|\le 1$ and since $\phi_h(z)\neq 0$, $|z|\le 1$, the same is true of $\varphi_h(z)$. Define $\bar{\varphi}_h(z)=\{\bar{\phi}_h(z)\}^{-1}$ similarly by replacing $\phi_h(z)$ by $\bar{\phi}_h(z)$. We now present some properties of the operator $\varphi_h(z)$ and its corresponding estimate $\bar{\varphi}_h(z)$.

Lemma 6.2 *Suppose that Assumptions 1 and 2 hold. Then*

$$|\bar{\varphi}_h(j)-\varphi_h(j)| \le \frac{h}{(1+1/h)^j}\,O\left\{\left(\frac{h}{\lambda_{\min}(\mathbf{\Gamma}_h)}\right)^{\frac{1}{2}}\left(\frac{\log T}{T}\right)^{\frac{1}{2}-d'}\right\}.$$

uniformly in j and $h\le H_T$. Moreover,

$$\sum_{j=0}^{\infty}|\bar{\varphi}_h(j)-\varphi_h(j)| = O\left\{\left(\frac{h^5}{\lambda_{\min}(\mathbf{\Gamma}_h)}\right)^{\frac{1}{2}}\left(\frac{\log T}{T}\right)^{\frac{1}{2}-d'}\right\}.$$

Proof: By definition

$$\bar{\varphi}_h(z)-\varphi_h(z)=\frac{\phi_h(z)-\bar{\phi}_h(z)}{\bar{\phi}_h(z)\phi_h(z)}$$

and the first part of the lemma follows directly from Cauchy's inequality for holomorphic functions and Theorem 3.1. We now have

$$\begin{aligned}\sum_{j=0}^{\infty}|\bar{\varphi}_h(j)-\varphi_h(j)| &\leq \sum_{j=0}^{\infty}\frac{h}{(1+1/h)^j}O\left\{\left(\frac{h}{\lambda_{\min}(\mathbf{\Gamma}_h)}\right)^{\frac{1}{2}}\left(\frac{\log T}{T}\right)^{\frac{1}{2}-d'}\right\}\\ &= O\left\{\left(\frac{h^5}{\lambda_{\min}(\mathbf{\Gamma}_h)}\right)^{\frac{1}{2}}\left(\frac{\log T}{T}\right)^{\frac{1}{2}-d'}\right\},\end{aligned}$$

as required. ∎

In the analysis of infinite autoregressions it is common practice to handle the truncation effect due to using an $AR(h)$ approximation by appealing to Baxter (1962)'s inequality, see also Berk (1974). Since under present assumptions an infinite autoregressive representation is not guaranteed to exist we cannot employ that technique here. We can, nevertheless, handle the consequences of using an $AR(h)$ approximation by using the following lemma.

Lemma 6.3 *Assume that the process $y(t)$ satisfies Assumption 2. Then for all $\delta > 0$ there exists an h sufficiently large such that $|\phi_h(e^{i\omega})k(e^{i\omega})-1| < \delta$ a.e. for $\omega \in (-\pi, \pi]$.*

Proof: Using the standard isometric isomorphism between the time and frequency domains we find that the lemma is an immediate consequence of the fact that as h increases $\epsilon_h(t)$ converges to $\varepsilon(t)$ in mean square. Indeed, let $\rho(z) = \sum_{j\geq 1}\rho(j)z^j = \phi_h(z)k(z)-1$. Then $\epsilon_h(t)-\varepsilon(t) = \sum_{j\geq 1}\rho(j)\varepsilon(t-j)$ and from Parseval's relation

$$\sum_{j\geq 1}\rho(j)^2 = \int_{-\pi}^{\pi}|\phi_h(e^{i\omega})k(e^{i\omega})-1|^2 d\omega = 2\pi\sigma^{-2}E[(\epsilon_h(t)-\varepsilon(t))^2]\,.$$

Since the mean squared difference $E[(\epsilon_h(t)-\varepsilon(t))^2]$ can be made arbitrarily small by taking h sufficiently large, we can conclude, via Arzelà's Theorem and Munroe (1953, Theorem 25.7), that $|\phi_h(e^{i\omega})k(e^{i\omega})-1| < \delta$ a.e. for $\omega \in (-\pi, \pi]$. ∎

Proof of Theorem 4.1: From the definition of Mallow's metric and Assumption 4, and applying the Cauchy-Schwartz inequality (twice), we have

$$\begin{aligned}\{\eta(F_{\mathbf{S}_T^*}, F_{\mathbf{S}_T})\}^2 &\leq E[E^*[\|\mathbf{S}_T^*-\mathbf{S}_T\|^2]]\\ &\leq \frac{1}{T}\sum_{t=1}^{T}E[E^*[B_t(y_t,y_t^*)^2]]\cdot\frac{1}{T}\sum_{t=1}^{T}E[E^*[(y(t)-y^*(t))^2]]\\ &\leq \limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}E[E^*[B_t(y_t,y_t^*)^2]]\cdot\frac{1}{T}\sum_{t=1}^{T}E[E^*[(y(t)-y^*(t))^2]]\,.\end{aligned}$$

By successive substitution into the recursions

$$y^*(t) = \epsilon_h^*(t) - \sum_{j=1}^{h}\bar{\phi}_h(j)y^*(t-j) \text{ and } y(t) = \epsilon_h(t) - \sum_{j=1}^{h}\phi_h(j)y(t-j)$$

we obtain the representations

$$y^*(t) = \sum_{j=0}^{\infty}\bar{\varphi}_h(j)\epsilon_h^*(t-j) \text{ and } y(t) = \sum_{j=0}^{\infty}\varphi_h(j)\epsilon_h(t-j)\,,$$

from which it follows that

$$\begin{aligned} y(t) - y^*(t) &= \sum_{j=0}^{\infty}(\varphi_h(j) - \bar{\varphi}_h(j))\epsilon_h(t-j) + \sum_{j=0}^{\infty}\bar{\varphi}_h(j)(\epsilon_h(t-j) - \epsilon_h^*(t-j)) \\ &= u(t) + v(t)\,, \text{ say.} \end{aligned}$$

Thus we are faced with the task of evaluating $E[E^*[(u(t)+v(t))^2]]$.

