CONSISTENT ESTIMATION AND ORDER SELECTION FOR NONSTATIONARY AUTOREGRESSIVE PROCESSES WITH STABLE INNOVATIONS

BY PETER BURRIDGE AND DANIELA HRISTOVA

University of York and City University

First version received November 2007

Abstract. A possibly nonstationary 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, $\alpha$, is established. We also establish consistency of lag-order selection criteria in the nonstationary 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 numbers. C13; C22.

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 associated with the presence of autoregressive unit roots, building on the seminal contribution of Chan and Wei (1988). In particular, lag-length selection for unit-root testing is discussed 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, equivalently, to select the length that minimizes an information criterion (IC). When the innovations are drawn from a distribution within the domain of attraction of the normal distribution, (or, as in Pötscher, 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 Tsay (1984), Paulsen (1984), Pötscher (1989) and Ng and Perron (2001), for the nonstationary case. A recent synthesis and extension is provided by Nielsen (2006), who shows that the usual lag-selection methods are robust in general vector autoregressions in the presence
of deterministic components, regardless of the location of the characteristic roots, when moment conditions are satisfied. However, many macroeconomic and financial series, notably stock returns, appear to violate such conditions, having heavy-tailed distributions as described in Adler et al. (1998), for example, and this raises the question of the applicability of existing consistency results.

The question is potentially important: recently, Charemza et al. (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 augmented Dickey–Fuller (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, the so-called $m$-estimators can be very much more efficient than least squares, especially for $\alpha$ 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 the empirical analysis of economic time series, and in our own previous work we found that estimated $\alpha$ 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.

Bhansali (1988) establishes the consistency of lag-length selection in the stationary heavy-tailed case via the final prediction error (FPE) criterion of Bhansali and Downham (1977). In a major contribution, Knight (1989) obtains somewhat more general results, showing 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, provided the upper bound on the true order does not grow too rapidly with the sample size. He also proves consistency of the ordinary least squares (OLS) parameter estimator in this context. In the present article, we study the least squares estimator of possibly nonstationary processes with no more than one unit autoregressive root. The article 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 are presented in Sections 3 and 4; some simulations illustrate finite-sample behaviour in Section 5.
and conclusions and a brief discussion appear in Section 6; proofs are 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 difference stationary process. Further, suppose $X_t$ has no deterministic components such as level or trend. With $L$ the usual lag operator, then, $X_t$ is an AR($m$) process,$$egin{align*}
\Phi(L)X_t &= u_t, \\
\Phi(L) &= 1 - \sum_{j=1}^{m} \Phi_j L^j,
\end{align*}
$$with $m \in \mathbb{N}$, $m \geq 1$, and $\Phi(L) = (1 - \rho L)\phi(L)$, where all the zeros of $\phi(L)$, $\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_{m-1} L^{m-1}$, 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 $Z_t$,$$egin{align*}
Z_t &= (1 - L)X_t.
\end{align*}
$$The true order, $m$, is unknown but bounded by some known finite integer, $K$. We comment on the situation in which $K$ is allowed to increase with sample size in Section 6.

The disturbances $u_t$ are independently and identically distributed (i.i.d.) 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,

$$
\lim_{x \to \infty} \frac{L(sx)}{L(x)} = 1, \quad \forall s > 0
$$

(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

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

© 2008 The Authors
Journal compilation © 2008 Blackwell Publishing Ltd.
JOURNAL OF TIME SERIES ANALYSIS Vol. 29, No. 4
for some $0 < a < \infty$, which corresponds to taking $L(x) = \text{constant}$. Innovations with such properties will be denoted by $u_t \sim \text{i.i.d. SP}(x)$ in the rest of the article.

Although, in the cases of interest, $x \in (0,2)$, 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 2, for those properties we require.

2.2. The results in general terms

For processes of 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 article, we seek results for processes of type (1) with $\rho = 1$ for which the YW estimator is poorly defined and becomes numerically unstable, and so cannot be applied. The OLS 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 nonstationary 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 2, we find a rate of convergence for the least squares estimator of $\Phi(L)$ in eqn (1) in the case in which $\rho = 1$, while in Theorem 4 we establish consistency of the lag order selected by the AIC criterion in the nonstationary $\alpha$-stable case, thus complementing the results of Knight (1989).

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, beginning with the definition of the relevant IC. Akaike’s (1974) criterion is

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

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...
\[
\text{BIC} = \ln \hat{\sigma}_u^2 + \frac{k \ln n}{n}, \tag{4}
\]

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

\[
\text{HQIC} = \ln \hat{\sigma}_u^2 + \frac{ck \ln(\ln n)}{n} \tag{5}
\]

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 \(\sigma_u^2\) 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 YW equations via the Levinson–Durbin recursion, due to Durbin (1960):

\[
\hat{\sigma}_u^2(k) = \hat{\sigma}_u^2(0) \prod_{j=1}^{k}(1 - \Phi_{jj}^2), \tag{7}
\]

in which \(\hat{\sigma}_u^2(0) = n^{-1} \sum X_t^2\) and \(\Phi_{jj}\) is the YW estimate of the \(j\)th partial autocorrelation coefficient. Hannan and Quinn use eqn (7) to define their information criterion, and the consistency proof for the stationary \(z\)-stable case devised by Knight also explicitly uses the YW estimators to form the AIC, and hence also exploits eqn (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, writing the \(j\)th partial autocorrelation as \(\Phi_{jj}^2\), we obtain

\[
\hat{\sigma}_u^2(k) = \hat{\sigma}_u^2(0) \prod_{j=1}^{k} \{1 - \Phi_{jj}^2 \delta_j\}, \tag{8}
\]

in which, as Theorem 3 shows, in all relevant cases, \((\delta_j - 1) = o_P(1)\), 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 by minimizing an IC defined in terms of the OLS estimator may be proved using the representation (eqn 8), consider a generic criterion that can be written,

\[
\text{IC}_{\text{OLS}}(k) = \ln \hat{\sigma}_u^2(k) + kC(n). \tag{9}
\]
The increment of IC(k) from \( k = m - j \) to \( k = m \) is

\[
\text{IC}_{\text{OLS}}(m) - \text{IC}_{\text{OLS}}(m - j) = \ln\left( \frac{\hat{\sigma}_\epsilon^2(m)}{\hat{\sigma}_\epsilon^2(m - j)} \right) + jC(n)
\]

\[
= \sum_{k=m-j+1}^{k=m} \{\ln(1 - \hat{\Phi}_{k,k}^2 \delta_k)\} + jC(n). \tag{10}
\]

