

ORIGINAL ARTICLE

## TESTING FOR A UNIT ROOT IN NONCAUSAL AUTOREGRESSIVE MODELS

PENTTI SAIKKONEN<sup>a</sup> AND RICKARD SANDBERG<sup>b\*</sup>

<sup>a</sup> *Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland*

<sup>b</sup> *Department of Economics, Center for Economic Statistics, Stockholm School of Economics, Stockholm, Sweden*

This work develops maximum likelihood-based unit root tests in the noncausal autoregressive (NCAR) model with a non-Gaussian error term formulated by Lanne and Saikkonen (2011, *Journal of Time Series Econometrics* 3, Issue 3, Article 2). Finite-sample properties of the tests are examined via Monte Carlo simulations. The results show that the size properties of the tests are satisfactory and that clear power gains against stationary NCAR alternatives can be achieved in comparison with available alternative tests. In an empirical application to a Finnish interest rate series, evidence in favour of an NCAR model with leptokurtic errors is found.

*Received 1 October 2013; Revised 26 February 2015; Accepted 8 May 2015*

**Keywords:** Maximum likelihood estimation; noncausal autoregressive model; non-Gaussian time series; unit root; bootstrap.

### 1. INTRODUCTION

Testing for the unit root hypothesis is an important part in the analysis of economic time series and has attracted an enormous amount of interest during the past decades. In this context, the most widely used model is the conventional (causal) autoregressive (AR) model where the current observation is expressed as a weighted average of past observations and an error term. An essential assumption of the conventional AR model is that the error term is unpredictable by the past of the considered time series. However, in (say) economic applications, this assumption may break down because the impact of omitted variables, interrelated with the considered (univariate) time series, is ignored. More specifically, if relevant variables are omitted, their impact goes (at least partly) to the error term of the model, and, as the considered time series may help to predict the omitted variables, the assumed unpredictability condition may break down. As economic variables are typically interrelated, this point appears particularly pertinent in economic applications. In cases like this, the noncausal AR (NCAR) model may provide a viable alternative, for it explicitly allows for the predictability of the error term by the past of the considered series.

Early studies of NCAR models and their extensions, noncausal and (potentially) noninvertible AR moving average (ARMA) models, were mainly motivated by applications to natural sciences and engineering [see, e.g. Breidt *et al.* (1991), Lii and Rosenblatt (1996), Huang and Pawitan (2000), Rosenblatt (2000), Breidt *et al.* (2001), & Wu and Davis (2010) and the references therein]. More recently, a slightly different formulation of the NCAR model was considered by Lanne and Saikkonen (2011) (hereafter L&S) and further studied by Lanne *et al.* (2012a, 2012b, 2012c), Lanne and Saikkonen (2013) & Gouriéroux and Zakoian (2013). These papers demonstrate that

---

\* Correspondence to: Rickard Sandberg, Department of Economics, Center for Economic Statistics, Stockholm School of Economics, PO Box 6501 (Sveavägen 65), 113 83 Stockholm, Sweden. E-mail: rickard.sandberg@hhs.se

the NCAR model can successfully describe and forecast many economic time series, and it often outperforms its conventional causal alternative in terms of model fit and forecasting accuracy.

(Basawa and Scott, 1983, Ch. 2)

Even though the properties of the stationary NCAR model are by now well understood and asymptotic distribution theory for various parameter estimators [typically maximum likelihood (ML) estimators] have been developed, the nonstationary case and tests for a unit root have not yet been studied in the literature. As unit root type nonstationarity appears quite common (particularly) in economic time series, and hence potential applications of the NCAR model, this work aims at proposing unit root tests in the context of the NCAR model of L&S. We develop Wald-type unit root tests by assuming that the possible unit root appears in the causal AR polynomial of the model, and to this end, we first derive asymptotic properties of a (local) ML estimator of the parameters of the model under the unit root hypothesis. As in the stationary case, a non-Gaussian error term is required to achieve identification [see, e.g. Brockwell and Davis (1987, pp. 124–125) and Rosenblatt (2000, pp. 10–11)]. This renders the estimation problem nonlinear, which, in turn, makes the derivation of limiting distributions straightforward than in the context of conventional unit root tests, where estimation is carried out by linear least squares (LS) techniques. To address this issue, we use ideas similar to those used in statistical models whose likelihood ratios satisfy the so-called locally asymptotically mixed normal (LAMN) condition (Basawa and Scott, 1983, Ch. 2). It turns out that the limiting distributions of our tests are not distribution free and appear, in general, very complicated depending on a number of nuisance parameters. To obtain tests with manageable limiting distributions, we assume that the error term of the model has a symmetric distribution. Then the limiting distributions of our tests only depend on a single nuisance parameter determined by the distribution of the error term, and this problem can be rather easily circumvented by using estimated critical values (described in Section 5.1). Extending this approach to skewed errors appears infeasible so that a bootstrap procedure (described in Section 5.2) is discussed in order to relax the symmetry assumption.

We examine the practical relevance of our asymptotic tests by means of Monte Carlo simulations. The results show that our tests perform satisfactorily in terms of size, and their power against correctly specified stationary NCAR alternatives is very good in comparison with conventional Dickey–Fuller (DF) tests, the  $M$ -tests of Lucas (1995) and the likelihood-based unit root tests of Rothenberg and Stock (1997). We also demonstrate that our bootstrap procedure works very well in cases where the error distribution is skewed. To illustrate the practical implementation of our tests, we present an application to a Finnish interest rate series for which a stationary NCAR model with Student’s  $t$ -distributed errors (symmetric or skewed) is found to provide a good description.

The plan of the paper is as follows. Section 2 defines the considered NCAR model and discusses the testing problem. Parameter estimation and related asymptotic results are presented in Section 3 and used in Section 4 to obtain our unit root tests. Section 5 reports the results of the Monte Carlo simulations, and Section 6 presents the empirical application. Section 7 concludes. Appendices A–C contain mathematical proofs and some technical details.

Finally, the following notation is used throughout the paper. The notation  $\xrightarrow{P}$  signifies convergence in probability, and  $\xrightarrow{d}$  is used for convergence in distribution and also for weak convergence in a function space. We write  $B(u) \sim BM(\cdot)$  for a Brownian motion  $B(u)$  with indicated variance or covariance matrix. Unless otherwise stated, all vectors will be treated as column vectors, and, for notational convenience, we shall write  $x = (x_1, \dots, x_n)$  for the (column) vector  $x$  where the components  $x_i$  may be either scalars or vectors (or both).

## 2. MODEL AND TESTING PROBLEM

Following L&S, we consider the NCAR model

$$\phi(B)\varphi(B^{-1})y_t = \epsilon_t, \quad t = 1, 2, \dots, \tag{1}$$

where  $\epsilon_t$  is a sequence of i.i.d. random variables with mean 0 and finite variance  $\sigma^2 > 0$ ,  $B$  is the usual backward shift operator ( $By_t = y_{t-k}$  for  $k = 0, \pm 1, \dots$ ), and  $\phi(B) = 1 - \phi_1 B - \dots - \phi_r B^r$  and  $\varphi(B^{-1}) = 1 - \varphi_1 B^{-1} - \dots - \varphi_s B^{-s}$ . L&S assume that the polynomials  $\phi(z)$  and  $\varphi(z)$  ( $z \in \mathbb{C}$ ) have their roots outside the unit circle

in which case the difference equation (1) has a stationary solution. In this paper, we allow for the possibility that, owing to a unit root in the causal AR polynomial  $\phi(z)$ , the process  $y_t$  is a nonstationary integrated process.

Thus, we assume that  $r > 0$  and proceed in the conventional way by writing the lag polynomial  $\phi(B)$  as

$$\phi(B) = \Delta - \phi B - \pi_1 \Delta B - \dots - \pi_{r-1} \Delta B^{r-1}, \tag{2}$$

where  $\Delta = 1 - B$  is the difference operator. Our focus is in testing for the unit root null hypothesis  $H_0 : \phi = 0$  against the stationary alternative  $H_1 : \phi < 0$ . At this point, we abstract from any deterministic terms such as a constant term or linear time trend in the process. These extensions will be discussed in Section 4.2.

Unless otherwise stated, we assume throughout the paper that the null hypothesis  $H_0$  holds and that the roots of the polynomials  $\pi(z) = 1 - \pi_1 z - \dots - \pi_{r-1} z^{r-1}$  and  $\varphi(z)$  lie outside the unit circle or, formally, that

$$\pi(z) \neq 0 \text{ for } |z| \leq 1 \text{ and } \varphi(z) \neq 0 \text{ for } |z| \leq 1. \tag{3}$$

Using equation (2), we can write equation (1) as

$$\Delta y_t = \phi y_{t-1} + \pi_1 \Delta y_{t-1} + \dots + \pi_{r-1} \Delta y_{t-r+1} + v_t, \quad t = 1, 2, \dots, \tag{4}$$

where the process  $v_t = \phi(B) y_t = \varphi(B^{-1})^{-1} \epsilon_t$  has the forward moving average representation

$$v_t = \sum_{j=0}^{\infty} \beta_j \epsilon_{t+j}, \quad \beta_0 = 1. \tag{5}$$

Here,  $\beta_j$  is the coefficient of  $z^{-j}$  in the Laurent series expansion of  $\varphi(z^{-1})^{-1}$ . By the latter condition in (3), this expansion is well defined for  $|z| \geq b_\varphi$  with some  $b_\varphi < 1$  and with the coefficients  $\beta_j$  decaying to zero at a geometric rate as  $j \rightarrow \infty$ . Equation (4) shows that our testing problem can be thought of as testing for a unit root in an AR( $r$ ) process with stationary errors following the purely noncausal AR(0,  $s$ ) process  $\varphi(B^{-1}) v_t = \epsilon_t$  [as in L&S, we use the acronym AR( $r, s$ ) for the model defined in equation (1)]. When  $r = 1$ , the lagged differences vanish from the right-hand side of equation (4), which becomes a special case of a first-order autoregression with general stationary (or short-memory) errors. Testing for a unit root in such contexts has been considered in a number of papers since the work of Phillips (1987) & Phillips and Perron (1988). That the errors in (4) are generated by a purely noncausal AR(0,  $s$ ) process distinguishes our formulation from its previous counterparts.

For later use, we also introduce the (causal) AR( $r$ ) process  $u_t = \varphi(B^{-1}) y_t$  or  $\phi(B) u_t = \epsilon_t$  ( $t = 1, 2, \dots$ ). Under the null hypothesis,  $\pi(B) \Delta u_t = \epsilon_t$ , and the former condition in (3) yields the conventional backward moving average representation

$$\Delta u_t = \sum_{j=0}^{\infty} \alpha_j \epsilon_{t-j}, \quad \alpha_0 = 1, \tag{6}$$

where the coefficients  $\alpha_j$  of the power series representation of  $\pi(z)^{-1}$  decay to zero at a geometric rate as  $j \rightarrow \infty$  for  $|z| \leq b_\pi$  and some  $b_\pi > 1$ . Thus,  $u_t$  is a nonstationary I(1) process.

Finally, note that equation (1) and the conditions in (3) imply that there exist initial values such that the differenced process  $\Delta y_t$  has the two-sided moving average representation

$$\Delta y_t = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j}, \tag{7}$$

where  $\psi_j$  is the coefficient of  $z^j$  in the Laurent series expansion of  $\pi(z)^{-1} \varphi(z^{-1})^{-1} \stackrel{\text{def}}{=} \psi(z)$  so that  $\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$  exists for  $b_\varphi \leq |z| \leq b_\phi$  with  $b_\varphi < 1 < b_\phi$  defined earlier and with  $\psi_j$  decaying to zero at a geometric rate as  $|j| \rightarrow \infty$ . The representation (7) implies that  $\Delta y_t$  is a stationary and ergodic process with finite second moments. Hence, the invariance principle and weak convergence results of sample covariance matrices given in Phillips (1988) apply to  $y_t$  for any (random or nonrandom) initial value  $y_0$ . This implies that the usual asymptotic results needed to develop limit theory for unit root tests are available. To simplify presentation, we assume that, under the null hypothesis, the processes  $\Delta y_t$  and  $\Delta u_t$  are stationary and not only asymptotically stationary.

We derive a unit root test in a likelihood framework similar to that in L&S [for the employed assumptions, see also Andrews *et al.* (2006)]. Thus, we impose the following assumption on the error term in (1).

**Assumption 1.** The zero mean error term  $\epsilon_t$  is a sequence of non-Gaussian i.i.d. random variables with a (Lebesgue) density  $\sigma^{-1} f(\sigma^{-1}x; \lambda)$ , which depends on the (finite and positive) error variance  $\sigma^2$  and (possibly) on the parameter vector  $\lambda$  ( $d \times 1$ ) taking values in an open set  $\Lambda \subseteq \mathbb{R}^d$ .

As discussed in Breidt *et al.* (1991), Rosenblatt (2000, pp. 10–11), L&S and others, causal and noncausal autoregressions are statistically indistinguishable if the error term (and hence the observed process) is Gaussian. This explains why Assumption 1 includes the requirement of non-Gaussian errors. Further assumptions on the density function  $f(x; \lambda)$  will be made later.

We close this section with a remark on the conceivable possibility of testing for a unit root in the noncausal polynomial  $\varphi(\cdot)$ . As equation (4) and the subsequent discussion indicate, a possible unit root in the causal polynomial  $\phi(\cdot)$  makes the testing problem conceptually very similar to its previous counterpart, where the existence of a unit root means that  $y_t$ , the value of the considered process at time  $t$ , can be expressed as a sum of the current and past values of a stationary process and an initial value  $y_0$ . If a unit root were in the noncausal polynomial  $\varphi(\cdot)$ , the counterpart of this would (presumably) be that  $y_t$  should be expressed as a sum of the current and future values of a stationary process. However, without truncation, such a sum does not converge and, therefore, cannot be used to define a process for all  $t > 0$ . For purposes of unit root testing, one could truncate the sum at the last value of the considered series,  $y_T$  say, although such an approach may not lend itself a natural interpretation. A potential technical difficulty is that conventional invariance principles are not directly applicable to the resulting process and its functions, such as the components of the score and Hessian of the log-likelihood function involving the unit root parameter, implying that the problem of testing for a unit root in the noncausal polynomial may lead to a rather involved asymptotic distribution theory. In this article, we therefore confine ourselves to the case where a unit root appears in the causal AR polynomial.

