Consistent estimation and order selection for non-stationary autoregressive processes with stable innovations.

Peter Burridge and Daniela Hristova
Department of Economics, City University London

May 1, 2007

Abstract

A possibly non-stationary autoregressive process, of unknown finite order, with possibly infinite-variance innovations is studied. The Ordinary Least Squares autoregressive parameter estimates are shown to be consistent, and their rate of convergence, which depends on the index of stability, α, is established. We also establish consistency of lag-order selection criteria in the non-stationary case. A small experiment illustrates the relative performance of different lag-length selection criteria in finite samples.

Keywords: consistent estimation, infinite variance innovations, unit root AR processes, consistent order-selection criteria.

JEL classification: C13, C22.

Acknowledgements: We thank colleagues at City University, and anonymous referees, for comments on an earlier version, and the Leverhulme Trust, Award Reference F/00353/D, for funding Hristova's work. Any mistakes or omissions are our sole responsibility.

1 Introduction

The twin problems of consistent parameter estimation and lag-length selection in autoregressive models have received intensive study. For example, Gonzalo and Pitarakis (2002) discuss the behaviour of well-known model selection methods in large systems, while many papers address the problems of implementing tests for autoregressive unit roots, as in Hall (1994), and Ng and Perron (2001). It is usual practice to base lag-length selection either on a sequence of $t-tests$, or to select the length that minimises an information criterion (IC). When the innovations are draws from a distribution within the domain of attraction of the Normal distribution, (or, as in Potscher (1989), Martingale differences with more than two finite moments), it is well known that the IC of Akaike

(1974) (AIC) is inconsistent, over-fitting in the limit, as shown by Shibata (1976), while, for example, those of Schwarz (1978), Rissanen (1978) (BIC), and Hannan and Quinn (1979) ($HQIC$) are consistent. Discussions of these and related results may be found in Hannan and Quinn (1979), for the stationary case, and Potscher (1989), and Ng and Perron (2001), for the non-stationary case. However, many macroeconomic and financial series, notably stock returns, appear to have heavy-tailed distributions as described in Adler, Feldman and Taqqu (1998), for example, and this raises the question of the applicability of existing consistency results.

The question is potentially important: recently, Charemza, Hristova and Burridge (2005) have demonstrated the sensitivity of unit root test outcomes on inflation series to assumptions about the tails of the innovation distribution. They show that applying the ADF test to 93 inflation series, but treating the innovations as draws from a symmetric stable distribution with possibly infinite variance, reduces the number that appear stationary. This effect arises from the shift in the sampling distribution of the unit root test statistic identified by Chan and Tran (1989) and Phillips (1990), and quantified by Rachev et al (1998). However, that empirical result begs the question of lag length determination in the $\alpha - stable$ case. A further motivation is the possible use of a sieve-type bootstrap for inference, for which consistent parameter estimates would be required. It is thus important to establish the consistency of parameter estimation and lag order selection for processes with possibly heavy-tailed innovations under both stationarity and unit-root nonstationarity.

Before discussing least squares estimation in greater detail, it is worth noting that in the $\alpha - stable$ case, so-called $m - estimators$ can be very much more efficient than least squares, especially for α much less than 2. Some compelling numerical evidence on this point is presented by Calder and Davis (1998). Nevertheless, use of such estimators is not yet widespread in empirical analysis of economic time series, and in our own previous work we found that estimated α values on a sample of series were in many cases quite close to 2. With this limitation in mind, our objective is to fill a gap in the results that relate to the properties of least squares estimators in this setting.

A major contribution to the study of least squares estimators for stationary heavy-tailed processes, is that of Knight (1989), who shows that the order of autoregression can be consistently estimated by the AIC criterion, and hence also by $HQIC$ and BIC, calculated using the Yule-Walker estimator. He also proves consistency of the OLS parameter estimator in this context. In the present paper we study the least squares estimator of possibly non-stationary processes with no more than one unit autoregressive root. The paper is organized as follows. In Section 2 we define the processes and problem of interest, and give an informal description of our results. This is followed by a brief discussion of the proof strategy, in which the ingredients required for a proof of consistent lag-length selection are identified. A number of existing results, together with a formal statement of our main contribution, Theorems 5 and 9, are presented in Sections 3 and 4; a simplified lag-selection criterion is introduced in Section 5, and some simulations illustrate finite sample behaviour in

Section 6. Conclusions and brief discussion appear in Section 7, and proofs in the Appendices.

2 Definitions and the nature of our results

2.1 Definitions

Suppose the process X_t is a finite-degree autoregression with innovations following a stable law, and that X_t has at most one autoregressive unit root, i.e. X_t is either a stationary or a difference stationary process. Further, suppose X_t has no deterministic components such as level or trend. With L the usual $lag-$operator, let X_t be an $AR(m)$ process,

$$\begin{aligned} \Phi(L)X_t &= u_t, \\ \Phi(L) &= 1 - \sum_{j=1}^{m} \Phi_j L^j, \end{aligned} \tag{1}$$

with $m \in \mathbb{Z}^+$ and $\Phi(L) = (1-\rho L)\phi(L)$, where all the zeros of $\phi(L)$, $\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - ... - \phi_{m-1}L^{m-1}$, are outside the unit circle. For $|\rho| < 1$ the process X_t is stationary, while for $\rho = 1$ X_t is a non-stationary $AR(m)$ process with one unit root and stationary first difference Z_t,

$$Z_t = (1-L)X_t. \tag{2}$$

The true order, m, is unknown but bounded by some finite integer, K.

The disturbances u_t are iid random variables in the domain of attraction of a stable law with index of stability $\alpha \in (0,2)$, that is

$$\Pr[|u_1| > x] = x^{-\alpha}L(x)$$

and

$$\lim_{x\to\infty} \frac{\Pr[u_1 > x]}{\Pr[|u_1| > x]} = \mu \in [0,1],$$

with $L(x)$ a non-negative function, slowly varying at infinity, $\lim_{x\to\infty} \frac{L(sx)}{L(x)} = 1$, $\forall s > 0$ (see Feller, 1971, p.276). We will suppose that the distribution of u_t has "Pareto-like tails" (see, for example, Davis and Resnick, 1985) such that

$$\begin{aligned} a_n &= \inf\{x : \Pr[|u_1| > x] \leq n^{-1}\} \\ &= an^{1/\alpha}, \end{aligned}$$

which corresponds to taking $L(x) =$ constant. Innovations with such properties will be denoted by $u_t \tilde{} iid\ SP(\alpha)$ in the rest of the paper.

Although, in the cases of interest, $\alpha \in (0,2)$, the innovation variance is not finite, $E\{u_t^2\} = \infty$, we may still employ *sample* second moments which are perfectly well defined functions of the observations, with properties investigated extensively by Davis and Resnick (1985, 1986); see Lemma 4, below, for those properties we require.

2.2 The results in general terms

For processes of the type (1) with $|\rho| < 1$, consistency of the autoregressive order selected by minimizing a version of the AIC criterion, expressed as a function of the Yule-Walker (YW) estimate for the "innovation variance", is established by Knight (1989). In this paper we seek results for processes of the type (1) with $\rho = 1$ for which the Yule-Walker estimator is poorly defined and becomes numerically unstable, and so cannot be applied. The ordinary least squares parameter estimates do not suffer from this deficiency, so we shall work with information criteria defined in terms of the least squares estimator. Broadly speaking, we will show that the possible presence of the autoregressive unit root has little effect on the behaviour of lag-length selection criteria, while the effect of a unit root on the rate of convergence of the least squares estimator of the autoregressive coefficients is similar to its effect in the Gaussian case. The significance of these results is that they allow unit-root non-stationary processes driven by heavy-tailed innovations to be estimated by least squares in the same way as neighbouring stationary processes, thus opening the door to sieve bootstrap methods.

In Theorem 5 we find a rate of convergence for the least squares estimator of $\Phi(L)$ in (1) in the case in which $\rho = 1$ while in Theorem 9 we establish consistency of the lag order selected by the AIC criterion in the non-stationary $a - stable$ case, thus complementing the results of Knight (1989).

The results of using a lag-selection criterion are usually presented in such a way that the question of control of the significance level of the implicit hypothesis test remains unexplored. With that in mind we introduce an alternative definition of an IC that facilitates some size control and illustrate its merits in Section 5.

2.3 Proof strategy for consistent lag-length selection

Since our objective is to establish consistency outside the circumstances to which Knight's (1989) results apply, it is natural to consider whether this can be achieved by the same approach. We must first clarify what is to be proved. Let us begin with the definition of the relevant IC. Akaike's (1974) criterion is

$$AIC = \ln \hat{\sigma}_u^2 + \frac{2k}{n} \tag{3}$$

which was shown by Shibata (1976) to lead to overfitting in the stationary Gaussian case, because the *penalty* for increasing k was too small. Subsequently, Rissanen (1978) and Schwarz (1978) introduced

$$BIC = \ln \hat{\sigma}_u^2 + \frac{k \ln n}{n} \tag{4}$$

which, while consistent in a stationary Gaussian setting, was found not to embody the "best" obtainable rate, and so was further modified by Hannan and Quinn (1979) to

$$HQIC = \ln \hat{\sigma}_u^2 + \frac{c.k \ln(\ln n)}{n} \tag{5}$$

for some $c > 2$. There are numerous other lag selection criteria available, but these three are the ones most often reported. In each of these criteria, $\hat{\sigma}_u^2$ is an estimate of the variance of u, when this exists, obtained from the estimated model. In Akaike's formulation, $\hat{\sigma}_u^2$ is the MLE of σ_u^2 for Normal u, that is,

$$\hat{\sigma}_u^2 = n^{-1} \sum \hat{u}_t^2 \tag{6}$$

but *in stationary cases* a more convenient estimator for the purpose of establishing consistency was found to be that obtained from the Yule-Walker equations via the Levinson-Durbin recursion, due to Durbin (1960):

$$\breve{\sigma}_u^2(k) = \breve{\sigma}_u^2(0) \prod_{j=1}^{j=k} (1 - \breve{\Phi}_{j,j}^2) \tag{7}$$

in which $\breve{\sigma}_u^2(0) = n^{-1} \sum X_t^2$ and $\breve{\Phi}_{j,j}$ is the Yule-Walker estimate of the jth partial autocorrelation coefficient. Hannan and Quinn use (7) to define their information criterion, and the consistency proof for the stationary $\alpha - stable$ case devised by Knight also explicitly uses the YW estimators to form the AIC, and hence also exploits (7). This raises our first problem, which is that the YW estimator is not well defined in the unit root nonstationary case. However, we can show that for the OLS estimator

$$\hat{\sigma}_u^2(k) = \hat{\sigma}_u^2(0) \prod_{j=1}^{j=k} \{1 - \hat{\Phi}_{j,j}^2\}\{1 + \delta_j\}, \tag{8}$$

in which, as Theorem 9 shows, in all relevant cases, δ_j is of smaller order in probability than $\hat{\Phi}_{j,j}^2$ both for $j \leq m$ and for $j > m$, which is of great importance, as we now explain.