Now, provided \( \Phi_{m,m} \) is a consistent estimator of \( \Phi_{m,m} \), \( 0 < |\Phi_{m,m}| < 1 \), and \( \delta_k - 1 = o_p(1) \), 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 \( jC(n) \) is \( o(1) \) 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,

\[
\text{IC}_{\text{OLS}}(m + j) - \text{IC}_{\text{OLS}}(m) = \sum_{k=m+1}^{k=m+j} \{\ln(1 - \hat{\Phi}_{k,k}^2 \delta_k)\} + jC(n). \tag{11}
\]

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

\[
\Pr\{\text{IC}_{\text{OLS}}(k) < \text{IC}_{\text{OLS}}(k - 1)\}
\]

for some such \( k \), that is, it is smaller than

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

Now suppose that for any \( k > m \), \( \hat{\Phi}_{k,k}^2 \to 0 \) (proved in Theorem 2) while \( (\delta_k - 1) = o_p(1) \), (proved in Theorem 3), then for large enough \( n \) we may write

\[
\ln(1 - \hat{\Phi}_{k,k}^2 \delta_k) = \ln(1 - \hat{\Phi}_{k,k}^2 (\delta_k - 1)\Phi_{k,k}^2)
\]

\[
\simeq \ln(1 - \hat{\Phi}_{k,k}^2)
\]

\[
\simeq -\hat{\Phi}_{k,k}^2
\]

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

3. Convergence of the OLS Estimator

We start with a lemma of Tiao and Tsay (1983, Lemma 2.3, 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 \( Y_t = (X_t, X_{t-1}, \ldots, X_{t-p} + 1)' \). If \( X_t \) is not a purely deterministic process, then, for
\[ n \geq 2p, \quad \Phi_n = \sum_{i=p+1}^{n} Y_{i-1}Y'_{i-1} \] is a symmetric and positive definite matrix with probability 1.

Next, we record results that give rates 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 \(\Phi\) with the number of lags included in the estimated model:

**Theorem 1.** Let \(X_t\) be an autoregressive process given by eqn (1) with \(|\rho| < 1\) and \(u_t \sim \text{i.i.d. SP}(\alpha)\) for \(\alpha \in (0,2)\). Let \(\hat{\Phi}_{t,p}\) denote the OLS estimator of \(\Phi_{t,p}, i = 1, \ldots, p\), where \(\Phi_{t,p} = 0\) for \(i > m\). Then for \(p \geq m\)

\[
\begin{align*}
(i) & \quad n^{1/\gamma}(\hat{\Phi}_{t,p} - \Phi_{t,p}) \longrightarrow 0 \quad \text{a.s. for any } \gamma > \alpha \quad (13) \\
(ii) & \quad \left(\frac{n}{\ln(n)}\right)^{1/\alpha} (\hat{\Phi}_{t,p} - \Phi_{t,p}) = O_p(1). \quad (14)
\end{align*}
\]

**Remark 1.** Part (i) is proved by Hannan and Kanter (1977, p. 412), while the stronger part (ii) is obtained by Davis and Resnick (1986 Thm 4.4 and corollaries).

Lemma 2 collects some auxiliary results in the orders of magnitude of various sample second moments. We now let \(X_t\) be a unit-root nonstationary autoregressive process.

**Lemma 2.** Let \(X_t\) be generated by eqn (1) with \(\rho = 1\) and \(Z_t\) be defined by eqn (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 \(V_\alpha(r)\) for the Lévy process on \([0,1]\) to which normalized partial sums of \(u_t\) converge. Then

\[
\begin{align*}
(i) & \quad \sum_{t=1}^{n} Z_{t-j}u_t = O_p(a^2(n \ln n)^{1/2}) \\
(ii) & \quad \sum_{t=1}^{n} X_{t-1}u_t = O_p(a_n^2) \\
(iii) & \quad a_n^{-2} \sum_{t=1}^{n} Z_{t-j}Z_{t-1} \Longrightarrow \left(\sum_{l=0}^{\infty} b_l b_{l+j-l}\right) \int_{0}^{1} (dU_\alpha)^2 = O_p(1) \\
(iv) & \quad n^{-1}a_n^{-2} \sum_{t=1}^{n} X_{t-1}^2 \Longrightarrow \omega^2 \int_{0}^{1} U_\alpha^2 = O_p(1)
\end{align*}
\]
\begin{align*}
&\text{(v)}\quad \sum_{t=1}^{n} X_{t-1} Z_{t-1} = O_p(a_n^2) \\
&\text{(vi)}\quad \left( \sum_{t=1}^{n} X_{t-1}^2 \right)^{-1} \left( \sum_{t=1}^{n} X_{t-1} Z_{t-1} \right) = O_p(n^{-1})
\end{align*}

where \( a_n = an^{1/\alpha} \) for some \( a > 0 \), 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 2.** Let \( X_t \) be generated by eqn (1) with \( \rho = 1 \), and \( \alpha \in (0,2) \) and let \( \hat{\Phi}_p = [\hat{\Phi}_1, \ldots, \hat{\Phi}_{p,m}]' \), \( p \geq m, m \geq 1 \), denote the \( p \) element vector of OLS autoregressive coefficient estimates. Then, defining the trailing \((p-m)\) elements of \( \Phi_p \) to be zero, so that \( \Phi_p = [\Phi_m, 0]' \),

\begin{align*}
&\text{(a) if } p = 1 \\
&\quad n(\hat{\Phi}_{1,1} - 1) = O_p(1) \\
&\text{(b) if } p > 1 \\
&\quad (n/\ln n)^{\delta/\alpha}(\hat{\Phi}_{i,p} - \Phi_{i,p}) = o_p(1), \quad i = 1, \ldots, p \\
&\quad \forall \delta \in (0,1) \cap (0,\alpha], \alpha \in (0,2) \\
&\text{(c) if } \alpha \geq 1 \text{ and } \delta = 1 \\
&\quad (n/\ln n)^{1/\alpha}(\hat{\Phi}_p - \Phi_p) = O_p(1).
\end{align*}