### 3. PARAMETER ESTIMATION

#### 3.1. Approximate likelihood function

To obtain our tests, we first discuss the likelihood function based on the observed time series  $\{y_1, \dots, y_T\}$  generated by the AR( $r, s$ ) process (1). Proceeding in the same way as in Section 3.1 of L&S suggests approximating the log-likelihood function by

$$l_T(\theta) = \sum_{t=r+1}^{T-s} g_t(\theta), \tag{8}$$

where

$$\begin{aligned} g_t(\theta) &= \log f(\sigma^{-1}(\Delta u_t(\varphi) - \phi u_{t-1}(\varphi) - \pi_1 \Delta u_{t-1}(\varphi) - \dots - \pi_{r-1} \Delta u_{t-r+1}(\varphi)); \lambda) - \log \sigma \\ &= \log f(\sigma^{-1}(v_t(\phi, \pi) - \varphi_1 v_{t+1}(\phi, \pi) - \dots - \varphi_s v_{t+s}(\phi, \pi)); \lambda) - \log \sigma. \end{aligned}$$

Here,  $u_t(\varphi)$  and  $v_t(\phi, \pi)$  signify the series  $u_t = \varphi(B^{-1})y_t$  and  $v_t = \phi(B)y_t$ , respectively, treated as functions of the parameters  $\varphi = (\varphi_1, \dots, \varphi_s)$  and  $(\phi, \pi) = (\phi, \pi_1, \dots, \pi_{r-1})$ , and the parameter vector  $\theta = (\phi, \pi, \varphi, \sigma, \lambda)$   $((r + s + 1 + d) \times 1)$  contains the parameters of the model. Maximizing  $l_T(\theta)$  over permissible values of  $\theta$  gives an (approximate) ML estimator of  $\theta$ . In what follows, we drop the word ‘approximate’ from the ML estimator and related quantities.

Earlier, we assumed unrealistically that the orders of the model,  $r$  and  $s$ , are known. As in Breidt *et al.* (1991) and L&S, we specify these orders in practice as follows. First, we fit a conventional causal AR model by LS and determine its order by using conventional procedures such as model selection criteria and residual diagnostics. We deem a causal model adequate when its residuals show no signs of autocorrelation. Owing to the aforementioned identifiability issue, we also need to check for the non-Gaussianity of the residuals because otherwise there is no point to consider noncausal models. If non-Gaussianity is supported by the data, a non-Gaussian error distribution is adopted, and all causal and noncausal models of the selected order are estimated. Of these models, the one that maximizes the likelihood function is selected, and its adequacy is evaluated by conventional diagnostic tools.

In practice, a purely noncausal model ( $r = 0, s > 0$ ) may turn out to be the most appropriate choice, but owing to the assumption  $r > 0$ , it is not in accordance with the assumed formulation. If one wants to perform a formal test in a case like this, one may augment the model with a first-order causal polynomial and base the test on the AR(1,  $s$ ) model.

### 3.2. Score vector and Hessian matrix

As our goal is to derive a Wald-type test for the unit root hypothesis, we have to assume that the likelihood function satisfies conventional differentiability conditions similar to those used in the related previous work of Andrews *et al.* (2006) and L&S. Thus, we impose the following assumption.

**Assumption 2.** For all  $(x, \lambda) \in (\mathbb{R}, \Lambda)$ ,  $f(x; \lambda) > 0$  and  $f(x; \lambda)$  is twice continuously differentiable with respect to  $(x, \lambda)$  and an even function of  $x$ , that is,  $f(x; \lambda) = f(-x; \lambda)$ .

Unlike the aforementioned previous authors, we require that the function  $f(\cdot; \lambda)$  is even. As will be discussed in Section 4.1, this assumption is imposed to simplify the limiting distribution of the obtained unit root test. However, in Appendix B, we derive the asymptotic distribution of our unit root test when this assumption is relaxed. These derivations make evident that this limiting distribution is of no or only little practical use. For cases where a skewed error distribution is expected to be plausible, a bootstrap procedure is suggested to obtain an approximation to the asymptotic distribution of our test. An example of such a bootstrap procedure is outlined in Section 5.2.

For the derivation of the Wald-type test, we need to estimate the unrestricted model and derive the limiting distribution of the ML estimator of  $\phi$  under the null hypothesis. Because the data are assumed to be generated by a nonstationary I(1) process, the derivation of the limiting distribution of the ML estimator involves features different from those in the previous literature on stationary NCAR models. Moreover, as the estimation problem is nonlinear, the presence of an I(1) process implies that methods used in the context of conventional unit root tests based on linear LS estimation are not directly applicable. Therefore, we use ideas similar to those developed for likelihood-based statistical models whose estimation theory is nonstandard in the sense that the information matrix is random even asymptotically. Such nonergodic models are discussed in Basawa and Scott (1983) & Jeganathan (1995) among others, and to facilitate their treatment, we introduce the notation  $\theta_0$  for the true value of  $\theta$  and similarly for its components. As the null hypothesis is assumed to hold, the true value of  $\phi$  is zero.

We shall now derive weak limits of (appropriately standardized versions of) the score vector and Hessian matrix associated with the log-likelihood function evaluated at the true parameter value. We use a subscript to signify a partial derivative indicated by the subscript; for instance,  $g_{\theta,t}(\theta) = \partial g_t(\theta) / \partial \theta$ ,  $f_x(x; \lambda) = \partial f(x; \lambda) / \partial x$ , and  $f_\lambda(x; \lambda) = \partial f(x; \lambda) / \partial \lambda$ . Denote  $V_{t+1} = (v_{t+1}, \dots, v_{t+s})$  and  $\Delta U_{t-1} = (\Delta u_{t-1}, \dots, \Delta u_{t-r+1})$ , where  $v_t$  are  $\Delta u_t$  have the representations (5) and (6) with the coefficients replaced by their true values  $\beta_{0,j}$  and  $\alpha_{0,j}$  so

that the latter, for example, is obtained from  $\pi_0(z)^{-1} = \sum_{j=0}^{\infty} \alpha_{0,j} z^j$ . The first and second partial derivatives of  $g_t(\theta)$ , the log-likelihood function based on a single observation, are presented in Appendix A. When evaluated at the true parameter value, the vector of first partial derivatives is

$$g_{\theta,t}(\theta_0) = \begin{bmatrix} g_{\phi,t}(\theta_0) \\ g_{\pi,t}(\theta_0) \\ g_{\varphi,t}(\theta_0) \\ g_{\sigma,t}(\theta_0) \\ g_{\lambda,t}(\theta_0) \end{bmatrix} = \begin{bmatrix} -\sigma_0^{-1} e_{x,t} u_{t-1} \\ -\sigma_0^{-1} e_{x,t} \Delta U_{t-1} \\ -\sigma_0^{-1} e_{x,t} V_{t+1} \\ -\sigma_0^{-2} (e_{x,t} \epsilon_t + \sigma_0) \\ e_{\lambda,t} \end{bmatrix},$$

where  $e_{x,t} = f_x(\sigma_0^{-1} \epsilon_t; \lambda_0) / f(\sigma_0^{-1} \epsilon_t; \lambda_0)$  and  $e_{\lambda,t} = f_{\lambda}(\sigma_0^{-1} \epsilon_t; \lambda_0) / f(\sigma_0^{-1} \epsilon_t; \lambda_0)$ .

To obtain the weak limit of the score, we have to assume that the error density  $f(x; \lambda)$  satisfies regularity conditions such as those employed by Andrews *et al.* (2006) and L&S. Rather than presenting the needed conditions explicitly, we simplify the presentation by using suitable ‘high level’ assumptions that can be verified by using the regularity conditions given in the aforementioned papers. To this end, it is convenient to write  $\theta = (\phi, \vartheta) = (\phi, \vartheta_1, \vartheta_2)$ , where  $\vartheta_1 = (\pi, \varphi)$  and  $\vartheta_2 = (\sigma, \lambda)$ . The score of  $\vartheta$  (evaluated at  $\theta_0$ ) is clearly a stationary and ergodic process similar to the score in L&S. We make the following assumption.

**Assumption 3.** (i)  $E[e_{x,t}] = 0$  and  $E[e_{x,t}^2] = \mathcal{J}$ , where  $\mathcal{J} = \int (f_x(x; \lambda_0)^2 / f(x; \lambda_0)) dx > 1$  is finite. Moreover,  $\text{Cov}[\epsilon_t, e_{x,t}] = -\sigma_0$ .  
 (ii) The score vector  $g_{\vartheta,t}(\theta_0) = (g_{\vartheta_1,t}(\theta_0), g_{\vartheta_2,t}(\theta_0))$  has zero expectation and finite positive definite covariance matrix  $\Sigma = \text{diag}(\Sigma_1, \Sigma_2)$ , where  $\Sigma_i = \text{Cov}[g_{\vartheta_i,t}(\theta_0)]$  ( $i = 1, 2$ ) and the partition is conformable to that of  $g_{\vartheta,t}(\theta_0)$ .

Part (i) of this assumption can be verified by using the definition of  $e_{x,t}$ , the regularity conditions in Andrews *et al.* (2006) and L&S, and direct calculation. Specifically, the expression of  $\text{Cov}[\epsilon_t, e_{x,t}]$  is obtained from the definition of  $e_{x,t}$  and condition (A2) of these papers, whereas condition (A5) implies that the inequality  $\mathcal{J} > 1$  holds if and only if the distribution of  $\epsilon_t$  is non-Gaussian. This inequality and the explicit expressions of the matrices  $\Sigma_1$  and  $\Sigma_2$  obtained from L&S can further be used to verify the positive definiteness of the covariance matrix  $\Sigma_1$  in part (ii), whereas, owing to the generality of the error distribution, the positive definiteness of  $\Sigma_2$  has to be assumed. The other conditions in part (ii) can be verified by using the regularity conditions imposed on the density function  $f(x; \lambda)$  in the aforementioned papers.

Assumption 3(i) and a standard functional central limit theorem for i.i.d. sequences yield

$$T^{-1/2} \sum_{t=1}^{[Tu]} (e_{x,t}, \epsilon_t) \xrightarrow{d} (B_{e_x}(u), B_{\epsilon}(u)) \sim \text{BM} \left( \begin{bmatrix} \mathcal{J} & -\sigma_0 \\ -\sigma_0 & \sigma_0^2 \end{bmatrix} \right), \tag{9}$$

where the covariance matrix is positive definite when  $\epsilon_t$  is non-Gaussian. Using Assumptions 1–3, we can further derive the limiting distribution of the score vector of  $\theta$ . The result is presented in the Lemma 1.

**Lemma 1.** Suppose that Assumptions 1–3 hold. Then,

$$T^{-1} \sum_{t=r+1}^{T-s} g_{\phi,t}(\theta_0) \xrightarrow{d} Z_1 = -\frac{1}{\sigma_0 \pi_0(1)} \int_0^1 B_{\epsilon}(u) dB_{e_x}(u) \tag{10}$$

and

$$T^{-1/2} \sum_{t=r+1}^{T-s} g_{\vartheta,t}(\theta_0) \xrightarrow{d} Z_2 \sim N(0, \Sigma). \tag{11}$$

Moreover, joint weak convergence applies with  $Z_1$  and  $Z_2$  independent.

The proof of this lemma is presented in Appendix B. As discussed therein, the requirement that the function  $f(\cdot; \lambda)$  is even is needed to establish the independence statement (further discussion on this issue will be given at the end of Section 4.1).

Next, consider the Hessian matrix associated with the log-likelihood function  $l_T(\theta)$ . Expressions for the required second partial derivatives are obtained from Appendix A. Similar to the first partial derivatives, we use notations such as  $g_{\theta\theta,t}(\theta) = \partial^2 g_t(\theta) / \partial\theta\partial\theta'$ ,  $f_{xx}(x; \lambda) = \partial^2 f(x; \lambda) / \partial x^2$  and  $f_{x\lambda}(x; \lambda) = \partial^2 f(x; \lambda) / \partial x \partial \lambda$ . We also define

$$e_{xx,t} = \frac{f_{xx}(\sigma_0^{-1}\epsilon_t; \lambda_0)}{f(\sigma_0^{-1}\epsilon_t; \lambda_0)} - e_{x,t}^2$$

and

$$e_{\lambda x,t} = \frac{f_{\lambda x}(\sigma_0^{-1}\epsilon_t; \lambda_0)}{f(\sigma_0^{-1}\epsilon_t; \lambda_0)} - \frac{f_{\lambda}(\sigma_0^{-1}\epsilon_t; \lambda_0)}{f(\sigma_0^{-1}\epsilon_t; \lambda_0)} e_{x,t},$$

and make the following assumption.

**Assumption 4.**  $E[e_{xx,t}] = -E[e_{x,t}^2]$  and  $E[g_{\vartheta\vartheta,t}(\theta_0)] = -\Sigma$  with  $\Sigma$  given in Assumption 3(ii). Moreover,  $E[e_{xx,t}\epsilon_t] = 0$  and  $E[e_{\lambda x,t}] = 0$ .

Similar to Assumption 3, this assumption can be verified by using the regularity conditions in Andrews *et al.* (2006) and L&S. The first moment equality is obtained from Assumption (A3) of these papers, whereas the second one states that the negative of the Hessian matrix of the log-likelihood function with respect to the short-run parameter  $\vartheta$  equals the covariance matrix of the score of  $\vartheta$ , a fact that can be established by direct calculation (see L&S). As for the last two moment conditions, both  $e_{xx,t}\epsilon_t$  and  $e_{\lambda x,t}$  are odd functions of  $\epsilon_t$  so that, given Assumption 2, only finiteness of the expectations is required. This in turn can be obtained from condition (A7) of Andrews *et al.* (2006) and L&S.

Now we can prove Lemma 2.

**Lemma 2.** Suppose that Assumptions 1–4 hold. Then,

$$-T^{-2} \sum_{t=r+1}^{T-s} g_{\phi\phi,t}(\theta_0) \xrightarrow{d} \frac{\mathcal{J}}{\sigma_0^2 \pi_0 (1)^2} \int_0^1 B_{\epsilon}^2(u) d(u) \stackrel{\text{def}}{=} g_{\phi\phi}(\theta_0), \tag{12}$$

$$-T^{-1} \sum_{t=r+1}^{T-s} g_{\vartheta\vartheta,t}(\theta_0) \xrightarrow{p} \Sigma, \tag{13}$$

and

$$-T^{-3/2} \sum_{t=r+1}^{T-s} g_{\phi\vartheta,t}(\theta_0) \xrightarrow{P} 0. \tag{14}$$

Moreover, the weak convergences in (12) and in Lemma 1 hold jointly, and  $g_{\phi\phi}(\theta_0)$  and  $Z_2$  are independent.

Using the limits obtained in Lemmas 1 and 2, we define  $Z = (Z_1, Z_2)$  and  $G(\theta_0) = \text{diag}(g_{\phi\phi}(\theta_0), \Sigma)$ , and we also introduce the matrix  $D_T = \text{diag}(T, T^{1/2}I_{r+s+d})$ . The following proposition is an immediate consequence of Lemmas 1 and 2.

**Proposition 1.** Suppose that Assumptions 1–4 hold. Then,

$$S_T(\theta_0) \stackrel{\text{def}}{=} D_T^{-1} \sum_{t=r+1}^{T-s} g_{\theta,t}(\theta_0) \xrightarrow{d} Z \tag{15}$$

and

$$G_T(\theta_0) \stackrel{\text{def}}{=} -D_T^{-1} \sum_{t=r+1}^{T-s} g_{\theta\theta,t}(\theta_0) D_T^{-1} \xrightarrow{d} G(\theta_0), \tag{16}$$

where the weak convergences in (15) and (16) hold jointly with  $(Z_1, G(\theta_0))$  and  $Z_2$  independent.