Consider first $E[E^*[v(t)^2]]$. By construction $\epsilon_h(t) - \epsilon_h^*(t)$ are *i.i.d.* with respect to P^* and $\epsilon_h(t) - \epsilon_h^*(t) = \epsilon_h(\tau) - \epsilon_h^*(\tau)$ with probability $1/T$, $\tau \in \{1, \dots, T\}$. Hence

$$E^*[v(t)^2] = T^{-1}\sum_{\tau=1}^{T}(\epsilon_h(\tau) - \epsilon_h^*(\tau))^2 \cdot \sum_{j=0}^{\infty}|\bar{\varphi}_h(j)|^2\,.$$

The first term in the product on the right hand side is

$$\begin{aligned} T^{-1}\sum_{\tau=1}^{T}(\epsilon_h(\tau) - \epsilon_h^*(\tau))^2 &= T^{-1}\sum_{\tau=1}^{T}(\epsilon_h(\tau) - \bar{\sigma}_h\tilde{\epsilon}_h(\tau))^2 \\ &= T^{-1}\sum_{\tau=1}^{T}\left(\epsilon_h(\tau) - \bar{\epsilon}_h(\tau) + \bar{\epsilon}_h(\tau)\left(1 - (\frac{\bar{\sigma}_h}{s_{\bar{\epsilon}_h}})\right) + \bar{\epsilon}_h(\frac{\bar{\sigma}_h}{s_{\bar{\epsilon}_h}})\right)^2 \\ &= T^{-1}\sum_{\tau=1}^{T}(\epsilon_h(\tau) - \bar{\epsilon}_h(\tau))^2 + o(1)\,, \end{aligned}$$

where the final line is a consequence of the fact that $1 - (\bar{\sigma}_h/s_{\bar{\epsilon}_h})$ and $\bar{\epsilon}_h$ are both $o(1)$. To show that $1 - (\bar{\sigma}_h/s_{\bar{\epsilon}_h})$ and $\bar{\epsilon}_h$ are $o(1)$ first note that

$$\bar{\epsilon}_h = T^{-1}\sum_{t=1}^{T}\epsilon_h(t) + T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{h}\{\bar{\phi}_h(j) - \phi_h(j)\}y(t-j) = o(1)\,.$$

Hence $s_{\bar{\epsilon}_h}^2 = T^{-1}\sum_{t=1}^{T}\bar{\epsilon}_h(t)^2 + o(1)$ and by Lemma 6.1 this equals $T^{-1}\sum_{t=1}^{T}\epsilon_h(t)^2 + o(1) = \sigma_h^2 + o(1)$. Thus $\bar{\sigma}_h^2/s_{\bar{\epsilon}_h}^2 \sim 1$ and $1 - (\bar{\sigma}_h/s_{\bar{\epsilon}_h}) = o(1)$. To bound the second term, recall that $\phi_h(z) \neq 0$, $|z| \leq 1$. This implies that constants $C < \infty$ and $\zeta < 1$ exist such that $|\varphi_h(j)| < C\zeta^j$ for $j = 1, 2, \dots$ and hence that $\sum_{j=0}^{\infty}|\varphi_h(j)| < \infty$. Using Lemma 6.2 we are lead to the conclusion that

$$\begin{aligned} \sum_{j=0}^{\infty}|\bar{\varphi}_h(j)| &\leq \sum_{j=0}^{\infty}|\varphi_h(j)| + \sum_{j=0}^{\infty}|\bar{\varphi}_h(j) - \varphi_h(j)| \\ &= O(1) + O\left\{\left(\frac{h^5}{\lambda_{\min}(\mathbf{\Gamma}_h)}\right)^{\frac{1}{2}}\left(\frac{\log T}{T}\right)^{\frac{1}{2}-d'}\right\}, \end{aligned}$$

and from Lemma 6.1 it follows that

$$E[E^*[v(t)^2]] = O\left\{\left(\frac{h}{\lambda_{\min}(\mathbf{\Gamma}_h)}\right)\left(\frac{\log T}{T}\right)^{1-2d'}\right\}. \tag{6.3}$$

Now consider $E[E^*[u(t)^2]]$. Since $u(t)$ is a constant relative to the probability measure

P^* we have $E[E^*[u(t)^2]] = E[u(t)^2 E^*[1]]$ and

$$E[u(t)^2] = \frac{\sigma^2}{2\pi}\int_{-\pi}^{\pi} |\varphi_h(e^{i\omega}) - \bar{\varphi}_h(e^{i\omega}))|^2 |\phi_h(e^{i\omega})k(e^{i\omega})|^2 d\omega\,.$$

From Lemma 6.3 however

$$\begin{aligned} |\phi_h(e^{i\omega})k(e^{i\omega})| &\leq 1 + |\phi_h(e^{i\omega})k(e^{i\omega}) - 1| \\ &\leq 1+\delta \qquad \text{a.e.} \quad (-\pi, \pi] \end{aligned}$$

for any $\delta > 0$ for all h sufficiently large and

$$E[u(t)^2] \leq \frac{\sigma^2}{2\pi}\int_{-\pi}^{\pi} |\varphi_h(e^{i\omega}) - \bar{\varphi}_h(e^{i\omega}))|^2 (1+\delta)^2 d\omega = \frac{(\sigma(1+\delta))^2}{2\pi}\sum_{j=0}^{\infty} |\bar{\varphi}_h(j) - \varphi_h(j)|^2$$

as $h = h_T^{AIC} \to \infty$ as $T \to \infty$. Using Lemma 6.2 we can therefore conclude that

$$E[u(t)^2] = O\left\{\left(\frac{h^5}{\lambda_{\min}(\mathbf{\Gamma}_h)}\right)\left(\frac{\log T}{T}\right)^{1-2d'}\right\}. \tag{6.4}$$

