**Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag**

Serena Ng, Pierre Perron

*Journal of the American Statistical Association*, Volume 90, Issue 429 (Mar., 1995), 268-281.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199503%2990%3A429%3C268%3AURTIAM%3E2.0.CO%3B2-M

# Unit Root Tests in ARMA Models With Data-Dependent Methods for the Selection of the Truncation Lag

Serena NG and Pierre PERRON*

We analyze the choice of the truncation lag in the context of the Said–Dickey test for the presence of a unit root in a general autoregressive moving average model. It is shown that a deterministic relationship between the truncation lag and the sample size is dominated by data-dependent rules that take sample information into account. In particular, we study data-dependent rules that are not constrained to satisfy the lower bound condition imposed by Said–Dickey. Akaike's information criterion falls into this category. The analytical properties of the truncation lag selected according to a class of information criteria are compared to those based on sequential testing for the significance of coefficients on additional lags. The asymptotic properties of the unit root test under various methods for selecting the truncation lag are analyzed, and simulations are used to show their distinctive behavior in finite samples. Our results favor methods based on sequential tests over those based on information criteria, because the former show less size distortions and have comparable power.

KEY WORDS: Dickey–Fuller $t$ test; General to specific; Information criteria; Order selection; Said–Dickey test.

## 1. INTRODUCTION

Testing for the presence of a unit root in a time series of data has become a common starting point of applied work in macroeconomics. Except in very special cases, one often assumes that the series to be tested is driven by serially correlated innovations and tests for the presence of a unit root using statistics that take serial dependence into account. One such statistic that has become very popular is the augmented Dickey–Fuller $t$ test due to Dickey and Fuller (1979) and Said and Dickey (1984). Their test, hereafter referred to as $t_\rho$, is based on estimates from an augmented autoregression. The test is valid for stationary and invertible autoregressive moving-average (ARMA) noise functions of unknown order provided that the truncation lag, $k$, is chosen in relationship to the sample size, $T$, to satisfy lower and upper bound conditions.

An issue that arises with the implementation of $t_\rho$ is the choice of $k$. Work of Schwert (1989), Agiakloglou and Newbold (1991), and Harris (1992) have found the order of the autoregression to have important size and power implications. This article provides a formal analysis of the relevance of $k$ in the test procedure. One of our objectives is to show, via simulations, that a deterministic rule that relates $k$ to $T$ is inferior to a data-dependent rule that takes sample information into account. Another objective is to clarify the role of the lower bound and the upper bound on $k$ in the limiting behavior of the statistic $t_\rho$. We study the asymptotic properties of $t_\rho$ and of the estimates from the augmented autoregression with $k$ chosen using different data-dependent rules. Among these are information-based model selection rules, such as the Akaike information criterion (AIC) and the Schwartz criterion, and sequential testing for the significance of the coefficients on lags, such as $F$ or $t$ tests. We show that with parameter values for which size problems surface, information-based rules tend to select values of $k$ that are consistently smaller than those chosen through sequential testing for the significance of coefficients on additional lags, and the size distortions associated with the former method are correspondingly larger. Thus the choice of the data-dependent rule has bearing on the size and power of the test. These issues are of particular relevance in finite samples.

The article is structured as follows. Section 2 puts forth the Said and Dickey framework, the role of the upper and lower bound conditions on $k$, and the implications for $t_\rho$ with and without the lower bound. Section 3 provides a discussion of procedures typically used to select $k$. Formal definitions of "deterministic" and "adaptive" rules are given. Sections 4 and 5 analyze the properties of $t_\rho$ with $k$ chosen according to information criteria and sequential testing for the significance of coefficients on lags. Section 6 presents implications of these results. We conclude with suggestions for procedures to select $k$ and directions for future research. Proofs of theorems are given in the Appendix.

## 2. THE SAID–DICKEY APPROACH

### 2.1 The Test Statistic

Suppose the data-generating process (DGP) for $\{y_t\}$ is given by

$$y_t = \rho y_{t-1} + u_t, \tag{1}$$

$$u_t = \sum_{i=1}^{p} \alpha_i u_{t-i} + e_t + \sum_{j=1}^{q} \theta_j e_{t-j},$$

where $e_t \sim$ iid $(0, \sigma_e^2)$ with bounded fourth moment. Assuming that $\{u_t\}$ is stationary and invertible with autoregressive and moving-average polynomials that do not share common roots, $\{y_t\}$ evolves according to

$$\Delta y_t = (\rho - 1) y_{t-1} + \sum_{i=1}^{\infty} d_i u_{t-i} + e_t, \tag{2}$$

where the coefficients $d_i$ $(i = 1, \ldots, \infty)$ are functions of the parameters $\{\alpha_i, \theta_j;\ i = 1, \ldots, p, j = 1, \ldots, q\}$. The true

* Serena Ng is Assistant Professor and Pierre Perron is Professor, Département de Sciences Économiques et C.R.D.E., Université de Montréal, Québec H3C 3J7, Canada. The authors acknowledge grants from the Social Science and Humanities Research Council of Canada. The second author also thanks the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche du Québec and the National Science Foundation for financial support.

order of the autoregression is infinity when $q > 0$. The null hypothesis of interest is $\rho = 1$, in which case a unit root is said to exist and the DGP is an ARIMA$(p, 1, q)$. Because $\Delta y_t = u_t$ under the null hypothesis, (2) can also be seen as an autoregression in $\Delta y_t$ augmented by $y_{t-1}$, namely

$$\Delta y_t = (\rho - 1)y_{t-1} + \sum_{i=1}^{\infty} d_i \Delta y_{t-i} + e_t. \tag{3}$$

When the orders $p$ and $q$ are unknown, as is often the case in practice, Said and Dickey (1984) suggested approximating the infinite autoregression by a truncated version whose order is a function of the number of observations, $T$:

$$\Delta y_t = d_0 y_{t-1} + \sum_{i=1}^{k} d_i \Delta y_{t-i} + e_{tk}, \tag{4}$$

where $d_0 = \rho - 1$, and for future reference, we denote $\mathbf{d}(k) = (d_1, \ldots, d_k)$. The ordinary least squares (OLS) estimates are similarly defined as $\hat{d}_0 = \hat{\rho} - 1$ and $\hat{\mathbf{d}}(k) = (\hat{d}_1, \ldots, \hat{d}_k)$. The order of truncation, $k$, is assumed to satisfy some conditions to ensure consistency of the least squares estimates. More precisely, Said and Dickey (1984) assumed the following:

A1. $k$ is chosen as a function of $T$ such that

$$k^3/T \to 0 \quad \text{and} \quad k \to \infty \quad \text{as} \quad T \to \infty.$$

A2. There exist $c > 0$ and

$$r > 0 \quad \text{such that} \quad ck > T^{1/r}.$$

Assumption A1 is based on the work of Berk (1974) who showed consistency of the parameter estimates in an autoregression of the form (4) but without the level regressor, $y_{t-1}$, and when the process is stationary. The assumption is imposed to ensure that the number of regressors does not increase so fast as to induce excess variability in the estimators. Assumption A2, often an overlooked condition, is a lower-bound condition that restricts $k$ to be at least a polynomial rate in $T$. It rules out values of $k$ that are proportional to log $T$. Intuitively, A2 prohibits $k$ from being so small as to provide an inadequate approximation to the true model. It is more restrictive than the following assumption:

A2′. $k$ satisfies $k^{1/2} \sum_{i=k+1}^{\infty} |d_i| \to 0$ and

$$k \to \infty \quad \text{as} \quad T \to \infty.$$

Assumption A2′ was used by Berk (1974), and in related work by Lewis and Reinsel (1985), to show consistency of the OLS estimates in an autoregression applied to a stationary process. Note that A2′ is satisfied for any $\{u_t\}$ that is a stationary and invertible ARMA process as long as $k \to \infty$ as $T \to \infty$, irrespective of the rate at which $k$ increases. Of particular importance is the fact that unlike A2, A2′ allows $k$ to grow at a logarithmic rate. Berk (1974) and Lewis and Reinsel (1985) strengthened Assumption A2′ to the following:

A2″. $k$ satisfies $T^{1/2} \sum_{i=k+1}^{\infty} |d_i| \to 0$ as

$$k \to \infty \quad \text{and} \quad T \to \infty,$$

to ensure $\sqrt{T}$ consistency of $\hat{\mathbf{d}}(k)$. Note that A2″ implicitly rules out $k$ growing at a log$(T)$ rate and is basically equivalent to A2. Consistency of $\hat{\mathbf{d}}(k)$ may be achieved at a rate slower than $\sqrt{T}$ if A2′ is satisfied but not A2″.

The foregoing discussion applies when the DGP is an infinite autoregression, as would be the case if moving-average components were present. When dealing with a finite autoregression, A2″ is automatically satisfied. In fact, $k$ need not grow to infinity as long as it is selected to be larger than the true order. Hence most of the results that follow also apply to the case of a finite autoregression. (For a more specific treatment of this case, see Hall 1994.)

Said and Dickey's result states that when $k$ satisfies A1 and A2, the least squares estimates $\hat{\mathbf{d}}(k)$ are $\sqrt{T}$-consistent, and the coefficient on $y_{t-1}$ provides a basis for testing the unit root hypothesis. The limiting distribution for the $t$ statistic on $\hat{d}_0 = (\hat{\rho} - 1)$ for testing $\rho = 1$ is such that

$$t_\rho \Rightarrow \left( \int_0^1 W(r)\, dW(r) \right) \left( \int_0^1 W(r)^2\, dr \right)^{-1/2}, \tag{5}$$

where $W(r)$ is a standard Brownian motion in the space $C[0,1]$. Percentiles of this distribution were given by Fuller (1976). The result stated in (5) extends naturally to the inclusion of deterministic components in (4). In that case the Wiener process is replaced by its detrended counterpart.

## 2.2 A Useful Result

Of interest are the properties of the test statistic when $k$ is chosen as a function of $T$ to satisfy A1 but not necessarily A2, because such procedures are commonly used in applied work. The following lemma considers the validity of Said and Dickey's (1984) result when the lower-bound condition A2 is not imposed.

*Lemma 2.1.* Let $\{y_t\}$ be given by (1). Let $t_\rho$ be obtained from the truncated autoregression (4) with $k$ chosen such that A1 is satisfied. Then (a) the asymptotic distribution of $t_\rho$ continues to be given by (5) without A2, and (b) $\hat{\mathbf{d}}(k) = (\hat{d}_1, \ldots, \hat{d}_k)$ is not in general $\sqrt{T}$-consistent for $d(k) = (d_1, \ldots, d_k)$ if A2 or A2″ does not hold. In that case there exists a $\lambda$, with $|d_j| \le C_1 \lambda^j$ for some constant $C_1$ and $0 < \lambda < 1$, such that $\lambda^{-k}(\hat{d}_i - d_i) = O_p(1)$, $(i = 1, \ldots, k)$.