To see how consistency of lag-length selection using an IC defined in terms of the OLS estimator may be proved using the approximation, (8), consider a generic criterion that can be written,

$$IC_{OLS}(k) = \ln \hat{\sigma}_u^2(k) + kC(n). \tag{9}$$

The increment of $IC(k)$ from $k = m - j$ to $k = m$ is

$$\begin{aligned} IC_{OLS}(m) - IC_{OLS}(m-j) &= \ln(\frac{\hat{\sigma}_u^2(m)}{\hat{\sigma}_u^2(m-j)}) + jC(n) \qquad (10) \\ &= \sum_{k=m-j+1}^{k=m} \{\ln(1 - \hat{\Phi}_{k,k}^2) + \ln(1 + \delta_k)\} + jC(n). \end{aligned}$$

Now, provided $\hat{\Phi}_{m,m}$ is a consistent estimator of $\Phi_{m,m}$, and $0 < |\Phi_{m,m}| < 1$, the first term on the RHS is bounded above in the limit by $\ln(1 - \Phi_{m,m}^2)$, which

is negative. Hence, for large enough n the increment will be negative provided $\sum_{k=m-j+1}^{k=m} \ln(1+\delta_k) + jC(n)$ is $o_p(1)$ and it will follow that in the limit, $\hat{k}$ cannot be smaller than m.

To complete the argument, consider the possibility of over-fitting, that is, $m < k \leq K$. The increment of interest is now,

$$\begin{aligned} & IC_{OLS}(m+j) - IC_{OLS}(m) \qquad (11) \\ = & \sum_{k=m+1}^{k=m+j} \{\ln(1-\hat{\Phi}^2_{k,k}) + \ln(1+\delta_k)\} + jC(n). \end{aligned}$$

The probability that a k is chosen such that $m < k \leq K$ is clearly smaller than

$$\Pr\{IC_{OLS}(k) < IC_{OLS}(k-1)\}$$

for *some* such k, that is, it is smaller than

$$\Pr\{\min[\ln[(1-\hat{\Phi}^2_{k,k})(1+\delta_k)] + C(n)] < 0\}. \qquad (12)$$

Now suppose that for any $k > m$, $\hat{\Phi}^2_{k,k} \to 0$ while $\delta_k / \hat{\Phi}^2_{k,k} = o_p(1)$, then for large enough n we may write

$$\begin{aligned} \ln[(1-\hat{\Phi}^2_{k,k})(1+\delta_k)] &= \ln(1-\hat{\Phi}^2_{k,k}) + \ln(1+\delta_k) \\ &\simeq -\hat{\Phi}^2_{k,k} \end{aligned}$$

and we see that to prove that (12) converges to zero, it will suffice to establish that $\max_{m<k\leq K(n)} \hat{\Phi}^2_{k,k}/C(n) = o_p(1)$.

3 Convergence of the OLS Estimator

We start with a lemma of Tiao and Tsay (1983, p.857), which ensures the existence of the OLS coefficient estimates in an autoregression of arbitrary order l.

Lemma 1 *For an $ARIMA(l,d,q)$ process X_t and a positive integer p, let $Y_t = (X_t, X_{t-1}, ..., X_{t-p+1})'$. If X_t is not a purely deterministic process, then, for $n \geq 2p$, $A_n = \sum_{t=p+1}^{n} Y_{t-1}Y'_{t-1}$ is a symmetric and positive definite matrix with probability 1.*

Next, we record a result from Hannan and Kanter (1977) which gives an *a.s.* rate of convergence of the least squares estimator in the stationary $SP(\alpha)$ case. It is convenient now to add an index, p to the coefficients to indicate variation in the number of elements of Φ with the number of lags included in the estimated model:

Theorem 2 *Let X_t be an autoregressive process given by (1) with $|\rho| < 1$ and $u_t \sim iid\ SP(\alpha)$ for $\alpha \in (0,2)$. Let $\hat{\Phi}_{i,p}$ denote the ordinary least squares estimator of $\Phi_{i,p}$ where $\Phi_{i,p} = 0$ for $i > m$. Then for $p \geq m$ and $\gamma > \alpha$*

$$\begin{aligned} n^{1/\gamma}(\hat{\Phi}_{i,p} - \Phi_{i,p}) &\longrightarrow 0\ a.s. \\ i &= 1, ..., p. \end{aligned} \tag{13}$$

A more precise result is available from Davis and Resnick (1986):

Theorem 3 *Let X_t be an autoregressive process given by (1) with $|\rho| < 1$ and $u_t \sim iid\ SP(\alpha)$ for $\alpha \in (0,2)$. Then for $p \geq m$*

$$\left(\frac{n}{\ln(n)}\right)^{1/\alpha} (\hat{\Phi}_{i,p} - \Phi_{i,p}) = O_p(1) \tag{14}$$

where again $\Phi_{i,p} = 0$ for $i > m$.

The next Lemma collects some auxiliary results on the orders of magnitude of various sample second moments. We now let X_t be a unit root non-stationary autoregressive process.

Lemma 4 *Let X_t be generated by (1) with $\rho = 1$ and Z_t be defined by (2). Further, write*

$$Z_t = \phi(L)^{-1} u_t = \sum_{j=1}^{\infty} b_j u_{t-j}$$

with $b_0 = 1$ and the b_j satisfying Phillips' (1990, condition (25), p.50). Further, following Phillips (1990), write $\omega = \sum_{j=1}^{\infty} b_j$, and write $U_\alpha(r)$ for the Levy process on $[0,1]$ to which normalised partial sums of u_t converge. Then

$$\begin{aligned}
(i) \quad \sum_{t=1}^{n} Z_{t-j} u_t &= O_p(a^2 (n \ln n)^{1/\alpha}) \\
(ii) \quad \sum_{t=1}^{n} X_{t-1} u_t &= O_p(a_n^2) \\
(iii) \quad a_n^{-2} \sum_{t=1}^{n} Z_{t-i} Z_{t-l} &\Longrightarrow (\sum_{l=0}^{\infty} b_l b_{l+|j-i|}) \int_0^1 (dU_\alpha)^2 = O_p(1) \\
(iv) \quad n^{-1} a_n^{-2} \sum_{t=1}^{n} X_{t-1}^2 &\Longrightarrow \omega^2 \int_0^1 U_\alpha^2 = O_p(1) \\
(v) \quad \sum_{t=1}^{n} X_{t-1} Z_{t-l} &= O_p(a_n^2) \\
(vi) \quad (\sum_{t=1}^{n} X_{t-1}^2)^{-1} (\sum_{t=1}^{n} X_{t-1} Z_{t-l}) &= O_p(n^{-1}),
\end{aligned}$$

where $a_n = an^{1/\alpha}$ and i, j, l are positive integers.

We can now establish an extension of the Davis and Resnick (1986) and Hannan and Kanter (1977) results, showing that the OLS coefficient estimates are consistent, and giving their convergence rate in the unit root non-stationary case.

Theorem 5 *Let X_t be generated by (1) with $\rho = 1$, and $\alpha \in (0,2)$ and let $\widehat{\mathbf{\Phi}}_p = [\hat{\Phi}_{1,p}, ..., \hat{\Phi}_{p,p}]'$, $p \geq m$, denote the p - element vector of OLS autoregressive coefficient estimates. Then, defining the trailing $(p-m)$ elements of $\mathbf{\Phi}_p$ to be zero, so that $\mathbf{\Phi}_p = [\mathbf{\Phi}_m', \mathbf{0}']'$,*

(a) if $m > 1$

$$(n/\ln n)^{\delta/\alpha}(\hat{\Phi}_{i,p} - \Phi_{i,p}) = o_p(1), \qquad i = 1, ..., p$$

for any $\delta \in (0,1) \cap (0,\alpha]$, $\alpha \in (0,2)$.

(b) if $m = 1$ and $p = 1$

$$n(\hat{\Phi}_{1,1} - 1) = O_p(1)$$

(c) if $m = 1$ and $p > 1$

$$(n/\ln n)^{\delta/\alpha}(\hat{\Phi}_{i,p} - \Phi_{i,p}) = o_p(1), \qquad i = 1, ..., p$$

for any $\delta \in (0,1) \cap (0,\alpha]$, $\alpha \in (0,2)$.

(d) if $m > 1$ and $\alpha \geq 1$ taking $\delta = 1$ then

$$(n/\ln n)^{1/\alpha}(\widehat{\mathbf{\Phi}}_p - \mathbf{\Phi}_p) = O_p(1).$$

Corollary 6 *For $m > 1$ and $\alpha < 1$ taking $\delta = \alpha$ we have*

$$(n/\ln n)(\widehat{\mathbf{\Phi}}_p - \mathbf{\Phi}_p) = o_p(1).$$

Remark 7 *In Part (d) we obtain the same rate of convergence in distribution as Davis and Resnick (14). We have also established that, as in the stationary case considered by Hannan and Kanter (1977), for $k > m$*

$$n^{1/\gamma}\hat{\Phi}_{k,p} = o_p(1)$$

for all $\gamma > \alpha$, when $\alpha \geq 1$, while for $\alpha < 1$ we still require $\gamma > 1$. To see this, observe that

(i) for $\alpha \geq 1$, since $n^{-\lambda}\ln(n) = o(1)$ for any $\lambda \in (0,1)$ we must have $n^{(1-\lambda)/\alpha}\hat{\Phi}_{k,p} = o_p(1)$, by Part (d); put $\gamma = \alpha/(1-\lambda) > \alpha$ and observe that, in particular, we may take $\gamma = 2$.

(ii) for $0 < \alpha < 1$, observe that for $\gamma > 1$

$$n^{1/\gamma}\hat{\Phi}_{k,p} = o_p(1)$$

by Corollary 6.

4 Consistency of Information Criteria for Lag Length Selection

We are now ready to find the convergence rate for the approximation, (8). Since this does not seem to have been given in the form in which it is most useful for present purposes, we make it the subject of the next theorem.

Theorem 8 *Consider data arrays defined as*

$$\begin{aligned} \mathbf{X}_0 &= [X_1, ..., X_n]' \\ \mathbf{X}_1 &= [X_0, ..., X_{n-1}]' \\ &\vdots \\ \mathbf{X}_K &= [X_{-(K-1)}, ..., X_{n-K}]'. \end{aligned} \tag{15}$$

Let the residual sum of squares from least squares regression of $\mathbf{X}_0$ on $\mathcal{X}_K$=$[\mathbf{X}_1 \vdots \mathbf{X}_2 \vdots \cdots \vdots \mathbf{X}_K]$ be denoted, RSS_U and that for regression on the reduced set, $\mathcal{X}_{K-1} = [\mathbf{X}_1 \vdots \mathbf{X}_2 \vdots \cdots \vdots \mathbf{X}_{K-1}]$ be denoted RSS_R. Further, write the least squares estimator of the last coefficient in the first regression as $\hat{\Phi}_{K,K}$. Then

(a)

$$\begin{aligned} RSS_R - RSS_U &= \hat{\Phi}^2_{K,K} / \left(\mathcal{X}'_K \mathcal{X}_K\right)^{-1}_{KK} \\ &= \hat{\Phi}^2_{K,K} \times RSS^{\dagger}, say, \end{aligned}$$

where

$$RSS^{\dagger} = \mathbf{X}'_K \mathbf{X}_K - \mathbf{X}'_K \mathcal{X}_{K-1} (\mathcal{X}'_{K-1} \mathcal{X}_{K-1})^{-1} \mathcal{X}'_{K-1} \mathbf{X}_K.$$

(b)

$$\begin{aligned} RSS_U &= RSS_R - \hat{\Phi}^2_{K,K} RSS^{\dagger} \\ &= RSS_R (1 - \hat{\Phi}^2_{K,K}) \{1 + \delta_K\}, \end{aligned}$$

where

$$\delta_K = \frac{\hat{\Phi}^2_{K,K} (RSS_R - RSS^{\dagger})}{RSS_R (1 - \hat{\Phi}^2_{K,K})}.$$

(c) Suppose X_t is a stationary $AR(m)$ and $0 < \alpha \leq 2$, then

$$\delta_k / \hat{\Phi}^2_{k,k} = o_p(1).$$

(d) Suppose ΔX_t is a stationary $AR(m-1)$ and $0 < \alpha \leq 2$, then

$$\delta_k / \hat{\beta}^2_{k,k} = o_p(1)$$

where $\hat{\beta}^2_{k,k}$ is the least squares estimator of the last coefficient in the augmented Dickey-Fuller type equation, (19) of Appendix A.