Evaluating $E[E^*[(u(t)+v(t))^2]]$ using Minkowski's inequality in conjunction with (6.3) and (6.4) now yields the first statement of the theorem. The second statement follows by noting from Poskitt (2004, Lemma 5.7) that the presence of spectral zeroes of the type characterized by Assumption 3 implies that $1/\lambda_{\min}(\mathbf{\Gamma}_h)$ is of order $O\{h^{2q}\}$ at most where $q \geq 0$. Thus $h^5/\lambda_{\min}(\mathbf{\Gamma}_h) \leq O\{(\log T)^{a(5+2q)}\}$ for all $h \leq M_T = [c(\log T)^a]$, and $(\log T)^{a(5+2q)}/T^{\beta} \to 0$ as $T \to \infty$ for all $\beta > 0$. ∎

Proof of Theorem 4.2: From equation (4.4) it is sufficient for us to establish that $\|\mathbf{K}_r^* - \mathbf{K}_r\| = O(T^{-\frac{1}{2}+d'+\beta})$ for any $\beta > 0$. This will follow if we can show that for all $\boldsymbol{\theta}$ in a neighbourhood of the origin $|\log\psi_T^*(\boldsymbol{\theta}) - \log\psi_T(\boldsymbol{\theta})| = O\{T^{-\frac{1}{2}+d'+\beta}\}$ uniformly in $\boldsymbol{\theta}$, which is equivalent to showing that $|\psi_T^*(\boldsymbol{\theta}) - \psi_T(\boldsymbol{\theta})| = O\{T^{-\frac{1}{2}(1-2d')+\beta}\}$.

Lemma 6.4 *Suppose that the process $y(t)$ satisfies Assumptions 1, 2 and 3, and that the statistic $\mathbf{S}_T$ satisfies Assumption 4. Then $|\psi_T^*(\boldsymbol{\theta}) - \psi_T(\boldsymbol{\theta})| = O\{T^{-\frac{1}{2}(1-2d')+\beta}\}$ for all $\beta > 0$ uniformly in $\boldsymbol{\theta}$.*

Proof: Consider the linear combinations $l_T = \boldsymbol{\lambda}'\mathbf{S}_T$ and $l_T^* = \boldsymbol{\lambda}'\mathbf{S}_T^*$ where $\boldsymbol{\lambda}$ is any fixed vector of unit length. Then $|\psi_T^*(\boldsymbol{\theta}) - \psi_T(\boldsymbol{\theta})| = |\psi_T^*(t) - \psi_T(t)|$ for $\boldsymbol{\theta} = t\boldsymbol{\lambda}$, where $\psi_T^*(t)$ and $\psi_T(t)$ denote the characteristic functions of l_T^* and l_T respectively. By Theorem 25.6 and Exercise 26–k of Munroe (1953), however, $|\psi_T^*(t) - \psi_T(t)| \leq V(F_T^* - F_T)$ where $V(F_T^* - F_T)$ is the total variation of $F_T^* - F_T$, the difference in the distribution functions of l_T^* and l_T. By definition

$$V(F_T^* - F_T) = \sup\sum_{i=1}^{M} |(F_T^*(x_i) - F_T(x_i)) - (F_T^*(x_{i-1}) - F_T(x_{i-1}))|$$

where the supremum is taken over all possible finite partitions of $\mathbb{R}$, namely, $-\infty < x_0 < x_1 < \ldots < x_M < \infty$ with $M < \infty$. The result $|\psi_T^*(\boldsymbol{\theta}) - \psi_T(\boldsymbol{\theta})| = O\{T^{-\frac{1}{2}(1-2d')+\beta}\}$ follows, *a là* the Cramèr-Wold device, by establishing that $V(F_T^* - F_T) = O\{T^{-\frac{1}{2}(1-2d')+\beta}\}$.

Consider then the events $\{l_T \leq x+\epsilon\}$, $\{l_T^* \leq x\}$ and $\{|l_T^* - l_T| > \epsilon\}$ where $x \in \mathbb{R}$ and $\epsilon > 0$. Since $\{l_T^* \leq x\}$ equals the union of $\{l_T^* \leq x\} \cap \{|l_T^* - l_T| \leq \epsilon\}$ and $\{l_T^* \leq x\} \cap \{|l_T^* - l_T| > \epsilon\}$, and $\{l_T^* \leq x\} \cap \{|l_T^* - l_T| \leq \epsilon\} \subseteq \{l_T \leq x+\epsilon\}$, it follows that $F_T^*(x)$ is bounded above by $F_T(x+\epsilon) + P(P^*(|l_T^* - l_T| > \epsilon))$. Similarly, $F_T^*(x)$ can be bounded below by $F_T(x-\epsilon) - P(P^*(|l_T^* - l_T| > \epsilon))$. The distribution $F_T(x) = \int_{\lambda'\mathbf{s} \leq x} dF_{\mathbf{S}_T}(\mathbf{s})$ is absolutely continuous with respect to Lebesgue measure and differentiable for all x and by the first Mean Value Theorem the difference

$$F_T(x+\epsilon) - F_T(x-\epsilon) \leq 2m\epsilon$$

where $m = \sup_x \lim_{\epsilon \to 0}(F_T(x+\epsilon) - F_T(x-\epsilon))/2\epsilon$. Thus

$$|F_T^*(x) - F_T(x)| \leq 2m\epsilon + P(P^*(|l_T^* - l_T| > \epsilon))$$

uniformly in x. Applying Markov's inequality we have

$$P(P^*(|l_T^* - l_T| > \epsilon)) \leq \frac{E[E^*[|l_T^* - l_T|^2]]}{\epsilon^2}$$

and by the Cauchy-Schwartz inequality $E[[E^*[|l_T^* - l_T|^2]] \leq E[E^*[\|\mathbf{S}_T^* - \mathbf{S}_T\|^2]]$. As already shown, under Assumptions 1 and 2 $E[E^*[\|\mathbf{S}_T^* - \mathbf{S}_T\|^2]] = O\left\{\left(h^5/\lambda_{\min}(\mathbf{\Gamma}_h)\right)(\log T/T)^{1-2d'}\right\}$. Set $\epsilon = T^\beta/T^{\frac{1}{2}-d'}$ where $\beta > 0$, and recall that under Assumption 3 the ratio $h^5/\lambda_{\min}(\mathbf{\Gamma}_h) = O\{(\log T)^{a(5+2q)}\}$ where $q \geq 0$ for all $h \leq [c(\log T)^a]$. Then $P(P^*(|l_T^* - l_T| > T^\beta/T^{\frac{1}{2}-d'})) \leq O(T^{-2\beta}(\log T)^{1+a(5+2q)-2d'})$ and we find that

$$\sup_x |F_T^*(x) - F_T(x)| \leq T^{-\frac{1}{2}+d'+\beta}(2m + o(T^{\frac{1}{2}-d'-\beta}))\,.$$