**Corollary 1.** For \( p > 1 \) and \( \alpha < 1 \) taking \( \delta = \alpha \) we have

\begin{align*}
&\quad (n/\ln n)(\hat{\Phi}_p - \Phi_p) = o_p(1).
\end{align*}

**Remark 2.** Part (a) is proved in Chan and Tran (1989, Thm 2, p. 358). In part (c), we obtain the same rate of convergence in distribution as Davis and Resnick [see eqn (14)]. We have also established that, as in the stationary case considered by Hannan and Kanter (1977), for \( k > m \)

\begin{align*}
&\quad n^{1/\gamma}\hat{\Phi}_{k,p} = o_p(1)
\end{align*}

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

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

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

\[
n^{1/\gamma} \Phi_{k,p} = o_p(1)
\]

by Corollary 1.

We have been unable to obtain a limit distribution for \( \alpha \in (0,1) \), but part (b) gives us convergence in probability for this case.

4. CONSISTENCY OF INFORMATION CRITERIA FOR LAG-LENGTH SELECTION

We are now ready to find the convergence rate for the relation (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 Theorem 3.1

**Theorem 3.** Consider data arrays defined, for any \( K > 0 \), as

\[
\begin{align*}
X_0 &= [X_1, \ldots, X_n]' \\
X_1 &= [X_0, \ldots, X_{n-1}]' \\
&\vdots \\
X_K &= [X_{-(K-1)}, \ldots, X_{n-K}]'.
\end{align*}
\]

Let \( \text{RSS}_K \) denote the residual sum of squares from OLS regression of \( X_0 \) on \( X_K = [X_1; X_2; \cdots; X_K] \), and that for regression on the reduced set, \( X_{K-1} = [X_1; X_2; \cdots; X_{K-1}] \) be denoted \( \text{RSS}_{K-1} \). Further, write the least squares estimator of the last coefficient in the first regression as \( \hat{\Phi}_{K,K} \). Then

(a)

\[
\text{RSS}_K = \text{RSS}_{K-1}(1 - \hat{\Phi}_{K,K}^2 \delta_K),
\]

where

\[
\delta_K = \frac{\text{RSS}_{K-1}^\dagger}{\text{RSS}_{K-1}}
\]

and

\[
\text{RSS}_{K-1}^\dagger = X_K'X_K - X_K'X_{K-1}(X_{K-1}'X_{K-1})^{-1}X_{K-1}'X_K.
\]

(b) Suppose \( X_t \) is a stationary AR(m) and \( 0 < \alpha \leq 2 \), then, for each \( K > 0 \),

\[
(\delta_K - 1) = o_p(1)
\]
Suppose \( \Delta X_t \) is a stationary AR\((m - 1)\) and 0 < \( \alpha \leq 2 \), then, for each \( K > 0 \),

\[
(\delta_K - 1) = o_p(1).
\]

Theorem 4 is our main result.

**Theorem 4.** Let \( X_t \) be a non-stationary AR\((m)\), \( m \geq 1 \), defined by eqn (1), with \( \rho = 1 \) and let \( K > m \), be an upper bound. If \( \hat{m} \) minimizes the criterion, \( IC_{\text{OLS}}(k) \), defined in eqn (9) and

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

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

then \( \hat{m} \to_p m \).

**Remark 3.** By Remark 2, \( C(n) = n^{-1}c \) for some \( c > 0 \) (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. In particular, notice that it is the more rapid convergence to zero of higher order estimated partial correlations in the \( \alpha < 2 \) case that permits a relaxation of the penalty term in the information criterion.

**Remark 4.** 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 \( \Phi_{1,1} \) in the non-stationary \( m = 1 \) case is faster than in other cases, it is not this rate that determines the appropriate rate for \( C(n) \).

**Remark 5.** In the Gaussian case, the best obtainable rate of decrease of the penalty term, found by Hannan and Quinn (1979), is obtained by applying the law of the iterated logarithm (LIL) to \( \Phi_{k,p}^2 \) for \( k,p > m \). That is, instead of requiring the sufficient condition (a) to hold, we could seek a \( C(n) \) for which \( Pr[\Phi_{k,p}^2 < C(n)] \to 1 \), say, which is obviously weaker. We have been unable to find LIL-type results for the infinite-variance case of sufficient generality, however, so this problem remains open.

### 5. 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 nonstationary 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 nonstationary cases, but more marked differences as \( \alpha \) varies across its range.

The criteria illustrated are defined as

1. AIC\((k) = \ln \hat{\sigma}_u^2 + 2k/n\),
2. BIC\((k) = \ln \hat{\sigma}_u^2 + k \ln(n)/n\).

We report results selectively for three DGPs in addition to a white noise process, \( u_t \):

\( P1: \) root(0) \( \Leftrightarrow X_t = u_t \)
\( P2: \) roots(1,0.6) \( \Leftrightarrow X_t = 1.6X_{t-1} - 0.60X_{t-2} + u_t \)
\( P3: \) roots(0.6,0.2) \( \Leftrightarrow X_t = 0.8X_{t-1} - 0.12X_{t-2} + u_t \)
\( P4: \) roots(1,0.6,0.2) \( \Leftrightarrow X_t = 1.8X_{t-1} - 0.92X_{t-2} + 0.120X_{t-3} + u_t \)

We report our numerical results very selectively in order to highlight important features. Sample size, and the index of stability are selected from \( n \in [100, 250, 500, 10,000] \) 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 10,000 replications.