The next theorem is our main result.

Theorem 9 *Let X_t be a non-stationary AR(m), $m \geq 1$, defined by (1), with $\rho = 1$ and let $K > m$, be an upper bound. If $\widehat{m}$ minimizes the criterion, $IC_{OLS}(k)$, defined in (9) and*
(a) $\hat{\Phi}^2_{k,p}/C(n) = o_p(1)$ for $k, p > m$, and
(b) $C(n) = o(1)$

then $\widehat{m} \longrightarrow_p m$.

Remark 10 *By Remark 7, $C(n) = c.n^{-1}$ (as in AIC) satisfies the conditions of the theorem when $0 < \alpha < 2$, since the rate of convergence we have obtained for $k > m$ is the same as Knight (1989) obtained for the stationary case.*

Remark 11 *We can also take $C(n) = \ln(n)/n$ for all $\alpha \in (0, 2)$, which would match the Schwarz BIC criterion, but obviously other choices are possible. Observe that although the convergence rate for $\hat{\Phi}_{1,1}$ in the non-stationary $m = 1$ case is faster than in other cases, it is not this rate that determines the appropriate rate for $C(n)$.*

5 A Modified Criterion

Theorem 12 *Letting $\widehat{\Phi}_{i,j}$ denote the least squares estimator of the lag-i coefficient in an autoregression of length, j, then $\widehat{m}$ defined by*

$$\begin{aligned} \widehat{m} &= \arg \min_{k,\ 0 \leq k \leq K} LC(k) \qquad (16) \\ LC(k) &= -\sum_{j=1}^{j=k} \widehat{\Phi}^2_{j,j} + (k+1)C(n), \\ LC(0) &= C(n) \end{aligned}$$

is a consistent estimator of m for penalty functions, $C(n)$, satisfying the conditions of Theorem 9, and with $m \leq K$, an upper bound.

The new criterion defined by (16) behaves in a similar manner to the usual criteria, but enjoys two advantages: firstly, its consistency is very simple to establish, and secondly, and more interestingly, $C(n)$ may be chosen to deliver, implicitly, a test of the "null hypothesis", $m = 0$, with controllable significance level, as illustrated numerically in the next section.

6 Finite Sample Performance

To illustrate the finite-sample properties of the lag-order selection criteria discussed above we conduct a small Monte Carlo experiment. We illustrate both the small and large sample properties of the various criteria for stationary and unit-root non-stationary processes with innovations in the domain of attraction of a stable law with $\alpha \in (1, 2]$, that is, including the finite variance case, $\alpha = 2$. In particular, in line with what the theory predicts, we find no significant difference between stationary and non-stationary cases, but more marked differences as α varies across its range.

The criteria illustrated are defined as

(i) $AIC(k) = \ln \hat{\sigma}_u^2 + 2k/n$,

(ii) $BIC(k) = \ln \hat{\sigma}_u^2 + k \ln(n)/n$,

(iii) $LCA(k) = -\sum_{j=1}^{j=k} \widehat{\Phi}_{j,j}^2 + 2(k+1)/n$,

(iv) $LCB(k) = -\sum_{j=1}^{j=k} \widehat{\Phi}_{j,j}^2 + (k+1) \ln(n)/n$,

and

(v) $LCB^*(l, k) = -\sum_{j=1}^{j=k} \widehat{\Phi}_{j,j}^2 + (k+1) C_l(n)$

where $C_l(n) = (1 + 0.1l) \ln(n)/n \qquad (l = -9, ..., 0, ...7)$,

in which the latter three are as defined at (16) for $k = 0$. We report results selectively for three DGPs in addition to a white noise process, u_t :

$P0 : root(0) \iff X_t = u_t$

$P2 : roots(1, 0.6) \iff X_t = 1.6X_{t-1} - 0.60X_{t-2} + u_t$

$P3 : roots(0.6, 0.2) \iff X_t = 0.8X_{t-1} - 0.12X_{t-2} + u_t$

$P4 : roots(1, 0.6, 0.2) \iff X_t = 1.8X_{t-1} - 0.92X_{t-2} + 0.120X_{t-3} + u_t$

We report our numerical results[1] very selectively in order to highlight important features. Sample size, and the index of stability are selected from

$n \in [100, 250, 500, 10000]$ and $\alpha \in [0.75, 1.00, 1.50, 1.75, 2.00]$,

and the maximum lag length is $K = 9$. Except where stated otherwise the tables are based on 10000 replications.

Table 1 reports results for the white noise process, $P0$, to act as a baseline. It illustrates three things. (i) We find in this table, and generally, that AIC and LCA give virtually identical results, as do BIC and LCB, as is only to be expected given that $\ln\{1 - \hat{\Phi}_{k,k}^2\} \approx -\hat{\Phi}_{k,k}^2$ for $k > m$. (ii) The inconsistency of AIC/LCA for $\alpha = 2$ is plain to see in the final column of the upper panel, while, as expected, these criteria also perform poorly for α close to, but less than 2. (iii) Assuming that the BIC/LCB is adopted, then at sample size, 100,

[1] All calculations were performed in GAUSS 5.0 using the Kiss+Monster random number generator and an implementation of the algorithm of Chambers, Mallows and Stuck (1976), coded in GAUSS by J. Huston McCulloch.

we have, in effect, approximately a 5% chance of rejecting the true hypothesis, $H_0 : m = 0$ whatever the value of α.

The latter point deserves further comment. Practitioners may often adopt the use of information criteria for lag-length selection as an apparently simpler alternative to the use of sequences of Student - t tests, or similar. If so, they may be mistaken in believing that the awkward question of the appropriate significance level for such tests has thereby been avoided. Consider Table 1 again; if the probabilities estimated in the first column of the lower panel were thought to be too low or too high, then a simple expedient would be to adjust the level of $C(n)$, by an appropriate amount, as in $C_l(n)$ defined above at (v). Indeed, for LCB we have $C(100) = \ln(100)/100 = .046$, while for $\alpha = 2$ we know that in the white noise case, $\hat{\Phi}_{1,1} \approx N(0, 1/100)$, approximately, so that $\hat{\Phi}^2_{1,1} \approx .1 \times \chi^2_1$ and $\Pr\{\chi^2_1 > 4.6\} = 3.2\%$, which thus puts an approximate lower bound under the probability of over-fitting in this instance. Obviously, this bound can be made as small as the investigator wants by choosing a level for $C(100)$, or indeed $C(n)$ for any specific n as in $C_l(n)$. Of course, such adjustments have knock-on effects on the performance of the model selection criteria when the process is *not* while noise, and we illustrate this in Tables 2 and 3.

Table 2 ($n = 100$) and Table 3 ($n = 250$) show the effect on the probability of correct model choice of varying the level of $C(n)$ when the data are generated by processes $P0$, $P2$ or $P3$. In each table the bold row corresponding to $l = 0$ shows the performance of LCB as defined in (iv), while the remaining rows show the performance of LCB^* as defined in (v).

Consider first the white noise process, $P0$. In this case, the most striking feature of the tables is the reversal of the relative performance over different α values as we move from low $C_l(n)$ values at the top of the tables, where performance degrades severely as α increases, to high $C_l(n)$ values at the foot of the tables, where performance improves slightly as α increases. These effects reflect the changes in the shape of the sampling distribution of $\hat{\Phi}$ in the white noise case as we vary the tail index, α.

For the process, $P2$, results were similar to those obtained for the pure $AR(1)$ with a root of 0.6 which are not shown, the only difference being the obvious one that the lag length chosen is 1 higher. Since erroneous lag length choices can be either too high or too low, for $P2$ and $P3$ the tables reveal the significance of the presence of the small lag- 3 coefficient, 0.12, in the latter. For $P2$, the pattern is remarkably similar to that for $P0$, because the large lag-2 coefficient, 0.6, is almost always detected, and any over-fitting that arises is in each case attributable to the sampling variation in the estimation of coefficients that are 0 in the model. For $P3$ performance below the bold line is poor, and degrades as $C_l(n)$ increases, because of underfitting; at very low $C_l(n)$ values the probability of correct model selection is comparable to the other two models. It may be salutary to observe that even in the familiar case, with $\alpha = 2$, there is only a 1-in-3 chance of selecting the correct lag length with the BIC criterion for this model.

Finally, a striking feature of the heavy-tailed case is the consistency of the AIC criterion, a property it does not enjoy when $\alpha = 2$. Nonetheless, as Table 4 illustrates, for a process with a small coefficient on the longest lag, the AIC still outperforms BIC even at sample size $n = 500$; for $n = 10,000$ however, and α close to or equal to 2 the situation is reversed.

7 Discussion and Conclusion

The $a.s.$ convergence rate found by Hannan and Kanter (1977) for the least squares estimator of the coefficients in the stationary AR process with infinite variance innovations was $n^{1/\gamma}$ for $\gamma > \alpha$. This result was developed further by Davis and Resnick (1986, Theorem 4.4), who established a limit law for the sample autocorrelation function, obtaining Theorem 3 above as a corollary. We note that our convergence rate in Theorem 5(d) matches theirs, but we have not been able to establish a corresponding rate for $0 < \alpha < 1$. To extend the lag-selection results to the unit-root non-stationary setting we have found it necessary to look quite closely at the approximation used by Shibata (8), and we sharpen this in Theorem 8. We also introduce a modified definition of a lag selection criterion, expressing this directly in terms of the estimated coefficient on the longest lag; this facilitates the interpretation of the numerical evidence on the performance of various criteria. We note also, that such a definition frees the development from undue dependence on a particular estimation method since all that is required is the rate of convergence of whatever estimator is adopted. Indeed, if an M-estimator in the class investigated by Davis, Knight and Liu (1992) were chosen, it would seem natural to define a lag-selection criterion in the form

$$LC^*(k) = (k+1)C(n) - \sum_{j=1}^{j=k} |\hat{\Phi}_{j,j}|^\gamma$$

for suitable γ, and then to deduce consistency of $LC^*(k)$ by a parallel argument to that adopted here. However, to the best of our knowledge, consistency of the M-estimator of Φ has not yet been proved in the general non-stationary heavy-tailed case.