Now, for every set of disjoint intervals $(x_{i-1}, x_i]$, $i = 1, \ldots, M$, we have

$$\sum_{i=1}^{M} |(F_T^*(x_i) - F_T(x_i)) - (F_T^*(x_{i-1}) - F_T(x_{i-1}))| \leq 2M \sup_x |F_T^*(x) - F_T(x)|\,.$$

We can therefore conclude that $V(F_T^* - F_T) = O\{T^{-\frac{1}{2}(1-2d')+\beta}\}$, as required. ∎

The heuristics behind Lemma 6.4 is straightforward; convergence of Mallow's metric implies convergence in distribution and hence, via an analogy of the Cramér-Levy continuity theorem, convergence of the characteristic function. Lemma 6.4 implies that $\|\mathbf{K}_r^* - \mathbf{K}_r\| = O(T^{-\frac{1}{2}+d'+\beta})$ for all $r \geq 1$ and Theorem 4.2 follows directly. ∎

Proof of Theorem 4.3: The events $\{T^{-\frac{1}{2}}(\mathbf{S}_T^* - \mathbf{S}_T) \leq \mathbf{s}\}$ and $\{T^{-\frac{1}{2}}(\mathbf{S}_T - E[\mathbf{S}_T]) \leq \mathbf{s}\}$ are equivalent to $\{\mathbf{Z}_T^* \leq \mathbf{z}^* + \boldsymbol{\zeta}_T\}$, where $\mathbf{z}^* = T^{\frac{1}{2}}{\mathbf{V}^*}_T^{-\frac{1}{2}}\mathbf{s}$ and $\boldsymbol{\zeta}_T = {\mathbf{V}^*}_T^{-\frac{1}{2}}(\mathbf{S}_T - E^*[\mathbf{S}_T^*])$, and $\{\mathbf{Z}_T \leq \mathbf{z}\}$, where $\mathbf{z} = T^{\frac{1}{2}}\mathbf{V}_T^{-\frac{1}{2}}\mathbf{s}$, respectively. Thus

$$\mathrm{P}^*(T^{-\frac{1}{2}}(\mathbf{S}_T^* - \mathbf{S}_T) \leq \mathbf{s}) - \mathrm{P}(T^{-\frac{1}{2}}(\mathbf{S}_T - E[\mathbf{S}_T]) \leq \mathbf{s}] = \mathrm{P}^*(\mathbf{Z}_T^* \leq \mathbf{z}^* + \boldsymbol{\zeta}_T) - \mathrm{P}(\mathbf{Z}_T \leq \mathbf{z})$$

and

$$|\mathrm{P}^*(\mathbf{Z}_T^* \leq \mathbf{z}^* + \boldsymbol{\zeta}_T) - \mathrm{P}(\mathbf{Z}_T \leq \mathbf{z})| \leq I_1 + I_2 + I_3$$

where

$$\begin{aligned} I_1 &= |\mathrm{P}^*(\mathbf{Z}_T^* \le \mathbf{z}^* + \boldsymbol{\zeta}_T) - \mathrm{P}^*(\mathbf{Z}_T^* \le \mathbf{z}^*)|\,, \\ I_2 &= |\mathrm{P}^*(\mathbf{Z}_T^* \le \mathbf{z}^*) - \mathrm{P}(\mathbf{Z}_T \le \mathbf{z}^*)| \qquad \text{and} \\ I_3 &= |\mathrm{P}(\mathbf{Z}_T \le \mathbf{z}^*) - \mathrm{P}(\mathbf{Z}_T \le \mathbf{z})|\,. \end{aligned}$$

To evaluate the magnitude of I_1 note that the derivatives of $G(\mathbf{z})$ and $g(\mathbf{z})$ of all orders, which can be expressed in terms of the covariant Hermite polynomials, are uniformly bounded in $\mathbb{R}^m$. Applying the Mean Value Theorem to (4.3) it follows that there exists a constant $m^* < \infty$ such that $\sup_{\mathbf{z}^*} I_1 \le m^* \|\boldsymbol{\zeta}_T\| + o(T^{-\frac{1}{2}})$. But $\|\boldsymbol{\zeta}_T\| \le \|(T^{\frac{1}{2}} \mathbf{V^*}_T^{-\frac{1}{2}})\| \cdot \|T^{-\frac{1}{2}}(\mathbf{S}_T - E^*[\mathbf{S}_T^*])\|$. Recognizing that $\mathbf{S}_T$ is a constant relative to the measure P^* we have

$$\begin{aligned} E[\|T^{-\frac{1}{2}}(\mathbf{S}_T - E^*[\mathbf{S}_T^*])\|^2] &= E[\|T^{-\frac{1}{2}} E^*[(\mathbf{S}_T - \mathbf{S}_T^*)]\|^2] \\ &\le T^{-1} E[E^*[\|(\mathbf{S}_T - \mathbf{S}_T^*)\|^2]] \\ &= T^{-1} O\left\{ \frac{(\log T)^{1+a(5+2q)-2d'}}{T^{1-2d'}} \right\}. \end{aligned}$$

Given that $T^{-1}\mathbf{V}_T^* = O(1)$ it follows from Markov's inequality that, for all $\beta > 0$, $\|\boldsymbol{\zeta}_T\|$ will exceed $T^{-(1-d')+\beta}$ with a probability that is bounded above by $O\{(\log T)^{1+a(5+2q)-2d'}/T^{2\beta}\}$. We can therefore conclude that I_1 is $O_p(T^{-(1-d')+\beta})$ uniformly in $\mathbf{z}^*$.

As in Theorem 4.2, we have $\|\mathbf{K}_r^* - \mathbf{K}_r\| = O(T^{-(\frac{1}{2}-d')+\beta})$ for all $r \ge 1$ and $\beta > 0$, and from equation (4.4) it follows that I_2 is $O(T^{-(1-d')+\beta})$ uniformly in $\mathbf{z}^*$.