Lemma 2.1 states that although $\sqrt{T}$ consistency of the coefficients on $\Delta y_{t-i}$ is not assured without A2, $\hat{d}_0$ is still consistent for $d_0$ at rate $T$, and $t_\rho$ attains the same limiting distribution as defined in (5) with Assumption A1 alone. The proof of consistency of $\hat{d}_0$ and $\hat{\mathbf{d}}(k)$ under A1 and A2 was given by Said and Dickey (1984). The lower-bound condition enters the analysis only when considering the properties of coefficients pertaining to $\Delta y_{t-i}$. Specifically, $\sqrt{T}$ consistency of $\hat{\mathbf{d}}(k)$ requires, from lemma 2 of Berk (1974), that

$$E\left( (T-k)^{-1} \sum_{j=1}^{k} \left\{ \sum_{t=k+1}^{T} u_{t-j}(e_{tk} - e_t) \right\}^2 \right)$$

$$\le k(T-k) \sum_{i=k+1}^{\infty} d_i^2 \to 0. \tag{6}$$

Because $(e_{tk} - e_t)$ is the error in approximating an infinite autoregression by a truncated autoregression, it is larger the

Table 1. Size and Power of Unit Root Tests, Moving-Average Case, $T = 100$ (5,000 Replications)

| $\rho$ | $\theta$ | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ | $k = 6$ | $k = 7$ | $k = 8$ | $k = 9$ | $k = 10$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.0 | .8 | .123 | .031 | .075 | .037 | .062 | .041 | .062 | .041 | .051 | .041 |
| 1.0 | .5 | .103 | .047 | .065 | .052 | .057 | .054 | .053 | .053 | .051 | .048 |
| 1.0 | .3 | .073 | .051 | .055 | .056 | .053 | .047 | .050 | .048 | .045 | .046 |
| 1.0 | .0 | .049 | .048 | .046 | .044 | .046 | .044 | .045 | .044 | .044 | .041 |
| 1.0 | −.3 | .091 | .062 | .057 | .052 | .052 | .048 | .049 | .048 | .048 | .045 |
| 1.0 | −.5 | .214 | .099 | .068 | .055 | .051 | .051 | .052 | .050 | .049 | .048 |
| 1.0 | −.8 | .880 | .640 | .434 | .283 | .200 | .132 | .110 | .082 | .074 | .060 |
| .95 | .8 | .307 | .044 | .176 | .059 | .129 | .064 | .106 | .067 | .085 | .064 |
| .95 | .5 | .237 | .074 | .130 | .089 | .103 | .086 | .086 | .079 | .077 | .072 |
| .95 | .3 | .162 | .088 | .101 | .092 | .092 | .085 | .087 | .081 | .075 | .069 |
| .95 | .0 | .117 | .111 | .108 | .102 | .099 | .093 | .083 | .082 | .080 | .068 |
| .95 | −.3 | .219 | .139 | .122 | .108 | .100 | .091 | .094 | .089 | .087 | .081 |
| .95 | −.5 | .477 | .246 | .157 | .115 | .104 | .087 | .085 | .081 | .078 | .069 |
| .95 | −.8 | .997 | .941 | .782 | .584 | .444 | .329 | .255 | .203 | .164 | .130 |
| .85 | .8 | .788 | .201 | .524 | .212 | .379 | .203 | .278 | .166 | .207 | .149 |
| .85 | .5 | .708 | .316 | .425 | .297 | .308 | .252 | .239 | .203 | .191 | .169 |
| .85 | .3 | .598 | .393 | .395 | .334 | .306 | .267 | .241 | .214 | .196 | .171 |
| .85 | .0 | .510 | .436 | .399 | .343 | .316 | .271 | .251 | .218 | .193 | .166 |
| .85 | −.3 | .746 | .540 | .452 | .376 | .344 | .287 | .266 | .233 | .209 | .176 |
| .85 | −.5 | .961 | .779 | .614 | .482 | .423 | .353 | .315 | .275 | .242 | .208 |
| .85 | −.8 | 1.000 | 1.000 | .996 | .956 | .886 | .759 | .671 | .565 | .483 | .399 |

NOTE: DGP: $y_t = \rho y_{t-1} + u_t$, $u_t = e_t + \theta e_{t-1}$. Regression: $\Delta y_t = \delta_0 y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta y_{t-i} + v_t$.

smaller is $k$; the role of the lower bound is thus intuitive in this context. Sufficient conditions for (6) to hold are provided by either A2, as invoked by Said and Dickey (1984), or A2″, as used by Berk (1974) and by Lewis and Reinsel (1985). Recall, however, that $k$ growing at a logarithmic rate is ruled out by either A2 or A2″.

To see the ramifications of this condition, suppose that $\{u_t\}$ is an MA(1) with coefficient $\theta$, and hence $d_i = -(-\theta)^i$. The condition (6) is equivalent to requiring that $\log(k) + \log(T - k) + k \log \theta^2$ diverges to $-\infty$. Now take $k = b \log(T)$ for some constant $b > 0$. Clearly, $k^3/T \to 0$ and A1 is satisfied, but the condition for $\sqrt{T}$ consistency is (approximately) $1 + b \log(\theta^2) < 0$. This condition fails when $|\theta| > \exp(-1/2b)$. Hence for any fixed rule satisfying $k = b \log(T)$, there will exist a range of values of $\theta$ such that (6) does not hold. In that case $\hat{\mathbf{d}}(k)$ will achieve consistency not at rate $\sqrt{T}$ but rather at the slower rate of $T^{(1-a)/2}$, with $a = 1 + b \log(\theta^2)$ in the case of an MA(1) (or, equivalently, $|\theta|^{-k}(\hat{d}_i - d_i) = O_p(1)$ as stated in lemma 2.1). As we will see in subsequent sections, this logarithmic rate is of special interest.

The result that the estimates for the coefficients on $\Delta y_{t-i}$ might achieve consistency at a rate slower than $\sqrt{T}$ extends to the case when $\{u_t\}$ satisfies a general ARMA($p$, $q$) model, using the fact that the coefficients $d_i$ are such that $|d_i| < C_1 \lambda^i$, $0 < \lambda < 1$ for some constant $C_1$ (see, for example, Fuller 1976). The important point is that $(\hat{\rho} - 1)$ will continue to be order $T$-consistent even without the lower-bound condition. The asymptotic equivalence of $t_\rho$ with and without A2 follows from this result and the result that consistency of the least squares estimates is enough to ensure the consistency of $\hat{\sigma}_k^2$ for $\sigma_e^2$.

Although all estimates from the regression (4) will be consistent whether or not A2 holds, the lower-bound condition on $k$ is important. The coefficients on the stationary regressors will converge at a rate slower than $\sqrt{T}$ when the lower-bound condition is not satisfied. Therefore, choices of $k$ that satisfy A2 will yield coefficient estimates on the stationary differences that achieve consistency at a faster rate and can be expected to lead to unit root tests having better finite-sample properties than those choices that do not satisfy A2.

## 3. SELECTION OF $k$

This section consists of three parts. First, in Section 3.1 we use simulations to show that any a priori rule that presets the value of $k$ is likely to result in size distortions and/or power loss, unless that value of $k$ happens to be chosen appropriately. This is so, even if $k$ is chosen to be a fixed function of $T$. In Section 3.2 we discuss the specifics of two data-dependent rules whereby the relationship between $k$ and $T$ depends on the given sample of data. In Section 3.3 we further restrict our analysis to data-dependent rules that satisfy A1 only and analyze the limiting distribution of $t_\rho$ when such data-dependent rules are used.

### 3.1 Rules of Thumb

*3.1.1 Fixing $k$.* Although the asymptotic distribution of $t_\rho$ is derived under the assumption that $k$ increases at an appropriate rate with $T$, the theoretical conditions A1 and A2 provide little practical guidance for choosing $k$. The common practice is to fix $k$ at a value independent of $T$. Using (1) as the DGP, we considered numerous parameterizations of $\alpha_i$ and $\theta_j$, with $k$ fixed to be 1 through 10. As the results reported in Table 1 (moving-average case) indicate, the properties of the statistic can be quite different, depending on the chosen value of $k$. For example, when $\theta = -.8$, fixing $k$ to be 4 yields an exact size of 28% instead of the 5% nominal size, noting that the exact size worsens to .939 when $\theta$ is $-.95$. But size distortions are much smaller as $k$ becomes larger. Although size distortions are much smaller when $\theta$ is positive, $t_\rho$ is oversized when $k$ is odd but undersized when $k$ is even.

Although in autoregressive models (see Table 2) the exact size of the test for all choices of $k$ is generally close to the nominal size (provided that $k$ is larger than the true order), the choice of $k$ has implications for power. As is evident from Table 2, an overparameterized model is associated with lower power. Thus, although a liberal choice of $k$ will reduce size distortions in moving-average models, it will generally yield lower power.

We also performed similar simulations for $T = 200$ and $T = 500$. As expected, power increases for every value of $k$ in both the moving-average and the autoregressive cases. With respect to the size of the test, the results for the autoregressive case are qualitatively the same as when $T = 100$. For positive moving-average models, the zig-zag pattern of size distortions as $k$ alternates between odd and even persists even when $T$ is 500. But for negative moving-average models, size distortions increase with $T$ for a given value of $k$. For example, with $\theta = -.8$ and $k = 3$, the exact size increases from .455 to .598 as $T$ increases from 100 to 500.

*3.1.2 Choosing k as a fixed function of T.* Any rule that defines $k$ as a deterministic function of $T$ fits into this category. A rule often used in unit root tests is due to Schwert (1989). For given constants $c$ and $d$, the truncation lag, $k$, is chosen according to

$$k = \text{int}\{c(T/100)^{1/d}\}.$$

Values of $c = 4$ and 12 and $d = 4$ were used in Schwert's extensive Monte Carlo analysis. He found that the size of the test is significantly better with $c = 12$ the closer the moving-average coefficient, $\theta$, is to $-1$. Problems encountered in fixing $k$ arbitrarily will also arise if $k$ is chosen as a deterministic function of $T$, because one is faced with a given sample size in practice. In general, there is no way to assure that arbitrarily chosen values of $c$ and $d$ are adequate for a given data series unless $c$ and $d$ happen to be chosen correctly.

The simulations highlight the fact that conditions on $k$ appropriate for asymptotic inference are not necessarily good practical guidelines for selecting $k$. Indeed, the value of $k$ that ensures an exact size close to the nominal size and also produces high power is highly dependent on the actual DGP; that is, the values of the AR and MA parameters. Rules of thumb ignore such sample information and are the main reason why fixing $k$ is to be avoided as a matter of practice.

## 3.2 Data-Dependent Rules

*3.2.1 Information-Based Rules.* The order of an autoregressive process is often chosen by minimizing an objective function that trades off parsimony against reductions in the sum of squared residuals. Following Hannan and Deistler (1988), we consider an objective function of the general form

$$I_k = \log \hat{\sigma}_k^2 + kC_T/T, \tag{7}$$

where $C_T$ is a sequence that satisfies $C_T > 0$, $C_T/T \to 0$. The familiar Akaike information criterion (AIC) (Akaike 1974) is obtained as a special case with $C_T = 2$. Another popular criterion is that of Schwartz (1978) with $C_T = \log T$. Other criteria, such as the Bayesian information criterion (BIC), can be shown to fall within the class of $I_k$. For econometric applications, the AIC and the Schwartz criteria are more common and will be considered in subsequent sections.

*3.2.2 Sequential Tests for the Significance of the Coefficients on Lags.* The premise of a sequential test is a general-to-specific modeling strategy that chooses between a model with $m$ lags and a model with $r = m + n$ lags. Let $\hat{\mathbf{d}}(m, r)$ denote the vector of coefficients $(\hat{d}_{m+1}, \ldots, \hat{d}_r)$ obtained by applying OLS to (4), with $\hat{\sigma}_r^2 = (T - r)^{-1} \sum_{t=r+1}^{T} \hat{e}_{tr}^2$ and

$$\mathbf{M}_r = \sum_{t=r+1}^{T} (y_{t-1}, \Delta y_{t-1}, \ldots, \Delta y_{t-r})' \times (y_{t-1}, \Delta y_{t-1}, \ldots, \Delta y_{t-r}).$$

Let $\mathbf{M}_r^{-1}(n)$ be the lower-right $(n \times n)$ block of $\mathbf{M}_r^{-1}$. We define the Wald test for the null hypothesis that the coefficients on the last $n$ lags are jointly equal to 0 as

$$J(m, r) = \hat{\mathbf{d}}(m, r)'(\mathbf{M}_r^{-1}(n))^{-1}\hat{\mathbf{d}}(m, r)/\hat{\sigma}_r^2. \tag{8}$$

We now provide a formal definition of the procedure for choosing $\hat{k}$ from a set of possible values $\{0, 1, \ldots, k \max\}$, where $k$ max is selected a priori.