For the stationary case, with $\alpha \in (0, 2)$, Knight (1989) establishes that with the upper bound on k depending on n, such that $K(n) = o(n^{1-\alpha/2})$, then the Yule-Walker estimator satisfies

$$n \max_{m<k\leq K(n)} \breve{\Phi}_{k,k}^2 \to_p 0$$

so that

$$n \min_{m<k\leq K(n)} \ln(1 - \breve{\Phi}_{k,k}^2) \to_p 0. \tag{17}$$

which is enough to give consistency of the Akaike criterion, $C(n) = 2/n$,

because

$$\begin{aligned} \Pr[\hat{k} &> m] \leq \Pr[\min_{m<k\leq K(n)} \ln(1-\breve{\Phi}_{k,k}^2) < -2/n] \\ &= \Pr[n \min_{m<k\leq K(n)} \ln(1-\breve{\Phi}_{k,k}^2) < -2] \qquad (18) \\ &\to 0 \text{ by (17).} \end{aligned}$$

For $\alpha \geq 1$, we find the first factor on the right-hand side of (37) is block diagonal, so the same limiting distribution for $\hat{\Phi}$ is obtained as in the stationary case. Consequently, Knight's result (1989, Theorem 5(a)) may be applied, and our consistency result will hold if $K = K(n) = o(n^{1-\alpha/2})$. More generally, it is striking that the presence of the unit root has no significant effect on the convergence rate of either the parameter estimation or the lag selection other than in the very special case of $m = 1$. Thus we may be confident in applied work in adopting the same lag selection strategy as in stationary cases.

We have assumed that no deterministic component is present, which amounts to assuming that the innovations are centred on zero, as seems natural. For more on the implications of relaxing this assumption, see Davis and Resnick (1986, p. 553), or Knight (1989, p.826). Of course, in practical situations one may not know whether the data are non-stationary, or indeed whether the innovations have heavy tails, and as the experiments demonstrate, *a safe choice is to adopt a lag selection criterion proportional to* $C(n) = \ln(n)/n$ *which is consistent in all the cases considered.* If it is particularly important not to under fit, then one should not adopt such a rule uncritically.

TABLES

$P0$: white noise								
	$n =$ **100**		**250**		**500**		**10 000**	
$\boldsymbol{\alpha}$	AIC	LCA	AIC	LCA	AIC	LCA	AIC	LCA
0.75	86	86	89	89	91	91	97	97
0.90	85	84	87	86	89	88	96	96
1.0	84	83	85	85	87	87	94	94
1.10	83	83	85	85	87	87	93	93
1.25	82	81	82	82	84	84	91	91
1.50	79	79	80	80	81	81	86	86
1.75	77	77	77	77	77	77	80	80
2.00	71	72	72	72	72	72	72	72
	BIC	LCB	BIC	LCB	BIC	LCB	BIC	LCB
0.75	95	94	96	96	96	96	99	99
0.90	95	94	96	96	96	96	98	98
1.0	95	94	95	95	96	96	98	98
1.10	95	95	96	96	96	96	98	98
1.25	96	95	96	96	96	96	98	98
1.50	96	96	97	97	97	97	98	98
1.75	97	96	98	98	98	98	99	99
2.00	96	96	98	98	99	99	99.7	99.7

Table 1 Percentage probabilities of selecting the true order, $m = 0$.

	$P0$: white noise or $P2$: AR(2) with roots 1 and 0.6											
	$P3$: AR(2) with roots 0.6 and 0.2											
	$\alpha = 0.75$			$\alpha = 1.0$			$\alpha = 1.5$			$\alpha = 2.0$		
l	PO	$P2$	$P3$	PO	$P2$	$P3$	PO	$P2$	$P3$	PO	$P2$	$P3$
-9	60	62	60	45	47	44	20	21	18	4	6	4
-8	72	75	61	64	66	50	43	46	31	25	27	17
-7	79	81	35	74	77	35	63	64	33	49	50	27
-6	84	86	18	81	83	25	75	75	30	68	66	31
-5	87	88	13	86	87	19	83	84	28	79	78	30
-4	90	90	10	89	90	15	89	89	25	85	85	29
-3	92	92	9	90	92	13	91	91	20	90	89	26
-2	92	93	7	93	93	10	94	94	18	93	92	23
-1	93	94	5	94	95	9	95	95	15	95	95	20
0	**94**	**95**	**5**	**95**	**95**	**7**	**96**	**96**	**13**	**97**	**96**	**17**
1	95	95	4	95	96	6	96	97	10	97	97	15
2	95	96	4	95	97	5	97	98	8	98	98	12
3	96	96	3	96	96	5	98	98	7	99	98	11
4	96	96	3	96	97	4	98	98	7	99	99	9
5	96	96	2	97	97	3	98	99	5	99	99	8
6	97	97	2	97	98	3	99	98	5	100	99	6
7	97	97	2	97	98	3	99	99	4	100	99	5

Table 2 Percentage probabilities of selecting the true order, $m = 0$, or $m = 2$ with $n = 100$ and using various $C_l(n)$ values, in the LCB^* criterion (v). Figures rounded to nearest 1 percent.

	$P0$: white noise or $P2$: AR(2) with roots 1 and 0.6											
	$P3$: AR(2) with roots 0.6 and 0.2											
	$\alpha = 0.75$			$\alpha = 1.0$			$\alpha = 1.5$			$\alpha = 2.0$		
l	PO	$P2$	$P3$	PO	$P2$	$P3$	PO	$P2$	$P3$	PO	$P2$	$P3$
-9	74	75	75	60	62	61	29	31	31	7	10	9
-8	82	84	81	75	77	74	56	58	52	35	37	30
-7	87	89	84	82	84	77	74	74	61	61	62	46
-6	89	91	82	86	89	75	83	84	60	77	77	51
-5	91	93	78	90	91	68	88	89	57	86	85	52
-4	92	93	65	92	93	57	92	93	51	91	90	49
-3	93	94	24	93	94	33	94	95	43	94	94	45
-2	95	96	15	94	95	24	95	96	36	96	95	41
-1	95	96	11	94	96	18	96	97	32	97	97	37
0	**96**	**97**	**9**	**96**	**97**	**14**	**97**	**98**	**26**	**98**	**98**	**32**
1	96	97	7	96	97	12	98	98	22	99	99	28
2	97	98	6	97	98	10	98	99	19	99	99	25
3	97	97	5	97	98	8	98	99	16	99	99	21
4	97	98	4	97	98	7	98	99	14	100	100	19
5	97	98	4	98	98	6	99	99	11	100	100	16
6	98	98	3	98	98	5	99	99	9	100	100	14
7	98	98	3	98	98	4	99	99	8	100	100	12

Table 3 Percentage probabilities of selecting the true order, $m = 0$, or $m = 2$ with $n = 250$ and using various $C_l(n)$ values, in the LCB^* criterion (v). Rounded to nearest 1 percent.

$P4: X_t = 1.8X_{t-1} - 0.92X_{t-2} + 0.12X_{t-3} + u_t$				
$\boldsymbol{\alpha}$	$\hat{m}$	**100**	**500**	**10 000**[a]
		AIC	*AIC*	*AIC*
0.75	**2**	70	1	0.1
0.75	**3**	21	91	98
1.00	**2**	63	2	0
1.00	**3**	25	89	96
1.75	**2**	52	7	0
1.75	**3**	31	73	82
2.00	**2**	48	9	0
2.00	**3**	31	65	75
		BIC	*BIC*	*BIC*
0.75	**2**	91	18	0.1
0.75	**3**	7	78	99
1.00	**2**	89	28	0
1.00	**3**	9	70	99
1.75	**2**	83	42	0
1.75	**3**	15	57	99
2.00	**2**	81	42	0
2.00	**3**	17	56	99.6

a Based on 1000 replications

Table 4 Percentage probabilities of selecting the true order, $m = 3$, and underfitting, $\hat{m} = 2$

8 Appendix A: Proofs

PROOF OF LEMMA 1: See Tiao and Tsay (1983, Lemma 2.3, p.857). ■

PROOF OF THEOREM 2: See Hannan and Kanter (1977, p.412). ■

PROOF OF THEOREM 3: See Davis and Resnick (1986, Theorem 4.4 and corollaries). ■

PROOF OF LEMMA 4:

Part (*i*). Since for $j > 0$ the random variables Z_{t-j} and u_t are independent the required norming sequence for their product is as given by Phillips (1990 Appendix A, p.58).

Part (*ii*). Write $u_t = \phi(L)Z_t$ and apply Phillips (1990, Thm 2.1, p.50) to each term.

Part (*iii*). Follows from Phillips (1990, equations (40,41), p.53)

Parts (*iv, v*). Follow from Phillips (1990, Thm 2.1, p.50), after fixing the typographical error in his equation (28).

Part (*vi*). Follows directly from Parts (*iv*) and (*v*). ■

PROOF OF THEOREM 5:

Part (b). See Chan and Tran (1989, Theorem 2 p.358). ■

Parts (a) and (c). Process (1) with $\rho = 1$ is equivalent to

$$\{(1-\beta L) - (1-L)\sum_{j=1}^{m-1}\eta_j L^j\}X_t = u_t,$$

and this equation leads to the familiar Dickey-Fuller style regression

$$X_t = \widehat{\beta}X_{t-1} + \sum_{j=1}^{m-1}\widehat{\eta}_j \Delta X_{t-j} + \widehat{u}_t \tag{19}$$

in which $\widehat{\beta} = \sum_{i=1}^{m}\widehat{\Phi}_{i,m}$, $\widehat{\eta}_j = -\sum_{i=j+1}^{m}\widehat{\Phi}_{i,m}$, $j = 1, 2, ..., m-1$, and $\widehat{\Phi}_{i,m}$, $i = 1, 2, ..., m$, as defined above (16).

We are interested in establishing a convergence rate for the least squares estimator of $\boldsymbol{\beta}' = (\beta, \boldsymbol{\eta}')$. To do so, we will find a normalising matrix, $\mathbf{\Lambda}_n^*$, say, such that $\mathbf{\Lambda}_n^*[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}]$ is $(O_p(1), \mathbf{o}_p(\mathbf{1}))'$. We first introduce some more notation.

As in Appendix A, suppose the sample available runs from $X_{-(m-1)}$ to X_n; as in Lemma 4, define the stationary process, $Z_{t-j} = \Delta X_{t-j}$, and then introduce the m element random vector,

$$Y_t = [X_{t-1}, Z_{t-1}, ..., Z_{t-m+1}]'$$

with corresponding sample sum of squares and cross-products matrix,

$$\mathbf{M}_n = \sum_{t=1}^{n} Y_t Y_t^{'}.$$

Associated with the linear process, Z_t, are the following objects; for the process itself, since $\rho = 1$, we have:

$$Z_t = \eta(L)^{-1}u_t = \phi(L)^{-1}u_t = \sum_{j=1}^{\infty} b_j u_{t-j}$$

and write

$$\omega = \sum_{j=1}^{\infty} b_j, \quad \sigma^2 = \sum_{j=1}^{\infty} b_j^2$$

and $U_\alpha(r)$ for the Levy process on $[0,1]$ to which normalised partial sums of u_t converge, as in Lemma 4.