Using the Mean Value Theorem in conjunction with (4.2) we can, as with I_1, bound I_3 by $m\|\mathbf{z} - \mathbf{z}^*\| + o(T^{-\frac{1}{2}})$ for some $m < \infty$. Now, $\mathbf{z} - \mathbf{z}^* = T^{\frac{1}{2}}(\mathbf{V}_T^{-\frac{1}{2}} - \mathbf{V^*}_T^{-\frac{1}{2}})\mathbf{s}$ and from Lemma 6.4 we have $T^{-1}\|\mathbf{V}_T - \mathbf{V}_T^*\| = O(T^\beta/T^{(3/2-d')})$ for all $\beta > 0$. Thus, if $\|\mathbf{s}\| \le C(T^\beta \log T/T^{(1-d')})^{\frac{1}{2}}$ where $C < \infty$ we can deduce that $\|\mathbf{z} - \mathbf{z}^*\| = O\{(T^\beta/T^{(1-d')})(\log T/T^{\frac{1}{2}})^{\frac{1}{2}}\}$. On the other hand, $\min\{\|\mathbf{z}\|, \|\mathbf{z}^*\|\} > C'(T^\beta \log T/T^{(1-d')})^{\frac{1}{2}}$ for any given $C' < \infty$ whenever $\|\mathbf{s}\| > C(T^\beta \log T/T^{(1-d')})^{\frac{1}{2}}$ and C is sufficiently large. From (4.2), however, it follows that $\mathrm{P}(\|\mathbf{Z}_T\| > C''(\log T)^{\frac{1}{2}}) = o(T^{-\frac{1}{2}})$ for any $C'' > 0$ and $\mathrm{P}(\|\mathbf{Z}_T\| > C'(T^\beta \log T/T^{(1-d')})^{\frac{1}{2}}) \le \mathrm{P}(\|\mathbf{Z}_T\| > C''(\log T)^{\frac{1}{2}})$ for all $C'' < C'(T^\beta/T^{(1-d')})^{\frac{1}{2}}$. Hence we are lead to the conclusion that, bar a set whose probability is $o(T^{-\frac{1}{2}})$, the term $I_3 = o(T^{-(1-d')+\beta)})$. Bringing this bound on I_3 together with those on I_1 and I_2 completes the proof. ∎

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of Institute of Statistical Mathematics* **21** 243–247.

Andrews, D. W. K. (2004). The block-block bootstrap: Improved asymptotic refinements. *Econometrica* **72** 673–700.

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. *Math. Scand.* **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*. Chapman and Hall, New York.

Berk, K. N. (1974). Consistent autoregressive spectral estimation. *Annals of Statistics* **2** 489–502.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics* **9** 1196–1217.

Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli* **3** 123–148.

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

Choi, E. and Hall, P. G. (2000). Bootstrap confidence regions from autoregressions of arbitrary order. *Journal of the Royal Statistical Society* **B 62** 461–477.

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

Durbin, J. (1980). Approximations for densities of sufficient estimators. *Biometrika* **67** 311–333.

Giraitis, L. and Robinson, P. M. (2003). Edgeworth expansions for semiparametric whittle estimation of long memory. *Annals of Statistics* **31** 1325–1375.

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 Deistler, M. (1988). *The Statistical Theory of Linear Systems*. Wiley, New York.

Horowitz, J. L. (2003). Bootstrap methods for markov processes. *Econometrica* **71** 1049–1082.

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.

Künsch, H. R. (1989). The jacknife and the bootstrap for general stationary observations. *Annals of Statistics* **17** 1217–1241.

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

Lieberman, O. and Phillips, P. C. B. (2004). Expansions for the distribution of the maximum likelihood estimator of the fractional difference parameter. *Econometric Theory* **20** 464–484.

Lieberman, O., Rousseau, J. and Zucker, D. M. (2001). Valid edgeworth expansion for the sample autocorrelation function under long range dependence. *Econometric Theory* **17** 257–275.

Lieberman, O., Rousseau, J. and Zucker, D. M. (2003). Valid asymptotic expansions for the maximum likelihood estimator of the parameter of a stationary, Gaussian, strongly dependent process. *Annals of Statistics* **31** 586–612.

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

Munroe, M. E. (1953). *Introduction to Measure and Integration*. Addison-Wesley, Reading.

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

Paulsen, J. and 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. (2004). Autoregressive approximation in nonstandard situations: The non-invertible and fractionally integrated cases. Tech. rep., Department of Econometrics and Business Statistics, Monash University, Victoria 3800: `http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/2005/`.

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.

Taniguchi, M. (1984). Validity of edgeworth expansions for statistics of time series. *Journal of Time Series Analysis* **5** 37–51.

Tjøstheim, D. and Paulsen, J. (1983). Bias of some commonly-used time series estimators. *Biometrika* **70** 389–400.

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

Wold, H. (1938). *The Analysis of Stationary Time Series*. Almquist and Wicksell, Uppsala, 2nd ed.