In the next section, we derive the limiting distribution of the ML estimator of the parameter  $\theta$  by using Proposition 1 and arguments similar to those used by Basawa and Scott (1983, Ch. 2.4) in the context of statistical models whose likelihood ratios satisfy the LAMN condition.

### 3.3. Limiting distribution of the ML estimator

To obtain the limiting distribution of the ML estimator of the parameter  $\theta$ , we have to supplement the assumptions made so far by conditions on the standardized Hessian matrix  $G_T(\theta) \stackrel{\text{def}}{=} -D_T^{-1} \sum_{t=r+1}^{T-s} g_{\theta\theta,t}(\theta) D_T^{-1}$ . A sufficient ‘high level’ condition, used by Basawa and Scott (1983, pp. 33–34) in a more general form, requires that, for all  $c > 0$ ,

$$\sup_{\theta \in N_{T,c}} \|G_T(\theta) - G_T(\theta_0)\| \xrightarrow{P} 0, \tag{17}$$

where  $N_{T,c} = \{\theta : D_T \|\theta - \theta_0\| \leq c\}$ . As discussed in Appendix C, this condition can be verified by using assumptions similar to those used by Lii and Rosenblatt (1996) in the context of (stationary) noncausal and noninvertible ARMA models and by Meitz and Saikkonen (2013) in the context of a (stationary) noninvertible ARMA model with conditionally heteroskedastic errors. Proposition 1 combined with condition (17) enables us to establish the limiting distribution of the ML estimator of  $\theta$  under the unit root hypothesis.

**Proposition 2.** Suppose that Assumptions 1–4 and condition (17) hold. Then, with probability approaching one, there exists a sequence of local maximizers of the log-likelihood function  $\hat{\theta}_T = (\hat{\phi}_T, \hat{\vartheta}_T)$  such that

$$\left( D_T(\hat{\theta}_T - \theta_0), G_T(\theta_0) \right) \xrightarrow{d} \left( G(\theta_0)^{-1} Z, G(\theta_0) \right).$$

Moreover,  $G_T(\hat{\theta}_T) - G_T(\theta_0) \xrightarrow{P} 0$ .



Proposition 2 can be proved along the same lines as Theorems 1 and 2 of Basawa and Scott (1983, pp. 56–59). An outline of the needed arguments is provided in Appendix B. Now, all ingredients for the derivation of our unit root tests are available.

#### 4. TEST PROCEDURES

##### 4.1. Test statistic

With Proposition 2 at hand, it is straightforward to derive Wald-type unit root tests. As we are interested in one-sided (stationary) alternatives, we use a ‘*t*-ratio’ type test statistic defined as

$$\tau_T \stackrel{def}{=} \frac{T \hat{\phi}_T}{\sqrt{G_T^{1,1}(\hat{\theta}_T)}},$$

where  $G_T^{1,1}(\hat{\theta}_T)$  abbreviates the (1,1)-element of  $G_T(\hat{\theta}_T)^{-1}$ . The following proposition presents the asymptotic distribution of  $\tau_T$ .

**Proposition 3.** Suppose that Assumptions 1–4 and condition (17) hold. Then

$$\tau_T \xrightarrow{d} \left( \mathcal{J} \int_0^1 W_\epsilon^2(u) \, d(u) \right)^{-1/2} \left( \int_0^1 W_\epsilon(u) \, dW_\epsilon(u) - (\mathcal{J} - 1)^{1/2} \int_0^1 W_\epsilon(u) \, dW(u) \right) \stackrel{def}{=} \tau(\mathcal{J}), \quad (18)$$

where  $W_\epsilon(u) = \sigma_0^{-1} B_\epsilon(u) \sim \text{BM}(1)$  and  $W(u) \sim \text{BM}(1)$  is independent of  $W_\epsilon(u)$ .

To see how this result can be obtained, note that Proposition 2 and the continuous mapping theorem yield

$$\tau_T \xrightarrow{d} - \left( \mathcal{J} \int_0^1 B_\epsilon^2(u) \, d(u) \right)^{-1/2} \int_0^1 B_\epsilon(u) \, dB_{e_x}(u).$$

The stated result is obtained by replacing the Brownian motion  $B_{e_x}(u)$  on the right-hand side by the expression

$$B_{e_x}(u) = -\sigma_0^{-1} B_\epsilon(u) + (\mathcal{J} - 1)^{1/2} W(u) = -W_\epsilon(u) + (\mathcal{J} - 1)^{1/2} W(u),$$

obtained via a Cholesky decomposition of the covariance matrix in (9).

Proposition 3 implies that the limiting distribution of test statistic  $\tau_T$  is free of nuisance parameters except for the parameter  $\mathcal{J}$ . For subsequent analysis and discussions, we notice that for Student’s *t*-distributed errors with  $\lambda > 2$  degrees of freedom,

$$\mathcal{J} = \frac{\lambda(\lambda + 1)}{(\lambda - 2)(\lambda + 3)}. \quad (19)$$

Of course, the obtained limiting distribution is of limited practical use because it depends on the nuisance parameter  $\mathcal{J}$ . Fortunately, this problem is rather easily circumvented and is further discussed in Section 5.1.

The distribution of the limiting variable  $\tau(\mathcal{J})$  is a weighted average of a standard normal distribution and a Dickey–Fuller type of distribution. More specifically, letting  $\mathcal{J} \rightarrow \infty$  in (18), a standard normal distribution is obtained, as

$$\lim_{\mathcal{J} \rightarrow \infty} \tau(\mathcal{J}) = \left( \int_0^1 W_\epsilon^2(u) d(u) \right)^{-1/2} \int_0^1 W_\epsilon(u) dW(u) = \eta \sim N(0, 1),$$

where the second equality holds true because  $\int_0^1 W_\epsilon(u) dW(u)$  is a scale mixture of normal distributions and can be written as  $\int_0^1 W_\epsilon(u) dW(u) = \left( \int_0^1 W_\epsilon^2(u) d(u) \right)^{1/2} \eta$ . On the other hand, letting  $\mathcal{J} \rightarrow 1$  in (18), the Dickey–Fuller type of distribution is obtained as

$$\lim_{\mathcal{J} \rightarrow 1} \tau(\mathcal{J}) = \left( \int_0^1 W_\epsilon^2(u) d(u) \right)^{-1/2} \int_0^1 W_\epsilon(u) dW_\epsilon(u).$$

That the limiting distribution of  $\tau_T$  is relatively simple, depending only on the nuisance parameter  $\mathcal{J}$ , is achieved by assuming that the function  $f(\cdot; \lambda)$  is even. This assumption is used to establish the independence of  $g_{\phi\phi}(\theta_0)$  and  $Z_2$  in Lemma 2 and further the independence of  $(Z_1, G(\theta_0))$  and  $Z_2$  in Proposition 1, and it is also used to justify the block diagonality of  $G(\theta_0)$  (see the proof of Lemma 2 for some details). If these results do not hold, the limiting distribution of  $\tau_T$  will be a considerably more complicated function of the short-run parameters of the model (Appendix B), making the implementation of the resulting test very difficult.

#### 4.2. Tests allowing for deterministic terms

The result of Proposition 3 only applies to mean-zero data. To accommodate series with trend components, we consider the model

$$x_t = \mu + \delta t + y_t, \quad t = 1, 2, \dots,$$

where  $x_t$  is the observed time series and  $y_t$  is a noncausal  $AR(r, s)$  process. The trend coefficients  $\mu$  and  $\delta$  are estimated by LS to obtain the estimates  $\hat{\mu}$  and  $\hat{\delta}$  after which the test statistic  $\tau_T$  introduced in the preceding section is formed by using  $y_t = x_t - \hat{\mu}$  in the case of demeaned data and  $y_t = x_t - \hat{\mu} - \hat{\delta}t$  in the case of detrended data. As in other unit root tests, the distribution of the resulting test statistic depends on the trend component chosen, and therefore, we denote the test statistic by  $\tau_T(m)$ , where  $m = 0$ ,  $m = 1$  and  $m = 2$  refer to mean-zero, demeaned, and detrended data, respectively. The result of Proposition 3 applies even for  $\tau_T(1)$  and  $\tau_T(2)$  as long as the Brownian motion  $W_\epsilon(u)$  is replaced by a corresponding detrended Brownian motion [see, e.g. Park and Phillips (1988)].

### 5. SIMULATION STUDIES

#### 5.1. Estimated critical values

The problem of the nuisance parameter  $\mathcal{J} (\in (0, \infty))$  appearing in the limiting distribution of test statistic  $\tau_T(m)$  is addressed next. We shall first illustrate how the value of the parameter  $\mathcal{J}$  affects the distribution of  $\tau(\mathcal{J})$  [see (18)]. It turns out to be convenient to study this effect by using the correlation between the two Brownian motions  $B_\epsilon(u)$  and  $B_{e_x}(u)$ , that is,  $\rho = \mathcal{J}^{-1/2} \in (0, 1)$  [see (9)]. Figure 1 displays the 1% (dotted lines), 5% (dashed lines) and 10% (dashed-dotted lines) percentiles of the distribution of  $\tau(\mathcal{J})$  as a function of  $\rho$ .

TESTING FOR A UNIT ROOT IN NCAR MODELS

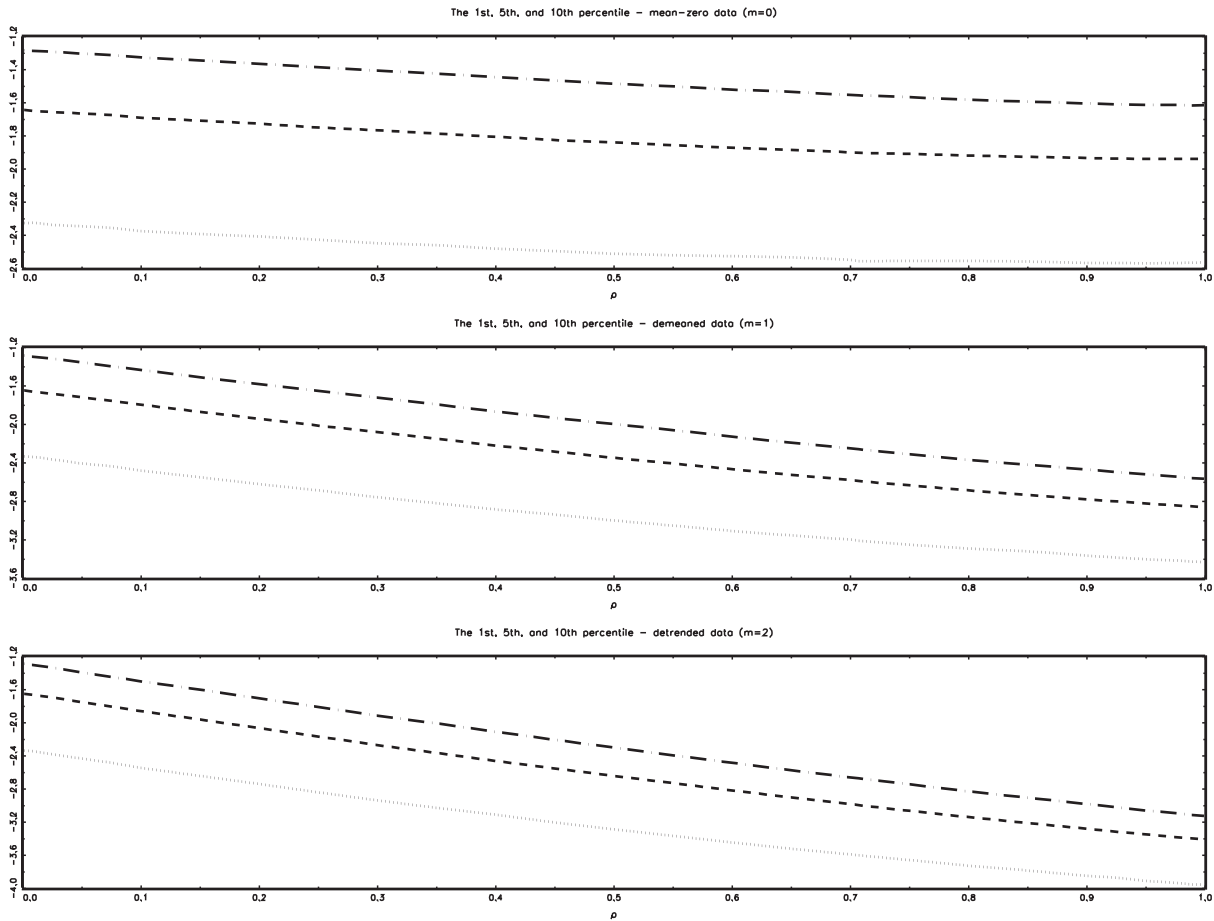


Figure 1. Percentiles of the distribution of  $\tau(\mathcal{J})$  as a function of  $\rho = \mathcal{J}^{-1/2}$ . Notes: first percentiles (dotted lines), fifth percentiles (dashed lines), and 10th percentiles (dash-dotted lines) for the asymptotic distribution of the  $\tau_T(m)$  statistic. The Brownian motions appearing in the limiting distribution of test statistic  $\tau_T(m)$  are approximated using (appropriately scaled) sums of normal IID(0, 1) variables with  $T = 5000$  and 500,000 replications.

In Figure 1, a monotonically decreasing relationship between the percentiles and  $\rho$  is seen. As already mentioned, the Dickey–Fuller distributions and the standard normal distribution are obtained as limiting cases by letting  $\mathcal{J} \rightarrow 1$  ( $\rho \rightarrow 1$ ) and  $\mathcal{J} \rightarrow \infty$  ( $\rho \rightarrow 0$ ), respectively. Thus, in Figure 1, the 1%, 5% and 10% critical values for the DF statistics and a standard normal variate are found at the leftmost and rightmost sides, respectively.

Owing to the monotonicity of the percentiles in  $\rho$ , it is obvious that if the value of  $\mathcal{J}$  were known, Figure 1 could be used to determine (conventional) critical values. Taking a more rigorous approach, we proceed instead with curve estimation of the percentiles by fitting a second-order polynomial  $cv_{\alpha,m}(\rho) = b_0 + b_1\rho + b_2\rho^2$  for  $\alpha \in \{0.01, 0.05, 0.10\}$  and  $m \in \{0, 1, 2\}$ . The curve estimates, obtained by LS, yield the coefficients in Table I that can be used to compute asymptotic critical values.