Table I reports results for the white noise process, \( P1 \), to act as a baseline. It illustrates two things. (i) The inconsistency of AIC for \( \alpha = 2 \) is plain to see in the final column of the upper panel, while, as expected, this criterion also performs poorly for \( \alpha \) close to, but less than 2. (ii) Assuming that the BIC is adopted, then at sample size 100,

### Table I

**Percentage Probabilities of Selecting the True Order, \( m = 0 \) (P1: white noise)**

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( n = 100 )</th>
<th>250</th>
<th>500</th>
<th>10,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>86</td>
<td>89</td>
<td>91</td>
<td>97</td>
</tr>
<tr>
<td>0.90</td>
<td>85</td>
<td>87</td>
<td>89</td>
<td>96</td>
</tr>
<tr>
<td>1.0</td>
<td>84</td>
<td>85</td>
<td>87</td>
<td>94</td>
</tr>
<tr>
<td>1.10</td>
<td>83</td>
<td>85</td>
<td>87</td>
<td>93</td>
</tr>
<tr>
<td>1.25</td>
<td>82</td>
<td>82</td>
<td>84</td>
<td>91</td>
</tr>
<tr>
<td>1.50</td>
<td>79</td>
<td>80</td>
<td>81</td>
<td>86</td>
</tr>
<tr>
<td>1.75</td>
<td>77</td>
<td>77</td>
<td>77</td>
<td>80</td>
</tr>
<tr>
<td>2.00</td>
<td>71</td>
<td>72</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>BIC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>95</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>0.90</td>
<td>95</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>1.0</td>
<td>95</td>
<td>95</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>1.10</td>
<td>95</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>1.25</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>98</td>
</tr>
<tr>
<td>1.50</td>
<td>96</td>
<td>97</td>
<td>97</td>
<td>98</td>
</tr>
<tr>
<td>1.75</td>
<td>97</td>
<td>98</td>
<td>98</td>
<td>99</td>
</tr>
<tr>
<td>2.00</td>
<td>96</td>
<td>98</td>
<td>99</td>
<td>99.7</td>
</tr>
</tbody>
</table>
we have, in effect, approximately a 5% chance of rejecting the true hypothesis, $H_0: m = 0$ whatever the value of $\alpha$. The latter point deserves further comment. Practitioners may often adopt the use of lag-selection criteria as an apparently simpler alternative to the use of sequences of Students $t$-tests or similar. If so, they may be mistaken in believing that the awkward question of the appropriate significance level has thereby been avoided. Consider Table I 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. Such adjustments would have knock-on effects on the performance of the model selection criteria when the process is not while noise, of course.

Table II shows the performance of the BIC criterion in samples of size $n = 100$ and $n = 250$ from models 1, 2 and 3. It is salutary to observe that the model with a small root, and thus a small coefficient on its longest lag, $P_3$, is chronically under-fitted. There is only a one-in-three chance of getting the

<table>
<thead>
<tr>
<th>TABLE II</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PERCENTAGE PROBABILITIES OF SELECTING THE TRUE ORDER, $m = 0$, OR $m = 2$ WITH $n = 100$ OR 250 USING THE BIC CRITERION $P_1$: WHITE NOISE OR $P_2$: AR(2) WITH ROOTS $1$ AND $0.6$; $P_3$: AR(2) WITH ROOTS $0.6$ AND $0.2$.</strong></td>
</tr>
<tr>
<td>$\alpha$</td>
</tr>
<tr>
<td>0.75</td>
</tr>
<tr>
<td>1.00</td>
</tr>
</tbody>
</table>

Figures rounded to nearest 1%.

<table>
<thead>
<tr>
<th>TABLE III</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PERCENTAGE PROBABILITIES OF SELECTING THE TRUE ORDER, $m = 3$, AND UNDER-FITTING, $m = 2$ ($P_4$: $X_t = 1.8X_{t-1} - 0.92X_{t-2} + 0.12X_{t-3} + u_t$)</strong></td>
</tr>
<tr>
<td>$\alpha$</td>
</tr>
<tr>
<td>AIC</td>
</tr>
<tr>
<td>0.75</td>
</tr>
<tr>
<td>1.00</td>
</tr>
<tr>
<td>1.75</td>
</tr>
<tr>
<td>2.00</td>
</tr>
<tr>
<td>BIC</td>
</tr>
<tr>
<td>0.75</td>
</tr>
<tr>
<td>1.00</td>
</tr>
<tr>
<td>1.75</td>
</tr>
<tr>
<td>2.00</td>
</tr>
</tbody>
</table>

*based on 1000 replications.
model order right with the BIC even for samples of 250 observations in the \( \alpha = 2 \) case. That matters are worse in the infinite-variance cases may seem at odds with the theory presented earlier; however, the explanation is that the coefficient is much more accurately estimated in such cases, and its true magnitude, 0.12, is too small to be detected against the \( C(n) \) criterion in this case, a point reinforced by the results in Table III, where the better performance of AIC for these cases is revealed.

6. 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, Thm 4.4), who established a limit law for the sample autocorrelation function, obtaining Theorem 1 (ii) as a corollary. We note that our convergence rate in Theorem 2 matches theirs. 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 (see eqn 8), and we sharpen this in Theorem 3.

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 YW estimator satisfies

\[
 n \max_{m < k \leq K(n)} \Phi_{k,k}^2 \to \rho \quad 0,
\]

so that

\[
 n \min_{m < k \leq K(n)} \ln(1 - \Phi_{k,k}^2) \to \rho \quad 0,
\]

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

\[
 \Pr[k > m] \leq \Pr[\min_{m < k \leq K(n)} \ln(1 - \Phi_{k,k}^2) < -2/n]
 = \Pr[n \min_{m < k \leq K(n)} \ln(1 - \Phi_{k,k}^2) < -2]
 \to 0 \text{ by eqn (16)}.
\]
convergence rate for $\hat{\Phi}$ in that case. 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 distribution of the initial observation is centred on zero. 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 nonstationary, 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.