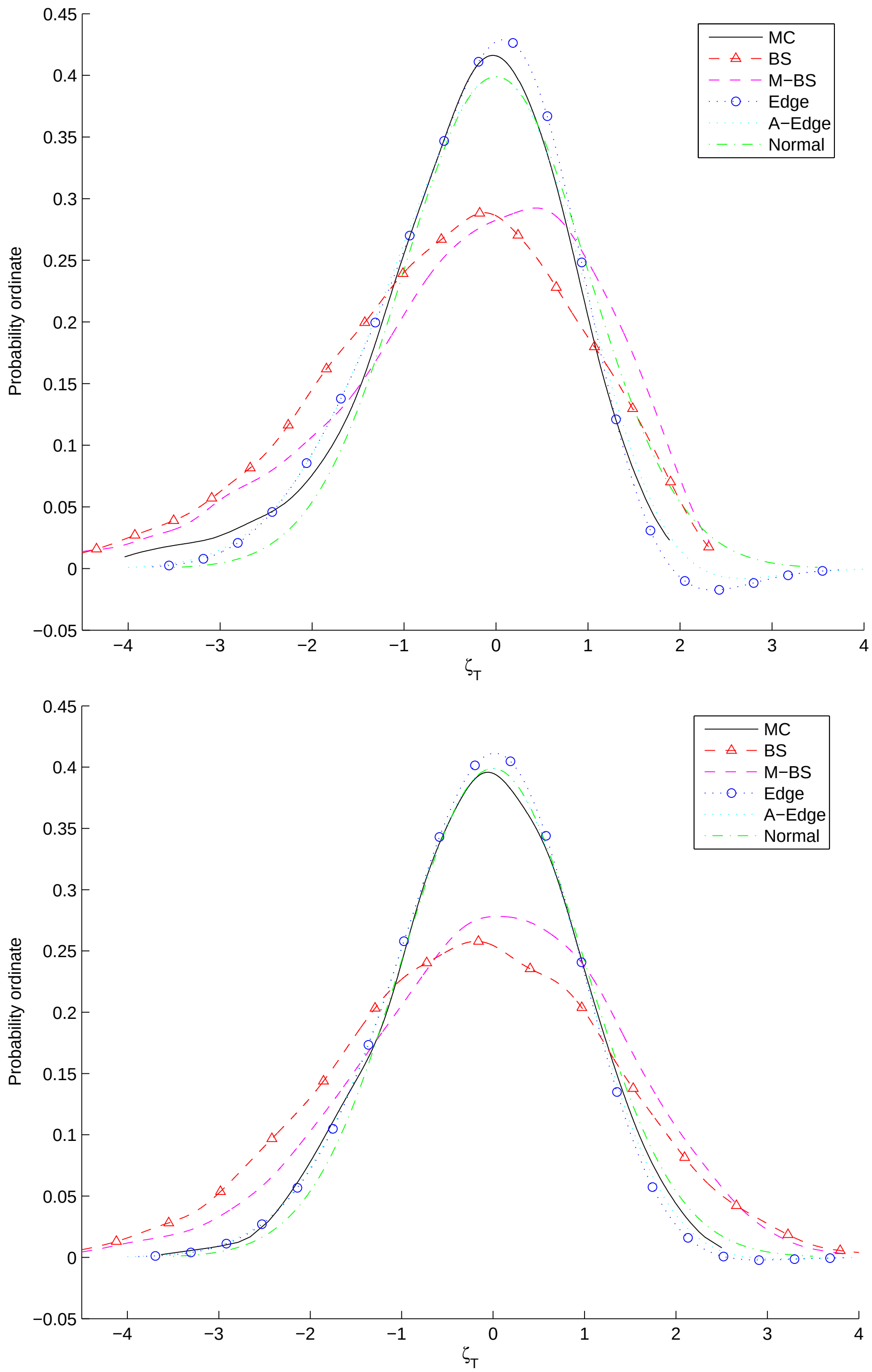


Figure 1: Probability densities for $d = 0.2$, $T = 40$, top panel, and $T = 160$, bottom panel: Exact (Monte–Carlo) (black), Edgeworth (blue), Approximate–Edgeworth (cyan), Normal (green), Model Bootstrap (magenta), Sieve Bootstrap (ζ_T^*) (red)