To exemplify how Table I can be used, assume that we wish to test the unit root hypothesis in the NCAR model at a 10% significance level in the case of demeaned data with  $\mathcal{J} = 2$  ( $\rho = 1/\sqrt{2}$ ). Then, the estimated asymptotic critical value equals  $cv_{0.10,1}(2) = -1.276 - 1.584 \times (1/\sqrt{2}) + 0.289 \times (1/\sqrt{2})^2 = -2.252$ . To this end, the value of  $\mathcal{J}$  is in practice obviously not known and must be estimated. In the case of Student’s  $t$ -distributed errors,

Table I. Coefficients to compute asymptotic critical values  $cv_{\alpha,m}(\rho)$  of test statistic  $\tau_T(m)$

Case	Significance level ( $\alpha$ ) (%)	$b_0$	$b_1$	$b_2$	$R^2$
Mean-zero data ( $m = 0$ )	1	-2.321	-0.492	0.251	0.998
	5	-1.639	-0.495	0.187	0.999
	10	-1.276	-0.480	0.131	0.999
Demeaned data ( $m = 1$ )	1	-2.322	-1.578	0.474	1.00
	5	-1.639	-1.591	0.367	1.00
	10	-1.276	-1.584	0.289	1.00
Detrended data ( $m = 2$ )	1	-2.324	-2.201	0.575	1.00
	5	-1.640	-2.230	0.462	1.00
	10	-1.276	-2.231	0.381	1.00

Note: For each significance level and each trend specification, the coefficients  $b_0$ ,  $b_1$  and  $b_2$  are obtained from the regression of  $cv_{\alpha,m}(\rho)$  on  $(1, \rho, \rho^2)$  (using LS).  $R^2$  is the regression coefficient of determination.

we can use equation (19) with the estimator  $\hat{\lambda}$  used in place of  $\lambda$ . More generally, in cases where the distribution of the error term comprises less straightforward calculations of  $\mathcal{J}$ , we may, by virtue of Assumption 3(i), use the estimator

$$\hat{\mathcal{J}} = \frac{1}{T-r-s} \sum_{t=r+1}^{T-s} \left[ \frac{f_x(\hat{\sigma}^{-1}\hat{\epsilon}_t; \hat{\lambda})}{f(\hat{\sigma}^{-1}\hat{\epsilon}_t; \hat{\lambda})} \right]^2, \tag{20}$$

where  $\hat{\epsilon}_t = \Delta \hat{u}_t - \hat{\phi} \hat{u}_{t-1} - \hat{\pi}_1 \Delta \hat{u}_{t-1} - \dots - \hat{\pi}_{r-1} \Delta \hat{u}_{t-r+1}$  with  $\hat{u}_t = \hat{\phi}(B^{-1})y_t$ .

### 5.2. Bootstrapped $p$ -values and critical values

As already mentioned, if the symmetry condition in Assumption 2 is relaxed, the limiting distributions of our unit root tests depend on several nuisance parameters in a very complex way (for details, see Appendix B). In this section, we discuss a bootstrap procedure that can be used to obtain approximations to the critical values and  $p$ -values of our tests that do not rely on the symmetry condition of Assumption 2. Our approach closely follows the bootstrap procedure described in Caner and Hansen (2001).

The bootstrap distribution of test statistic  $\tau_T$  ( $= \tau_T(1)$ ) is obtained by the following simple steps:

- (i) Use the observed time series  $\{y_1, \dots, y_T\}$  and the assumed distribution for the error term  $\epsilon_t$  to compute  $\hat{\theta} = (\hat{\phi}, \hat{\pi}, \hat{\phi}, \hat{\sigma}, \hat{\lambda})$ , the unrestricted ML estimate of  $\theta$ , and furthermore, the value of the unit root test statistic  $\tau_T$ .
- (ii) Generate  $T^b$  random draws  $\{\epsilon_1^b, \epsilon_2^b, \dots, \epsilon_{T^b}^b\}$  from the estimated error distribution with density  $\hat{\sigma}^{-1} f(\hat{\sigma}^{-1}\epsilon_t; \hat{\lambda})$  and insert these draws and the estimate  $(0, \hat{\pi}, \hat{\phi})$  into the NCAR specification (1) to yield

$$\hat{\phi}(B) \hat{\phi}(B^{-1}) y_t^b = \epsilon_t^b, \quad t = 1, 2, \dots, T^b, \tag{21}$$

where  $\hat{\phi}(B) = \Delta - \hat{\pi}_1 \Delta B - \dots - \hat{\pi}_{r-1} \Delta B^{r-1} = 1 - \hat{\phi}_1 B - \dots - \hat{\phi}_r B^r$ , and the last equation defines the coefficients  $\hat{\phi}_1, \dots, \hat{\phi}_r$ , and  $\hat{\phi}(B^{-1}) = 1 - \hat{\phi}_1 B^{-1} - \dots - \hat{\phi}_s B^{-s}$ . The reason for defining the lag polynomial  $\hat{\phi}(B)$  in this way is to ensure that the bootstrap samples obey the null hypothesis of a unit root. A bootstrap sample  $\{y_1^b, y_2^b, \dots, y_{T^b}^b\}$  is obtained via equation (21) by generating first the ‘noncausal’ part as

$$v_t^b = \hat{\varphi}_1 v_{t+1}^b + \dots + \hat{\varphi}_s v_{t+s}^b + \epsilon_t^b, \quad t = T^b, T^b - 1, \dots, 1,$$

where  $v_{T^b+1}^b = \dots = v_{T^b+s}^b = 0$ , and thereafter the ‘causal’ part as

$$y_t^b = \hat{\phi}_1 y_{t-1}^b + \dots + \hat{\phi}_r y_{t-r}^b + v_t^b, \quad t = 1, 2, \dots, T^b,$$

where  $y_{-r+1}^b = \dots = y_0^b = 0$ .

- (iii) Use the bootstrap sample  $\{y_1^b, y_2^b, \dots, y_{T^b}^b\}$  to compute the value of our unit root test statistic denoted by  $\tau_{T^b}^b$ .
- (iv) Repeat the resampling scheme in (ii) and (iii) BR times to yield the bootstrap distribution of the test statistic  $\tau_T$ , from which, for example, approximate bootstrap  $p$ -values can be computed as the average number of times  $\tau_{T^b}^b$  is smaller than  $\tau_T$ .

In practice, the number of bootstrap replications BR is set relatively large in order to obtain reasonable approximations. The number of bootstrap draws  $T^b$  may be set equal to the (effective) sample size. However, to eliminate the effects of the terminal and starting values, one may generate 200 extra observations (say) and discard the first and last 100 observations at the end and beginning of each realization. The properties of this bootstrap procedure in the case of symmetric and skewed errors are examined in the next section.

### 5.3. Empirical size and power simulations

We examine finite sample properties of the  $\tau_T(m)$ -test for  $m \in \{0, 1, 2\}$  by means of simulation experiments. The nominal significance level employed is 5%, and the benchmark process is a noncausal AR process as defined in (1) with  $r = s = 1$ , and with the i.i.d. error term  $\epsilon_t$  having Student’s  $t$ -distribution with degrees of freedom  $\lambda$  equal to 3 and standard deviation  $\sigma$  equal to 0.1. Realizations  $\{y_1, \dots, y_T\}$  from this process are generated as described in step (ii) of the bootstrap scheme (see the preceding section). To eliminate effects of the terminal and starting values, 100 observations at the end and beginning of each realization are discarded. Finally, in all experiments, the true order of the process is assumed known (i.e.  $r = s = 1$ ), and the estimation of the parameter  $\hat{\theta} = (\hat{\phi}, \hat{\phi}_1, \hat{\sigma}, \hat{\lambda})$  is carried out in GAUSS 12 using the Berndt–Hall–Hall–Hausman algorithm in the Constrained Maximum Likelihood (CML) library.

In the first experiment, the empirical size of the  $\tau_T(m)$ -test is examined in the case of Student’s  $t$ -distributed errors when the parameter  $\varphi_1$  is varied and estimated (asymptotic) critical values based on different estimates of  $\mathcal{J}$  are used. The parameter values and sample sizes considered are  $\phi_1 = 1$  ( $\phi = 0$ ),  $\varphi_1 \in \{0.1, 0.5, 0.9\}$  and  $T \in \{100, 250\}$ , respectively. Moreover, all the results in this experiment are based on 10,000 realizations of the  $\{y_1, \dots, y_T\}$  process, and for each realization, 5% critical values are obtained by the second-order polynomials in Table II using (19) with  $\hat{\lambda}$  ( $\hat{\mathcal{J}}_1$ ), the estimate in (20) ( $\hat{\mathcal{J}}_2$ ) and  $\mathcal{J} = 2$  (the true value) as estimates. The outcomes of this experiment are reported in Table II.

In Table II, the reported estimates for  $\mathcal{J}$  are (for each sample size) based on the average number of replications in the case of demeaned data with  $\varphi_1 = 0.5$ . It is seen that these estimates are close to the true value even for moderate samples sizes. For the other cases, the estimates of  $\mathcal{J}$  are similar and therefore omitted. It is further noticed that the empirical size is close to the nominal size for most of the cases considered, and the influence of the parameter  $\varphi_1$  appears to be modest. One exception, though, is for  $T = 100$  and  $\varphi_1 = 0.9$ , where the test is somewhat over-sized so that some cautiousness is required. Taking the results in Table II together, it appears that the asymptotic distributions of the  $\tau_T(m)$ -test, also with the values of  $\mathcal{J}$  estimated, yield reasonable approximations to the finite sample distributions even for relatively small sample sizes, various trend components and a wide range of parameter values for  $\varphi_1$ .

In the second experiment, the empirical size of the  $\tau_T(m)$ -test is examined in the case where the error term has a skewed Student’s  $t$ -distribution but the regular Student’s  $t$ -distribution is (incorrectly) assumed in the test. In this experiment and the subsequent experiments, critical values are based on the estimate  $\hat{\mathcal{J}}_2$  in (20). The skewed  $t$ -distribution employed is the one of Azzalini and Genton (2008), which, in addition to the parameters  $\lambda$  and  $\sigma$ ,

Table II. Empirical size of the  $\tau_T(m)$ -test in the case of a symmetric error distribution

Sample size	$T$	Mean-zero data ( $m = 0$ )			Demeaned data ( $m = 1$ )			Detrended data ( $m = 2$ )		
		$\varphi_1$			$\varphi_1$			$\varphi_1$		
		0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
100	$\mathcal{J} = 2$	0.053	0.053	0.083	0.059	0.059	0.086	0.058	0.056	0.098
	$\widehat{\mathcal{J}}_1 \approx 2.127$	0.052	0.052	0.083	0.054	0.054	0.080	0.056	0.059	0.097
	$\widehat{\mathcal{J}}_2 \approx 2.122$	0.052	0.052	0.083	0.054	0.053	0.080	0.055	0.059	0.097
250	$\mathcal{J} = 2$	0.054	0.047	0.045	0.063	0.061	0.055	0.057	0.052	0.056
	$\widehat{\mathcal{J}}_1 \approx 2.183$	0.053	0.046	0.044	0.058	0.058	0.054	0.056	0.052	0.057
	$\widehat{\mathcal{J}}_2 \approx 2.181$	0.053	0.046	0.044	0.058	0.058	0.054	0.056	0.052	0.057

Note: The results are based on 10,000 replications, and the nominal size of the tests is 5%. Reported estimated values for  $\mathcal{J}$  are based on the average value over the number of replications for each sample size in the case of demeaned data with  $\varphi_1 = 0.5$ .

also includes a skewness parameter  $\alpha_s$ .<sup>1</sup> The skewness parameter  $\alpha_s$  is assumed to take on the values  $\alpha_s = 0$  (symmetric errors),  $\alpha_s = 0.66$  (errors with skewness 1.33) and  $\alpha_s = 2$  (errors with skewness 3). The setup for this experiment is the same as in the first experiment except that, to conserve space, the results with zero mean data are excluded (these results are available upon request from the authors). As before, we let  $\sigma = 0.1$  but choose  $\lambda = 4$  to make the conventional skewness measure well defined. Finally, we also report the empirical size of the bootstrap version of our tests based on the correctly specified skewed error distribution. The bootstrap version of our test, denoted by  $\tau_T^b(m)$ , is based on 500 bootstrap replications and on 1000 Monte Carlo replications. The outcomes of this experiment are reported in Table III.

The results in Table III indicate that, except for the case  $T = 100$  and  $\varphi_1 = 0.90$ , the  $\tau_T(m)$ -test is not very sensitive to violations of the symmetry condition in Assumption 2. From Table III, it is also seen that the performance of the  $\tau_T^b(m)$  test is very satisfactory for all sample sizes and all values of  $\alpha_s$  considered. Thus, one could consider using it always in combination with a distribution allowing for skewed errors. However, limited simulation experiments (results available upon request from the authors) indicate that in the case of symmetric errors, this leads to a slight loss of power compared with using the  $\tau_T$ -test that assumes symmetric errors.

In our third Monte Carlo experiment, the power of the  $\tau_T(m)$ -test is examined. The data are generated as described in our earlier first experiment with  $\varphi_1 = 0.5$  and  $\phi_1 \in [0.6, 1.0]$  ( $\phi \in [-0.4, 0]$ ). The sample sizes considered are  $T \in \{100, 250\}$ . For comparison, we also report the outcomes of the conventional Dickey–Fuller unit root  $t$ -test based on an AR(2) process, the  $t$ -type unit root test of Lucas (1995) based  $M$ -estimation in an AR(1) model and an assumption of strictly stationary strong-mixing errors, and the unit root test of Rothenberg and Stock (1997) based on the ML estimation of an AR(2) model and an assumption of Student’s  $t$ -distributed errors. These tests are denoted by  $\tau_{DF}(m)$ ,  $M(m)$  and  $RS(m)$ , respectively.<sup>2</sup> The  $\tau_{DF}(m)$ -test is a natural alternative to our test in that it is widely used among practitioners, and it has also been shown to be rather robust against various misspecifications. The  $M(m)$ -test can also be viewed as a natural alternative, for

<sup>1</sup> The density of the skewed  $t$ -distribution of Azzalini and Genton (2008) parameterized to have mean zero and variance  $\sigma^2$  takes the following form:

$$f(z_t; -m(\alpha_s, \lambda) s^{-1}(\alpha_s, \lambda), \sigma s^{-1}(\alpha_s, \lambda), \alpha_s, \lambda) = 2s(\alpha_s, \lambda) \sigma^{-1} t(z_t, \lambda) T\left(\alpha_s \xi \sqrt{(\lambda + 1)/(\lambda + \xi^2)}; \lambda + 1\right),$$

where  $m(\alpha_s, \lambda) = \alpha_s(1 + \alpha_s^2)^{-1/2} (\lambda/\pi)^{-1/2} \Gamma((\lambda - 1)/2)/\Gamma(\lambda/2)$ ,  $s^2(\alpha_s, \lambda) = (\lambda/2)\Gamma((\lambda - 2)/2)/\Gamma(\lambda/2) - m^2(\alpha_s, \lambda)$ , and  $z_t = s(\alpha_s, \lambda) \sigma^{-1}(\epsilon_t + \sigma m(\alpha_s, \lambda) s^{-1}(\alpha_s, \lambda))$ . Furthermore,  $t$  and  $T$  denote the Student’s  $t$  density and distribution function, respectively.