*Definition 3.1.* The general-to-specific modeling strategy chooses $\hat{k}$ to be either i) $m + 1$ if, at significance level $\alpha$, $J(m, r)$ is the first statistic in the sequence $J(i, i + n)$, $\{i = k \max - 1, \ldots, 1\}$, which is significantly different from

*Table 2. Size and Power of Unit Root Tests, Autoregressive Case, T = 100 (5,000 Replications)*

| $\rho$ | $\phi_1$ | $\phi_2$ | $\phi_3$ | $\phi_4$ | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ | $k = 6$ | $k = 7$ | $k = 8$ | $k = 9$ | $k = 10$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.0 | .6 | .0 | .0 | .0 | .058 | .055 | .057 | .053 | .053 | .053 | .058 | .055 | .055 | .052 |
| 1.0 | −.6 | .0 | .0 | .0 | .054 | .054 | .054 | .052 | .052 | .047 | .045 | .043 | .044 | .043 |
| 1.0 | .4 | .2 | .0 | .0 | .033 | .052 | .052 | .051 | .050 | .052 | .050 | .048 | .047 | .046 |
| 1.0 | .3 | .3 | .25 | .14 | .062 | .034 | .038 | .051 | .054 | .052 | .050 | .047 | .048 | .052 |
| .95 | .6 | .0 | .0 | .0 | .391 | .346 | .320 | .281 | .260 | .225 | .210 | .191 | .178 | .154 |
| .95 | −.6 | .0 | .0 | .0 | .078 | .075 | .073 | .068 | .070 | .061 | .061 | .058 | .056 | .055 |
| .95 | .4 | .2 | .0 | .0 | .125 | .354 | .328 | .288 | .275 | .237 | .225 | .204 | .194 | .165 |
| .95 | .3 | .3 | .25 | .14 | .137 | .650 | .865 | .903 | .876 | .814 | .763 | .675 | .608 | .522 |
| .85 | .6 | .0 | .0 | .0 | .976 | .938 | .883 | .805 | .720 | .626 | .560 | .479 | .435 | .357 |
| .85 | −.6 | .0 | .0 | .0 | .252 | .224 | .207 | .188 | .171 | .159 | .154 | .134 | .129 | .122 |
| .85 | .4 | .2 | .0 | .0 | .818 | .933 | .889 | .799 | .718 | .626 | .560 | .479 | .435 | .357 |
| .85 | .3 | .3 | .25 | .14 | .688 | .964 | .991 | .993 | .984 | .960 | .908 | .827 | .748 | .647 |

NOTE: DGP: $y_t = \rho y_{t-1} + \sum_{i=1}^{4} \phi_i \Delta y_{t-i} + u_t$. Regression: $\Delta y_t = \delta_0 y_{t-1} + \sum_{i=1}^{4} \delta_i \Delta y_{t-i} + v_t$.

zero, or ii) 0 if $J(i, i+n)$ is not significantly different from zero for all $i = k\max - 1, k\max - 2, \ldots, 1$.

The idea is to start with the most general model with $k\max + n - 1$ lags and test whether the coefficients of the last lags are significant. If they are, then $\hat{k} = k\max$; otherwise, the next step is to estimate an autoregression of order $k\max - 2 + n$ and perform the joint test again. This procedure is repeated until a rejection occurs or the sequential testing leads to the boundary of zero lags. This procedure has been analyzed by Hall (1994) in the case of a pure autoregressive process.

The $J(m, r)$ statistic specializes to a $t$ statistic on the last lag if the test is performed with $n = 1$. This special case of the general-to-specific procedure has been used by Perron (1989). (See also Perron 1990 and Perron and Vogelsang 1992 for simulation results for unit root tests allowing a break in the trend function and the noise function assumed to be an ARMA process.) Although in principle one can start with $k\min$ lags and gradually increase $k$ until the next included lag is insignificant, Hall (1994) found that a specific-to-general approach is not generally asymptotically valid in the pure AR case. He also found the finite sample properties of statistics associated with a specific-to-general approach to be inferior to those based on a general-to-specific scheme in more general ARMA models. In subsequent analyses only the general-to-specific approach will be analyzed.

### 3.3 Rules Satisfying the Upper-Bound Condition

We now restrict our attention to deterministic and data-dependent rules that satisfy Assumption A1. Formal definitions for the rules considered are as follows.

*Definition 3.2: Deterministic Rules.* Let $\tilde{K} = (\tilde{k}_1, \tilde{k}_2, \ldots)$ be the set of points in $\tilde{K}_\infty = \times_{T=1}^{\infty} S_T$, where $S_T = \{0, 1, \ldots, [T/2]\}$, with $\tilde{k}_T \to \infty$ and $\tilde{k}_T^3/T \to 0$ as $T \to \infty$.

Simply put, $\tilde{K}$ is the collection of deterministic rules that satisfy the conditions of Lemma 2.1. Our definition of deterministic rules is adapted from Eastwood and Gallant (1991), who studied the selection of the truncation point in a univariate Fourier series expansion fitted by least squares. In our context Schwert's rule of thumb is, for example, an element of $\tilde{K}$.

*Definition 3.3: Adaptive Rules.* An adaptive truncation rule is a sequence of random variables $\hat{K}_\infty = (\hat{k}_1, \hat{k}_2, \ldots)$. We say that an adaptive truncation rule maps into the set of deterministic rules $\tilde{K}$ if there exists a deterministic rule $\tilde{k}_T$ such that $\tilde{K} = (\tilde{k}_1, \tilde{k}_2, \ldots)$ is a subset of $\tilde{K}_\infty$ and $\hat{k}_T - \tilde{k}_T \to_p 0$.

The following lemma considers the limiting distribution of $t_\rho$ when it is based on adaptive rules that map into the set of deterministic rules $\tilde{K}$.

*Lemma 3.4.* Suppose that we have an adaptive truncation rule $\hat{K}_\infty = (\hat{k}_1, \hat{k}_2, \ldots)$ that maps into the set of deterministic rules $\tilde{K}$ stated in Definition 3.2, and let $t_\rho(\hat{k}_T)$ be the $t$ statistic for testing $\rho = 1$ in regression (4) estimated with $\hat{k}_T$ lags. Then $t_\rho(\hat{k}_T) \Rightarrow \int_0^1 W(r)\, dW(r) \left(\int_0^1 W(r)^2\, dr\right)^{-1/2}$.

The proof is analogous to theorem 5 of Eastwood and Gallant (1991) and thus is omitted. The importance of Lemma 3.4 is that the limiting distribution of $t_\rho$ is the same whether one uses a deterministic rule in $\tilde{K}$ or an adaptive rule that maps into $\tilde{K}$. The issue then becomes which of the selection procedures delivers better finite-sample properties in testing for the presence of a unit root.

Deterministic rules are useful for analytical purposes, because they help establish the properties of $t_\rho$ under adaptive rules. But as seen from the results reported earlier, size and power will be affected whenever $k$ is fixed in a deterministic way unless the rule happens to be chosen correctly. Adaptive rules take sample information into account and are thus likely to dominate deterministic rules. In the next two sections our analysis will be further restricted to adaptive rules only.

## 4. ADAPTIVE RULE 1: INFORMATION CRITERIA

This section presents properties of $\hat{k}$ and $t_\rho$ when an information criterion as defined in (7) is used to select the truncation lag in regression (4). A related issue has been studied by Hannan and Deistler (1988) in the context of stationary variables with the autoregression

$$x_t = \sum_{i=1}^{k} \delta_i x_{t-i} + e_{tk}. \tag{9}$$

The next lemma summarizes a result of theirs that is relevant to our analysis.

*Lemma 4.1.* Let $x_t$ be a stationary and invertible ARMA process with finite fourth moment and $\tilde{\sigma}_k^2 = (T-k)^{-1} \sum_{t=k+1}^{T} \tilde{e}_{tk}^2$ with $\tilde{e}_{tk}$ the OLS residuals from regression (9). Let $C_T$ be a function of $T$ such that $C_T > 0$ and $C_T/T \to 0$, and $\tilde{k}_T = \arg\min_k(\log(\tilde{\sigma}_k^2) + kC_T/T)$. Then $\lim_{T\to\infty} \tilde{k}_T/b \log T = 1$ for some constant $b$.

The result that the AIC with $C_T = 2$ chooses a value of $k$ that is proportional to $\log T$ in a univariate Gaussian ARMA model is due to Shibata (1980). Hannan and Deistler (1988) provided a unified asymptotic framework to show that the feature of $\log T$ proportionality is generic to information-based rules applied, in particular, to stationary and invertible ARMA models. The logarithmic rule also extends to multivariate and/or autorecessive moving-average with (lagged) exogenous terms (ARMAX) models, as Hannan and Deistler (1988) have shown. Their result is useful in studying the properties of $\hat{k}_T$ within the context of an augmented autoregression of the form (4) derived for an autoregressive integrated moving-average (ARIMA) $(p, 1, q)$ process. The following lemma shows that their result extends to this latter case.

*Lemma 4.2.* Let $y_t$ satisfy (1) and define $\hat{\sigma}_k^2 = (T-k)^{-1} \sum_{t=k+1}^{T} \hat{e}_{tk}^2$, where $\hat{e}_{tk}$ are the least squares residuals from the augmented autoregression (4). Let $\tilde{\sigma}_k^2 = (T-k)^{-1} \sum_{t=k+1}^{T} \tilde{e}_{tk}^2$, where $\tilde{e}_{tk}$ are the OLS residuals from the restricted regression (9) with $x_t = \Delta y_t$. Then $\tilde{\sigma}_k^2 = \hat{\sigma}_k^2 + o_p(T^{-1/2})$, provided $k$ satisfies A1.

Lemma 4.2 implies that the difference between the residual sum of squares from an augmented autoregression and a

restricted one is $o_p(T^{-1/2})$ uniformly in $k$. Hence the information criteria and the corresponding values of $k$ that minimize such criteria are asymptotically the same in both cases. Thus the AIC and Schwartz criteria, when applied to the augmented autoregression defined in (4), also select truncation lags proportional to log $T$ under the null hypothesis of a unit root. The implication for the unit root test is summarized in the following theorem.

*Theorem 4.3.* If $k$ is selected using an information criterion in the class $I_k$ as defined in (7), then $t_\rho$ has a limiting distribution defined by (5) under the null hypothesis of a unit root.

The order of truncation selected by the AIC or the Schwartz criteria is proportional to log $T$. Because such a rule satisfies A1 that $k^3/T \to 0$ and $k \to \infty$ as $T \to \infty$, it is also an adaptive rule that maps into the set of deterministic rules $\tilde{K}$. Thus the result follows directly from Lemmas 2.1 and 3.4. Theorem 4.3 still holds when the DGP is a finite instead of an infinite autoregression, provided that the information criterion does not asymptotically underparameterize the model (see Hall 1994). Note, however, that the information criterion is not an adaptive rule that maps into the set of deterministic rules that satisfy both A1 and A2.

### 4.1 A Special Case: An MA(1)

Because the truncation lag selected from regression (4) when the series is an ARIMA($p$, 1, $q$) and the truncation lag selected on the basis of (9) when the series is a stationary and invertible ARMA process have the same asymptotic properties, we can, for simplicity, use the restricted framework to provide more insight about the properties of the truncation lag selected using information criteria. Specifically, we consider an MA(1) process defined as

$$x_t = e_t + \theta e_{t-1} = \sum_{i=1}^{\infty} \phi_i x_{t-i} + e_t,$$

where $\phi_i = -(-\theta)^i$. The true order of the autoregression is infinity for all values of $\theta \neq 0$. The estimated regression is

$$x_t = \sum_{i=1}^{k} \phi_i x_{t-i} + e_{tk}.$$

It is straightforward to show that $\tilde{\sigma}_k^2$ is approximately related to $k$ by

$$\tilde{\sigma}_k^2 \simeq \sigma_e^2(1 - \theta^{2(k+2)})(1 - \theta^{2(k+1)})^{-1}.$$

Minimizing the AIC, $\log \tilde{\sigma}_k^2 + 2k/T$, the solution is asymptotically equivalent to

$$\tilde{k}(\text{AIC}) \approx (\log(T) + \log[(\theta^2 - 1)\log \theta^2] - \log 2) \times (|\log \theta^2|)^{-1}. \quad (10)$$

Table 3 presents the approximation to $\tilde{k}$(AIC) provided by (10) for various values of $|\theta|$ and $T$. For small $\theta$, $\phi_i$ is small and declines geometrically as $i$ increases. One might then expect the AIC to choose a low order, because extra parameters have little information content but reduce the degrees of freedom. Table 3 shows that indeed for $|\theta| \leq .4$, low values of $k$ are selected by AIC. But as $|\theta|$ gets large, $\phi_i$ will remain nonnegligible even for $i$ quite large. Increasing the length of the autoregression should, in principle, improve the approximation to the DGP. But the $k$ selected by AIC increases only at a logarithmic rate. Except when $T$ becomes impracticably large, the AIC will abandon information at large lags in favor of a very parsimonious model. Hence, in practice one can expect the chosen $k$ to be no higher than 5 even with $T$ as large as 500 when $|\theta|$ is close to 1.

*Table 3. Approximation to the Selected Truncation Lag Using AIC in the MA(1) Model*

| $\lvert\theta\rvert$ | .2 | .4 | .6 | .8 |
|---|---|---|---|---|
| $T$ = 100 | 1 | 2 | 3 | 3 |
| $T$ = 10,000 | 3 | 4 | 7 | 13 |
| $T$ = 1,000,000 | 4 | 7 | 12 | 23 |

NOTE: DGP: $x_t = e_t + \theta e_{t-1}$. Regression: $x_t = \sum_{i=1}^{k} \delta_i x_{t-i} + v_t$.