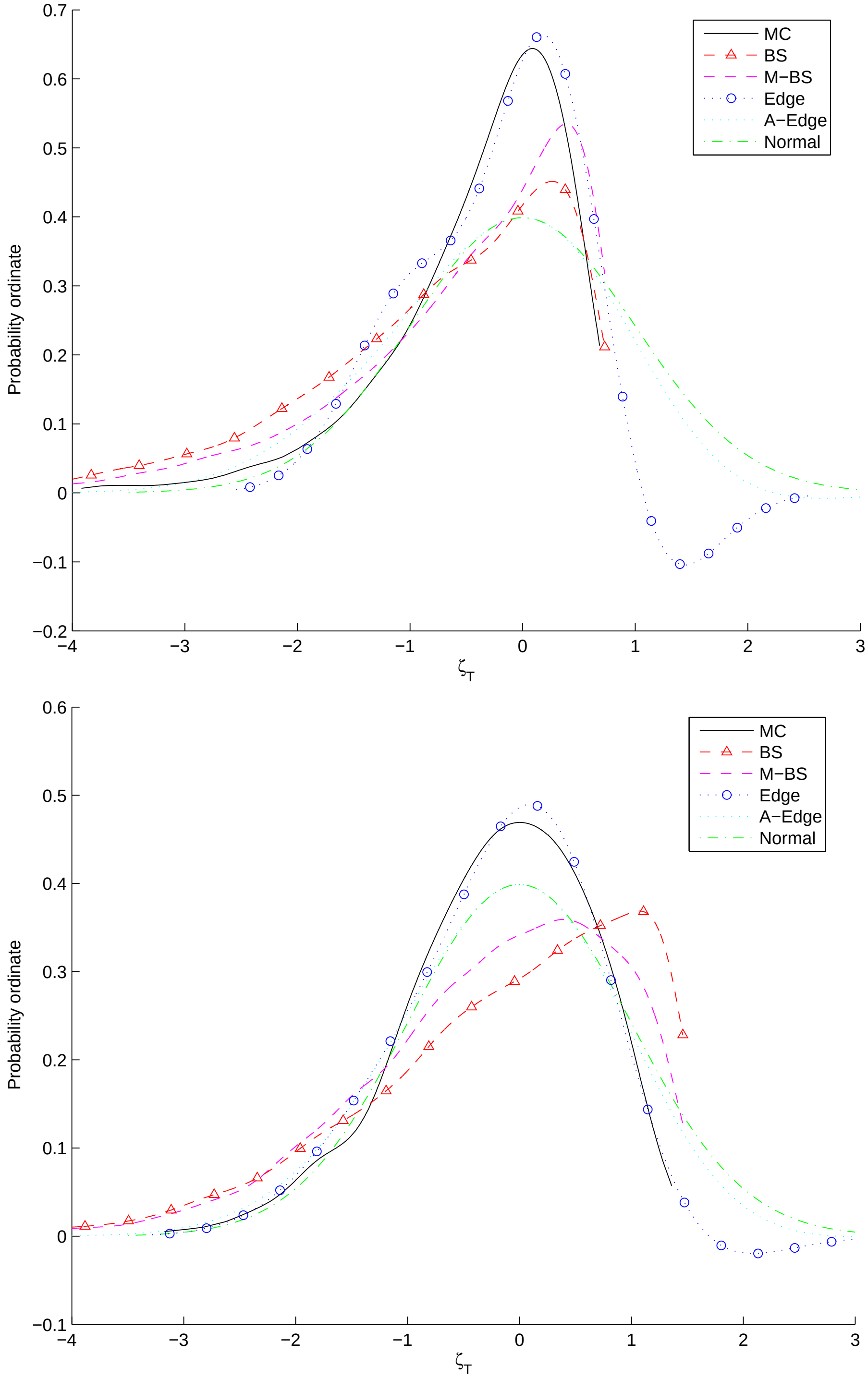


Figure 2: Probability densities for $d = 0.4$, $T = 40$, top panel, and $T = 160$, bottom panel: Exact (Monte–Carlo) (black), Edgeworth (blue), Approximate–Edgeworth (cyan), Normal (green), Model Bootstrap (magenta), Sieve Bootstrap (ζ_T^*) (red)

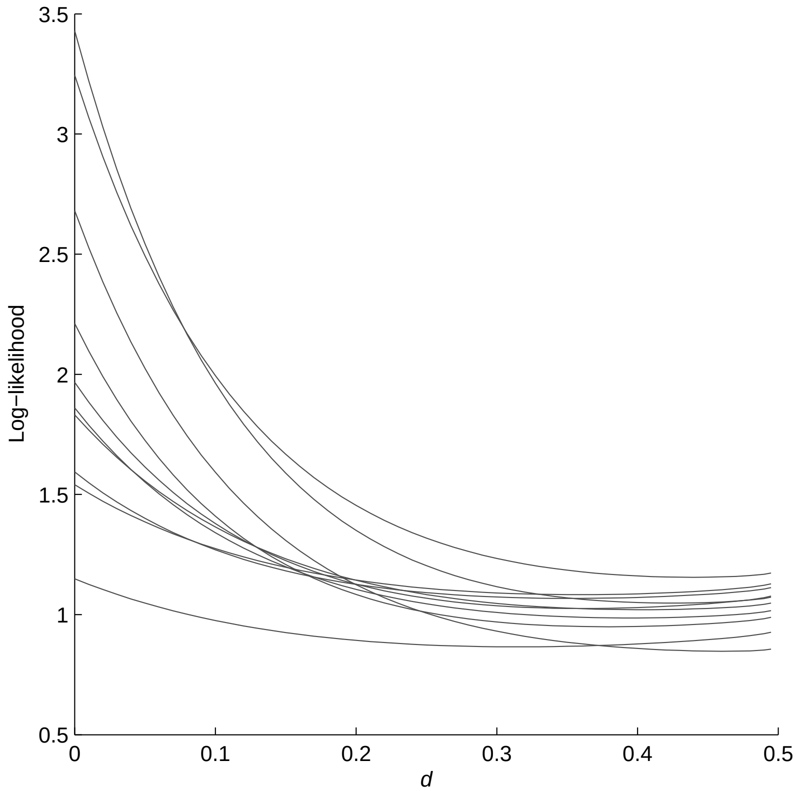


Figure 3: Log-likelihood function for 10 realizations of process $(1-z)^d y(t) = \varepsilon(t)$ when $d = 0.4$, $T = 160$.