The error in the least squares estimator is

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{M}_n^{-1} \sum_{t=1}^{t=n} Y_t u_t = \mathbf{M}_n^{-1}\mathbf{C}_n \quad \text{say.} \tag{20}$$

A difficulty now emerges. In the well-known $\alpha = 2$ case, one proceeds to the statement,

$$\mathbf{T}_n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \{\mathbf{T}_n^{-1/2}\mathbf{M}_n\mathbf{T}_n^{-1/2}\}^{-1}\mathbf{T}_n^{-1/2}\mathbf{C}_n$$

via the transformation, $\mathbf{T}_n^{-1/2} = \begin{bmatrix} n^{-1} & \mathbf{0}' \\ \mathbf{0} & n^{-1/2}\mathbf{I}_{m-1} \end{bmatrix}$, in which the objects on the RHS converge in distribution. Unfortunately, in the present setting, the normalisation required for $\mathbf{C}_n$ is not the square root of that required for $\mathbf{M}_n$. We therefore proceed as follows.

Define the $m \times m$ diagonal matrices,

$$\boldsymbol{\Lambda}_n = \begin{bmatrix} a_n n^{1/2} & \mathbf{0}' \\ \mathbf{0} & a_n \times \mathbf{I}_{m-1} \end{bmatrix}, \tag{21}$$

$$\boldsymbol{\Lambda}_n^* = \begin{bmatrix} n & \mathbf{0} \\ \mathbf{0} & \left(\frac{n}{\ln(n)}\right)^{\delta/\alpha} \times \mathbf{I}_{m-1} \end{bmatrix} \tag{22}$$

and

$$\boldsymbol{\Upsilon}_{\delta,n} = \begin{bmatrix} a_n^2 & \mathbf{0} \\ \mathbf{0} & a_n^2 \left(\frac{n}{\ln(n)}\right)^{-\delta/\alpha} \times \mathbf{I}_{m-1} \end{bmatrix},$$

where $\delta > 0$ is a fixed real number. Observe that $\boldsymbol{\Upsilon}_{\delta,n}^{-1} = \boldsymbol{\Lambda}_n^* \boldsymbol{\Lambda}_n^{-2}$ and define

$$\mathbf{D}_n = \boldsymbol{\Lambda}_n^{-1}\mathbf{M}_n\boldsymbol{\Lambda}_n^{-1}.$$

It follows from (20) that

$$\boldsymbol{\Lambda}_n\boldsymbol{\Lambda}_n^{-1}\mathbf{M}_n\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\Lambda}_n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{C}_n$$

hence

$$\boldsymbol{\Upsilon}_{\delta,n}^{-1}\boldsymbol{\Lambda}_n\mathbf{D}_n\boldsymbol{\Lambda}_n(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})=\boldsymbol{\Upsilon}_{\delta,n}^{-1}\mathbf{C}_n$$

and

$$(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})=\left(\boldsymbol{\Lambda}_n^{-1}\mathbf{D}_n^{-1}\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\Upsilon}_{\delta,n}\right)\left(\boldsymbol{\Upsilon}_{\delta,n}^{-1}\mathbf{C}_n\right)$$

giving us finally

$$\begin{aligned}\boldsymbol{\Lambda}_n^*(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})&=\boldsymbol{\Lambda}_n^*\boldsymbol{\Lambda}_n^{-1}\mathbf{D}_n^{-1}\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\Upsilon}_{\delta,n}(\boldsymbol{\Upsilon}_{\delta,n}^{-1}\mathbf{C}_n)\\&=\left(\boldsymbol{\Upsilon}_{\delta,n}^{-1}\boldsymbol{\Lambda}_n\mathbf{D}_n^{-1}\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\Upsilon}_{\delta,n}\right)(\boldsymbol{\Upsilon}_{\delta,n}^{-1}\mathbf{C}_n).\end{aligned}$$

We remark that what follows is necessarily more complicated than in the case of Normal shocks because of the asymmetry in the first factor on the RHS. We shall show that the RHS is $(O_p(1),\mathbf{o_p(1)})'$, that is,

$$\begin{bmatrix}n & \mathbf{0}\\ \mathbf{0} & \left(\frac{n}{\ln(n)}\right)^{\delta/\alpha}\times\mathbf{I}_{m-1}\end{bmatrix}[\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}]=\begin{bmatrix}O_p(1)\\ \mathbf{o_p(1)}\end{bmatrix}_{m\times 1}\tag{23}$$

for any $\delta\in(0,1)\cap(0,\alpha]$, $\alpha\in(0,2)$, and thus the result will hold. For $\alpha\geq 1$ we have a slightly better convergence in distribution result, Corollary 6. The technical details follow.

(i) Consider the $m\times 1$ vector $\boldsymbol{\Upsilon}_{\delta,n}^{-1}\mathbf{C}_n$. Using an obvious partitioning, for the first element we have

$$\boldsymbol{\Upsilon}_{\delta,n,11}^{-1}\mathbf{C}_{n,1}=a_n^{-2}\sum X_{t-1}u_t=a_n^{-2}\sum X_{t-1}\left\{\phi(L)Z_t\right\}\Longrightarrow O_p(1),\tag{24}$$

by Lemma 4 Part (ii).

A typical element of $\boldsymbol{\Upsilon}_{\delta,n,22}^{-1}\mathbf{C}_{n,2}$ is

$$\begin{aligned}a_n^{-2}\left(\frac{n}{\ln(n)}\right)^{\delta/\alpha}\sum Z_{t-j}u_t&=a^{-2}n^{-2/\alpha}\left(\frac{n}{\ln(n)}\right)^{\delta/\alpha}\sum Z_{t-j}u_t\\&=\left(\frac{\ln n}{n}\right)^{\frac{1-\delta}{\alpha}}\left(a^{-2}(n\ln n)^{-1/\alpha}\sum Z_{t-j}u_t\right)\\\Longrightarrow O_p\left(\left(\frac{\ln n}{n}\right)^{\frac{1-\delta}{\alpha}}\right)&=\begin{array}{l}o_p(1)\text{ for }\forall\delta\in(0,1)\\ O_p(1)\text{ for }\delta=1\end{array},\end{aligned}\tag{25}$$

by Lemma 4 Part (i). It follows from (24) and (25) that

$$\begin{aligned}\boldsymbol{\Upsilon}_{\delta,n}^{-1}\mathbf{C}_n&=\left[\begin{array}{c}O_p(1)\\ \mathbf{o}_p(\mathbf{1})\end{array}\right]_{m\times 1}\\\boldsymbol{\Upsilon}_{1,n}^{-1}\mathbf{C}_n&=\left[\begin{array}{c}O_p(1)\\ \mathbf{O}_p(\mathbf{1})\end{array}\right]_{m\times 1}.\end{aligned}\tag{26}$$

(ii) Consider the product $\boldsymbol{\Upsilon}_{\delta,n}^{-1}\boldsymbol{\Lambda}_n\mathbf{D}_n^{-1}\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\Upsilon}_{\delta,n}$.
The matrix $\mathbf{D}_n=\boldsymbol{\Lambda}_n^{-1}\mathbf{M}_n\boldsymbol{\Lambda}_n^{-1}$, is of the form

$$\mathbf{D}_n = \begin{bmatrix} \frac{\sum X_{t-1}^2}{na_n^2} & \frac{\sum X_{t-1}Z_{t-1}}{n^{1/2}a_n^2} & \cdots \frac{\sum X_{t-1}Z_{t-(m-1)}}{n^{1/2}a_n^2} \\ \frac{\sum X_{t-1}Z_{t-1}}{n^{1/2}a_n^2} & \frac{\sum Z_{t-1}^2}{a_n^2} & \cdots \frac{\sum Z_{t-1}Z_{t-(m-1)}}{a_n^2} \\ ... & & \\ \frac{\sum X_{t-1}Z_{t-(m-1)}}{n^{1/2}a_n^2} & \frac{\sum Z_{t-1}Z_{t-(m-1)}}{a_n^2} & \cdots \frac{\sum Z_{t-(m-1)}^2}{a_n^2} \end{bmatrix}.$$

Using Parts (iii), (iv) and (v) of Lemma 4 we establish the following. For the top left element,

$$\mathbf{D}_{n,11} = n^{-1}a_n^{-2}\sum X_{t-1}^2 \Longrightarrow \omega^2 \int_0^1 U_\alpha^2 = O_p(1). \tag{27}$$

For the bottom right block, a typical element is

$$\mathbf{D}_{n,22}\{i,j\} = a_n^{-2}\sum Z_{t-i}Z_{t-j} \Longrightarrow (\sum_{l=0}^{\infty} b_l b_{l+|j-i|}) \int_0^1 (dU_\alpha)^2 = O_p(1) \tag{28}$$

for $i, j = 1, ..., m-1$. Thus for the entire block we have, say,

$$\mathbf{D}_{n,22} \Longrightarrow \int_0^1 (dU_\alpha)^2 \times \mathbf{B}$$

in which the matrix, $\mathbf{B}$, is symmetric and positive definite.
Finally, for the off-diagonal block, a typical element is

$$\mathbf{D}_{n,12}\{1,j\} = n^{-1/2}\left(a_n^{-2}\sum X_{t-1}Z_{t-j}\right) \Longrightarrow O_p(n^{-1/2}) = o_p(1) \tag{29}$$

for $j = 1, ..., m-1$.Now let

$$\mathbf{D}_n^{-1} = \begin{bmatrix} \overline{d}_{11} & \overline{\mathbf{D}}_{12} \\ \overline{\mathbf{D}}_{21} & \overline{\mathbf{D}}_{22} \end{bmatrix}.$$

Hence,

$$\begin{aligned} \mathbf{D}_n^* &= \mathbf{\Lambda}_n \mathbf{D}_n^{-1} \mathbf{\Lambda}_n^{-1} \\ &= \begin{bmatrix} n^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-1} \end{bmatrix} \begin{bmatrix} \overline{d}_{11} & \overline{\mathbf{D}}_{12} \\ \overline{\mathbf{D}}_{12}' & \overline{\mathbf{D}}_{22} \end{bmatrix} \begin{bmatrix} n^{-1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-1} \end{bmatrix} \\ &= \begin{bmatrix} \overline{d}_{11} & n^{1/2}\overline{\mathbf{D}}_{12} \\ n^{-1/2}\overline{\mathbf{D}}_{12}' & \overline{\mathbf{D}}_{22} \end{bmatrix}. \end{aligned}$$