## 5. ADAPTIVE RULE 2: TESTING FOR THE SIGNIFICANCE OF COEFFICIENTS ON LAGS

This section analyzes the properties of $\hat{k}$ and $t_\rho$ when $\hat{k}$ is chosen by the $J(m, r)$ statistic described in Section 3.2 to sequentially test for the significance of coefficients on additional lags. The following lemma is useful in establishing the limiting distribution of $t_\rho$.

*Lemma 5.1.* Let $\{y_t\}$ be generated by (1) and suppose that Assumptions A1 and A2″ hold. Let $\hat{\mathbf{d}}(k)$ be obtained from the augmented autoregression (4), and let $J(k-n, k)$ be as defined in (8). Then $J(k-n, k)$ is asymptotically distributed as $\chi^2$ with $n$ degrees of freedom.

Berk (1974) proved consistency and asymptotic normality of the coefficients in the restricted regression under A1 and A2″ (see also Lewis and Reinsel 1985). The crucial element in the proof of Lemma 5.1 is the fact that when $\{\Delta y_t\}$ is a stationary and invertible ARMA process, the coefficients $\mathbf{d}(k)$ converge to 0 at a rate that yields an asymptotic equivalence between the Wald test that $\hat{\mathbf{d}}(k) = \mathbf{d}(k)$ and the Wald test that $\hat{\mathbf{d}}(k) = 0$. Indeed, Lemma 5.1 requires Assumption A2″ to ensure that $\sqrt{T}\mathbf{d}(k) \to 0$, which in turn ensures asymptotic normality of $\sqrt{T}\hat{\mathbf{d}}(k)$.

### 5.1 A Special Case: An MA(1) and the $t$ Test

We now specialize the sequential procedure described in Section 3.2.2 to the case where $n = 1$. The square root of the statistic $J(k-1, k)$ then simplifies to a $t$ test for the significance of the coefficient on the last lag in an autoregression of order $k$:

$$t_{\hat{d}_k} = \sqrt{T}\hat{d}_k(\hat{\sigma}_k^2 T \mathbf{M}_k^{-1}(1))^{-1/2}.$$

The sequential procedure chooses a value of $\hat{k}$ if $t_{\hat{d}_k}$ is significant at some prespecified level $\alpha$ in an estimated autoregression of order $\hat{k}$, whereas the $t$ statistics $t_{\hat{d}_k}$ are insignificant in estimated autoregressions of order $k$ for all $k$ in the range $(\hat{k}, k \max]$. We can show that if $\Delta y_t$ is an MA(1) (i.e., $\Delta y_t = e_t + \theta e_{t-1}$), then

$$\hat{d}_k \simeq \theta^{k-1}(1-\theta^2)(1-\theta^{2(k+1)})^{-1}$$

and

$$t_{\hat{d}_k} \simeq \sqrt{T}\theta^{k-1}(1-\theta^2)((1-\theta^{2k})(1-\theta^{2(k+2)}))^{-1/2}.$$

These results show that both $t_{\hat{d}_k}$ and $\sqrt{T}\hat{d}_k$ will converge to zero if $k$ increases at a polynomial rate. Given the result of Lewis and Reinsel (1985, thm. 4) that $t_{\hat{d}_k=d_k} \equiv (\hat{d}_k - d_k)/(\hat{\sigma}_k^2\mathbf{M}_k^{-1}(1))^{1/2}$ is asymptotically distributed $N(0, 1)$ if $k$ increases at a polynomial rate satisfying A1 and A2″, $t_{\hat{d}_k}$ can be shown to have the same asymptotic distribution under these restrictions on the rate of increase of $k$.

It is of some interest to note that the foregoing results also imply that a specific-to-general procedure starting from any lower bound $k$ min that tests for the significance of the coefficient on the last lag would select a $\hat{k}$ that increases to infinity at a logarithmic rate when $\{\Delta y_t\}$ contains a moving-average component. Hence such a specific-to-general procedure would have the same asymptotic properties as a selection rule based on an information criterion.

Note that the asymptotic normality result of Berk (1974) and Lewis and Reinsel (1985) we used to prove Lemma 5.1 requires that $\hat{k}$ increase at some polynomial rate, or at least at a rate that ensures A2″ is satisfied. A logarithmic rate is not sufficient. We now show that the truncation lag selected by a general-to-specific procedure will be of an order higher than log $T$ provided that $k$ max increases at a rate faster than log $T$. In fact the selected truncation lag will grow at the same rate as $k$ max.

*Lemma 5.2.* If $\hat{k}$ is selected by means of the general-to-specific strategy described in Definition 3.1 and $k$ max increases at a rate such that A1 and A2″ are satisfied, then $\hat{k}$ increases at the same rate as $k$ max.

The intuition behind the result stated in Lemma 5.2 is as follows. Under the assumptions of Lemma 5.1, $J(k-n, k)$ is asymptotically distributed as a $\chi^2$ random variable with $n$ degrees of freedom. Thus the limiting probability that $J(k-n, k)$ is statistically significant is $\alpha$, the size of the test. For a given $\hat{k} < k$ max to be chosen, it must be the case that all prior statistics in the sequential procedure $(J(i-n, i); i = k \text{ max} - 1, \ldots, \hat{k} - 1)$ are statistically insignificant. This event occurs for large samples with probability $\alpha(1-\alpha)^{k\max-\hat{k}}$. Because $\hat{k} \leq k$ max and $k$ max $\to \infty$, this probability vanishes as $T \to \infty$ unless $\hat{k}$ increases at the same rate as $k$ max.

The importance of Lemma 5.2 is that if $k$ max is chosen to increase at a polynomial rate, then $\hat{k}$ will also increase at a polynomial rate. This implies that Assumption A2 or A2″ can be satisfied with judicious choice of $k$ max, thereby ensuring that the results of Lemma 5.1 hold. Lemma 5.2 allows us to state the following theorem concerning the limiting behavior of the unit root test under this truncation lag selection rule.

*Theorem 5.3.* If $k$ max satisfies A1 and A2″ and $\hat{k}$ is chosen from the general-to-specific sequential procedure stated in Definition 3.1, then $t_\rho(\hat{k})$ has the same limiting distribution as (5).

Because $\hat{k}$ maps into a deterministic rule in the set $\tilde{K}$ by Lemma 5.2, the result follows from Lemma 3.4. In fact $\hat{k}$ maps into the set of deterministic rules that satisfies A1 and A2, because $k$ increases at a polynomial rate under the conditions of Theorem 5.3.

## 6. FINITE-SAMPLE SIMULATIONS

The results of the preceding sections can be summarized as follows. An information criterion will choose values of $k$ that are proportional to log $T$, a rate ruled out by A2. But the $k$ selected using the $J(m, r)$ statistic to test for the significance of lags will increase at the same rate as the prespecified $k$ max, itself increasing at a polynomial rate. Because a logarithmic rate of increase is slow compared to a polynomial rate, an information criterion will choose values of $k$ that are generally much smaller than those chosen by a general-to-specific $t$ test, for example. Although the log proportionality rule might fail the lower-bound condition, the limiting distribution of $t_\rho$ is unaffected. In such a case the estimates of the coefficients on $\Delta y_{t-i}$ in the augmented autoregression will be consistent at a rate slower than $\sqrt{T}$ for some DGP's. In the MA(1) case, a large value of $|\theta|$ is more likely to be associated with a slower rate of consistency for $\hat{\mathbf{d}}(k)$. We now examine the implications of these results in finite samples.

The results we report are based on 5,000 simulations for different values of $\theta_j$ and $\alpha_i$. For each parameterization, the selected values of $k$ and the corresponding values of $t_\rho$ are recorded. The simulations were performed on a 486/25 MHz PC with code compiled using the Borland C (Version 3) compiler. Random numbers were generated using the ran1( ) function from Press, Teukolsky, Vetterling, and Flannery (1988), with time (in seconds) as seed. We considered $T$ = 100, 200, and 500. For a given $T$, different values for $k$ max and $k$ min were examined. We focus on results for $T$ = 100 with $k$ max = 10 and $k$ min = 0 without loss of generality and discuss results for other configurations where appropriate. The complete set of results is available on request.

We select, for presentation, results based on two information criteria: the AIC and Schwartz. The results for the BIC and the Hannan–Quinn criteria show no appreciable difference. For the general-to-specific strategy, we considered the $t$ as well as the $F$ test, but only present results for the $t$ test at the 5% and 10% levels. In general, a tighter model is selected using a lower significance level.

### 6.1 Frequency Distribution of $\hat{k}$

We first examine the number of times that $k = i$ $(i = 1, \ldots, 10)$ is being selected by each procedure during the 5,000 simulations. Tables 4 and 5 are the frequency counts. As we can see, both information criteria consistently select values of $k$ less than 3. Although the $k$'s selected for autoregressive models seem appropriate given that the DGP's considered are of order no higher than 4, the information criteria yield very parsimonious models when the DGP is driven by a moving-average process. Although the true order of autoregression is infinity in those cases, the AIC and Schwartz criteria continue to choose values of 2 and 3 for $k$. When $\theta$ is large, the coefficients in the autoregression die off only

Table 4. Frequency Count of Selected Lag Lengths k, Moving-Average Case, $T = 100$

| | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ | $k = 6$ | $k = 7$ | $k = 8$ | $k = 9$ | $k = 10$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\theta = .8$ | | | | | |
| $t_{sig}(10)$ | .001 | .028 | .115 | .131 | .149 | .113 | .135 | .104 | .121 | .102 |
| $t_{sig}(5)$ | .013 | .103 | .210 | .157 | .147 | .092 | .097 | .063 | .065 | .053 |
| AIC | .016 | .149 | .325 | .219 | .150 | .069 | .039 | .020 | .009 | .005 |
| Schwartz | .151 | .381 | .326 | .100 | .035 | .004 | .001 | .001 | .000 | .000 |
| | | | | | $\theta = .3$ | | | | | |
| $t_{sig}(10)$ | .304 | .081 | .067 | .055 | .059 | .062 | .074 | .079 | .086 | .086 |
| $t_{sig}(5)$ | .468 | .071 | .047 | .040 | .038 | .039 | .043 | .031 | .049 | .039 |
| AIC | .676 | .131 | .037 | .013 | .007 | .002 | .002 | .001 | .000 | .000 |
| Schwartz | .611 | .037 | .004 | .001 | .000 | .000 | .000 | .000 | .000 | .000 |
| | | | | | $\theta = -.5$ | | | | | |
| $t_{sig}(10)$ | .220 | .189 | .075 | .061 | .058 | .068 | .069 | .077 | .077 | .083 |
| $t_{sig}(5)$ | .387 | .212 | .054 | .041 | .039 | .041 | .041 | .041 | .041 | .039 |
| AIC | .484 | .331 | .073 | .025 | .011 | .007 | .003 | .001 | .000 | .000 |
| Schwartz | .621 | .169 | .016 | .003 | .000 | .000 | .000 | .000 | .000 | .000 |
| | | | | | $\theta = -.8$ | | | | | |
| $t_{sig}(10)$ | .074 | .113 | .109 | .116 | .076 | .097 | .075 | .092 | .080 | .085 |
| $t_{sig}(5)$ | .123 | .162 | .121 | .100 | .061 | .065 | .045 | .047 | .045 | .041 |
| AIC | .197 | .225 | .146 | .091 | .033 | .019 | .008 | .003 | .002 | .001 |
| Schwartz | .264 | .172 | .056 | .019 | .002 | .000 | .000 | .000 | .000 | .000 |

NOTE: DGP: $y_t = y_{t-1} + u_t$; $u_t = e_t + \theta e_{t-1}$. Regression: $\Delta y_t = \delta_0 y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta y_{t-i} + v_t$.

slowly. Truncating the autoregression at a low order will yield a more parsimonious model but with a loss of information. The cost of parsimony will be judged in terms of the size and power of $t_\rho$ in the next subsection.

In the moving-average case, the values of $k$ selected by a general-to-specific modeling strategy are quite evenly distributed over the range [2, $k$ max = 10], with some mass concentrated at $k = 1$. This result follows directly from Lemma 5.1. A further implication of this lemma is that the chosen value of $k$ will be closer to $k$ max the more liberal the size of the test. Thus the frequency of $k$ chosen to be 5 and above is higher under the 10% $t$ test than under the 5% $t$ test.