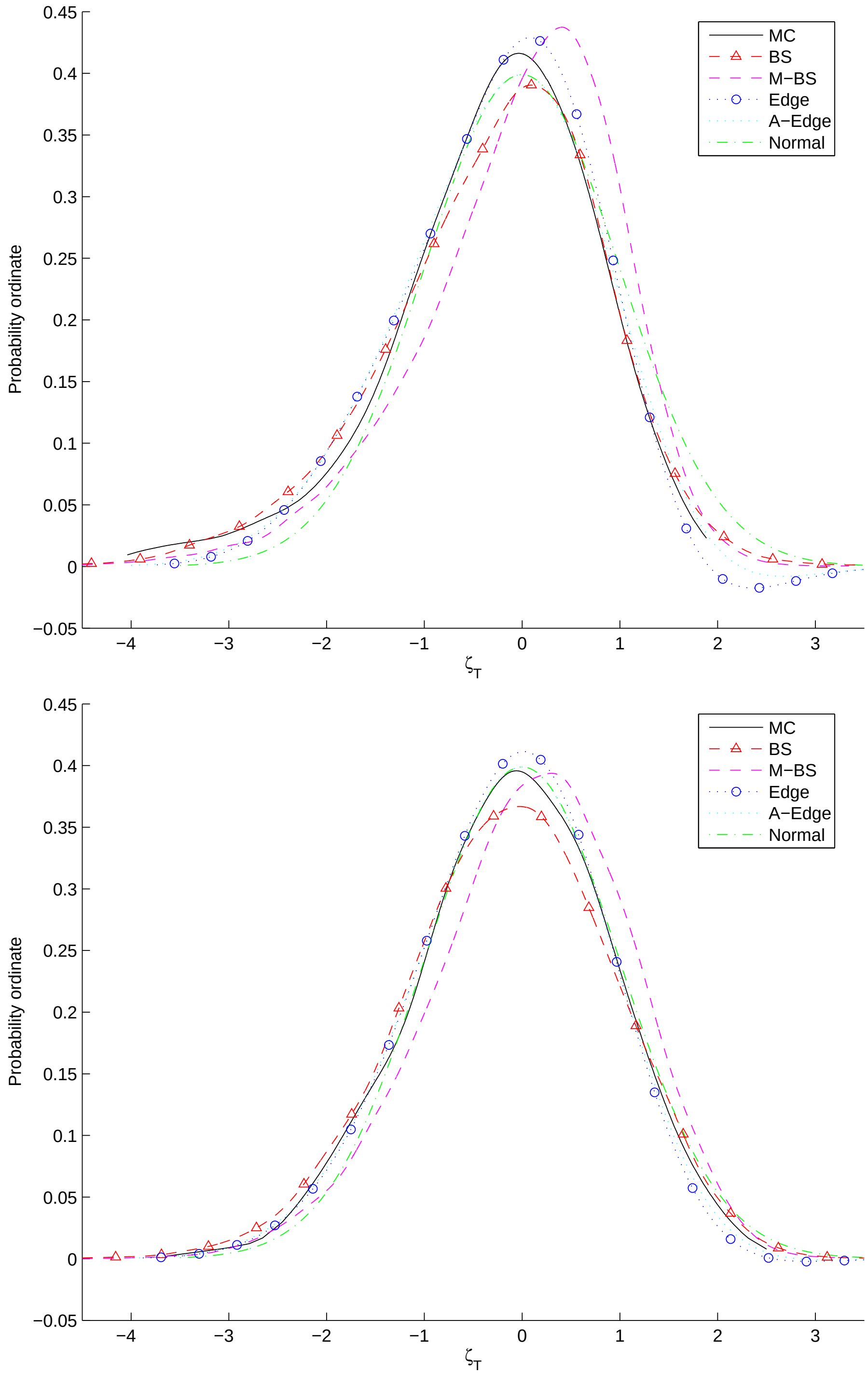
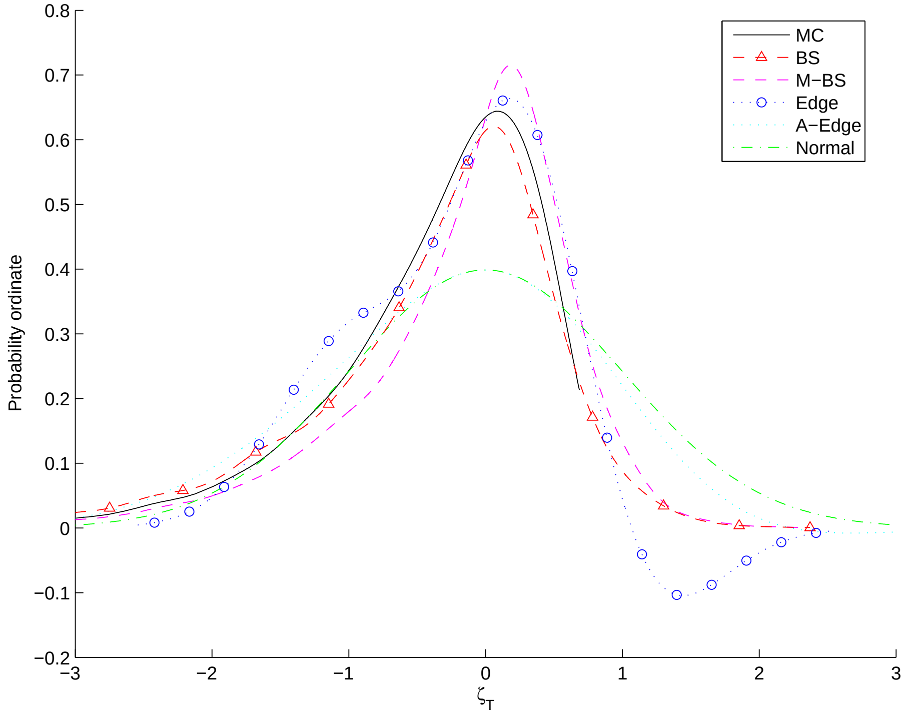


Figure 4: Probability densities for $d = 0.2$, $T = 40$, top panel, and $T = 160$, bottom panel: Exact (Monte–Carlo) (black), Edgeworth (blue), Approximate–Edgeworth (cyan), Normal (green), Model Bootstrap (magenta), Sieve Bootstrap ($\tilde{\zeta}_T^*$) (red)



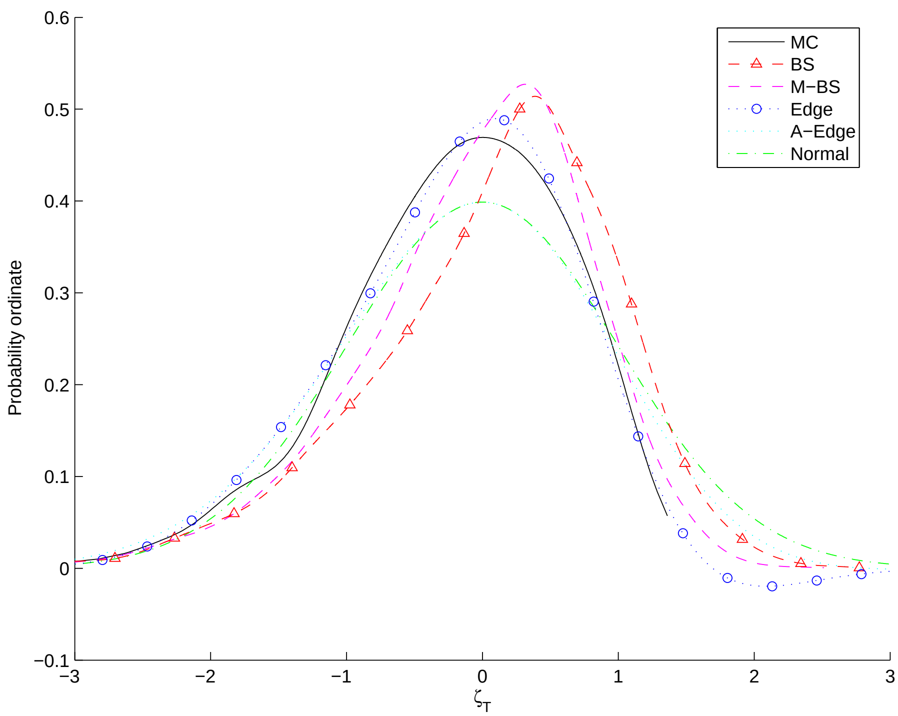


Figure 5: Probability densities for $d = 0.4$, $T = 40$, top panel, and $T = 160$, bottom panel: Exact (Monte–Carlo) (black), Edgeworth (blue), Approximate–Edgeworth (cyan), Normal (green), Model Bootstrap (magenta), Sieve Bootstrap ($\tilde{\zeta}_T^*$) (red)