APPENDIX: PROOFS

Proof of Lemma 2. 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, eqns 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). $\square$

Proof of Theorem 2. Part (b). Process (2) with $\rho = 1$ is equivalent to

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

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

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

in which

$$\hat{\beta} = \sum_{i=1}^{m} \hat{\Phi}_{i,m}, \quad \hat{\eta}_j = - \sum_{i=j+1}^{m} \hat{\Phi}_{i,m}, \quad j = 1, 2, \ldots, m - 1,$$

and $\hat{\Phi}_{i,m}, i = 1, 2, \ldots, m$, is as defined in the theorem statement.

We are interested in establishing a convergence rate for the least squares estimator of $\beta' = (\beta, \eta')$. To do so, we will find a normalizing matrix, $A_n^*$, say, such that $A_n^*[\hat{\beta} - \beta]$ is $(O_p(1), o_p(1))'$. We first introduce some more notation.
Suppose the sample available runs from $X_{(m-1)}$ to $X_n$; as in Lemma 2, define the stationary process, $Z_{t-j} = \Delta X_{t-j}$, and then introduce the $m$ element random vector,

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

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

$$M_n = \sum_{t=1}^n Y_t Y_t'.$$

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 $U_\omega(r)$ for the Lévy process on $[0,1]$ to which normalized partial sums of $u_t$ converge, as in Lemma 2.

The error in the least squares estimator is

$$\hat{\beta} - \beta = M_n^{-1} \sum_{t=1}^n Y_t u_t = M_n^{-1} C_n, \quad \text{say.}$$

(19)

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

$$T_n^{1/2} (\hat{\beta} - \beta) = (T_n^{-1/2} M_n T_n^{-1/2})^{-1} T_n^{-1/2} C_n$$

via the transformation, $T_n^{-1/2} = \text{diag}(n^{-1}, n^{-1/2} I_{m-1})$, in which the objects on the RHS converge in distribution. Unfortunately, in the present setting, the normalization required for $C_n$ is not the square root of that required for $M_n$. We therefore proceed as follows.

Define the $m \times m$ diagonal matrices,

$$\Lambda_n = \begin{bmatrix} a_n n^{1/2} & 0 \\ 0 & a_n I_{m-1} \end{bmatrix},$$

(20)

$$\Lambda_n^* = \begin{bmatrix} n & 0 \\ 0 & \left( \frac{n}{\ln(n)} \right)^{\delta/2} I_{m-1} \end{bmatrix}$$

(21)

and

$$Y_{\delta,n} = \begin{bmatrix} a_n^2 & 0 \\ 0 & a_n^2 \left( \frac{n}{\ln(n)} \right)^{-\delta/2} I_{m-1} \end{bmatrix}.$$
where $\delta > 0$ is a fixed real number. Observe that $Y_{\delta,n}^{-1} = \Lambda_n^{-1} \Lambda_{n-2}$ and define
\[ D_n = \Lambda_n^{-1} M_n \Lambda_n^{-1}. \]
It follows from eqn (19) that
\[ \Lambda_n A_n^{-1} M_n A_n^{-1} A_n (\beta - \beta) = C_n. \]
Hence,
\[ Y_{\delta,n}^{-1} A_n D_n A_n (\beta - \beta) = Y_{\delta,n}^{-1} C_n \]
and
\[ (\beta - \beta) = (A_n^{-1} D_n^{-1} A_n^{-1} Y_{\delta,n}) (Y_{\delta,n}^{-1} C_n) \]
giving us finally
\[ A_n^*(\beta - \beta) = \Lambda_n^* \Lambda_n^{-1} D_n^{-1} \Lambda_n^{-1} Y_{\delta,n} (Y_{\delta,n}^{-1} C_n) \]
\[ = \left( Y_{\delta,n}^{-1} A_n D_n^{-1} A_n^{-1} Y_{\delta,n} \right) (Y_{\delta,n}^{-1} C_n) . \tag{22} \]

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 $O_p(1)$, that is,
\[ n a_n^{-2} \sum X_{t-1} u_t = O_p(1) \]
\[ \text{for any } \delta \in (0,1) \cap (0,2), \ z \in (0,2), \end{align} \]
and thus the result will hold. The technical details follow.

(i) Consider the $m \times 1$ vector $Y_{\delta,n}^{-1} C_n$. Using an obvious partitioning, for the first element we have
\[ Y_{\delta,n,1}^{-1} C_{n,1} = a_n^{-2} \sum X_{t-1} u_t = O_p(1) \tag{24} \]
by Lemma 2 part (ii).

A typical element of $Y_{\delta,n,22}^{-1} C_{n,2}$ is
\[ a_n^{-2} \left( \frac{n}{\ln(n)} \right)^{\delta/2} \sum Z_{t-j} u_t = a_n^{-2} \left( \frac{n}{\ln(n)} \right)^{\delta/2} \sum Z_{t-j} u_t \]
\[ = \left( \frac{\ln n}{n} \right)^{(1-\delta)/2} a_n^{-2} \left( n \ln n \right)^{-1/2} \sum Z_{t-j} u_t \]
\[ = O_p \left( \frac{\ln n}{n} \right)^{(1-\delta)/2} = \begin{cases} a_p(1), & \forall \delta \in (0,1), \ 
O_p(1), & \text{for } \delta = 1 \end{cases} \tag{25} \]
by Lemma 2 part (i). It follows from eqns (24) and (25) that
\[ Y_{\delta,n}^{-1}C_n = \begin{bmatrix} O_p(1) \\ O_p(1) \end{bmatrix}_{m \times 1}, \quad \forall \delta \in (0, 1) \]

\[ Y_{1,n}^{-1}C_n = \begin{bmatrix} O_p(1) \\ O_p(1) \end{bmatrix}_{m \times 1}. \]  

(ii) Consider the product \( Y_{\delta,n}^{-1}A_n D_n^{-1}A_n^{-1} Y_{1,n} \).