## 6.2 Size and Power

Having confirmed that information criteria choose values of $k$ that tend to be small, we now proceed to show that in many cases, the method used to choose $k$ can have size and power implications. The results are reported in Tables 6 and 7 for $T = 100$, with the power of the test evaluated at $\rho = .95$ and .85. Turning first to moving-average models (Table 6), we see that for positive values of $\theta$, the size of the test is similar for all methods of selecting $k$. When $\theta = .8$, the 10% $t$ test picks $k$ to be 5 or smaller 40% of the time, whereas the AIC picks $k$ to be in the same range twice as often (see Table 4). Although such variations in the choice of $k$ appear to

Table 5. Frequency Count of Selected Lag Lengths k, Autoregressive Case, $T = 100$

| | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ | $k = 6$ | $k = 7$ | $k = 8$ | $k = 9$ | $k = 10$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\phi_1 = .6, \phi_2 = \phi_3 = \phi_4 = .0$ | | | | | |
| $t_{sig}(10)$ | .407 | .045 | .047 | .051 | .057 | .064 | .071 | .078 | .091 | .089 |
| $t_{sig}(5)$ | .644 | .037 | .033 | .033 | .037 | .042 | .040 | .043 | .047 | .044 |
| AIC | .878 | .075 | .027 | .010 | .004 | .002 | .001 | .001 | .001 | .000 |
| Schwartz | .976 | .020 | .003 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| | | | | | $\phi_1 = -.6, \phi_2 = \phi_3 = \phi_4 = .0$ | | | | | |
| $t_{sig}(10)$ | .407 | .049 | .051 | .050 | .060 | .069 | .068 | .076 | .081 | .090 |
| $t_{sig}(5)$ | .652 | .036 | .037 | .035 | .041 | .040 | .036 | .040 | .040 | .044 |
| AIC | .866 | .086 | .027 | .011 | .006 | .003 | .000 | .000 | .000 | .000 |
| Schwartz | .978 | .018 | .002 | .001 | .000 | .000 | .000 | .000 | .000 | .000 |
| | | | | | $\phi_1 = .4, \phi_2 = .2, \phi_3 = \phi_4 = .0$ | | | | | |
| $t_{sig}(10)$ | .200 | .242 | .055 | .057 | .060 | .058 | .073 | .070 | .092 | .089 |
| $t_{sig}(5)$ | .391 | .285 | .038 | .039 | .038 | .038 | .043 | .035 | .042 | .044 |
| AIC | .455 | .443 | .057 | .021 | .008 | .003 | .002 | .001 | .001 | .000 |
| Schwartz | .680 | .277 | .008 | .001 | .000 | .000 | .000 | .000 | .000 | .000 |
| | | | | | $\phi_1 = .30, \phi_2 = .30, \phi_3 = .25, \phi_4 = .14$ | | | | | |
| $t_{sig}(10)$ | .008 | .095 | .299 | .149 | .059 | .063 | .067 | .078 | .084 | .092 |
| $t_{sig}(5)$ | .023 | .192 | .388 | .137 | .034 | .037 | .042 | .047 | .043 | .046 |
| AIC | .026 | .196 | .494 | .204 | .035 | .016 | .007 | .003 | .001 | .001 |
| Schwartz | .082 | .369 | .419 | .076 | .005 | .002 | .000 | .000 | .000 | .000 |

NOTE: DGP: $y_t = y_{t-1} + \sum_{i=1}^{k} \phi_i \Delta y_{t-i} + u_t$. Regression: $\Delta y_t = \delta_0 y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta y_{t-i} + v_t$.

Table 6. Size and Power of Unit Root Tests; Moving-Average Case, $T = 100$, $k$ max $= 10$

| $\rho$ | $\theta$ | $t_{sig}(10)$ | $t_{sig}(5)$ | AIC | Schwartz |
|---|---|---|---|---|---|
| 1.0 | .80 | .069 | .073 | .068 | .071 |
| 1.0 | .50 | .083 | .087 | .082 | .088 |
| 1.0 | .30 | .075 | .077 | .070 | .069 |
| 1.0 | .00 | .063 | .059 | .052 | .046 |
| 1.0 | −.30 | .097 | .126 | .127 | .174 |
| 1.0 | −.50 | .116 | .158 | .167 | .244 |
| 1.0 | −.80 | .304 | .424 | .561 | .733 |
| .95 | .80 | .136 | .151 | .146 | .158 |
| .95 | .50 | .164 | .184 | .170 | .196 |
| .95 | .30 | .158 | .162 | .152 | .144 |
| .95 | .00 | .153 | .151 | .140 | .126 |
| .95 | −.30 | .228 | .292 | .294 | .393 |
| .95 | −.50 | .254 | .336 | .377 | .510 |
| .95 | −.80 | .534 | .704 | .877 | .963 |
| .85 | .80 | .347 | .387 | .405 | .451 |
| .85 | .50 | .445 | .510 | .520 | .586 |
| .85 | .30 | .465 | .513 | .536 | .505 |
| .85 | .00 | .486 | .540 | .580 | .575 |
| .85 | −.30 | .555 | .682 | .758 | .859 |
| .85 | −.50 | .627 | .753 | .860 | .936 |
| .85 | −.80 | .825 | .908 | .996 | 1.000 |

NOTE: DGP: $y_t = \rho y_{t-1} + u_t$, $u_t = e_t + \theta e_{t-1}$. Regression: $\Delta y_t = \delta_0 y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta y_{t-i} + v_t$.

yield small size differences, power is slightly higher the more parsimonious the model. It is well known that the Schwartz criterion imposes a heavy penalty for overparameterization. Thus for positive moving-average models, the Schwartz criterion tends to yield higher power for a given size.

The result that stands out in Table 6 is the large size distortions when $\theta$ is negative. The problem of size distortion with unit root tests in the presence of negative moving errors is well documented (e.g., Schwert 1989). Although Schwert used deterministic rules to select $k$, he also noted that the exact size depends on the choice of $k$. Our results confirm that the more conservative the criterion for selecting the truncation order, the larger the size distortions associated with $t_\rho$. For example, size distortions associated with the conservative Schwartz criterion are significantly larger than those associated with the 10% $t$ test, the most liberal criterion considered. From the frequency counts, we see that the Schwartz criterion chooses values of $k$ less than 3 90% of the time, whereas the 10% $t$ test chooses values of $k$ *greater* than 3 with a probability of .9.

Table 7 indicates that for autoregressive models, all methods produce estimates of $k$ that are as large as the true order with high probability. Accordingly, all selection procedures produce an exact size close to the nominal size. The 10% $t$ test tends to have lower power, however. According to the frequency counts, the $t$ test tends to overparameterize autoregressive models. For example, the 10% $t$ test selects $k$ greater than 4 more than 40% of the time when the DGP is a fourth-order autoregression. Thus underparameterization is associated with larger size distortions, and overparameterization with power loss when $T = 100$.

The size of the test for moving-average models with $T = 200$ is reported in Table 8. Note that size distortions in negative moving-average models persist as $T$ increases. The Schwartz criterion continues to be associated with significantly larger size distortions than the 10% $t$ test. But in cases for which size distortion is not an issue, as in autoregressive models, the discrepancies in power across selection procedures vanish almost completely when $T = 500$. We report in Table 9 the size and power for autoregressive models at $T = 200$. Compared to the results for $T = 100$, power is higher throughout, and the differences in power across selection procedures are smaller. Thus discrepancies in power across selection procedures are small for typical sample sizes encountered in economic analyses, but size distortions are not. A $t$ or an $F$ test, therefore, has an advantage over information criterion in that it produces tests with more accurate size without much loss of power.

### 6.3 The Choice of $k$ and Size Distortions

When $\theta$ in the noise function is large and negative, $y_t$ is close to having a common factor and behaves more like a white noise than an integrated process. The asymptotic properties of the normalized least squares estimator in this case have been shown by Nabeya and Perron (1994) and Perron (1992) to be different from those derived under standard assumptions. In view of those results, one would conjecture that there is also a discrepancy between the finite-sample distribution of $t_\rho$ and its approximate distribution as defined by (5). But as we can see, the extent of size distortions

Table 7. Size and Power of Unit Root Tests; Autoregressive Case, $T = 100$, $k$ max $= 10$

| $\rho$ | $\phi_1$ | $\phi_2$ | $\phi_3$ | $\phi_4$ | $t_{sig}(10)$ | $t_{sig}(5)$ | AIC | Schwartz |
|---|---|---|---|---|---|---|---|---|
| 1.0 | .6 | .0 | .0 | .0 | .078 | .075 | .066 | .060 |
| 1.0 | −.6 | .0 | .0 | .0 | .068 | .066 | .066 | .060 |
| 1.0 | .4 | .2 | .0 | .0 | .066 | .062 | .055 | .047 |
| 1.0 | .3 | .3 | .25 | .14 | .066 | .062 | .058 | .052 |
| .95 | .6 | .0 | .0 | .0 | .371 | .399 | .404 | .394 |
| .95 | −.6 | .0 | .0 | .0 | .101 | .099 | .087 | .080 |
| .95 | .4 | .2 | .0 | .0 | .346 | .336 | .338 | .267 |
| .95 | .3 | .3 | .25 | .14 | .822 | .840 | .886 | .837 |
| .85 | .6 | .0 | .0 | .0 | .782 | .870 | .960 | .972 |
| .85 | −.6 | .0 | .0 | .0 | .269 | .274 | .268 | .256 |
| .85 | .4 | .2 | .0 | .0 | .763 | .824 | .899 | .867 |
| .85 | .3 | .3 | .25 | .14 | .901 | .937 | .976 | .947 |

NOTE: DGP: $y_t = \rho y_{t-1} + \sum_{i=1}^{4} \phi_i \Delta y_{t-i} + u_t$. Regression: $\Delta y_t = \delta_0 y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta y_{t-i} + v_t$.

*Table 8. Size of Unit Root Tests; Moving-Average Case, T = 200, k max = 12*

| $\rho$ | $\theta$ | $t_{sig}(10)$ | $t_{sig}(5)$ | AIC | Schwartz |
|---|---|---|---|---|---|
| 1.0 | .80 | .056 | .060 | .059 | .063 |
| 1.0 | .50 | .061 | .064 | .056 | .064 |
| 1.0 | .30 | .061 | .064 | .061 | .066 |
| 1.0 | .00 | .064 | .066 | .059 | .057 |
| 1.0 | −.30 | .067 | .076 | .076 | .102 |
| 1.0 | −.50 | .085 | .110 | .121 | .168 |
| 1.0 | −.80 | .177 | .250 | .366 | .557 |

NOTE: DGP: $y_t = \rho y_{t-1} + u_t$, $u_t = e_t + \theta e_{t-1}$. Regression: $\Delta y_t = \delta_0 y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta y_{t-i} + v_t$.

varies with $k$. This suggests that $k$ affects the adequacy of (5) as an approximating distribution. The question is, how?

Using a local asymptotic framework, Pantula (1991) parameterized $\theta$ as $-(1 - T^{-\eta})$ and showed that the limit of $t_\rho$ is given by (5) only if $0 < \eta < .25$, but diverges to $-\infty$ at rate $T^\eta/k$ if $.25 < \eta < .625$, with limiting distribution given by

$$kT^{-\eta}t_\rho \Rightarrow -\left(\int_0^1 W(r)^2\,dr\right)^{-1}. \qquad (11)$$

Because $k = O(T^{1/4})$ by assumption, and (11) is valid for $\eta > .25$, the limiting distribution of $t_\rho$ will always tend to $-\infty$. But the larger the rate of increase of $k$, the slower the rate of divergence and the smaller the discrepancies between the exact and the approximate distributions of $t_\rho$. Consequently, even though $\eta$ is .35 when $T = 100$ and $\theta = -.8$, size distortions are noticeably smaller at larger values of $k$ when critical values from (5) are used for hypothesis testing.

To reinforce the importance of a large $k$ when $\theta$ is large and negative, we report in Table 10 the size of the test at selected parameter values for $T = 200$ and $T = 500$ when a different lower bound, $k$ min, is prescribed. We set $k$ min and $k$ max to 4 and 12 when $T = 200$, and to 6 and 14 when $T = 500$. Evidently, the larger the $k$ min the larger the $k$ and the smaller the size distortions.