<sup>2</sup> Following Lucas (1995), we use the Huber  $\psi$ -function  $\psi(x) = \min\{c, \max(-c, x)\}$  with  $c = 1.345$  to obtain the M-estimator. Furthermore, to operationalize the  $M(m)$ -test, nuisance parameters are estimated by the Newey–West estimator with the lag-truncation parameter set at  $\lceil 4(T/100)^{2/9} \rceil$ . Finally, in the computations of the M-estimator, a scale-free version is used (Lucas, 1995, p. 337), and an iterative weighted LS algorithm similar to the one described in Van Dijk *et al.* (1999, p. 219) is applied.

TESTING FOR A UNIT ROOT IN NCAR MODELS

Table III. Empirical size of the  $\tau_T(m)$ -test and its bootstrap version in the case of skewed errors

Sample size	Test	$\alpha_s = 0$			$\alpha_s = 0.66$			$\alpha_s = 2$		
		$\varphi_1$			$\varphi_1$			$\varphi_1$		
$T$		0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
Demeaned data ( $m = 1$ )										
100	$\tau_T(1)$	0.051	0.053	0.075	0.052	0.057	0.079	0.047	0.056	0.109
	$\tau_T^b(1)$	0.050	0.048	0.051	0.053	0.047	0.051	0.051	0.055	0.053
250	$\tau_T(1)$	0.054	0.046	0.046	0.059	0.043	0.050	0.056	0.047	0.060
	$\tau_T^b(1)$	0.051	0.053	0.049	0.048	0.049	0.052	0.049	0.049	0.054
Detrended data ( $m = 2$ )										
100	$\tau_T(2)$	0.045	0.053	0.087	0.049	0.056	0.071	0.042	0.051	0.075
	$\tau_T^b(2)$	0.047	0.049	0.053	0.052	0.051	0.050	0.046	0.048	0.051
250	$\tau_T(2)$	0.052	0.051	0.056	0.048	0.050	0.048	0.051	0.049	0.059
	$\tau_T^b(2)$	0.045	0.053	0.050	0.047	0.053	0.049	0.054	0.055	0.051

Note:  $\tau_T^b$  signifies the bootstrap version of our test in the case of skewed  $t$ -distributed errors. All results are based on 1000 replications, and the number of bootstrap replications for the  $\tau_T^b$  test is 500. Nominal sizes of the tests are 5%. Estimated critical values for the  $\tau_T$ -test are based on the estimate  $\hat{\mathcal{J}}_2$  in (20).

it is designed to be robust against innovation outliers (fat-tailed distributions). Finally, the  $RS(m)$  is a natural alternative in the sense that it explicitly assumes nonnormal errors. The results of this experiment are summarized in Figure 2.

Figure 2 shows that, in general, the  $\tau_T(m)$ -test is more powerful than the three alternatives considered, and in some cases, its superiority is quite substantial. For instance, in the case of detrended data with  $T = 250$  and  $\phi_1 = 0.95$ , the differences in power between the  $\tau_T(2)$ -test and the  $\tau_{DF}(2)$ ,  $M(2)$  and  $RS(2)$ -tests are (approximately) as large as 0.40, 0.25 and 0.15 units, respectively. The good performance of the  $\tau_T(m)$ -test is of course not surprising because, unlike the other tests considered, the  $\tau_T(m)$ -test is based on the correctly specified NCAR model. In practice, its application requires choosing two orders,  $r$  and  $s$ , as well as specifying the error distribution, which involves pretesting, not taken into account, in our power simulations. This should be kept in mind when one compares the power of the  $\tau_T(m)$ -test with the considered alternatives, especially with the Dickey–Fuller test whose application only requires choosing one AR order.

We also examined the power of the bootstrap version of our tests with the errors having both symmetric and skewed  $t$ -distribution. Results of these experiments are available upon request from the authors. Here, we only note that, overall, the results were similar to those obtained in the symmetric case in Figure 2.

6. EMPIRICAL APPLICATION

We provide an empirical illustration of our test by analysing a Finnish interest rate series (government bonds). These data range from 1988:Q1 to 2012:Q4 (quarterly observations) and yield a sample size of 100 observations. The series, obtained from International Monetary Fund’s *International Financial Statistics*, is shown in Figure 3.

For interest rate series (in general), it is most natural to use demeaned data. But, as the Finnish interest rate series is trending in the sample, we will also consider the case of detrended data. As a first step in our analysis, we fit an  $AR(p)$  model to the data by LS and thereafter check if the residual series appears non-Gaussian. For the case of demeaned data, both the Akaike and Bayesian information criteria select an  $AR(3)$  model, whereas for the case of detrended data, an  $AR(2)$  model is selected by both the Akaike and Bayesian information criteria (the maximum lag considered was  $p_{\max} = \lceil 4(T/100)^{2/9} \rceil = 4$ ). Even though the null hypothesis of no fourth-order remaining serial correlation is not rejected by the Ljung–Box (LB) test for the two residual series

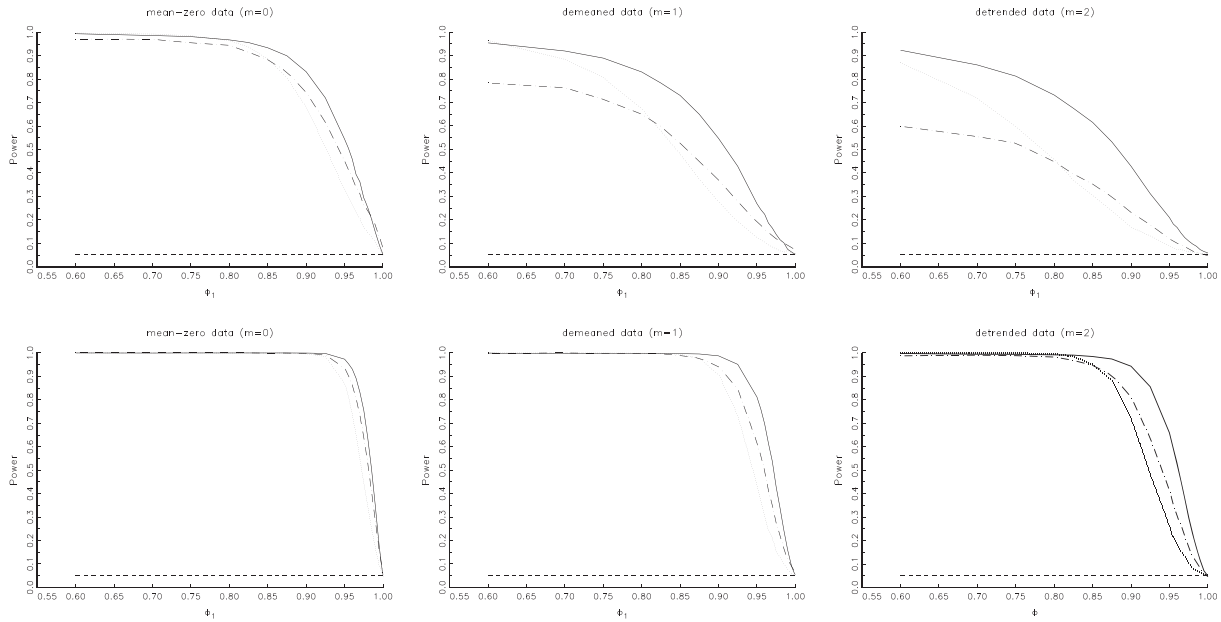


Figure 2. Empirical power of the tests  $\tau_T(m)$ ,  $\tau_{DF}(m)$ ,  $M(m)$  and  $RS(m)$ . Notes:  $\tau_T(m)$ -test, solid line;  $\tau_{DF}(m)$ -test, dotted line;  $M(m)$ -test, dashed-dotted line;  $RS(m)$ -test, dashed line; and nominal size, short-dashed line. The results are based on 10,000 replications, and the nominal size of the tests is 5%. Estimated critical values for the  $\tau_T$ -test are based on the estimate  $\hat{J}_2$  in (20).

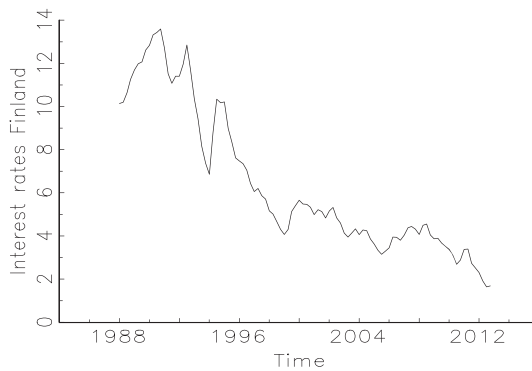


Figure 3. Finnish government bonds

( $p$ -values: 0.492 and 0.811 for demeaned and detrended residual series, respectively), we find that the normality assumption is strongly rejected by the Lomnicki, Jarque and Bera (LJB) test ( $p$ -values:  $<0.001$  and  $<0.001$  for demeaned and detrended residual series, respectively), and some evidence of fourth-order autoregressive conditional heteroskedastic effects are also found by the McLeod–Li (McL) test ( $p$ -values: 0.089 and 0.250 for demeaned and detrended residual series, respectively).<sup>3</sup> In addition, quantile–quantile plots of the residuals of the AR(3) and AR(2) models (not shown here) indicate that a normal distribution is not appropriate because

<sup>3</sup> The skewness part of the LJB test is significant at 7.6% and 8.8 % levels for demeaned and detrended data, respectively, indicating that the rejection of Gaussian errors mainly stems from the kurtosis part of the LJB test.



TESTING FOR A UNIT ROOT IN NCAR MODELS

Table IV. Unit root testing for demeaned and detrended Finnish interest rate series

$T^* = 97$		$m = 1$			
Model	Test	Outcome	$cv_{0.05,1}$	$cv_{0.05,1}(\hat{\rho}_1)$	$cv_{0.05,1}(\hat{\rho}_2)$
AR(3)	$\tau_{DF}$	-1.041	-2.860		
AR(1)	$M$	0.038	-3.060		
AR(3)- $t$	$RS$	0.213	-2.750		
AR(2,1)- $t$	$\tau_T$	-5.101***		-2.558	-2.577
AR(1,2)- $t$	$\tau_T$	-5.997***		-2.542	-2.531
AR(2,1)-St	$\tau_T$	-4.811(<0.001)			
AR(1,2)-St	$\tau_T$	-5.845(<0.001)			
$T^* = 98$		$m = 2$			
Model	Test	Outcome	$cv_{0.05,2}$	$cv_{0.05,2}(\hat{\rho}_1)$	$cv_{0.05,2}(\hat{\rho}_2)$
AR(2)	$\tau_{DF}$	-3.213*	-3.410		
AR(1)	$M$	-1.435	-3.660		
AR(2)- $t$	$RS$	-1.620	-3.280		
AR(1,1)- $t$	$\tau_T$	-5.379***		-3.072	-3.048
AR(1,1)-St	$\tau_T$	-6.482(<0.001)			

Note:  $T^*$  is the effective sample size. AR( $r, s$ ) abbreviates an AR model with  $r$ th-order and  $s$ th-order polynomials  $\phi(B)$  and  $\varphi(B^{-1})$ , respectively.  $N, t$  and St refer to Gaussian,  $t$ -distributed and skewed  $t$ -distributed errors, respectively. \*\*\* and \* denote significance at the 1% and 10% levels, respectively.  $cv_{0.05,m}$  ( $m = 1, 2$ ) is the 5% critical value for the  $\tau_{DF}$ -test, the  $M$ -test, and the  $RS$ -test, and  $cv_{0.05,m}(\hat{\rho}_1)$  and  $cv_{0.05,m}(\hat{\rho}_2)$  ( $m = 1, 2$ ) are the 5% estimated critical values obtained by letting  $\hat{\rho}_1 = \hat{\mathcal{J}}_1^{-1/2}$  and  $\hat{\rho}_2 = \hat{\mathcal{J}}_2^{-1/2}(\hat{\mathcal{J}}_1)$  and  $\hat{\mathcal{J}}_2$  are estimators of  $\mathcal{J}$  using (19) and (20), respectively). In the case of skewed errors, bootstrap  $p$ -values (using 500 bootstrap replications) are reported in parentheses for the  $\tau_T$ -test.

excess kurtosis in the data is left unexplained. Taking these results together, it seems worthwhile to proceed with estimation and unit root testing of NCAR specifications, and to capture the leptokurtic behaviour of the residuals series, we will adopt  $t$ -distributed errors. More specifically, for demeaned and detrended data, we consider an AR( $r, s$ ) model with  $r + s = p = 3$  and  $r + s = p = 2$ , respectively, and conduct unit root testing for the AR(1, 2) and AR(2, 1) specifications in the former case and for the AR(1, 1) specification in the latter case. As will be discussed later, these noncausal models are supported by the specification strategy discussed in Section 3.1. For comparison, we also employ the  $\tau_{DF}$ -test, the  $M$ -test and the  $RS(m)$ -test based on AR(3) and AR(2) models in the case of demeaned and detrended data, respectively. Finally, for the aforementioned NCAR specifications, we will also report the outcomes of our unit root tests when the errors are assumed to have the skewed version of Student's  $t$ -distribution discussed in the preceding section. The outcomes of these unit root tests as well as various estimation results and LB, LJB and McL misspecification tests are reported in Tables IV and V.<sup>4</sup>

In Table IV we see that the  $\tau_T$ -tests based on NCAR specifications give rise to strong rejections (in the case of demeaned data the strongest rejection arises from the NCAR(1, 2) model). On the other hand, the  $\tau_{DF}$ -test fails to reject the unit root hypothesis in the case of demeaned data and yields a rejection only at the 10% level using detrended data, whereas the  $M$  and  $RS$ -tests strongly support the unit root hypothesis irrespective of detrending procedure.

By the results in Table V we conclude that in the case of demeaned data the log-likelihood (LL) is maximized for an AR(1, 2) model with  $t$ -distributed errors (in general), albeit closely followed by an AR(2, 1) model with  $t$ -distributed errors (in general). In the case of detrended data the LL is maximized for an AR(1, 1) model with  $t$ -distributed errors (in general). The maximized LL's of causal and pure noncausal models are substantially lower than those of the aforementioned three NCAR models. Moreover, according to the LB test and the McL test

<sup>4</sup> As discussed in L&S (p. 12), we use least absolute deviation estimators to find starting values for  $\phi, \pi$  and  $\varphi$  ( $\tilde{\phi}, \tilde{\pi}$  and  $\tilde{\varphi}$ , say) and, thereafter maximize  $l_T(\tilde{\phi}, \tilde{\pi}, \tilde{\varphi}, \sigma, \lambda)$  to also find starting values for  $\sigma$  and  $\lambda$ .