The matrix \( D_n = A_n^{-1} M_n A_n^{-1} \) is of the form

\[
D_n = \begin{bmatrix}
\sum \frac{X_{t-1}^2}{na_n^2} & \sum \frac{X_{t-1}Z_{t-1}}{n^{1/2}a_n} & \ldots & \sum \frac{X_{t-1}Z_{t-(m-1)}}{n^{1/2}a_n} \\
\sum \frac{X_{t-1}Z_{t-1}}{n^{1/2}a_n^2} & \sum \frac{Z_{t-1}^2}{a_n} & \ldots & \sum \frac{Z_{t-1}Z_{t-(m-1)}}{a_n} \\
\vdots & \vdots & \ddots & \vdots \\
\sum \frac{X_{t-1}Z_{t-(m-1)}}{n^{1/2}a_n^2} & \sum \frac{Z_{t-1}Z_{t-(m-1)}}{a_n} & \ldots & \sum \frac{Z_{t-(m-1)}^2}{a_n}
\end{bmatrix}.
\]

Using parts (iii), (iv) and (v) of Lemma 2 we establish the following. For the top left element,

\[ D_{n,11} = n^{-1}a_n^{-2} \sum X_{t-1}^2 \implies \sigma^2 \int_0^1 U_z^2 = O_p(1). \]

For the bottom right block, a typical element is

\[ D_{n,22}(i,j) = a_n^{-2} \sum Z_{t-i}Z_{t-j} \implies \left( \sum_{l=0}^{\infty} b_l b_{l+j-i} \right) \int_0^1 (dU_z)^2 = O_p(1) \]

for \( i, j = 1, \ldots, m-1 \). Thus, for the entire block we have, say,

\[ D_{n,22} \implies \int_0^1 (dU_z)^2 \times B, \]

in which the matrix, \( B \), is symmetric and positive definite.

Finally, for the off-diagonal block, a typical element is

\[ D_{n,12}(1,j) = n^{-1/2} \left( a_n^{-2} \sum X_{t-1}Z_{t-j} \right) = O_p(n^{-1/2}) = o_p(1) \]

for \( j = 1, \ldots, m-1 \).

In summary,

\[ D_n = \begin{bmatrix} O_p(1) & O_p(n^{-1/2}) \\ O_p(n^{-1/2}) & O_p(1) \end{bmatrix}_{m \times m}. \]

Partitioned inversion implies

\[ D_n^{-1} = \begin{bmatrix} O_p(1) & O_p(n^{-1/2}) \\ O_p(n^{-1/2}) & O_p(1) \end{bmatrix}_{m \times m}. \]
Then
\[
\Lambda_n^{-1}D_n^{-1}\Lambda_n^{-1} = \begin{bmatrix}
O_p(1) & O_p(1) \\
n^{-1}O_p(1) & O_p(1)
\end{bmatrix}_{m\times m}.
\]

Then
\[
Y_{\delta,n}^{-1}A_n^{-1}A_n^{-1}Y_{\delta,n} = \begin{bmatrix}
O_p(1) & O_p((\ln n/n)^{\delta/2}) \\
O_p(n^{-1}(n/\ln n)^{\delta/2}) & O_p(1)
\end{bmatrix}_{m\times m}
= \begin{bmatrix}
o_p(1) & o_p(1) \\
o_p(1) & o_p(1)
\end{bmatrix}_{m\times m}
\]  
(29)
\[
\forall \delta \in (0,1) \cap (0,2], x \in (0,2).
\]

It then follows from eqns (26) and (29) that
\[
\Lambda_n^*(\hat{\beta} - \beta) = \begin{bmatrix}
O_p(1) & O_p(1) \\
o_p(1) & o_p(1)
\end{bmatrix}_{m\times m} \begin{bmatrix}
o_p(1) \\
o_p(1)
\end{bmatrix}_{m\times 1} = \begin{bmatrix}
o_p(1) \\
o_p(1)
\end{bmatrix}_{m\times 1}.
\]
(30)

To complete the proof, we observe that \(\hat{\Phi} - \Phi\) is a linear combination of \(\hat{\beta} - \beta\) from which we deduce that
\[
\left(\frac{n}{\ln(n)}\right)^{\delta/2} = \begin{bmatrix}
\hat{\Phi}_1 - \Phi_1 \\
\hat{\Phi}_2 - \Phi_2 \\
\vdots \\
\hat{\Phi}_m - \Phi_m
\end{bmatrix} \left(\frac{n}{\ln(n)}\right)^{\delta/2} = \begin{bmatrix}
\hat{\beta} - \beta + \hat{\eta}_1 - \eta_1 \\
\hat{\eta}_2 - \eta_2 - \hat{\eta}_1 + \eta_1 \\
\vdots \\
\hat{\eta}_m - \eta_m - \hat{\eta}_{m-1} + \eta_{m-1}
\end{bmatrix}
\]
and by virtue of eqn (30) therefore,
\[
(n/\ln(n))^{\delta/2}(\hat{\Phi} - \Phi) = o_p(1)
\]
for all \(\delta \in (0,1) \cap (0,2], x \in (0,2)\).

Part (c). Put \(\delta = 1\) in eqn (21) and subsequently, so that in eqn (22), \(Y_{\delta,A}^{-1}C_n = O_p(1)\).

Observe that eqn (29) continues to hold provided \(x \in [\delta,2)\), so that eqn (30) is replaced by
\[
\Lambda_n^*(\hat{\beta} - \beta) = \begin{bmatrix}
O_p(1) & o_p(1) \\
o_p(1) & o_p(1)
\end{bmatrix}_{m\times m} \begin{bmatrix}
o_p(1) \\
o_p(1)
\end{bmatrix}_{m\times 1}
= \begin{bmatrix}
o_p(1) \\
o_p(1)
\end{bmatrix}_{m\times 1},
\]
(31)
which gives the result.

\textbf{Proof of Theorem 3.} 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 model (1). We let \(\Rightarrow\) denote convergence in probability or in distribution as is obvious from the context.