The importance of $k$ min and $k$ max in all selection procedures must be emphasized. If we raise the value of $k$ min and let the information criteria select $k$ from the range [$k$ min, $k$ max] and $k$ min > log $T$, then the criteria will choose $k$ min in large samples, because log $T$ is outside the permissible range. Loosely speaking, the choice of $k$ min can be seen as a practical way of imposing the lower-bound condition A2. On the other hand, the choice of $k$ max is more important in a general-to-specific model selection strategy. By Lemma 5.1, the test statistic will choose $k \in$ [$k$ min, $k$ max] with declining probability as $k$ moves away from $k$ max. Thus the larger the $k$ max, the higher the probability that a larger $k$ will be chosen. The larger the $k$, the better the size—at the expense, however, of power losses.

## 7. CONCLUSIONS

This article has analyzed issues related to the selection of the truncation lag in unit root tests of the type proposed by Dickey and Fuller (1979) and Said and Dickey (1984). We have focused on the implications of the lower-bound condition on $t_\rho$ used by Said and Dickey (1984). Procedures that do not satisfy this condition tend to select truncation lags that are too small for some parameter values. Information-based rules such as AIC and Schwartz fit into this category.

A general feature of our results is that an overly parsimonious model can have large size distortions, but an overparameterized model may have low power. But the size problem is more severe than power loss in the sense that discrepancies in power across selection procedures diminish as $T$ increases, but size distortions persist even for large sample sizes for some methods of selecting $k$. In this regard, a $t$ or an $F$ test for the significance of lags will have an advantage over information-based rules such as the AIC, because the former produces tests with more robust size properties across models.

There remain, of course, several avenues for further research that follow from the framework used in this article. First, given the problems associated with approximating a general ARMA process by a finite autoregression, one might be tempted to construct unit root tests from an estimated ARMA($p$, $q$) process whose order is selected using a consistent procedure, such as the one discussed by Dickey and Said (1981). But in view of the problems associated with maximum likelihood estimates of processes with moving-average components, it is not evident that this method can

*Table 9. Size and Power of Unit Root Tests; Autoregressive Case, T = 200, k max = 12*

| $\rho$ | $\phi_1$ | $\phi_2$ | $\phi_3$ | $\phi_4$ | $t_{sig}(10)$ | $t_{sig}(5)$ | AIC | Schwartz |
|---|---|---|---|---|---|---|---|---|
| 1.0 | .6 | .0 | .0 | .0 | .063 | .060 | .057 | .054 |
| 1.0 | −.6 | .0 | .0 | .0 | .063 | .064 | .058 | .056 |
| 1.0 | .4 | .2 | .0 | .0 | .062 | .061 | .056 | .048 |
| 1.0 | .3 | .3 | .25 | .14 | .076 | .072 | .070 | .059 |
| .95 | .6 | .0 | .0 | .0 | .738 | .815 | .897 | .908 |
| .95 | −.6 | .0 | .0 | .0 | .166 | .168 | .160 | .153 |
| .95 | .4 | .2 | .0 | .0 | .712 | .709 | .837 | .784 |
| .95 | .3 | .3 | .25 | .14 | .979 | .988 | 1.000 | .998 |
| .85 | .6 | .0 | .0 | .0 | .955 | .974 | 1.000 | 1.000 |
| .85 | −.6 | .0 | .0 | .0 | .608 | .603 | .738 | .749 |
| .85 | .4 | .2 | .0 | .0 | .954 | .975 | 1.000 | 1.000 |
| .85 | .3 | .3 | .25 | .14 | .994 | .997 | 1.000 | 1.000 |

NOTE: DGP: $y_t = \rho y_{t-1} + \sum_{i=1}^{4} \phi_i \Delta y_{t-i} + u_t$. Regression: $\Delta y_t = \delta_0 y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta y_{t-i} + v_t$.

*Table 10. Size of Unit Root Tests, Moving-Average Case; Different Choices of T, k max, and k min*

| $\rho$ | $\theta$ | $t_{sig}(10)$ | $t_{sig}(5)$ | AIC | Schwartz |
|---|---|---|---|---|---|
| | | *T = 200, k max = 12, k min = 4* | | | |
| 1.0 | .80 | .066 | .068 | .066 | .056 |
| 1.0 | .50 | .057 | .055 | .050 | .048 |
| 1.0 | .30 | .062 | .060 | .057 | .054 |
| 1.0 | .00 | .056 | .056 | .053 | .052 |
| 1.0 | −.30 | .061 | .059 | .051 | .050 |
| 1.0 | −.50 | .061 | .061 | .060 | .059 |
| 1.0 | −.80 | .168 | .228 | .281 | .343 |
| | | *T = 500, k max = 14, k min = 6* | | | |
| 1.0 | .80 | .050 | .048 | .050 | .046 |
| 1.0 | .50 | .053 | .052 | .048 | .047 |
| 1.0 | .30 | .057 | .057 | .060 | .060 |
| 1.0 | .00 | .052 | .052 | .053 | .052 |
| 1.0 | −.30 | .065 | .064 | .062 | .059 |
| 1.0 | −.50 | .059 | .060 | .058 | .056 |
| 1.0 | −.80 | .104 | .134 | .167 | .213 |
| | | *T = 500, k max = 14, k min = 9* | | | |
| 1.0 | .80 | .059 | .062 | .057 | .058 |
| 1.0 | .50 | .059 | .059 | .058 | .057 |
| 1.0 | .30 | .058 | .058 | .059 | .056 |
| 1.0 | .00 | .056 | .056 | .056 | .056 |
| 1.0 | −.30 | .052 | .053 | .052 | .052 |
| 1.0 | −.50 | .058 | .056 | .056 | .056 |
| 1.0 | −.80 | .093 | .103 | .108 | .116 |

NOTE: DGP: $y_t = \rho y_{t-1} + u_t$, $\mu_t = e_t + \theta e_{t-1}$. Regression: $\Delta y_t = \delta_0 y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta y_{t-i} + v_t$.

provide statistical improvement. A comparison of the various estimation methods in the context of unit root tests would be useful.

The second avenue is an extension of the results to the multivariate case whereby vector autoregressive processes are used to approximate more general multivariate linear processes. Although one expects, and preliminary work suggests, that the same qualitative results would hold, the analysis is not a straightforward extension because of possible cointegration among the variables.

The third topic concerns the issue of optimal lag selection. Our analysis has concentrated on two particular classes of lag length selection that are widely used in practice. None of these needs to be optimal. The difficulty, however, lies in finding the proper way to assess the procedures for selecting $k$, because the purpose of estimating these autoregressions is not in obtaining a particular estimate that is as precise as possible, but rather in obtaining the unit root test itself. The optimality criterion thus needs to be based on an appropriate trade-off between type I and type II errors in the application of the unit root test.

## APPENDIX: PROOFS

The following notation will be used in this Appendix. Unless otherwise stated, we shall let $C_1$ be an arbitrary constant (not necessarily the same throughout). Let $\mathbf{D}_T = \text{diag}\{(T-k)^{-1}, (T-k)^{-1/2}, \ldots, (T-k)^{-1/2}\}$, $\mathbf{U}_t' = (y_{t-1}, \mathbf{X}_t')$, and $\mathbf{X}_t' = (\Delta y_{t-1}, \ldots, \Delta y_{t-k})$. Let $\mathbf{M}_k = \sum_{t=k+1}^{T} \mathbf{U}_t\mathbf{U}_t'$ and $\mathbf{R}_k = \sum_{t=k+1}^{T} \mathbf{X}_t\mathbf{X}_t'$. Thus

$$\mathbf{R}_k = \sum_{t=k+1}^{T} (\Delta y_{t-1}, \Delta y_{t-2}, \ldots, \Delta y_{t-k})'(\Delta y_{t-1}, \Delta y_{t-2}, \ldots, \Delta y_{t-k})$$

and

$$\mathbf{D}_T\mathbf{M}_k\mathbf{D}_T = \begin{bmatrix} (T-k)^{-2}\sum_{t=k+1}^{T} y_{t-1}^2 & (T-k)^{-3/2}\sum_{t=k+1}^{T} y_{t-1}\mathbf{X}_t' \\ (T-k)^{-3/2}\sum_{t=k+1}^{T} y_{t-1}\mathbf{X}_t & (T-k)^{-1}\mathbf{R}_k \end{bmatrix}.$$

Note that from Said and Dickey (1984), the limit of $\mathbf{D}_T\mathbf{M}_k\mathbf{D}_T$ is block diagonal with the two blocks corresponding to the limits of $(T-k)^{-2}\sum_{t=k+1}^{T} y_{t-1}^2$ and $(T-k)^{-1}\mathbf{R}_k$. We also let $\mathbf{M}_k^{-1}(1)$ be the first diagonal element of the matrix $\mathbf{M}_k^{-1}$ and let $\mathbf{M}_k^{-1}(k)$ be the lower right $k \times k$ block of $\mathbf{M}_k^{-1}$.

For a matrix $\mathbf{C}$, the matrix norm is defined by $\|\mathbf{C}\| = \sup_{\|\mathbf{x}\|<1}\|\mathbf{C}\mathbf{x}\|$, where for a vector $\mathbf{z}$, $\|\mathbf{z}\| = (\mathbf{z}'\mathbf{z})^{1/2}$. Using this norm, lemma 3 of Berk (1974) showed that $k^{1/2}\|(T-k)\mathbf{R}_k^{-1} - \mathbf{\Gamma}^{-1}\| \to 0$, where $\mathbf{\Gamma}$ is a $k \times k$ matrix with typical elements $\mathbf{\Gamma}_{ij} = E(\Delta y_{t-i}\Delta y_{t-j})$.

### Proof of Lemma 2.1

The proof for consistency of the least squares estimates in the augmented autoregression (4) was given by Said and Dickey (1984) and will not be repeated here. Nevertheless, it is important to point out the two steps involved in the proof. The first step is to show that $k^{1/2}\|(\mathbf{D}_T\mathbf{M}_k\mathbf{D}_T)^{-1} - \mathbf{M}^{-1}\|$ converges to zero for some limiting block-diagonal matrix $\mathbf{M}$. For this step, Assumption A1 is sufficient and the argument follows from lemma 3 of Berk (1974). The second step is to show that $\|\mathbf{D}_T\sum_{t=k+1}^{T}\mathbf{U}_t e_{tk}\| = O_p(k^{1/2})$. The combination of the two steps implies that $Td_0 = O_p(1)$, $\sqrt{T}(\hat{d}_i - d_i) = O_p(1)$ $(i = 1, \ldots, k)$, and $\hat{\sigma}_k^2 \to \sigma_e^2$. Assumption A2 is used only in this second step and, more specifically, to ensure that

$$E\left((T-k)^{-1}\sum_{j=1}^{k}\left(\sum_{t=k+1}^{T}\Delta y_{t-j}(e_{tk} - e_t)\right)^2\right) \le C_1 \cdot k(T-k)\sum_{i=k+1}^{\infty} d_i^2 \to 0, \quad \text{(A.1)}$$

as $T \to \infty$ for some constant $C_1$. Note that A2″ is also sufficient to guarantee that A1 holds; however, A1 is sufficient but not necessary to ensure that $t_\rho$ has the limiting distribution given by (5). To see this, we first express $t_\rho$ as

$$t_\rho = \left((T-k)^{-1}\sum_{t=k+1}^{T} y_{t-1}e_{tk}\right)(\hat{\sigma}_k^2(T-k)^{-2}[\mathbf{M}_k^{-1}(1)]^{-1})^{-1/2}.$$

From Said and Dickey (1984), $T^{-2}[\mathbf{M}_k^{-1}(1)]^{-1} \Rightarrow \sigma_e^2\int_0^1 W(r)^2\,dr$ provided that A1 holds. Consider now the numerator,

$$(T-k)^{-1}\sum_{t=k+1}^{T} y_{t-1}e_{tk} = (T-k)^{-1}\sum_{t=k+1}^{T} y_{t-1}e_t + (T-k)^{-1}\sum_{t=k+1}^{T} y_{t-1}\sum_{i=k+1}^{\infty} d_i\Delta y_{t-i}. \quad \text{(A.2)}$$

It is straightforward to show that $(T-k)^{-1}\sum_{t=k+1}^{T} y_{t-1}e_t \Rightarrow \sigma_e^2\int_0^1 W(r)\,dW(r)$, provided that $k \to \infty$ and $k/T \to 0$ as $T \to \infty$.