Table V. NCAR estimation results for demeaned and detrended Finnish interest rate series

$m = 1$	AR( $r, s$ ) specifications									
	$T^* = 97$	AR(3,0)- $N$	AR(3,0)- $t$	AR(2,1)- $t$	AR(1,2)- $t$	AR(0,3)- $t$	AR(3,0)-St	AR(2,1)-St	AR(1,2)-St	AR(0,3)-St
$\phi$	-0.015 (0.013)	0.003 (0.013)	-0.472 (0.093)	-0.425 (0.071)	0.873 (0.068)	1.550 (0.118)	0.001 (0.014)	-0.463 (0.096)	-0.417 (0.071)	1.576 (0.106)
$\pi_1$	0.568 (0.099)	0.590 (0.091)	0.056 (0.091)	0.056 (0.067)	0.056 (0.067)	-0.695 (0.160)	0.580 (0.099)	0.058 (0.088)	0.052 (0.067)	-0.708 (0.142)
$\pi_2$	-0.180 (0.099)	-0.084 (0.104)	0.941 (0.025)	0.873 (0.068)	0.873 (0.068)	0.130 (0.093)	-0.082 (0.110)	0.938 (0.027)	0.874 (0.066)	0.124 (0.085)
$\varphi_1$						0.461 (0.069)			0.478 (0.132)	0.486 (0.095)
$\varphi_2$						0.461 (0.069)			2.964 (0.971)	3.737 (1.594)
$\varphi_3$						4.129 (1.731)			0.198 (0.516)	-0.534 (0.530)
$\sigma$	0.447 (0.170)	0.457 (0.085)	0.477 (0.126)	0.498 (0.155)	0.498 (0.155)	0.461 (0.069)	0.473 (0.111)	0.462 (0.110)	0.478 (0.132)	0.486 (0.095)
$\lambda$		3.585 (1.356)	3.017 (0.987)	2.831 (0.868)	2.831 (0.868)	4.129 (1.731)	3.381 (1.411)	3.165 (1.083)	2.964 (0.971)	3.737 (1.594)
$\alpha_s$							0.219 (0.513)	0.198 (0.516)	0.180 (0.482)	-0.534 (0.530)
LL	-59.556	-49.111	-46.052	-45.637	-45.637	-53.437	-49.032	-45.975	-45.564	-53.001
LB(4)	0.492	0.324	0.674	0.348	0.348	0.522	0.317	0.712	0.380	0.432
McL(4)	0.089	0.233	0.256	0.306	0.306	<0.001	0.228	0.260	0.300	<0.001
LJB	<0.001									

Table V. Continued

$m = 2$ $T^* = 98$	AR( $r,s$ ) specifications						
	AR(2,0)- $N$	AR(2,0)- $t$	AR(1,1)- $t$	AR(0,2)- $t$	AR(2,0)-St	AR(1,1)-St	AR(0,2)-St
$\phi$	-0.104 (0.032)	-0.052 (0.032)	-0.418 (0.078)		-0.050 (0.032)	-0.423 (0.065)	
$\pi_1$	0.505 (0.086)	0.563 (0.085)			0.565 (0.085)		
$\varphi_1$			0.806 (0.057)	1.445 (0.108)		0.802 (0.048)	1.538 (0.107)
$\varphi_2$				-0.521 (0.103)			-0.551 (0.102)
$\sigma$	0.433 (0.031)	0.442 (0.073)	0.440 (0.091)	0.451 (0.071)	0.442 (0.073)	0.441 (0.086)	0.486 (0.090)
$\lambda$		3.861 (1.581)	3.293 (1.058)	3.966 (1.596)	3.866 (1.563)	3.368 (1.055)	3.737 (1.491)
$\alpha_s$					-0.090 (0.504)	0.668 (0.544)	-0.487 (0.531)
LL	-57.055	-48.998	-43.480	-51.528	-48.982	-42.570	-54.603
LB(4)	0.811	0.358	0.503	0.584	0.350	0.510	0.254
McL(4)	0.250	0.257	0.338	<0.001	0.279	0.349	<0.001
LJB	<0.001						

Note:  $T^*$  designates the effective sample. AR( $r, s$ ) abbreviates an AR model with  $r$ th-order and  $s$ th-order polynomials  $\phi(B)$  and  $\varphi(B^{-1})$ , respectively.  $N, t$  and St refer to Gaussian,  $t$ -distributed and skewed  $t$ -distributed errors, respectively. The figures in parentheses are standard errors computed using (16) and taking the square root of the diagonal elements of the matrix  $D_T^{-1}G_T(\hat{\theta}_T)^{-1}D_T^{-1}$ . LL denotes the value of the log-likelihood evaluated at the ML estimates. LB(4) and McL(4) signify the LB test and the McLeod-Li test with four lags, and LJB is the Lomnicki, Jarque and Bera normality test. The  $p$ -value of these tests are reported.

these NCAR models are satisfactory, and an inspection of their quantile–quantile plots (not shown here) lends support to the choice of  $t$ -distributed errors (in general). The estimation accuracy of the parameters appears reasonable perhaps with the exception of the degrees of freedom parameter  $\lambda$ , whose small point estimates still clearly point to a non-Gaussian error distribution, and the skewness parameter  $\alpha_s$ , which is only significant for the NCAR(1,1) model in the case of detrended data. Altogether, the results in Tables IV and V suggest that a stationary NCAR process with non-Gaussian leptokurtic errors provides a reasonable approximation for the Finnish interest rate series.

### 7. CONCLUDING REMARKS

In this article, we develop likelihood-based unit root tests in the NCAR model of L&S. Using assumptions comparable with those used in the previous literature on noncausal autoregressions, we find that the limiting distributions of the obtained unit root tests become infeasible unless a symmetric error distribution is assumed. With this assumptions, the limiting distributions of our tests are shown to depend on a single nuisance parameter, the effect of which can readily be eliminated by using estimated critical values. For cases where a skewed error distribution is needed, a bootstrap procedure is suggested.

According to the Monte Carlo simulations performed, the size properties of our tests are satisfactory even when estimated critical values or critical values based on the suggested bootstrap procedure are used. Our simulations also demonstrate that, compared with conventional DF-tests, the  $M$ -tests of Lucas (1995) and the  $RS$ -tests of Rothenberg and Stock (1997), our tests have good power against correctly specified stationary NCAR alternatives.

In an application to a Finnish interest rate series, we demonstrate how the tests are implemented in practice. In this application, we find that NCAR models with  $t$ -distributed errors fit the data better than causal AR models and that our tests clearly reject the unit root hypothesis, whereas no evidence or only week evidence against a unit root is obtained by using DF-tests,  $M$ -tests and  $RS$ -tests. All in all, this application and our simulations demonstrate the gains that can be achieved in unit root testing with a careful investigation of the possibility that the considered series is better described by a NCAR model than its conventional causal alternative.

## ACKNOWLEDGEMENTS

The authors thank the Editor, a Co-Editor and two anonymous referees for useful comments. The first author acknowledges financial support from the Academy of Finland, the OP-Pohjola Group Research Foundation and the Finnish Cultural Foundation. Part of this research was carried out during his research visit to the Monetary Policy and Research Department of the Bank of Finland whose hospitality is gratefully acknowledged. The second author acknowledges financial support from Jan Wallander's and Tom Hedelius' Foundation, grant no. P2012-0085:1.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.

## REFERENCES

- Andrews B, Davis RA, Breidt FJ. 2006. Maximum likelihood estimation for all-pass time series models. *Journal of Multivariate Analysis* **97**: 1638–1659.
- Azzalini A, Genton MG. 2008. Robust likelihood methods based on the skew- $t$  and related distributions. *International Statistical Review* **76**: 106–129.
- Basawa IV, Scott DJ. 1983. *Asymptotic Optimal Inference for Nonergodic Models*. New York: Springer-Verlag.
- Breidt J, Davis RA, Lii KS, Rosenblatt M. 1991. Maximum likelihood estimation for noncausal autoregressive processes. *Journal of Multivariate Analysis* **36**: 175–198.
- Breidt J, Davis RA, Trindade AA. 2001. Least absolute deviation estimation for all-pass time series models. *The Annals of Statistics* **29**: 919–946.
- Brockwell PJ, Davis RA. 1987. *Time Series: Theory and Methods*. New York: Springer-Verlag.
- Caner M, Hansen BE. 2001. Threshold autoregression with a unit root. *Econometrica*: 1555–1596.
- Davidson J. 1994. *Stochastic Limit Theory*. Oxford: Oxford University Press.
- Gouriéroux C, Zakoian JM. 2013. Explosive bubble modelling by noncausal process. *CREST Working Paper*: 2013–04.
- Hansen BE. 1992. Convergence to stochastic integrals for dependent heterogeneous processes. *Econometric Theory* **8**: 489–500.
- Huang J, Pawitan Y. 2000. Quasi-likelihood estimation of noninvertible moving average processes. *Scandinavian Journal of Statistics* **27**: 689–710.
- Jeganathan P. 1995. Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* **11**: 818–887.
- Lanne M, Saikkonen P. 2011. Modeling expectations with noncausal autoregressions. *Journal of Time Series Econometrics* **3** (3): Article 2.
- Lanne M, Luoto J, Saikkonen P. 2012a. Optimal forecasting of noncausal autoregressive time series. *International Journal of Forecasting* **28**: 623–631.
- Lanne M, Luoma A, Luoto J. 2012b. Bayesian model selection and forecasting in noncausal autoregressive models. *Journal of Applied Econometrics* **27**: 812–830.
- Lanne M, Nyberg H, Saarinen E. 2012c. Forecasting U.S. macroeconomic and financial time series with noncausal and causal AR models: a comparison. *Economics Bulletin* **32**: 2849–2859.
- Lanne M, Saikkonen P. 2013. Noncausal vector autoregression. *Econometric Theory* **29**: 447–481.
- Lii K-S, Rosenblatt M. 1996. Maximum likelihood estimation for non-Gaussian nonminimum phase ARMA sequences. *Statistica Sinica* **6**: 1–22.
- Lucas A. 1995. Unit root tests based on M estimators. *Econometric Theory* **11**: 331–346.
- Meitz M, Saikkonen P. 2013. Maximum likelihood estimation of a noninvertible ARMA model with autoregressive conditional heteroskedasticity. *Journal of Multivariate Analysis* **114**: 227–255.
- Park JY, Phillips PCB. 1988. Statistical inference in regressions with integrated processes: part 1. *Econometric Theory* **4**: 468–497.
- Phillips PCB. 1987. Time series regression with a unit root. *Econometrica* **55**: 277–301.
- Phillips PCB. 1988. Weak convergence of sample covariance matrices to stochastic integrals via martingale approximations. *Econometric Theory* **4**: 528–533.
- Phillips PCB, Perron P. 1988. Testing for a unit root in time series regression. *Biometrika* **75**: 335–346.
- Rothenberg T.J, Stock JH. 1997. Inference in a nearly integrated autoregressive model with nonnormal innovations. *Journal of Econometrics* **80**: 269–286.

Rosenblatt M. 2000. *Gaussian and Non-Gaussian Linear Time Series and Random Fields*. New York: Springer-Verlag.  
 Saikkonen P. 1993. Continuous weak convergence and stochastic equicontinuity results for integrated processes with an application to the estimation of a regression model. *Econometric Theory* **9**: 155–188.  
 Scott DJ. 1973. Central limit theorems for martingales and for processes with stationary increments using a Skorokhod representation approach. *Advances in Applied Probability* **5**: 119–137.  
 Van Dijk D, Franses PH, Lucas A. 1999. Testing for smooth transition nonlinearity in the presence of outliers. *Journal of Business and Economic Statistics* **17**: 217–235.  
 Wu R, Davis RA. 2010. Least absolute deviation estimation for general autoregressive moving average time-series models. *Journal of Time Series Analysis* **31**: 98–112.

APPENDIX A: PARTIAL DERIVATIVES OF THE LOG-LIKELIHOOD FUNCTION

**First partial derivatives.** As in Section 3.2, we use a subscript to denote a partial derivative indicated by the subscript. For notational brevity, denote

$$e_{x,t}(\theta) = \frac{f_x(\sigma^{-1}\epsilon_t(\theta); \lambda)}{f(\sigma^{-1}\epsilon_t(\theta); \lambda)} \quad \text{and} \quad e_{\lambda,t}(\theta) = \frac{f_\lambda(\sigma^{-1}\epsilon_t(\theta); \lambda)}{f(\sigma^{-1}\epsilon_t(\theta); \lambda)},$$

where  $\epsilon_t(\theta) = \phi(B)\varphi(B^{-1})y_t$  is treated as a function of the parameters  $\phi$  and  $\varphi$  (for notational convenience, we do not make explicit that  $\epsilon_t(\theta)$  is independent of the parameters  $\sigma$  and  $\lambda$ ). Note also the alternative expressions

$$\begin{aligned} \epsilon_t(\theta) &= \Delta u_t(\varphi) - \phi u_{t-1}(\varphi) - \pi_1 \Delta u_{t-1}(\varphi) - \dots - \pi_{r-1} \Delta u_{t-r+1}(\varphi) \\ &= (v_t(\phi, \pi) - \varphi_1 v_{t+1}(\phi, \pi) - \dots - \varphi_s v_{t+s}(\phi, \pi)), \end{aligned}$$

where  $u_t(\varphi) = \varphi(B^{-1})y_t$  and  $v_t(\phi, \pi) = \phi(B)y_t$  [see (2)]. We also set  $U_{t-1}(\varphi) = (u_{t-1}(\varphi), \dots, u_{t-r+1}(\varphi))$  and  $V_{t+1}(\phi, \pi) = (v_{t+1}(\phi, \pi), \dots, v_{t+s}(\phi, \pi))$ . With straightforward differentiation, one now obtains from (8)

$$g_{\theta,t}(\theta) = \begin{bmatrix} g_{\phi,t}(\theta) \\ g_{\pi,t}(\theta) \\ g_{\varphi,t}(\theta) \\ g_{\sigma,t}(\theta) \\ g_{\lambda,t}(\theta) \end{bmatrix} = \begin{bmatrix} -\sigma^{-1}e_{x,t}(\theta)u_{t-1}(\varphi) \\ -\sigma^{-1}e_{x,t}(\theta)\Delta U_{t-1}(\varphi) \\ -\sigma^{-1}e_{x,t}(\theta)V_{t+1}(\phi, \pi) \\ -\sigma^{-2}(e_{x,t}(\theta)\epsilon_t(\theta) + \sigma) \\ e_{\lambda,t}(\theta) \end{bmatrix}$$

**Second partial derivatives.** To simplify notation, define

$$\begin{aligned} e_{xx,t}(\theta) &= \frac{f_{xx}(\sigma^{-1}\epsilon_t(\theta); \lambda)}{f(\sigma^{-1}\epsilon_t(\theta); \lambda)} - \left( \frac{f_x(\sigma^{-1}\epsilon_t(\theta); \lambda)}{f(\sigma^{-1}\epsilon_t(\theta); \lambda)} \right)^2 \\ e_{\lambda x,t}(\theta) &= \frac{f_{\lambda x}(\sigma^{-1}\epsilon_t(\theta); \lambda)}{f(\sigma^{-1}\epsilon_t(\theta); \lambda)} - \frac{f_\lambda(\sigma^{-1}\epsilon_t(\theta); \lambda)}{f(\sigma^{-1}\epsilon_t(\theta); \lambda)} \frac{f_x(\sigma^{-1}\epsilon_t(\theta); \lambda)}{f(\sigma^{-1}\epsilon_t(\theta); \lambda)} \\ e_{\lambda\lambda,t}(\theta) &= \frac{f_{\lambda\lambda}(\sigma^{-1}\epsilon_t(\theta); \lambda)}{f(\sigma^{-1}\epsilon_t(\theta); \lambda)} - \frac{f_\lambda(\sigma^{-1}\epsilon_t(\theta); \lambda)}{f(\sigma^{-1}\epsilon_t(\theta); \lambda)} \frac{f_\lambda(\sigma^{-1}\epsilon_t(\theta); \lambda)}{f(\sigma^{-1}\epsilon_t(\theta); \lambda)}, \end{aligned}$$

and let  $Y_t$  stand for the  $(r-1) \times s$  matrix with elements  $y_{t-i+j}$  ( $i = 1, \dots, r-1, j = 1, \dots, s$ ), whereas  $Y'_{1t}$  ( $1 \times s$ ) is used to signify the first row of  $Y_t$ .