Part (a). Apply the well-known formula for the increase in the residual sum of squares from imposition of a linear restriction, to obtain
\[ \text{RSS}_{K-1} - \text{RSS}_K = \Phi_{K,K}^2 / (\mathbf{X}_K' \mathbf{X}_K)^{-1}. \]

Applying the usual partitioned inversion formula, we find the (KK)th element of 
\((\mathbf{X}_K' \mathbf{X}_K)^{-1}\) is given by
\[
(\mathbf{X}_K' \mathbf{X}_K)^{-1}
= \{ \mathbf{X}_K' \mathbf{X}_K - \mathbf{X}_K' \mathbf{X}_{K-1} (\mathbf{X}_K' \mathbf{X}_{K-1})^{-1} \mathbf{X}_K' \mathbf{X}_K \}^{-1}
\]
hence the increase in the RSS is, exactly,
\[
\text{RSS}_{K-1} - \text{RSS}_K
= \Phi_{K,K}^2 \{ \mathbf{X}_K' \mathbf{X}_K - \mathbf{X}_K' \mathbf{X}_{K-1} (\mathbf{X}_K' \mathbf{X}_{K-1})^{-1} \mathbf{X}_K' \mathbf{X}_K \}
= \Phi_{K,K}^2 \text{RSS}_{K-1}, \quad \text{say. (32)}
\]

A rearrangement then yields
\[
\text{RSS}_K = \text{RSS}_{K-1} (1 - \Phi_{K,K}^2 \delta_K)
\]
with
\[
\delta_K = \frac{\text{RSS}_{K-1}^{1/2}}{\text{RSS}_{K-1}}.
\]

Part (b). 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:
\[
\sum_{j=0}^{\infty} c_j^2 < \infty. \quad \text{(33)}
\]

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
\[
\mathbf{R}_{K-1} = \begin{bmatrix}
0 & 0 & \cdots & 0 & 1 \\
0 & 0 & \ddots & 1 & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 1 & \ddots & 0 & 0 \\
1 & 0 & \cdots & 0 & 0
\end{bmatrix}
\]
and noting that
\[
\mathbf{R}_{K-1} = \mathbf{R}_{K-1}^{-1}
\]
\[
\mathbf{P}_{K-1} = \mathbf{R}_{K-1}^{-1} \mathbf{R}_{K-1} \mathbf{R}_{K-1}
\]
\[
\mathbf{P}_{K-1}^{-1} = \mathbf{R}_{K-1} \mathbf{R}_{K-1}^{-1} \mathbf{R}_{K-1}
\]
observe that because the averaged cross-products matrix converges to \(\sigma^2 \mathbf{P}_{K-1}\) we must have
\[
\begin{align*}
\begin{aligned}
    n^{-1} X_0' X_{K-1} &\Rightarrow \sigma^2 [\rho_1, \ldots, \rho_{K-1}] \\
    n^{-1} X_K' X_{K-1} &\Rightarrow \sigma^2 [\rho_{K-1}, \ldots, \rho_1]
\end{aligned}
\end{align*}
\] 

so that 

\[
\begin{align*}
    n^{-1} \text{RSS}^1_{K-1} &= n^{-1} \{ X_K' X_K - X_K' X_{K-1} (X_{K-1}' X_{K-1})^{-1} X_{K-1}' X_K \} \\
    &\Rightarrow \sigma^2 \{ \rho_0 - [\rho_{K-1}, \ldots, \rho_1] [P_{K-1}^{-1} [\rho_{K-1}, \ldots, \rho_1] \}' \},
\end{align*}
\]

while similarly 

\[
\begin{align*}
    n^{-1} \text{RSS}^2_{K-1} &= n^{-1} \{ X_0' X_0 - X_0' X_{K-1} (X_{K-1}' X_{K-1})^{-1} X_{K-1}' X_0 \} \\
    &\Rightarrow \sigma^2 \{ \rho_0 - [\rho_1, \ldots, \rho_{K-1}] [P_{K-1}^{-1} [\rho_1, \ldots, \rho_{K-1}] \}' \},
\end{align*}
\]

so that 

\[
\begin{align*}
    (\text{RSS}^1_{K-1} - \text{RSS}^2_{K-1}) / n &= o_p(1).
\end{align*}
\]

Thus, 

\[
\begin{align*}
    (\delta_K - 1) &= \frac{(\text{RSS}^1_{K-1} - \text{RSS}^2_{K-1}) / n}{\text{RSS}^2_{K-1} / n} \\
    &= \frac{o_p(1)}{O_p(1)} = o_p(1).
\end{align*}
\]

Now suppose that \( z < 2 \). If we strengthen eqn (33), following Phillips (1990, condition 25), to 

\[
\sum_{j=0}^{\infty} j |c_j|^\tau < \infty \quad \text{for some} \quad 0 < \tau < 1 \land z,
\]

(which is satisfied by the MA representation of a stationary AR process), then we obtain from Davis and Resnick (1985, Thm 4.2), or Lemma 2 (iii): 

\[
\begin{align*}
    a_n^{-2} X_0' [X_0, X_1, X_2, \ldots, X_K] &\Rightarrow v^2 [\rho_0, \ldots, \rho_K] \\
    a_n^{-2} X_K' [X_0, X_1, X_2, \ldots, X_K] &\Rightarrow v^2 [\rho_K, \ldots, \rho_0],
\end{align*}
\]

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

\[
\begin{align*}
\frac{X_{K-1}' X_{K-1}}{a_n^2} &\Rightarrow v^2 P_{K-1}.
\end{align*}
\]

Putting these results together, we obtain 

\[
\begin{align*}
    \frac{\text{RSS}_{K-1}}{a_n^2} &\Rightarrow v^2 \{ \rho_0 - [\rho_1, \ldots, \rho_{K-1}] [P_{K-1}^{-1} [\rho_1, \ldots, \rho_{K-1}] \}' \}
\end{align*}
\]

and 

\[
\begin{align*}
    \frac{\text{RSS}^1_{K-1}}{a_n^2} &\Rightarrow v^2 \{ \rho_0 - [\rho_{K-1}, \ldots, \rho_1] [P_{K-1}^{-1} [\rho_{K-1}, \ldots, \rho_1] \]' \}
\end{align*}
\]

© 2008 The Authors
Journal compilation © 2008 Blackwell Publishing Ltd.
JOURNAL OF TIME SERIES ANALYSIS Vol. 29, No. 4
in which the second factors on the RHS are identical because $P_{K-1}$ is a Toeplitz matrix, as previously noted. Thus, with the change of normalization,

$$a_n^{-2} (\text{RSS}_{K-1}^l - \text{RSS}_{K-1}) = o_p(1).$$

We see therefore, as before, that

$$(\delta_K - 1) = \frac{(\text{RSS}_{K-1}^l - \text{RSS}_{K-1})/a_n^2}{\text{RSS}_{K-1}/a_n}$$

$$= o_p(1).$$ (36)

Part (c). To deal with the unit-root nonstationary case, we apply the usual linear transformation eqn (18), 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 (b).