Consider now the second term in A2. We have

$$E\left[(T-k)^{-1}\sum_{t=k+1}^{T} y_{t-1}\sum_{i=k+1}^{\infty} d_i\Delta y_{t-i}\right]^2$$

$$= \sum_{i=k+1}^{\infty}\sum_{j=k+1}^{\infty} d_i d_j (T-k)^{-2} \times \sum_{t=k+1}^{T}\sum_{s=k+1}^{T} E[y_{t-1}\Delta y_{t-i} y_{s-1}\Delta y_{s-j}]$$

$$\le C_1 \sum_{i=k+1}^{\infty} d_i \sum_{j=k+1}^{\infty} d_j \le C_1 \sum_{i=k+1}^{\infty}\lambda^i \sum_{j=k+1}^{\infty}\lambda^j$$

$$= C_1\left(\sum_{i=k+1}^{\infty}\lambda^i\right)^2 = C_1\lambda^{2k}/(1-\lambda)^2 \to 0 \quad \text{as} \quad k\to\infty.$$

The first inequality follows from Said and Dickey (1984, p. 601) who stated that there exists a constant $C_1$ such that $(T-k)^{-2}\sum_{t=k+1}^{T}\sum_{s=k+1}^{T} E[y_{t-1}\Delta y_{t-i} y_{s-1}\Delta y_{s-j}] \le C_1$. The second inequality uses the fact that $\Delta y_t$ is a stationary and invertible ARMA process, and hence there exists $\lambda$, $0<\lambda<1$, such that $|d_i| < C_1\lambda^i$ for a different constant $C_1$. Therefore, $(T-k)^{-1}\sum_{t=k+1}^{T} y_{t-1}e_{tk} \Rightarrow \sigma_e^2 \int_0^1 W(r)\,dW(r)$ under the sole condition that $k/T\to 0$ and $k\to\infty$ as $T\to\infty$. Neither A2 nor A2″ is needed to establish the limiting distribution of (5).

To consider the properties of $\hat{\mathbf{d}}(k)$ without the lower-bound condition, it can be seen, from lemma 2 of Berk (1974) or lemma 5.2 of Said and Dickey (1984), that consistency of $\hat{\mathbf{d}}(k)$ still holds if

$$E\left((T-k)^{-2}\sum_{j=1}^{k}\left(\sum_{t=k+1}^{T} u_{t-j}(e_{tk}-e_t)\right)^2\right) \le C_1 k\sum_{i=k+1}^{\infty} d_i^2 \le C_1 k\lambda^{2k}/(1-\lambda^2)\to 0. \quad \text{(A.3)}$$

The condition (A3) is satisfied for any stationary and invertible ARMA process provided that $k\to\infty$, which is assured under A1. More generally, the rate at which $\hat{\mathbf{d}}(k)$ converges to $\mathbf{d}(k)$ can be found by writing

$$\hat{\mathbf{d}}(k)-\mathbf{d}(k) = ((T-k)\mathbf{M}_k^{-1}(k)-\boldsymbol{\Gamma}^{-1})(T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t e_{tk} + \boldsymbol{\Gamma}^{-1}(T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t e_t + \boldsymbol{\Gamma}^{-1}(T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t(e_{tk}-e_t). \quad \text{(A.4)}$$

Taking norms, the first term is $o_p(T^{-1/2})$ and the second is $O_p(k^{1/2}T^{-1/2})$ by the results of Said and Dickey (1984), whether or not A2 is satisfied. Using (A3), the third term is $O_p(k^{1/2}\lambda^k)$ for some $\lambda$ such that $|d_i| \le C_1\lambda^i$. If A2 or A2″ is satisfied, then the second term in (A.4) dominates, because the third term is $o_p(1)$. In that case, $\|\sqrt{T}(\hat{\mathbf{d}}(k)-\mathbf{d}(k))\| = O_p(k^{1/2})$ and $\sqrt{T}(\hat{d}_i - d_i) = O_p(1)$, $i=1,\ldots,k$. If A2″ is not satisfied, then the third term in (A.4) dominates and $\|\lambda^{-k}(\hat{\mathbf{d}}(k)-\mathbf{d}(k))\| = O_p(k^{-1/2})$ or $\lambda^{-k}(\hat{d}_i - d_i) = O_p(1)$ $(i=1,\ldots,k)$.

The proof of Lemma 2.1 is completed by showing $\hat{\sigma}_k^2 \to \sigma_e^2$ without any lower-bound condition. The result follows from consistency of the least squares estimates. The proof is standard and is omitted.

## Proof of Lemma 4.2

Let $\hat{\mathbf{d}}(k) = (\hat{d}_1,\ldots,\hat{d}_k)$ be obtained by applying OLS to the augmented autoregression (4), let $\tilde{\mathbf{d}}(k) = (\tilde{d}_1,\ldots,\tilde{d}_k)$ be obtained by applying OLS to (9) with $x_t = \Delta y_t$. We have $\hat{\mathbf{d}}(k) - \mathbf{d}(k) = \mathbf{M}_k^{-1}(k)\sum_{t=k+1}^{T}\mathbf{X}_t' e_{tk}$ and $\tilde{\mathbf{d}}(k) - \mathbf{d}(k) = \mathbf{R}_k^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t'(e_{tk} + d_0 y_{t-1}) \equiv \mathbf{R}_k^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t' e_{tk}$, because $d_0 = 0$ under the null hypothesis of a unit root. Hence

$$\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k) = (\mathbf{M}_k^{-1}(k)-\mathbf{R}_k^{-1})\sum_{t=k+1}^{T}\mathbf{X}_t' e_{tk}.$$

Note from lemma 5.2 of Said and Dickey (1984) that $\|(T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t' e_{tk}\| = O_p(k^{1/2}T^{-1/2})$. By partition inversion, $(T-k)\mathbf{M}_k^{-1}(k) = ((T-k)^{-1}\mathbf{R}_k - \mathbf{A})^{-1}$, where

$$\mathbf{A} = (T-k)^{-1}\left(\sum_{t=k+1}^{T} y_{t-1}\mathbf{X}_t\right)\left(\sum_{s=k+1}^{T} y_{s-1}\mathbf{X}_s'\right)\left(\sum_{t=k+1}^{T} y_{t-1}^2\right)^{-1} = (T-k)^{-3}\left(\sum_{t=k+1}^{T}\sum_{s=k+1}^{T} y_{t-1}y_{s-1}\mathbf{X}_t\mathbf{X}_s'\right)\Big/(T-k)^{-2}\sum_{t=k+1}^{T} y_{t-1}^2.$$

Note that $(T-k)^{-2}\sum_{t=k+1}^{T} y_{t-1}^2 = O_p(1)$, and by Said and Dickey (1984, p. 601), each element of the numerator of $\mathbf{A}$ is bounded by $C_1/(T-k)$ for some constant $C_1$. Because $\mathbf{A}$ is a $k\times k$ matrix and $E(\|\mathbf{A}\|^2) \le C_1 k^2/(T-k)$, we have that $k^{1/2}\|\mathbf{A}\|$ converges to zero provided that $k^3/T\to 0$. Thus

$$\begin{aligned}\|(T-k)\mathbf{M}_k^{-1}(k) - (T-k)\mathbf{R}_k^{-1}\| &= \|(T-k)\mathbf{M}_k^{-1}(k)((T-k)^{-1}\mathbf{R}_k \\ &\quad - ((T-k)\mathbf{M}_k^{-1}(k))^{-1})(T-k)\mathbf{R}_k^{-1}\| \\ &= \|(T-k)\mathbf{M}_k^{-1}(k)\mathbf{A}(T-k)\mathbf{R}_k^{-1}\| \\ &\le \|(T-k)\mathbf{M}_k^{-1}(k)\|\|\mathbf{A}\|\|(T-k)\mathbf{R}_k^{-1}\|.\end{aligned}$$

Because $\|(T-k)\mathbf{M}_k^{-1}(k)\|$ and $\|(T-k)\mathbf{R}_k^{-1}\|$ are $O_p(1)$ (see Said and Dickey 1984) and $k^{1/2}\|\mathbf{A}\|\to 0$, $k^{1/2}\|(T-k)\mathbf{M}_k^{-1}(k) - (T-k)\mathbf{R}_k^{-1}\|\to 0$. Combining these results, we have

$$T^{1/2}\|\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k)\| \le k^{1/2}\|(T-k)\mathbf{M}_k^{-1}(k) - (T-k)\mathbf{R}_k^{-1}\|k^{-1/2}T^{1/2}\left\|(T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t' e_{tk}\right\| \quad \text{(A.5)}$$

$\to 0$ as $T\to\infty$, provided that $k^3/T\to 0$.

We are now in a position to prove Lemma 4.2. Using the definitions of $\hat{e}_{tk}$ and $\tilde{e}_{tk}$, we have

$$\begin{aligned}\hat{\sigma}_k^2 &= (T-k)^{-1}\sum_{t=k+1}^{T}(\Delta y_t - \hat{d}_0 y_{t-1} - \hat{\mathbf{d}}(k)'\mathbf{X}_t)^2 \\ &= \tilde{\sigma}_k^2 + (T-k)^{-1}\hat{d}_0^2\sum_{t=k+1}^{T} y_{t-1}^2 - 2(T-k)^{-1}\hat{d}_0 \\ &\quad\times \sum_{t=k+1}^{T} y_{t-1}\tilde{e}_{tk} + (\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))'\left[(T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t\mathbf{X}_t'\right] \\ &\quad\times(\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k)) - 2(\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))'(T-k)^{-1} \\ &\quad\times\sum_{t=k+1}^{T}\mathbf{X}_t\tilde{e}_{tk} + 2(T-k)^{-1}\hat{d}_0(\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))'\sum_{t=k+1}^{T}\mathbf{X}_t y_{t-1}.\end{aligned}$$

We now consider each term individually.

1. $(T-k)^{-1}(T-k)^2\hat{d}_0^2(T-k)^{-2}\sum_{t=k+1}^{T} y_{t-1}^2 = O_p(T^{-1})$, because $(T-k)^2\hat{d}_0^2 = O_p(1)$ and $(T-k)^{-2}\sum_{t=k+1}^{T} y_{t-1}^2 = O_p(1)$.

2. $(T-k)^{-1}(T-k)\hat{d}_0(T-k)^{-1}\sum_{t=k+1}^{T} y_{t-1}\tilde{e}_{tk} = O_p(T^{-1})$. Because $T\hat{d}_0 = O_p(1)$, we need to show that $(T-k)^{-1}\sum_{t=k+1}^{T} y_{t-1}\tilde{e}_{tk}$ is $O_p(1)$. Using the fact that $\tilde{e}_{tk} = e_{tk} + (\tilde{\mathbf{d}}(k)-\mathbf{d}(k))'\mathbf{X}_t$, we have

$$(T-k)^{-1}\sum_{t=k+1}^{T} y_{t-1}\tilde{e}_{tk}$$

$$= (T-k)^{-1}\sum_{t=k+1}^{T} y_{t-1}e_{tk} + (T-k)^{-1}(\tilde{\mathbf{d}}(k)-\mathbf{d}(k))'$$

$$\times \sum_{t=k+1}^{T} y_{t-1}\mathbf{X}_t.$$

The first term is $O_p(1)$ (see the proof of Lemma 2.1). We now show that the second term vanishes. We have for $|d_i| \le C_1\lambda^i$, with $0 < \lambda < 1$,

$$\|\tilde{\mathbf{d}}(k)-\mathbf{d}(k)\|\|(T-k)^{-1}\sum_{t=k+1}^{T} y_{t-1}\mathbf{X}_t\|$$

$$= O_p(k^{1/2}\lambda^k)O_p(k^{1/2}) \quad \text{if A2 is not satisfied;}$$

$$= O_p(k^{1/2}T^{-1/2})O_p(k^{1/2}) \quad \text{if A2 is satisfied;}$$

and is $o_p(1)$, because $k^2/T \to 0$ in the latter case and $k \to \infty$ with $\lambda < 1$ in the former case.

3. Taking norms, for the third term we have

$$\left\| (\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))'\left((T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)(\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))\right\|$$

$$\le \|(\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))\|\left\|(T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t\mathbf{X}_t'\right\|\|(\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))\|$$

$$= o_p(T^{-1/2})\cdot O_p(1)\cdot o_p(T^{-1/2}) = o_p(T^{-1})$$

using (A.5). Hence the third term is $o_p(T^{-1})$.