Therefore

$$\begin{aligned}
\boldsymbol{\Upsilon}_{\delta,n}^{-1}\boldsymbol{\Lambda}_n\mathbf{D}_n^{-1}\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\Upsilon}_{\delta,n} &= \boldsymbol{\Upsilon}_{\delta,n}^{-1}\mathbf{D}_n^{*}\boldsymbol{\Upsilon}_{\delta,n} \\
&= \begin{bmatrix} a_n^{-2} & \mathbf{0} \\ \mathbf{0} & a_n^{-2}\left(\frac{\ln n}{n}\right)^{-\delta/\alpha}\times\mathbf{I}_{m-1}\end{bmatrix} \\
&\quad\times\begin{bmatrix} \overline{d}_{11} & n^{1/2}\overline{\mathbf{D}}_{12} \\ n^{-1/2}\overline{\mathbf{D}}_{12}' & \overline{\mathbf{D}}_{22}\end{bmatrix} \\
&\quad\times\begin{bmatrix} a_n^{2} & \mathbf{0} \\ \mathbf{0} & a_n^{2}\left(\frac{\ln n}{n}\right)^{\delta/\alpha}\times\mathbf{I}_{m-1}\end{bmatrix} \\
&= \begin{bmatrix} \overline{d}_{11} & n^{1/2}\left(\frac{\ln n}{n}\right)^{\delta/\alpha}\overline{\mathbf{D}}_{12} \\ n^{-1/2}\left(\frac{\ln n}{n}\right)^{-\delta/\alpha}\overline{\mathbf{D}}_{12}' & \overline{\mathbf{D}}_{22}\end{bmatrix} \qquad (30)
\end{aligned}$$

and we are interested in the limit of this matrix as $n\longrightarrow\infty$. It remains to find $\overline{d}_{11}$ and the elements of $\overline{\mathbf{D}}_{12}$ and $\overline{\mathbf{D}}_{22}$. Using the formula for the partitioned inversion of a nonsingular symmetric matrix, and writing $\overline{d}_{11}=\left(d_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}'\right)^{-1}$ we obtain:

$$\begin{aligned}
\mathbf{D}_n^{-1} &= \begin{bmatrix} \overline{d}_{11} & \overline{\mathbf{D}}_{12} \\ \overline{\mathbf{D}}_{12}' & \overline{\mathbf{D}}_{22}\end{bmatrix} = \begin{bmatrix} d_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{12}' & \mathbf{D}_{22}\end{bmatrix}^{-1} = \\
&\quad\begin{bmatrix} \overline{d}_{11} & -\overline{d}_{11}\mathbf{D}_{12}\mathbf{D}_{22}^{-1} \\ -\mathbf{D}_{22}^{-1}\mathbf{D}_{12}'\overline{d}_{11} & \mathbf{D}_{22}^{-1}+\mathbf{D}_{22}^{-1}\mathbf{D}_{12}'\overline{d}_{11}\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\end{bmatrix}.
\end{aligned}$$

Consider the elements of the partitioned matrix $\boldsymbol{\Upsilon}_{\delta,n}^{-1}\boldsymbol{\Lambda}_n\mathbf{D}_n^{-1}\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\Upsilon}_{\delta,n}$ given by (30); using (27, 29, and 28) we obtain:

$$\begin{aligned}
\overline{d}_{11} &\Longrightarrow \left[O_p(1)-[o_p(1)...o_p(1)]\times\mathbf{O}_p(\mathbf{1})\times[o_p(1)...o_p(1)]'\right]^{-1} \\
&= [O_p(1)-o_p(1)]^{-1}=O_p(1), \qquad (31)
\end{aligned}$$

$$\begin{aligned}
n^{1/2}\left(\frac{\ln n}{n}\right)^{\delta/\alpha}\overline{\mathbf{D}}_{12} &= (-1)n^{1/2}\left(\frac{\ln n}{n}\right)^{\delta/\alpha}\left(d_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{12}'\right)^{-1}\mathbf{D}_{12}\mathbf{D}_{22}^{-1} \\
&= (-1)\,\overline{d}_{11}\times\left(n^{1/2}\left(\frac{\ln n}{n}\right)^{\delta/\alpha}\mathbf{D}_{12}\right)\times\mathbf{D}_{22}^{-1} \\
&= (-1)\,\overline{d}_{11}\times\boldsymbol{\Gamma}\times\mathbf{D}_{22}^{-1},
\end{aligned}$$

where $\boldsymbol{\Gamma}=n^{1/2}\left(\frac{\ln n}{n}\right)^{\delta/\alpha}\mathbf{D}_{12}$ is a $1\times(m-1)$ matrix with a typical element $\boldsymbol{\Gamma}\{j\}$, $j=2,...,m$:

$$\begin{aligned}
\boldsymbol{\Gamma}\{j\} &= n^{1/2}\left(\frac{\ln n}{n}\right)^{\delta/\alpha}\mathbf{D}_{12}\{j\} \\
&\Longrightarrow \left(\frac{\ln n}{n}\right)^{\delta/\alpha}\times O_p(1)=O_p\left(\left(\frac{\ln n}{n}\right)^{\delta/\alpha}\right)=o_p(1)\text{ for }\forall\delta>0.
\end{aligned}$$

by (29).
Thus

$$\begin{aligned} n^{1/2}\left(\frac{\ln n}{n}\right)^{\delta/\alpha}\overline{\mathbf{D}}_{12} &= (-1)\,\overline{d}_{11}\times\mathbf{\Gamma}\times\mathbf{D}_{22}^{-1} \\ &\Longrightarrow O_p(1)\times\mathbf{o}_p(\mathbf{1})\times\mathbf{O}_p(\mathbf{1})=\mathbf{o}_p(\mathbf{1}). \end{aligned} \tag{32}$$

Similarly,

$$\begin{aligned} n^{-1/2}\left(\frac{\ln n}{n}\right)^{-\delta/\alpha}\overline{\mathbf{D}}'_{12} &= (-1)n^{-1/2}\left(\frac{\ln n}{n}\right)^{-\delta/\alpha}\mathbf{D}_{22}^{-1}\mathbf{D}'_{12}\bar{d}_{11} \\ &= (-1)\,\overline{d}_{11}\times\mathbf{D}_{22}^{-1}\times\left(\mathbf{D}'_{12}n^{-1/2}\left(\frac{\ln n}{n}\right)^{-\delta/\alpha}\right) \\ &= (-1)\,\overline{d}_{11}\times\mathbf{D}_{22}^{-1}\times\mathbf{\Delta}, \end{aligned}$$

where $\mathbf{\Delta}=n^{-1/2}\left(\frac{\ln n}{n}\right)^{-\delta/\alpha}\mathbf{D}'_{12}$. Now, as before, $\mathbf{D}'_{12}$ is $O_p(n^{-1/2})$, so we find that $\mathbf{\Delta}$ is $O_p(n^{\delta/\alpha-1}\left(\ln n\right)^{-\delta/\alpha})=o_p(1)$ for $\forall\delta\in(0,\alpha]$.
Thus

$$\begin{aligned} n^{-1/2}\left(\frac{\ln n}{n}\right)^{-\delta/\alpha}\overline{\mathbf{D}}'_{12} &= (-1)\,\overline{d}_{11}\times\mathbf{D}_{22}^{-1}\times\mathbf{\Delta} \\ &\Longrightarrow O_p(1)\times\mathbf{O}_p(\mathbf{1})\times\mathbf{o}_p(\mathbf{1})=\mathbf{o}_p(\mathbf{1}). \end{aligned} \tag{33}$$

Finally

$$\begin{aligned} \overline{\mathbf{D}}_{22} &= \mathbf{D}_{22}^{-1}+\mathbf{D}_{22}^{-1}\mathbf{D}'_{12}\left(d_{11}-\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}'_{12}\right)^{-1}\mathbf{D}_{12}\mathbf{D}_{22}^{-1} \\ &\Longrightarrow \mathbf{O}_p(\mathbf{1})+\mathbf{O}_p(\mathbf{1})\times\mathbf{o}_p(\mathbf{1})\times O_p(1)\times\mathbf{o}_p(\mathbf{1})\times\mathbf{O}_p(\mathbf{1}) \\ &= \mathbf{O}_p(\mathbf{1}). \end{aligned} \tag{34}$$

It follows from (30)-(34) that

$$\mathbf{\Upsilon}_{\delta,n}^{-1}\mathbf{\Lambda}_n\mathbf{D}_n^{-1}\mathbf{\Lambda}_n^{-1}\mathbf{\Upsilon}_{\delta,n}\Longrightarrow\left[\begin{array}{cc} O_p(1) & \mathbf{o}_p(\mathbf{1}) \\ \mathbf{o}_p(\mathbf{1}) & \mathbf{O}_p(\mathbf{1}) \end{array}\right]_{m\times m}. \tag{35}$$

It then follows from (26) and (35) that

$$\mathbf{\Lambda}_n^*(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\Longrightarrow\left[\begin{array}{cc} O_p(1) & \mathbf{o}_p(\mathbf{1}) \\ \mathbf{o}_p(\mathbf{1}) & \mathbf{O}_p(\mathbf{1}) \end{array}\right]_{m\times m}\times\left[\begin{array}{c} O_p(1) \\ \mathbf{o}_p(\mathbf{1}) \end{array}\right]_{m\times 1}=\left[\begin{array}{c} O_p(1) \\ \mathbf{o}_p(\mathbf{1}) \end{array}\right]_{m\times 1}. \tag{36}$$

To complete the proof we observe that by equating powers in

$$\Phi(L)=1-\sum_{j=1}^{m}\Phi_jL^j=\{(1-\beta L)-(1-L)\sum_{j=1}^{m-1}\eta_jL^j\}$$

we obtain

$$\begin{array}{ccc} \Phi_1 & = & \beta+\eta_1 \\ \Phi_2 & = & \eta_2-\eta_1 \\ \vdots & \vdots & \vdots \\ \Phi_{m-1} & = & \eta_{m-1}-\eta_{m-2} \\ \Phi_m & = & -\eta_{m-1} \end{array}$$

from which we deduce that

$$\left(\frac{n}{\ln(n)}\right)^{\delta/\alpha}\begin{pmatrix} \hat{\Phi}_1-\Phi_1 \\ \hat{\Phi}_2-\Phi_2 \\ \vdots \\ \hat{\Phi}_{m-1}-\Phi_{m-1} \\ \hat{\Phi}_m-\Phi_m \end{pmatrix} = \left(\frac{n}{\ln(n)}\right)^{\delta/\alpha}\begin{pmatrix} \hat{\beta}-\beta+\hat{\eta}_1-\eta_1 \\ \hat{\eta}_2-\eta_2-\hat{\eta}_1+\eta_1 \\ \vdots \\ \hat{\eta}_{m-1}-\eta_{m-1}-\hat{\eta}_{m-2}+\eta_{m-2} \\ -\hat{\eta}_{m-1}+\eta_{m-1} \end{pmatrix}$$

and by virtue of (36) therefore,

$$(n/\ln n)^{\delta/\alpha}(\widehat{\mathbf{\Phi}}-\mathbf{\Phi}) = o_p(1),$$

for any $\delta \in (0,1)\cap(0,\alpha]$, $\alpha \in (0,2)$.