So, now define the vectors,

$$Z_0 = [\Delta X_1, \ldots, \Delta X_n]'$$
$$X_1 = [X_0, \ldots, X_{n-1}]'$$
$$Z_1 = [\Delta X_0, \ldots, \Delta X_{n-1}]'$$
$$\vdots$$
$$Z_{K-1} = [\Delta X_{(K-2)}, \ldots, \Delta X_{n-K+1}]'$$
$$\tilde{X}_{K-1} = [X_1, Z_1, \ldots, Z_{K-2}].$$

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

$$Z_0 = \tilde{X}_{K-1} \beta + e$$ (37)

and

$$X_0 = \lambda_{K-1} \pi + e$$

are identical. However, eqn (37) has the advantages that the dependent variable is stationary and the regressor cross-products matrix is asymptotically nonsingular. We now introduce a norming matrix, as in eqn (20)

$$A_n = \begin{bmatrix} n^{1/2} a_n & 0' \\ 0 & a_n I \end{bmatrix}$$

and write

$$M_n = \tilde{X}_{K-2}^l \tilde{X}_{K-2}.$$

It follows from Lemma 2 and Davis and Resnick (1985, Thm 4.2) that

$$A_n^{-1} M_n A_n^{-1} = \begin{bmatrix} o_p(1) & o_p(1) \\ o_p(1) & w^2 P_{z,K-2} \end{bmatrix}$$
and from Davis and Resnick (1985, Thm 4.2)

\[
\frac{\mathbf{Z}_0' \mathbf{Z}_0}{a_n} \implies w^2 \rho_{z,0}, \text{ say},
\]

in which the stable random variable \(w^2\) and constants, \(\rho_{z,j}\) are defined with respect to the shocks driving \(Z_t\).

Thus, we have, noting the slight change in notation,

\[
a_n^{-2} \text{RSS}_{K-2} = \frac{\mathbf{Z}_0' \mathbf{Z}_0}{a_n} - \frac{\mathbf{Z}_0' \mathbf{X}_{K-2} A_n^{-1}}{a_n} \left( \Lambda_n^{-1} \mathbf{M}_n \Lambda_n^{-1} \right)^{-1} \Lambda_n^{-1} \mathbf{X}_{K-2}' \mathbf{Z}_0 \implies w^2 \left( \rho_{z,0} - \rho_{K-2}^{-1} P_{z,K-2} \rho_{K-2} \right)
\]

and similarly

\[
a_n^{-2} \text{RSS}_{K-2} = \frac{\mathbf{Z}_0' \mathbf{Z}_0}{a_n} - \frac{\mathbf{Z}_0' \mathbf{X}_{K-2} A_n^{-1}}{a_n} \left( \Lambda_n^{-1} \mathbf{M}_n \Lambda_n^{-1} \right)^{-1} \Lambda_n^{-1} \mathbf{X}_{K-2}' \mathbf{Z}_{K-1} \implies w^2 \left( \rho_{z,0} - \rho_{K-2}^{-1} P_{z,K-2} \rho_{K-2} \right),
\]

and once again we find that

\[
a_n^{-2} (\text{RSS}_{K-2} - \text{RSS}_{K-2}) \implies 0.
\]

It now follows that

\[
(\hat{\delta}_K - 1) = o_p(1)
\]
as in the stationary case.

\[\Box\]

**Proof of Theorem 4.** We have to show that if \(X_t\) is a nonstationary AR\((m)\), \(m \geq 1\), defined by eqn (1), with \(\rho = 1, 0 < \alpha < 2\), the upper bound, \(K > m\), and if \(\hat{m}\) minimizes the criterion, \(\text{IC}_{\text{OLS}}(k)\), defined in eqn (9) and

(a) \(\Phi_{k,p}^2/C(n) = o_p(1)\) for \(k, p > m\), and
(b) \(C(n) = o(1)\)

then \(\hat{m} \to_p m\).

Consider first the probability that \(\hat{m} < m\). By the argument following eqn (10) we have to show that \(\hat{\delta}_k - 1 = o_p(1)\) for each \(k < m\) and \(C(n) = o(1)\); the former is proved in Theorem 3, and the latter is ensured by condition (b). For the probability that \(\hat{m} > m\), note that from the discussion following eqn (12), given the result of Theorem 3, we need only show that \(\Phi_{k,p}^2/C(n) = o_p(1)\) for \(k, p > m\), which is condition (a). \[\Box\]
ACKNOWLEDGEMENTS

Colleagues at City University and the University of York commented on earlier versions; two anonymous referees made detailed suggestions which have greatly improved the paper. The Leverhulme Trust, Award Reference F/00353/D, funded D. Hristova’s work. We thank all these persons and organizations for their contributions and support. Any mistakes or omissions are our sole responsibility.

NOTES

1. We are grateful to a referee who suggested a shortening of our original formulation of this result.
2. All calculations were performed using Gauss 5.0 using the Kiss + Monster random number generator and an implementation of the algorithm of Chambers et al. (1976), coded in GAUSS by J. Huston McCulloch.

Corresponding author: Peter Burridge, Department of Economics and Related Studies, University of York, Heslington, York YO10 5DD, UK. E-mail: pb539@york.ac.uk

REFERENCES