4.

$$(\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))'(T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t\tilde{e}_{tk}$$

$$= (\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))'(T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_te_{tk}$$

$$+ (\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))'\left((T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)(\hat{\mathbf{d}}(k)-\mathbf{d}(k)).$$

Taking norms, for the first term we have

$$\left\|(\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))'(T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_te_{tk}\right\|$$

$$\le \|(\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))\|\left\|(T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_te_{tk}\right\|$$

$$= o_p(T^{-1/2})\cdot O_p(k^{1/2}T^{-1/2}) = o_p(k^{1/2}T^{-1}).$$

For the second term we have

$$\left\|(\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))'\left((T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)(\hat{\mathbf{d}}(k)-\mathbf{d}(k))\right\|$$

$$\le \|(\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))\|\left\|(T-k)^{-1}\sum_{t=k+1}^{T}\mathbf{X}_t\mathbf{X}_t'\right\|\|(\hat{\mathbf{d}}(k)-\mathbf{d}(k))\|$$

$$= o_p(T^{-1/2})\cdot O_p(1)\cdot O_p(k^{1/2}T^{-1/2})$$

$$= o_p(k^{1/2}T^{-1}) \quad \text{if A2 is satisfied;}$$

$$= o_p(T^{-1/2})\cdot O_p(1)\cdot O_p(k^{1/2}\lambda^k)$$

$$= o_p(k^{1/2}T^{-1/2}\lambda^k) \quad \text{if A2 is not satisfied.}$$

5. Because $T\hat{d}_0 = O_p(1)$, we consider

$$\left\|(\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k))(T-k)^{-2}\sum_{t=k+1}^{T}\mathbf{X}_ty_{t-1}\right\|$$

$$\le \|\hat{\mathbf{d}}(k)-\tilde{\mathbf{d}}(k)\|\left\|(T-k)^{-2}\sum_{t=k+1}^{T}\mathbf{X}_ty_{t-1}\right\|$$

$$= o_p(T^{-1/2})O_p(k^{1/2}T^{-1}) = o_p(k^{1/2}T^{-3/2}).$$

Collecting results from 1–5, we have

$$\hat{\sigma}_k^2 = \tilde{\sigma}_k^2 + o_p(k^{1/2}T^{-1}) \quad \text{if A2 is satisfied;}$$

$$\hat{\sigma}_k^2 = \tilde{\sigma}_k^2 + o_p(k^{1/2}T^{-1/2}\lambda^k) \quad \text{if A2 is not satisfied.}$$

Because $k/T \to 0$ and $k^{1/2}\lambda^k \to 0$ as $k \to \infty$ and $T \to \infty$, we have, whether or not A2 is satisfied,

$$\hat{\sigma}_k^2 = \tilde{\sigma}_k^2 + o_p(T^{-1/2}).$$

## Proof of Lemma 5.1

We first note (from the proof of Lemma 4.1) that $\hat{\mathbf{d}}(n) = \tilde{\mathbf{d}}(n) + o_p(1)$, where $\tilde{\mathbf{d}}(n)$ corresponds to the OLS estimates from the restricted regression without the lagged dependent variable. Using the block diagonality of $\mathbf{M}_k$, we have the following asymptotic relation:

$$J(k-n,k) = (T-k)\tilde{\mathbf{d}}(n)'((T-k)\hat{\sigma}_r^2\mathbf{R}_k^{-1}(n))^{-1}\tilde{\mathbf{d}}(n) + o_p(1),$$

where $\mathbf{R}_k^{-1}(n)$ is the lower $n \times n$ block of $\mathbf{R}_k^{-1}$. We now apply the following decomposition:

$$J(k-n,k)$$

$$= \sqrt{T}(\tilde{\mathbf{d}}(n)-\mathbf{d}(n))'((T-k)^{-1}\hat{\sigma}_r^2\mathbf{R}_k^{-1}(n))^{-1}\sqrt{T}(\tilde{\mathbf{d}}(n)-\mathbf{d}(n))$$

$$+ 2\sqrt{T}(\tilde{\mathbf{d}}(n)-\mathbf{d}(n))'((T-k)\hat{\sigma}_r^2\mathbf{R}_k^{-1}(n))^{-1}\sqrt{T}\mathbf{d}(n)'$$

$$+ \sqrt{T}\mathbf{d}(n)'((T-k)\hat{\sigma}_r^2\mathbf{R}_k^{-1}(n))^{-1}\sqrt{T}\mathbf{d}(n) + o_p(1).$$

By theorem 4 of Lewis and Reinsel (1984), the first term is asymptotically distributed as $\chi^2$ with $n$ degrees of freedom. To complete the proof, it remains to show that the other terms vanish as $T \to \infty$. We first note that $[(T-k)\hat{\sigma}_r^2\mathbf{R}_k^{-1}(n)]^{-1} \to_p \mathbf{R}$, say. Given that $\Delta y_t$ is a stationary and invertible ARMA process, a typical element of $\sqrt{T}\mathbf{d}(n)$, say $\sqrt{T}d_{k+i}$ $(i = 1, \ldots, n)$, is such that $|\sqrt{T}d_{k+i}| \le C_1\sqrt{T}\lambda^{k+i}$ for some $C_1$ and $0 < \lambda < 1$. Hence, under the conditions of A2″, $\sqrt{T}\mathbf{d}(n) \to 0$. It follows that the last term converges to zero. Finally, to show that the second term also vanishes, we simply note that under the conditions of A2″, $\sqrt{T}(\tilde{\mathbf{d}}(n) - \mathbf{d}(n)) = O_p(1)$.

## Proof of Lemma 5.2

Because $k$ max is assumed to increase in such a way that assumption A2″ is satisfied, the conditions of Lemma 5.1 hold and $J(k \text{ max}, k \text{ max} + n)$ is asymptotically distributed as a $\chi^2$ random variable with $n$ degrees of freedom. Let $\hat{k}_T$ be the estimate of $k$ selected by the sequential procedure described in Definition 3.1. Then

$$\lim_{T\to\infty} P[\hat{k}_T \ne k \text{ max}] = 1-\alpha,$$

and using the rules of conditional probability,

$$\lim_{T\to\infty} P[\hat{k}_T = k \text{ max} - 1 \,|\, \hat{k}_T \ne k \text{ max}] = \alpha.$$

This implies that

$$\lim_{T\to\infty} P[\hat{k}_T = k\max - 1 \cap \hat{k}_T \neq k\max]$$
$$= \lim_{T\to\infty} P[\hat{k}_T = k\max - 1 \mid \hat{k}_T \neq k\max]P[\hat{k}_T \neq k\max]$$
$$= \alpha(1-\alpha).$$

Now

$$\lim_{T\to\infty} P[\hat{k}_T = k\max - 2 \mid \hat{k}_T \neq k\max - 1 \cap \hat{k}_T \neq k\max] = \alpha$$
$$= \lim_{T\to\infty} \frac{P[\hat{k}_T = k\max - 2 \cap \hat{k}_T \neq k\max - 1 \cap \hat{k}_T \neq k\max]}{P(\hat{k}_T \neq k\max - 1 \cap \hat{k}_T \neq k\max)}$$

and

$$\lim_{T\to\infty} P(\hat{k}_T \neq k\max - 1 \cap \hat{k}_T \neq k\max)$$
$$= \lim_{T\to\infty} P(\hat{k}_T \neq k\max - 1 \mid \hat{k}_T \neq k\max)P(\hat{k}_T \neq k\max)$$
$$= (1-\alpha)^2.$$

This implies that

$$\lim_{T\to\infty} P[\hat{k}_T = k\max - 2 \cap \hat{k}_T \neq k\max - 1 \cap \hat{k}_T \neq k\max] = \alpha(1-\alpha)^2.$$

We can deduce, by recursion, that

$$\lim_{T\to\infty} P[\hat{k}_T = k\min \cap \hat{k}_T \neq k\min + 1 \cdots \cap \hat{k}_T \neq k\max] = \lim_{T\to\infty}[\alpha(1-\alpha)^{k\max - k\min}]. \quad (A.6)$$

Now suppose that $k$ min increases to infinity at a rate slower than $k$ max. From (A.6), the application of the sequential procedure implies the probability that $\hat{k}_T$ reaches $k$ min is zero in the limit, because $k$ max $- k$ min $\to \infty$. It follows that for any given $k$ min and $k$ max, $\hat{k}_T$ must be bounded away from $k$ min. Because $k$ min can be any arbitrary sequence, it follows that $\hat{k}_T$ has a zero probability of increasing at a rate slower than $k$ max. Thus with the inequality $k$ min $\leq k$ max, $\hat{k}_T$ must increase at the same rate as $k$ max.

*[Received August 1993. Revised February 1994.]*

## REFERENCES


Akaike, H. (1974), "A New Look at Statistical Model Identification," *IEEE Transactions on Automatic Control,* AC-19, 716–723.

Agiakloglou, C., and Newbold, P. (1992), "Empirical Evidence on Dickey–Fuller Type Tests," *Journal of Time Series Analysis,* 13, 471–483.

Banerjee, A., Lumsdaine, R. L., and Stock, J. H. (1992), "Recursive and Sequential Tests of the Unit Root and Tread Break Hypothesis," *Journal of Business & Economic Statistics,* 10, 271–287.

Berk, K. (1974), "Consistent Autoregressive Spectral Estimates," *The Annals of Statistics,* 2, 489–502.

Chan, N. H., and Wei, C. (1988), "Limiting Distribution of Least Squares Estimates of Unstable Autoregressive Processes," *The Annals of Statistics,* 16, 367–401.

Dickey, D. A., and Fuller, W. (1979), "Distribution of the Estimators for Autoregressive Time Series With a Unit Root," *Journal of the American Statistical Association,* 74, 427–431.

Dickey, D. A., and Said, S. E. (1981), "Testing ARMA($p$, 1, $q$) Versus ARMA($p+1, q$)," in *Proceedings of the Business and Economics Statistics Section, American Statistical Association,* pp. 318–322.

Eastwood, B. J., and Gallant, R. A. (1991), "Adaptive Rules for Seminonparametric Estimators That Achieve Asymptotic Normality," *Econometric Theory,* 7, 307–340.

Fuller, W. (1976), *Introduction to Statistical Time Series,* New York: John Wiley.

Hall, A. (1994), "Testing for a Unit Root in Time Series With Pretest Data Based Model Selection," *Journal of Business & Economic Statistics,* 12, 461–470.

Hannan, E. J., and Deistler, M. (1988), *The Statistical Theory of Linear Systems,* New York: John Wiley.

Harris, R. I. D. (1992), "Testing for Unit Roots Using the Augmented Dickey–Fuller Test," *Economic Letters,* 38, 381–386.

Lewis, R., and Reinsel, G. C. (1985), "Prediction of Multivariate Time Series by Autoregressive Model Fitting," *Journal of Multivariate Analysis,* 16, 393–411.

Nabeya, S., and Perron, P. (1994), "Local Asymptotic Distributions Related to the AR(1) Model With Dependent Errors," *Journal of Econometrics,* 62, 229–264.

Pantula, S. G. (1991), "Asymptotic Distributions of Unit-Root Tests When the Process is Nearly Stationary," *Journal of Business & Economic Statistics,* 9, 63–71.

Perron, P. (1989), "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis," *Econometrica,* 57, 1361–1401.

—— (1990), "Further Evidence of Breaking Trend Functions in Macroeconomic Time Series," unpublished manuscript, Princeton University.

—— (1992), "The Adequacy of Asymptotic Approximations in the Near-Integrated Autoregressive Model With Dependent Errors," *Journal of Econometrics,* in press.

Perron, P., and Vogelsang, T. J. (1992), "Nonstationarity and Level Shifts With an Application to Purchasing Power Parity," *Journal of Business & Economic Statistics,* 10, 301–320.

Press, W. H., Tuekolsky, S., Vetterling, W., and Flannery, B. (1988), *Numerical Recipies in C,* Cambridge, U.K.: Cambridge University Press.

Said, S., and Dickey, D. A. (1984), "Testing for Unit Roots in Autoregressive Moving-Average Models of Unknown Order," *Biometrika,* 71, 599–607.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* 6, 461–464.

Schwert, G. W. (1989), "Tests for Unit Roots: A Monte Carlo Investigation," *Journal of Business & Economic Statistics,* 7, 147–160.

Shibata, R. (1980), "Asymptotic Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process," *The Annals of Statistics,* 8, 147–164.