The following five 2nd partial derivatives involve the long-run parameter  $\phi$ :

$$\begin{aligned} g_{\phi\phi,t}(\theta) &= \sigma^{-2} e_{xx,t}(\theta) u_{t-1}^2(\varphi) \\ g_{\pi\phi,t}(\theta) &= \sigma^{-2} e_{xx,t}(\theta) u_{t-1}(\varphi) \Delta U_{t-1}(\varphi) \\ g_{\varphi\phi,t}(\theta) &= \sigma^{-2} e_{xx,t}(\theta) u_{t-1}(\varphi) V_{t+1}(\phi, \pi) + \sigma^{-1} e_{x,t}(\theta) Y_{1t} \\ g_{\sigma\phi,t}(\theta) &= \sigma^{-3} e_{xx,t}(\theta) \epsilon_t(\theta) u_{t-1}(\varphi) + \sigma^{-2} e_{x,t}(\theta) u_{t-1}(\varphi) \\ g_{\lambda\phi,t}(\theta) &= -\sigma^{-1} u_{t-1}(\varphi) e_{\lambda x,t}(\theta). \end{aligned}$$

The remaining partial derivatives are as in L&S expect that  $r$  is replaced by  $r - 1$  and differences of  $U_{t-1}(\varphi)$  and  $Y_t$  are used. We have

$$\begin{aligned} g_{\pi\pi,t}(\theta) &= \sigma^{-2} e_{xx,t}(\theta) \Delta U_{t-1}(\varphi) \Delta U'_{t-1}(\varphi) \\ g_{\varphi\varphi,t}(\theta) &= \sigma^{-2} e_{xx,t}(\theta) V_{t+1}(\phi, \pi) V'_{t+1}(\phi, \pi) \\ g_{\sigma\sigma,t}(\theta) &= 2\sigma^{-3} e_{x,t}(\theta) \epsilon_t(\theta) + \sigma^{-4} e_{xx,t}(\theta) \epsilon_t^2(\theta) + \sigma^{-2} \\ g_{\lambda\lambda,t}(\theta) &= e_{\lambda\lambda,t}(\theta) \\ g_{\pi\varphi,t}(\theta) &= \sigma^{-2} e_{xx,t}(\theta) \Delta U_{t-1}(\varphi) V'_{t+1}(\phi, \pi) + \sigma^{-1} e_{x,t}(\theta) \Delta Y_t \\ g_{\pi\sigma,t}(\theta) &= \sigma^{-3} e_{xx,t}(\theta) \epsilon_t(\theta) \Delta U_{t-1}(\varphi) + \sigma^{-2} e_{x,t}(\theta) \Delta U_{t-1}(\varphi) \\ g_{\varphi\sigma,t}(\theta) &= \sigma^{-3} e_{xx,t}(\theta) \epsilon_t(\theta) V_{t+1}(\phi, \pi) + \sigma^{-2} e_{x,t}(\theta) V_{t+1}(\phi, \pi) \\ g_{\lambda\pi,t}(\theta) &= -\sigma^{-1} e_{\lambda x,t}(\theta) \Delta U'_{t-1}(\varphi) \\ g_{\lambda\varphi,t}(\theta) &= -\sigma^{-1} e_{\lambda x,t}(\theta) V'_{t+1}(\phi, \pi) \\ g_{\lambda\sigma,t}(\theta) &= -\sigma^{-2} e_{\lambda x,t}(\theta) \epsilon_t(\theta). \end{aligned}$$

These partial derivatives form the matrix  $g_{\vartheta\vartheta,t}(\theta)$ .

#### APPENDIX B: PROOFS FOR SECTION 3

**Proof of Lemma 1.** Recall the notation  $\Delta u_t = \sum_{j=0}^{\infty} \alpha_{0,j} \epsilon_{t-j}$  and  $v_t = \sum_{j=0}^{\infty} \beta_{0,j} \epsilon_{t-j}$ . Using the expression of the components of  $g_{\theta,t}(\theta_0)$ , one obtains (10) as a straightforward application of the theorem in Phillips (1988), whereas (11) can be justified by a central limit theorem for stationary ergodic processes [see Breidt *et al.* (1991) or L&S]. The stated joint convergence holds because both  $g_{\phi,t}(\theta_0)$  and  $g_{\vartheta,t}(\theta_0)$  are generated by the same i.i.d. sequence  $\epsilon_t$ .

To establish the independence of the limits, let  $\mathcal{F}_t^\epsilon$  signify the  $\sigma$ -algebra generated by  $(\epsilon_t, \epsilon_{t-1}, \dots)$  and consider the random vector  $w_t = (e_{x,t}, \epsilon_t, g_{\vartheta,t}(\theta_0))$ . Using the definitions of  $e_{x,t}$  and  $g_{\vartheta,t}(\theta_0)$ , and Assumption 3, it is straightforward to check that  $(w_t, \mathcal{F}_t^\epsilon)$  is an  $L_2$ -mixingale with size  $-1$  [for the definition of an  $L_2$ -mixingale and its size, see Davidson (1994, p. 247)]. Thus, Theorem 3 of Scott (1973) and the Cramér-Wold device imply that a functional central limit theorem applies to  $w_t$  and the resulting limiting Brownian motion,  $Q(u)$  say, determines the limits in (10) and (11). Thus, it suffices to show that the first two components of  $Q(u)$  are independent of the remaining components, which in turn follows if the long-run covariance matrix between the component vectors  $(e_{x,t}, \epsilon_t)$  and  $g_{\vartheta,t}(\theta_0)$  of  $w_t$  is zero.

That the long-run covariance matrix between  $(e_{x,t}, \epsilon_t)$  and  $(g_{\pi,t}(\theta_0), g_{\varphi,t}(\theta_0))$  is zero is easy to see. For instance,  $g_{\varphi,t}(\theta_0) = -\sigma_0^{-1} e_{x,t} V_{t+1}$  is uncorrelated with  $(e_{x,t}, \epsilon_t)$  because  $V_{t+1}$  is independent of  $(e_{x,t}, \epsilon_t)$  and has zero mean, and  $g_{\varphi,t}(\theta_0)$  is also uncorrelated with  $(e_{x,t+k}, \epsilon_{t+k})$ ,  $k \neq 0$ , because  $e_{x,t}$  is independent of  $(V_{t+1}, e_{x,t+k}, \epsilon_{t+k})$  and has zero mean. This implies the desired result, and a similar reasoning applies to  $g_{\pi,t}(\theta_0) = -\sigma_0^{-1} e_{x,t} \Delta U_{t-1}$ . Note that here we do not need the assumption that the function  $f(\cdot; \lambda)$  is even.

The long-run covariance matrix between  $(e_{x,t}, \epsilon_t)$  and  $(g_{\sigma,t}(\theta_0), g_{\lambda,t}(\theta_0))$  reduces to the ordinary covariance matrix. Assumption 2 and the definition of  $e_{x,t}$  imply that  $e_{x,t}$  is an odd function of  $\epsilon_t$ . Thus, as  $E[g_{\sigma,t}(\theta_0) \epsilon_t] =$

$-\sigma_0^{-2} E [e_{x,t} \epsilon_t^2]$ , the fact that  $e_{x,t} \epsilon_t^2$  is an odd function of  $\epsilon_t$  yields  $E [g_{\sigma,t} (\theta_0) \epsilon_t] = 0$ . In the same way, it is seen that  $E [g_{\sigma,t} (\theta_0) e_{x,t}] = -\sigma_0^{-2} E [e_{x,t}^2 \epsilon_t] = 0$ . Furthermore,  $g_{\lambda,t} (\theta_0)$  is an even function of  $\epsilon_t$  so that  $E [g_{\lambda,t} (\theta_0) \epsilon_t] = 0$  and  $E [g_{\lambda,t} (\theta_0) e_{x,t}] = 0$ . Thus, we have shown that the long-run covariance matrix between  $(e_t, \epsilon_t)$  and  $g_{\vartheta,t} (\theta_0)$  is block diagonal.

In the derivations of the preceding paragraph, the assumption that the function  $f (x; \lambda)$  is even is needed because otherwise, for instance, the expectation

$$\begin{aligned} E [g_{\sigma,t} (\theta_0) e_{x,t}] &= -\sigma_0^{-3} \int x \left( \frac{f_x (\sigma_0^{-1} x; \lambda_0)}{f (\sigma_0^{-1} x; \lambda_0)} \right)^2 f (\sigma_0^{-1} x; \lambda_0) dx \\ &= -\sigma_0^{-2} \int z \frac{(f_x (z; \lambda_0))^2}{f (z; \lambda_0)} dz \end{aligned}$$

need not vanish. We now demonstrate how the result of the lemma changes when the function  $f (x; \lambda)$  is not even.

First, note that the aforementioned functional central limit theorem by Scott (1973) now reads as

$$T^{-1/2} \sum_{t=1}^{[Tu]} w_t \xrightarrow{d} Q (u) = (B_{e_x} (u), B_{\epsilon} (u), B_{\vartheta_1} (u), B_{\vartheta_2} (u)) \sim \text{BM} (\Omega),$$

where the covariance matrix  $\Omega$  is given by

$$\Omega = \text{Cov} [Q_1 (1)] = \begin{bmatrix} \mathcal{J} & -\sigma_0 & 0 & \omega'_{\vartheta_2, e_x} \\ -\sigma_0 & \sigma_0^2 & 0 & \omega'_{\vartheta_2, \epsilon} \\ 0 & 0 & \Sigma_1 & 0 \\ \omega_{\vartheta_2, e_x} & \omega_{\vartheta_2, \epsilon} & 0 & \Sigma_2 \end{bmatrix}$$

with  $\omega_{\vartheta_2, e_x} = E [g_{\vartheta_2,t} (\theta_0) e_{x,t}]$  and  $\omega_{\vartheta_2, \epsilon} = E [g_{\vartheta_2,t} (\theta_0) \epsilon_t]$  (possibly) nonzero. If desired, expressions of  $\omega_{\vartheta_2, e_x}$  and  $\omega_{\vartheta_2, \epsilon}$  in terms of the density function  $f (x; \lambda_0)$  and its partial derivatives can be obtained as performed earlier for  $E [g_{\sigma,t} (\theta_0) e_{x,t}]$ . The zeros in  $\Omega$  are obtained from Assumption 3(ii) and the aforementioned fact that the long-run covariance matrix between  $(e_{x,t}, \epsilon_t)$  and  $(g_{\pi,t} (\theta_0), g_{\varphi,t} (\theta_0))$  is zero even when the function  $f (x; \lambda)$  is not even.

From the preceding discussion, we can now conclude that the (joint) limit  $(Z_1, Z_2)$  [see (10) and (11)] satisfies

$$(Z_1, Z_2) \sim \left( -\frac{1}{\sigma_0 \pi_0 (1)} \int_0^1 B_{\epsilon} (u) dB_{e_x} (u), B_{\vartheta_1} (1), B_{\vartheta_2} (1) \right) \stackrel{\text{def}}{=} \tilde{Z} = (\tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3), \quad (\text{B1})$$

where the covariance structure of the involved Brownian motions is as shown earlier. Note that when the function  $f (x; \lambda)$  is even,  $\tilde{Z}_1 = Z_1$  and  $(\tilde{Z}_2, \tilde{Z}_3) = Z_2$  are independent, but this is generally not the case, although independence of  $\tilde{Z}_2$  and  $(\tilde{Z}_1, \tilde{Z}_3)$  clearly holds generally.

**Proof of Lemma 2.** First, conclude from the definitions that the elements of the matrix  $g_{\vartheta \vartheta,t} (\theta_0)$  (Appendix A) are (jointly) stationary and ergodic so that an application of a law of large numbers yields (13). Thus, it suffices to consider (12) and (14). We use the well-known fact that a functional central limit theorem applies to the process  $\Delta u_t$  [see (6)] and that  $T^{-1/2} u_{[T \cdot]} \xrightarrow{d} \pi_0 (1)^{-1} B_{\epsilon} (\cdot)$  [see the discussion at the beginning of the proof of the theorem in Phillips (1988)]. Similarly, it is seen that  $T^{-1/2} y_{[T \cdot]}$  obeys a functional central limit theorem.

As  $e_{xx,t} = e_{xx,t}(\theta_0)$ , the expression of  $g_{\phi\phi,t}(\theta)$  in Appendix A yields

$$g_{\phi\phi,t}(\theta_0) = \sigma_0^{-2} \frac{f_{xx}(\sigma_0^{-1}\epsilon_t; \lambda_0)}{f(\sigma_0^{-1}\epsilon_t; \lambda_0)} u_{t-1}^2 - \sigma_0^{-2} e_{x,t}^2 u_{t-1}^2,$$

where

$$E \left( \frac{f_{xx}(\sigma_0^{-1}\epsilon_t; \lambda_0)}{f(\sigma_0^{-1}\epsilon_t; \lambda_0)} \right) = 0$$

by Assumption 4. As  $T^{-1/2}u_{[T\cdot]}$  converges weakly, the continuous mapping theorem implies that the same is true for  $(T^{-1/2}u_{[T\cdot]})^2$ , so that from Theorem 3.3 of Hansen (1992), we obtain

$$T^{-2} \sum_{t=r+1}^{T-s} \frac{f_{xx}(\sigma_0^{-1}\epsilon_t; \lambda_0)}{f(\sigma_0^{-1}\epsilon_t; \lambda_0)} u_{t-1}^2 = o_p(1).$$

Hence,

$$\begin{aligned} T^{-2} \sum_{t=r+1}^{T-s} g_{\phi\phi,t}(\theta_0) &= -\sigma_0^{-2} \mathcal{J} T^{-2} \sum_{t=r+1}^{T-s} u_{t-1}^2 - \sigma_0^{-2} T^{-2} \sum_{t=r+1}^{T-s} (e_{x,t}^2 - \mathcal{J}) u_{t-1}^2 + o_p(1) \\ &= -\sigma_0^{-2} \mathcal{J} T^{-2} \sum_{t=r+1}^{T-s} u_{t-1}^2 + o_p(1), \end{aligned}$$

where the second equality can again be justified by Theorem 3.3 of Hansen (1992) and the assumption  $E[e_{x,t}^2] = \mathcal{J}$ . Thus, as  $T^{-1/2}u_{[T\cdot]} \xrightarrow{d} \pi_0(1)^{-1} B_\epsilon(\cdot)$ , an application of the continuous mapping theorem yields (12).