Part (d). Put $\delta = 1$ in (22) and subsequently, so that $\mathbf{\Upsilon}_{1,n}^{-1}\mathbf{C}_n = \mathbf{O}_p(1)$. Then, (32) continues to hold, while (33) continues to hold *provided* $\alpha \in [\delta, 2)$, so that (36) is replaced by

$$\begin{array}{rcl} \mathbf{\Lambda}_n^*(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) & \Longrightarrow & \left[\begin{array}{cc} O_p(1) & \mathbf{o}_p(\mathbf{1}) \\ \mathbf{o}_p(\mathbf{1}) & \mathbf{O}_p(\mathbf{1}) \end{array}\right]_{m\times m} \times \left[\begin{array}{c} O_p(1) \\ \mathbf{O}_p(\mathbf{1}) \end{array}\right]_{m\times 1} \\ & = & \left[\begin{array}{c} O_p(1) \\ \mathbf{O}_p(\mathbf{1}) \end{array}\right]_{m\times 1}, \end{array} \tag{37}$$

which gives the result. ■

PROOF OF COROLLARY 6: Immediate from the Theorem statement. We have been unable to obtain a normalization to $O_p(1)$ for $\alpha \in (0,1)$; this is because for (33) to hold we require $\delta \in (0,\alpha]$ which in turn forces (25) to be $o_p(1)$.

PROOF OF THEOREM 8:

The theorem establishes the order in probability of an approximate recursion for the residual sum of squares of the least squares estimator in the setting of the model, (1).

The data arrays are defined as

$$\begin{array}{rcl} \mathbf{X}_0 & = & [X_1,...,X_n]' \\ \mathbf{X}_1 & = & [X_0,...,X_{n-1}]' \\ & \vdots & \\ \mathbf{X}_K & = & [X_{-(K-1)},...,X_{n-K}]'. \end{array} \tag{38}$$

We define RSS_U to be the residual sum of squares from least squares regression of $\mathbf{X}_0$ on $\mathcal{X}_K$=$[\mathbf{X}_1 \vdots \mathbf{X}_2 \vdots \cdots \vdots \mathbf{X}_K]$ and RSS_R to be that obtained by regression on the reduced set, $\mathcal{X}_{K-1} = [\mathbf{X}_1 \vdots \mathbf{X}_2 \vdots \cdots \vdots \mathbf{X}_{K-1}]$. The least squares estimator of the last coefficient in the first regression is written $\hat{\Phi}_{K,K}$.

Part (a). Apply the well-known formula for the increase in residual sum of squares from imposition of a linear restriction, to obtain

$$RSS_R - RSS_U = \hat{\Phi}^2_{K,K} / \left(\mathcal{X}'_K \mathcal{X}_K\right)^{-1}_{KK}.$$

Applying the usual partitioned inversion formula we find the $(KK)th$ element of $\left(\mathcal{X}'_K \mathcal{X}_K\right)^{-1}$ is given by

$$\begin{aligned} & \left(\mathcal{X}'_K \mathcal{X}_K\right)^{-1}_{KK} \\ = \; & \{\mathbf{X}'_K \mathbf{X}_K - \mathbf{X}'_K \mathcal{X}_{K-1} (\mathcal{X}'_{K-1} \mathcal{X}_{K-1})^{-1} \mathcal{X}'_{K-1} \mathbf{X}_K\}^{-1} \end{aligned}$$

hence the increase in the RSS is, exactly,

$$\begin{aligned} & RSS_R - RSS_U \\ = \; & \hat{\Phi}^2_{K,K} \{\mathbf{X}'_K \mathbf{X}_K - \mathbf{X}'_K \mathcal{X}_{K-1} (\mathcal{X}'_{K-1} \mathcal{X}_{K-1})^{-1} \mathcal{X}'_{K-1} \mathbf{X}_K\} \\ = \; & \hat{\Phi}^2_{K,K} \times RSS^\dagger, say. \end{aligned} \tag{39}$$

Part (b). Now consider RSS_R; this is the residual sum of squares from regression of $\mathbf{X}_0$ on $\mathcal{X}_{K-1}$, and is thus

$$RSS_R = \mathbf{X}'_0 \mathbf{X}_0 - \mathbf{X}'_0 \mathcal{X}_{K-1} (\mathcal{X}'_{K-1} \mathcal{X}_{K-1})^{-1} \mathcal{X}'_{K-1} \mathbf{X}_0.$$

We may write

$$\begin{aligned} RSS_U &= RSS_R - \hat{\Phi}^2_{K,K} RSS^\dagger \\ &= RSS_R (1 - \hat{\Phi}^2_{K,K}) \{1 + \frac{\hat{\Phi}^2_{K,K} (RSS_R - RSS^\dagger)}{RSS_R (1 - \hat{\Phi}^2_{K,K})}\} \\ &= RSS_R (1 - \hat{\Phi}^2_{K,K}) \{1 + \delta_K\}, say. \end{aligned}$$

Parts (c, and d) relate to the order in probability of the object,

$$\delta_K = \frac{\hat{\Phi}^2_{K,K} (RSS_R - RSS^\dagger)}{RSS_R (1 - \hat{\Phi}^2_{K,K})}.$$

Part (c) Suppose first that $\{X_t\}$ is a zero-mean covariance stationary Gaussian process with moving average representation,

$$X_t = \sum_{j=0}^{\infty} c_j u_{t-j}$$

in which the $\{c_j\}$ satisfy:

$$\sum_{j=0}^{\infty} {c_j}^2 < \infty. \tag{40}$$

Then we may write the lag-j autocovariance as $\sigma^2\rho_j$, and let $\sigma^2\mathbf{P}_K$ denote the $K \times K$ covariance matrix with ij element, $\sigma^2\rho_{|i-j|}$. The key fact is that $\mathbf{P}_{K-1}$ is a Toeplitz matrix. Writing the row or column reversing transformation as

$$\mathbf{R}_{K-1} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \ddots & 1 & 0 \\ \vdots & \ddots & \diagup & \ddots & \vdots \\ 0 & 1 & \ddots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

and noting that

$$\begin{aligned} \mathbf{R}_{K-1} &= \mathbf{R}_{K-1}^{-1} \\ \mathbf{P}_{K-1} &= \mathbf{R}_{K-1}\mathbf{P}_{K-1}\mathbf{R}_{K-1} \\ \mathbf{P}_{K-1}^{-1} &= \mathbf{R}_{K-1}\mathbf{P}_{K-1}^{-1}\mathbf{R}_{K-1} \end{aligned}$$

observe that because the averaged cross-products matrix converges to $\sigma^2\mathbf{P}_{K-1}$ we must have

$$\begin{aligned} n^{-1}\mathbf{X}_0'\mathcal{X}_{K-1} &\implies \sigma^2[\rho_1, ..., \rho_{K-1}] \\ n^{-1}\mathbf{X}_K'\mathcal{X}_{K-1} &\implies \sigma^2[\rho_{K-1}, ..., \rho_1] \end{aligned} \tag{41}$$

so that

$$\begin{aligned} n^{-1}RSS^\dagger &= n^{-1}\{\mathbf{X}_K'\mathbf{X}_K - \mathbf{X}_K'\mathcal{X}_{K-1}(\mathcal{X}_{K-1}'\mathcal{X}_{K-1})^{-1}\mathcal{X}_{K-1}'\mathbf{X}_K\} \\ &\implies \sigma^2\{\rho_0 - [\rho_{K-1}, ..., \rho_1]\mathbf{P}_{K-1}^{-1}[\rho_{K-1}, ..., \rho_1]'\} \end{aligned}$$

while similarly

$$\begin{aligned} n^{-1}RSS_R &= n^{-1}[\mathbf{X}_0'\mathbf{X}_0 - \mathbf{X}_0'\mathcal{X}_{K-1}\{\mathcal{X}_{K-1}'\mathcal{X}_{K-1}\}^{-1}\mathcal{X}_{K-1}'\mathbf{X}_0] \\ &\implies \sigma^2\{\rho_0 - [\rho_1, ..., \rho_{K-1}]\mathbf{P}_{K-1}^{-1}[\rho_1, ..., \rho_{K-1}]'\} \end{aligned}$$

so that

$$(RSS_R - RSS^\dagger)/n = o_p(1).$$

Thus,

$$\begin{aligned} \frac{\delta_K}{\hat{\Phi}_{K,K}^2} &= \frac{(RSS_R - RSS^\dagger)}{RSS_R(1 - \hat{\Phi}_{K,K}^2)} \\ &= \frac{(RSS_R - RSS^\dagger)/n}{RSS_R(1 - \hat{\Phi}_{K,K}^2)/n} \\ &= \frac{o_p(1)}{O_p(1)} = o_p(1). \end{aligned}$$

Now suppose that $\alpha < 2$. If we strengthen (40), following Phillips (1990, condition (25)), to

$$\sum_{j=0}^{\infty} j|c_j|^{\tau} < \infty \qquad \text{for some } 0 < \tau < 1 \wedge \alpha, \tag{42}$$

(which is satisfied by the MA representation of a stationary AR process), then we obtain from Davis and Resnick (1985 Theorem 4.2), or Lemma 4 (iii) above:

$$a_n^{-2}\mathbf{X}_0'(\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_K) \Longrightarrow v^2(\rho_0, ..., \rho_K)$$

and

$$a_n^{-2}\mathbf{X}_K'(\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_K) \Longrightarrow v^2(\rho_K, ..., \rho_0)$$

in which $\rho_i = \sum_{j=0}^{\infty} c_j c_{j+i}$ and v^2 is a stable random variable with index of stability $\alpha/2$. Again writing $\mathbf{P}_{K-1}$ for the $(K-1) \times (K-1)$ matrix with ij element, $\rho_{|i-j|}$ we see that

$$\frac{\mathcal{X}_{K-1}'\mathcal{X}_{K-1}}{a_n^2} \Longrightarrow v^2\mathbf{P}_{K-1}.$$

Putting these results together we obtain

$$\frac{RSS_R}{a_n^2} \Longrightarrow v^2(\rho_0 - [\rho_1, ..., \rho_{K-1}]\mathbf{P}_{K-1}^{-1}[\rho_1, ..., \rho_{K-1}]')$$

and

$$\frac{RSS^{\dagger}}{a_n^2} \Longrightarrow v^2(\rho_0 - [\rho_{K-1}, ..., \rho_1]\mathbf{P}_{K-1}^{-1}[\rho_{K-1}, ..., \rho_1]')$$

in which the second factors on the RHS are identical because $\mathbf{P}_{K-1}$ is a Toeplitz matrix, as previously noted. Thus,

$$a_n^{-2}(RSS_R - RSS^{\dagger}) = o_p(1).$$

We see therefore, that

$$\begin{aligned} \frac{\delta_K}{\hat{\Phi}_{K,K}^2} &= \frac{(RSS_R - RSS^{\dagger})a_n^{-2}}{a_n^{-2}RSS_R(1 - \hat{\Phi}_{K,K}^2)} \\ &= o_p(1). \end{aligned} \tag{43}$$

Part (d) To deal with the unit root nonstationary case we apply the usual linear transformation, (19), to eliminate the asymptotic singularity of the regressor cross products matrix. We present the details for the case $0 < \alpha < 2$ as a minor modification yields the result for $\alpha = 2$ in the same way as in Part (c) above.

So, now define the vectors,

$$\begin{aligned}
\mathbf{Z}_0 &= [\Delta X_1, ..., \Delta X_n]' \\
\mathbf{X}_1 &= [X_0, ..., X_{n-1}]' \\
\mathbf{Z}_1 &= [\Delta X_0, ..., \Delta X_{n-1}]' \\
&\vdots \\
\mathbf{Z}_{K-1} &= [\Delta X_{-(K-2)}, ..., \Delta X_{n-K+1}]' \\
\tilde{\mathbf{X}}_{K-1} &= [\mathbf{X}_1, \mathbf{Z}_1, ..., \mathbf{Z}_{K-2}]
\end{aligned}$$

Then, the residual sum of squares from estimation of the equations,

$$\mathbf{Z}_0 = \tilde{\mathbf{X}}_{K-1}\boldsymbol{\beta} + \mathbf{e} \tag{44}$$

and

$$\mathbf{X}_0 = \mathcal{X}_{K-1}\boldsymbol{\pi} + \mathbf{e}$$

are identical. However, (44) has the advantages that the dependent variable is stationary and the regressor cross-products matrix is asymptotically non-singular. We now introduce a norming matrix, as in (21)

$$\boldsymbol{\Lambda}_n = \begin{bmatrix} n^{1/2}a_n & \mathbf{0}' \\ \mathbf{0} & a_n\mathbf{I} \end{bmatrix}$$

and write

$$\mathbf{M}_n = \tilde{\mathbf{X}}'_{K-2}\tilde{\mathbf{X}}_{K-2}.$$

It follows from Lemma 4 and Davis and Resnick (1985, Theorem 4.2) that

$$\boldsymbol{\Lambda}_n^{-1}\mathbf{M}_n\boldsymbol{\Lambda}_n^{-1} = \begin{bmatrix} O_p(1) & \mathbf{o}'_p(1) \\ \mathbf{o}_p(1) & w^2\mathbf{P}_{z,K-2} \end{bmatrix}$$

and

$$\begin{aligned}
a_n^{-1}\mathbf{Z}'_0\tilde{\mathbf{X}}_{K-2}\boldsymbol{\Lambda}_n^{-1} &= [o_p(1), w^2\rho_{z,1}, ..., w^2\rho_{z,K-2}] \\
&= [o_p(1), w^2\boldsymbol{\rho}'_{K-2}]
\end{aligned}$$

and from Davis and Resnick (1985, Theorem 4.2)

$$\frac{\mathbf{Z}'_0\mathbf{Z}_0}{a_n^2} \Longrightarrow w^2\rho_{z0} \qquad say,$$

in which the stable RV w^2 and constants, ρ_{zj} are defined with respect to the shocks driving Z_t.

Thus we have

$$\begin{aligned}
& a_n^{-2}RSS_R \\
= \quad & \frac{\mathbf{Z}'_0\mathbf{Z}_0}{a_n^2} - \frac{\mathbf{Z}'_0\tilde{\mathbf{X}}_{K-2}\boldsymbol{\Lambda}_n^{-1}}{a_n}\left[\boldsymbol{\Lambda}_n^{-1}\mathbf{M}_n\boldsymbol{\Lambda}_n^{-1}\right]^{-1}\frac{\boldsymbol{\Lambda}_n^{-1}\tilde{\mathbf{X}}'_{K-2}\mathbf{Z}_0}{a_n} \\
\Longrightarrow \quad & w^2[\rho_{z0} - \boldsymbol{\rho}'_{K-2}\mathbf{P}_{z,K-2}^{-1}\boldsymbol{\rho}_{K-2}]
\end{aligned}$$

and similarly

$$\begin{aligned} & a_n^{-2} RSS^{\dagger} \\ = \quad & \frac{\mathbf{Z}'_{K-1}\mathbf{Z}_{K-1}}{a_n^2} - \frac{\mathbf{Z}'_{K-1}\tilde{\mathbf{X}}_{K-2}\mathbf{\Lambda}_n^{-1}}{a_n}\left[\mathbf{\Lambda}_n^{-1}\mathbf{M}_n\mathbf{\Lambda}_n^{-1}\right]^{-1}\frac{\mathbf{\Lambda}_n^{-1}\tilde{\mathbf{X}}'_{K-2}\mathbf{Z}_{K-1}}{a_n} \\ \Longrightarrow \quad & w^2[\rho_{z0} - \boldsymbol{\rho}'_{K-2}\mathbf{P}^{-1}_{z,K-2}\boldsymbol{\rho}_{K-2}] \end{aligned}$$

and once again we find that

$$a_n^{-2}(RSS_R - RSS^{\dagger}) \Longrightarrow 0.$$

It now follows that

$$\begin{aligned} \frac{\delta_K}{\hat{\beta}^2_{K,K}} &= \frac{(RSS_R - RSS^{\dagger})a_n^{-2}}{a_n^{-2}RSS_R(1-\hat{\beta}^2_{K,K})} \\ &= o_p(1) \end{aligned} \tag{45}$$

as in the stationary case.

PROOF OF THEOREM 9:

We have to show that if X_t is a non-stationary $AR(m)$, $m \geq 1$, defined by (1), with $\rho = 1$, $0 < \alpha < 2$, the upper bound, $K > m$, and if $\widehat{m}$ minimizes the criterion, $IC_{OLS}(k)$, defined in (9) and

(a) $\tilde{\Phi}^2_{k,p}/C(n) = o_p(1)$ for $k, p > m$, and

(b) $C(n) = o(1)$

then $\widehat{m} \longrightarrow_p m$.

Consider first the probability that $\hat{m} < m$. By the argument below (10) we have to show that $\sum_{k=m-j+1}^{k=m} \ln(1+\delta_k) + jC(n)$ is $o_p(1)$. However, δ_k is $o_p(1)$ as is clear from (45), while $jC(n)$ is $o_p(1)$ by condition (b). For the probability that $\hat{m} > m$, note that from the discussion below (12), given the result of Theorem 8, we need only show that $\tilde{\Phi}^2_{k,p}/C(n) = o_p(1)$ for $k, p > m$, which is condition (a).

PROOF OF THEOREM 12:

We have to show that when $\alpha \in (1, 2)$, the lag length estimator,

$$\hat{m} = \arg \min_{k,\ 0 \leq k \leq K} LC(k)$$

in which $K \geq m$ is consistent when

$$\begin{aligned} LC(k) &= -\sum_{j=1}^{j=k} \widehat{\Phi}^2_{j,j} + (k+1)C(n) \\ LC(0) &= C(n), \end{aligned}$$

$C(n)$ satisfies the two conditions,

(a) $\hat{\Phi}^2_{k,p}/C(n) = o_p(1)$ for $k, p > m$, and

(b) $C(n) = o(1)$

and $\widehat{\Phi}^2_{j,j}$ is as defined above (13).

Observe that the increment from $k = j$ to $k = m$ is

$$LC(m) - LC(j) = (m-j)C(n) - \sum_{k=j}^{k=m} \widehat{\Phi}^2_{k,k}$$

and that

$$(m-j)C(n) = o(1)$$

while $\displaystyle\sum_{k=j}^{k=m} \widehat{\Phi}^2_{k,k}$ is eventually bounded away from zero by virtue of the fact that $\widehat{\Phi}^2_{m,m}$ converges to $\Phi^2_m > 0$. Hence the increment is a.s. negative for all large enough n proving that $\Pr\{\hat{m} < m\} \to 0$.

For $\hat{m} > m$ we observe that it will suffice to show that eventually, the increment,

$$LC(l) - LC(m) = (l-m)C(n) - \sum_{k=m}^{k=l} \widehat{\Phi}^2_{k,k}$$

is positive for every $l > m$. However, by condition (a)

$$\frac{LC(l) - LC(m)}{C(n)} \to (l-m) > 0. \blacksquare$$

References

[1] Adler, R.J., Feldman, R.E., and M.S. Taqqu (1998), *A practical guide to heavy tails*, Birkhauser, Boston.

[2] Akaike, H. (1974) A new look at statistical model identification. *IEEE Transactions on Automatic Control* AC-19 pp. 716-723.

[3] Calder, M. and R.A.Davis (1998), Inference for linear processes with stable noise. In Adler, R.J., Feldman, R.E., and M.S. Taqqu (1998), *A practical guide to heavy tails*, Birkhauser, Boston, pp 159-176.

[4] Chambers, J.M., C.L. Mallows and B.W. Stuck, 1976, A Method for Simulating Stable Random Variables, Journal of the American Statistical Association 71, 340-344.

[5] Chan, N.H. and Tran, L.T. (1989) On the first order autoregressive process with infinite variance. *Econometric Theory,* 5, 354-362.

[6] Charemza, W.W., Hristova, D. and P. Burridge (2005), Is Inflation Stationary?, *Applied Economics*, 37, 8, 901-903.

[7] Davis, R., Knight, K., and J.Liu (1992), M-Estimation for autoregressions with infinite variance. *Stochastic Processes and their Applications* 40 pp. 145-180.

[8] Davis, R. and S. Resnick (1986), Limit Theory for the Sample Covariance and Correlation Functions of Moving Averages. *The Annals of Statistics*, Vol 14 No 2 pp. 533-558.

[9] Durbin, J. (1960), The fitting of time series models, *Review of the International Statistical Institute*, 28, pp. 233-244.

[10] Feller, W. (1971), *An Introduction to Probability Theory and its Applications 2*, 2nd ed., Wiley, New York.

[11] Gonzalo, J. and J-Y. Pitarakis (2002), Lag Length Estimation in Large Dimensional Systems, *Journal of Time Series Analysis* 23, 4, 401-423.

[12] Hall, A.R. (1994), Testing for a Unit Root with Pretest Data Based Model Selection, *Journal of Business and Economic Statistics*, 12, 461-470.

[13] Hannan, E.J. and M. Kanter (1977), Autoregressive processes with infinite variance, *Journal of Applied Probability*, 14, 411-415.

[14] Hannan, E.J., and B. G. Quinn (1979) *The Determination of the Order of an Autoregression.* Journal of the Royal Statistical Society. Series B (Methodological), Vol. 41, No. 2, pp. 190-195.

[15] Knight, K. (1989), Consistency of Akaike's Information Criterion for Infinite Variance Autoregressive Processes, *The Annals of Statistics*, 17, 2, 824-840.

[16] Ng, S. and P. Perron (2001), Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power, *Econometrica,* 69, 6, 1519-1554.

[17] Phillips, P.C.B. (1990), Time Series Regression with a Unit Root and Infinite-Variance Errors, *Econometric Theory*, 6, 44-62.

[18] Phillips, P.C.B., and Ploberger, W. (1994), Posterior odds testing for a unit root with data-based model selection, *Econometric Theory* 10 pp. 774-808.

[19] Pötscher, B.M. (1989), Model selection under non-stationarity: autoregressive models and stochastic linear regression models, *The Annals of Statistics*, 17, 3, pp. 1257-1274.

[20] Rachev, S. T., Mittnik, S. and Kim, J.R. (1998) Time series with unit roots and infinite variance disturbances, *Applied Mathematics Letters*, 11, 69-74.

[21] Schwarz, G. (1978), Estimating the Dimension of a Model, *The Annals of Statistics*, 6, 2, 461-464.

[22] Shibata, R. (1976), Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika*, 63, 1, 117-126.

[23] Tiao, C.G. and R.S.Tsay (1983), Consistency properties of least squares estimates of autoregressive parameters in ARMA models, *The Annals of Statistics*, 11, 4, 1425-1433.