To establish (14), first, consider

$$g_{\pi\phi,t}(\theta_0) = \sigma_0^{-2} \frac{f_{xx}(\sigma_0^{-1}\epsilon_t; \lambda_0)}{f(\sigma_0^{-1}\epsilon_t; \lambda_0)} u_{t-1} \Delta U_{t-1} - e_{x,t}^2 u_{t-1} \Delta U_{t-1},$$

where

$$E \left( \frac{f_{xx}(\sigma_0^{-1}\epsilon_t; \lambda_0)}{f(\sigma_0^{-1}\epsilon_t; \lambda_0)} \Delta U_{t-1} \right) = 0 \quad \text{and} \quad E(e_{x,t}^2 \Delta U_{t-1}) = 0$$

because  $E(\Delta U_{t-1}) = 0$  and  $\Delta U_{t-1}$  is independent of  $(\epsilon_t, e_{x,t})$ . Thus, Theorem 3.3 of Hansen (1992) can again be used to obtain

$$T^{-3/2} \sum_{t=r+1}^{T-s} g_{\pi\phi,t}(\theta_0) = o_p(1).$$

As  $T^{-1/2}y_{[T\cdot]}$  converges weakly, arguments similar to those already used yield

$$\begin{aligned} T^{-3/2} \sum_{t=r+1}^{T-s} g_{\phi\phi,t}(\theta) &= \sigma_0^{-2} T^{-3/2} \sum_{t=r+1}^{T-s} e_{xx,t} u_{t-1} V_{t+1} + \sigma_0^{-1} T^{-3/2} \sum_{t=r+1}^{T-s} e_{x,t} Y_{1t} \\ &= o_p(1). \end{aligned}$$



Note, however, that the summand in the first term on the right-hand side of the former equation differs from its counterpart in  $g_{\pi\phi,t}(\theta_0)$  in that, unlike  $\Delta U_{t-1}$ , the zero mean stationary process  $V_{t+1}$  depends on  $\epsilon_{t+j}$ ,  $j > 0$ . Hence, the process  $e_{xx,t}V_{t+1}$  is not adapted to  $\mathcal{F}_t^e$  and does not satisfy the assumptions imposed on the sequence  $\{e_i\}$  in Hansen's (1992) Theorem 3.3. However, an inspection of the proof of that theorem reveals that this theorem applies even if  $\{e_i\}$  is a general mixingale such as  $e_{xx,t}V_{t+1}$  instead of the assumed special type of mixingale [that  $e_{xx,t}V_{t+1}$  is a mixingale, see Davidson (1994, p. 247)]. Note also that the preceding derivations apply even if the function  $f(x; \lambda)$  is not assumed even.

Finally, arguments similar to those used for  $g_{\pi\phi,t}(\theta_0)$  apply to  $g_{\sigma\phi,t}(\theta_0)$  and  $g_{\lambda\phi,t}(\theta_0)$ . Because  $E[e_{xx,t}\epsilon_t] = 0$ ,  $E[e_{x,t}] = 0$  and  $E[e_{\lambda x,t}] = 0$  by Assumption 4, Hansen's (1992) Theorem 3.3 can be used to obtain

$$T^{-3/2} \sum_{t=r+1}^{T-s} g_{\sigma\phi,t}(\theta_0) = o_p(1) \quad \text{and} \quad T^{-3/2} \sum_{t=r+1}^{T-s} g_{\lambda\phi,t}(\theta_0) = o_p(1).$$

Thus, we have established (14). Note that here we also need the assumption that the function  $f(x; \lambda)$  is even, which, for instance, guarantees that  $E[e_{xx,t}\epsilon_t] = 0$ , and hence a direct applicability of Hansen's (1992) Theorem 3.3.

The stated independence follows in the same way as in the proof of Lemma 1 because the long-run covariance matrix between  $\epsilon_t$  and  $g_{\vartheta,t}(\theta_0)$  is zero. This completes the proof of the lemma, but we still demonstrate how the result looks like when the assumption of  $f(x; \lambda)$  being even is not made. According to what was said earlier, it suffices to consider  $g_{\sigma\phi,t}(\theta_0)$  and  $g_{\lambda\phi,t}(\theta_0)$ . Using Hansen's (1992) Theorem 3.3 twice, we first obtain

$$\begin{aligned} T^{-3/2} \sum_{t=r+1}^{T-s} g_{\sigma\phi,t}(\theta_0) &= \sigma_0^{-3} T^{-3/2} \sum_{t=r+1}^{T-s} u_{t-1} e_{xx,t} \epsilon_t + \sigma_0^{-2} T^{-3/2} \sum_{t=r+1}^{T-s} u_{t-1} e_{x,t} \\ &= \sigma_0^{-3} T^{-3/2} \sum_{t=r+1}^{T-s} u_{t-1} E[e_{xx,t} \epsilon_t] + o_p(1) \\ &\xrightarrow{d} \sigma_0^{-3} \int_0^1 B_\epsilon(u) du E[e_{xx,t} \epsilon_t] \\ &\stackrel{\text{def}}{=} g_{\sigma\phi}(\theta_0). \end{aligned}$$

Then, by the same argument

$$T^{-3/2} \sum_{t=r+1}^{T-s} g_{\lambda\phi,t}(\theta_0) = -\sigma_0^{-1} T^{-3/2} \sum_{t=r+1}^{T-s} u_{t-1} e_{\lambda x,t} \xrightarrow{d} \sigma_0^{-3} \int_0^1 B_\epsilon(u) du E[e_{\lambda x,t}] \stackrel{\text{def}}{=} g_{\lambda\phi}(\theta_0).$$

From the preceding discussion, we also see that now the independence of  $g_{\phi\phi}(\theta_0)$  and  $(g_{\sigma\phi}(\theta_0), g_{\lambda\phi}(\theta_0))$  does generally not hold.

**Proposition 1.** As indicated after Proposition 1, the proof follows the arguments in Basawa and Scott (1983, pp. 56–59) with suitable modifications. In particular, our probability statements assume that the true parameter value  $\theta_0$  is a fixed point in the parameter space so that, instead of uniform and continuous weak convergence employed by Basawa and Scott (1983, pp. 56–59), we employ ordinary weak convergence. A minor modification is that in place of the scaling matrix  $I_n(\theta)$  in Assumptions (B.2) and (B.3) of Basawa and Scott (1983, pp. 33–34), we have the matrix  $D_T$  that is independent of the parameter  $\theta$ . It is straightforward to check that, as far as the proof is concerned, this replacement has no essential effect. The same applies to the fact that in our case the limits  $Z$  and  $G(\theta_0)$  (Proposition 1) are not independent, implying that the LAMN condition does not hold.

To see how the proof proceeds, note first that, with the preceding modifications, we can repeat the arguments in the proof of Theorem 1 of Basawa and Scott (1983, pp. 56–58) and conclude that, with probability approaching one, there exists a local ML estimator  $\hat{\theta}_T$  such that  $D_T(\hat{\theta}_T - \theta_0) = O_p(1)$ . Next, as in the proof of Theorem 2 of Basawa and Scott (1983, pp. 58–59), we find that

$$D_T(\hat{\theta}_T - \theta_0) - G_T^{-1}(\theta_0) S_T(\theta_0) = o_p(1).$$

The first result now follows from Proposition 1 and the continuous mapping theorem. The second one can be established by using arguments similar to those used to prove Proposition 3.2 of Saikkonen (1993, p. 161) modified as above to concern ordinary weak convergence instead of continuous weak convergence [see also Remark 3.1 of Saikkonen (1993, p. 161)].

We shall now demonstrate how the result is changed when the function  $f(x; \lambda)$  is not assumed even. From the definition of  $S_T(\theta_0)$ , Lemma 1 and (B1), we find that  $S_T(\theta_0) \xrightarrow{d} \tilde{Z}$ , whereas Lemma 2 and its extension discussed at the end of the proof of that lemma yield

$$G_T(\theta_0) \xrightarrow{d} \begin{bmatrix} g_{\phi\phi}(\theta_0) & 0 & g_{\vartheta_2\phi}(\theta_0)' \\ 0 & \Sigma_1 & 0 \\ g_{\vartheta_2\phi}(\theta_0) & 0 & \Sigma_2 \end{bmatrix} \stackrel{\text{def}}{=} \tilde{G}(\theta_0),$$

where  $g_{\vartheta_2\phi}(\theta_0) = (g_{\sigma\phi}(\theta_0), g_{\lambda\phi}(\theta_0))$  (see the end of the proof of Lemma 2). Thus, Propositions 1 and 2 hold with  $Z$  and  $G(\theta_0)$  replaced by  $\tilde{Z}$  and  $\tilde{G}(\theta_0)$ , respectively, but without the independence statement in Proposition 1.

From the preceding extension of Proposition 2 and the continuous mapping theorem, it follows that when the function  $f(x; \lambda)$  is not assumed even, the limiting distribution of test statistic  $\tau_T$  is given by

$$\begin{aligned} \tau_T \xrightarrow{d} & - \left( \frac{\mathcal{J}}{\sigma_0^2 \pi_0(1)^2} \int_0^1 B_\epsilon^2(u) \, d(u) - g_{\vartheta_2\phi}(\theta_0)' \Sigma_2^{-1} g_{\vartheta_2\phi}(\theta_0) \right)^{-1/2} \\ & \times \left( \frac{1}{\sigma_0 \pi_0(1)} \int_0^1 B_\epsilon(u) \, dB_{e_x}(u) + g_{\vartheta_2\phi}(\theta_0)' \Sigma_2^{-1} B_{\vartheta_2}(1) \right). \end{aligned}$$

Thus, when the function  $f(x; \lambda)$  is not required to be even, the limiting distribution of test statistic  $\tau_T$  becomes complicated. Not only are the Brownian motions  $B_\epsilon(u)$  and  $B_{e_x}(u)$  correlated, but the limiting distribution also depends on several nuisance parameters in a complicated way.

#### APPENDIX C: SUFFICIENT CONDITIONS FOR ASSUMPTION (17)

To provide a discussion on sufficient conditions for the ‘high level’ condition (17), we first note that the regularity conditions employed by Andrews *et al.* (2006) and L&S can be used to show that this condition holds for the lower right-hand block of the Hessian,  $\sum_{t=r+1}^{T-s} g_{\vartheta\vartheta,t}(\theta)$ . Thus, assuming these conditions, it suffices to consider the blocks involving the parameter  $\phi$ , that is,  $g_{\phi\phi,t}(\theta)$  and  $g_{\vartheta\phi,t}(\theta)$  (in what follows, the null hypothesis of a unit root will also be assumed).

We denote by  $\Lambda_0$  a neighbourhood of  $\lambda_0$  and assume that, for all  $x \in \mathbb{R}$ ,  $\Delta x \in \mathbb{R}$  and  $\lambda \in \Lambda_0$ , and for some  $C < \infty$  and  $d_1, d_2 > 0$ ,

$$|w(x + \Delta x; \lambda) - w(x; \lambda)| \leq C \left[ (1 + |x|^{d_1}) |\Delta x| + |\Delta x|^{d_2} \right] \tag{C1}$$

for the following choices of the function  $w(x; \lambda)$ :

$$w(x; \lambda) = \begin{cases} \frac{f_x(x; \lambda)}{f(x; \lambda)} \\ \frac{f_{xx}(x; \lambda)}{f(x; \lambda)} - \left( \frac{f_x(x; \lambda)}{f(x; \lambda)} \right)^2 \\ \frac{f_{\lambda x}(x; \lambda)}{f(x; \lambda)} - \frac{f_\lambda(x; \lambda)}{f(x; \lambda)} \frac{f_x(x; \lambda)}{f(x; \lambda)}. \end{cases}$$

This assumption is an analogue of Assumption B of Lii and Rosenblatt (1996) who used it with the first two choices of  $w(x; \lambda)$  and with  $f(x; \lambda)$  independent of  $\lambda$  in the context of (stationary) noncausal and noninvertible ARMA models. The third choice is a simplified analogue of Assumption 7 of Meitz and Saikkonen (2013) who developed an estimation theory for a (stationary) noninvertible ARMA model with conditionally heteroskedastic errors. Note that replacing the argument  $x$  with  $\sigma^{-1}\epsilon_t(\theta)$ , the three choices of  $w(x; \lambda)$  become  $e_{x,t}(\theta)$ ,  $e_{xx,t}(\theta)$  and  $e_{\lambda x,t}(\theta)$  (Appendix A), and for the second one, for instance, condition (C1) implies

$$|e_{xx,t}(\theta) - e_{xx,t}(\theta_0)| \leq C \left[ \left( 1 + |\sigma_0^{-1}\epsilon_t(\theta_0)|^{d_1} \right) |\sigma^{-1}\epsilon_t(\theta) - \sigma_0^{-1}\epsilon_t(\theta_0)| + |\sigma^{-1}\epsilon_t(\theta) - \sigma_0^{-1}\epsilon_t(\theta_0)|^{d_2} \right], \tag{C2}$$

where  $\epsilon_t(\theta_0) = \epsilon_t$ , and  $E[|\epsilon_t|^{2+d_1}] < \infty$  and  $E[|\epsilon_t|^{1+d_2}] < \infty$  is assumed.

It suffices to establish analogues of condition (17) for  $g_{\kappa\phi,t}(\theta)$ ,  $\kappa = \phi, \pi, \varphi, \sigma, \lambda$  or to show that, for all  $c > 0$ ,

$$\sup_{\theta \in N_{T,c}} \left\| T^{-b_\kappa} \sum_{t=r+1}^{T-s} [g_{\kappa\phi,t}(\theta) - g_{\kappa\phi,t}(\theta_0)] \right\| \xrightarrow{P} 0, \quad \kappa = \phi, \pi, \varphi, \sigma, \lambda, \tag{C3}$$

where  $b_\kappa = 2$  for  $\kappa = \phi$  and  $b_\kappa = 3/2$  otherwise. Establishing (C3) is straightforward but tedious. Details for  $\kappa = \phi, \varphi$  are published online only as Supporting Information, or are available from the authors on request (the other cases can be handled with similar arguments).