The Asymptotic Size and Power of the Augmented Dickey-Fuller Test for a Unit Root

Efstathios Paparoditis
Department of Mathematics and Statistics
University of Cyprus
P.O.Box 20537
CY 1678 Nicosia, CYPRUS

Dimitris N. Politis
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093-0112
USA

December 2013

Abstract

It is shown that the limiting distribution of the augmented Dickey-Fuller (ADF) test under the null hypothesis of a unit root is valid under a very general set of assumptions that goes far beyond the linear AR(∞) process assumption typically imposed. In essence, all that is required is that the error process driving the random walk possesses a spectral density that is strictly positive. Given that many economic time series are nonlinear, this extended result may have important applications. Furthermore, under the same weak assumptions, the limiting distribution of the ADF test is derived under the alternative of stationarity, and a theoretical explanation is given for the well-known empirical fact that the test's power is a decreasing function of the autoregressive order p used in the augmented regression equation. The intuitive reason for the reduced power of the ADF test as p tends to infinity is that the p regressors become asymptotically collinear.

Key words: Autoregressive Representation, Hypothesis Testing, Integrated Series, Unit Root.

1 Introduction

Testing for the presence of a unit root is a widely investigated problem in econometrics; cf. Hamilton (1994) or Patterson (2011) for extensive treatments of this topic. Given a stretch of time series observations $X_1, X_2, \ldots, X_n$, one of the commonly used tests for the null hypothesis of a unit root, is the so-called augmented Dickey-Fuller (ADF) test. This test decides about the presence of a unit root in the data generating mechanism by using the ordinary least squares (OLS) estimator $\widehat{\rho}_n$ of ρ, obtained by fitting the regression equation

$$X_t = \rho X_{t-1} + \sum_{j=1}^{p} a_{j,p} \Delta X_{t-j} + e_{t,p}, \tag{1.1}$$

to the observed stretch of data. In the above notation, $\Delta X_t = X_t - X_{t-1}$, while the order p is allowed to depend on n, $p = p(n)$, and its value is related to the assumptions imposed on the underlying process. In particular, under the null hypothesis $H_0 : \rho = 1$, it is commonly assumed that X_t is obtained by integrating a linear, infinite order autoregressive process, (AR(∞)), i.e., that

$$X_t = X_{t-1} + U_t, \qquad t = 1, 2, \ldots, \tag{1.2}$$

where $X_0 = 0$ and

$$U_t = \sum_{j=1}^{\infty} a_j U_{t-j} + e_t. \tag{1.3}$$

Here $\{e_t\}$ is a sequence of independent, identically distributed (i.i.d.) random variables having mean zero and variance $0 < \sigma_e^2 < \infty$. Stationarity and causality of $\{U_t\}$ is ensured by assuming that $\sum_{j=1}^{\infty} |j|^s |a_j| < \infty$ for some $s \geq 1$ and $\sum_{j=1}^{\infty} a_j z^j \neq 0$ for all $|z| \leq 1$.

To test H_0, Dickey and Fuller (1979) proposed the studentized statistic

$$t_n = \frac{\hat{\rho}_n - 1}{\widehat{Std}(\widehat{\rho}_n)}, \tag{1.4}$$

where $\widehat{Std}(\widehat{\rho}_n)$ denotes an estimator of the standard deviation of the OLS estimator $\widehat{\rho}_n$. The asymptotic distribution of t_n under H_0 is non-standard and is well known in the literature. Dickey and Fuller (1979) and Dickey and Fuller (1981) derived this distribution under the assumption that the order of the underlying autoregressive process is finite and known. Said and Dickey (1984) extended this result for the case where the innovation process $\{U_t\}$ driving the random walk (1.2) is an invertible autoregressive moving-average (ARMA) process, i.e., an AR(∞) process with exponentially decaying coefficients. Ng and Perron (1995) relaxed the assumptions needed on the rate at which the order $p(n)$ in (1.1) increases to infinity with n. Chang and Park (2002) established the same limiting distribution of t_n by further relaxing the assumptions regarding the rate at which p increases to infinity, by allowing for a polynomial decrease of the coefficients a_j in the AR(∞) representation (1.3) and by assuming a martingale difference structure instead of i.i.d. innovations e_t, that is, by assuming that $E(e_t|\mathcal{E}_{t-1}) = 0$ and $n^{-1}\sum_{t=1}^{n} E(e_t^2|\mathcal{E}_{t-1}) \to \sigma^2$, as $n \to \infty$, where $\mathcal{E}_l = \sigma(\{e_t : t \leq l\})$ is the σ-algebra generated by the random variables $\{e_l, e_{l-1}, \ldots\}$.

To derive the power behavior of the test under the alternative hypothesis $H_1 : \rho < 1$, the limiting distribution of t_n is required under the assumption that $\{X_t\}$ is a stationary process. Investigating the power of the ADF-test for fixed (stationary) alternatives, has attracted less interest in the literature. Nabeya and Tanaka (1990) and Perron (1991) analyzed the limiting power of unit root tests for sequences of local alternatives. For a first order autoregression, Abadir (1993) gives closed forms for the distribution of certain statistics leading to the derivation of the limiting distribution of unit root tests under the null and the alternative. For the ADF unit root test, Lopez (1997) considered the asymptotic distribution of $\hat{\rho}_n$ under the alternative that $\{X_t\}$ is a causal and invertible ARMA process. However, apart from the restrictive process set-up used to derive this distribution, the derivations seem to be incorrect since Lopez (1997) erroneously replaces regression equation (1.1) by a regression equation that contains only levels of the X_t's; see Remark 2.4 for details.

The aim of this paper is twofold. First we show that the established limiting distribution of the ADF-test under the null hypothesis of a unit root is valid under a most general set of assumptions regarding the innovation process $\{U_t\}$ driving the random walk (1.2). These assumptions go far beyond the AR(∞) linear process class (1.3). In particular, we prove validity of the limiting distribution of t_n under the general condition that the stationary process $\{U_t\}$ possesses a Wold-type, AR-representation with respect to white noise errors ε_t. This much wider class of stationary processes should not be confused with the linear AR(∞) class (1.3) driven by i.i.d. or by martingale difference innovations. In fact, this class consists of all zero mean, second order stationary (linear or nonlinear) processes having a continuous and strictly positive spectral density; cf. Pourahmadi (2001) and Section 2 for details.

Secondly, under the same set of general assumptions on the underlying stationary process $\{U_t\}$, we establish the limiting distribution of the ADF-test t_n under the alternative hypothesis in which $\{X_t\}$ is stationary. It turns out that under the alternative, the estimator $\widehat{\rho}_n$ is only $\sqrt{n/p_n}$–consistent, and therefore, its convergence rate is considerably smaller compared with the n–consistency of the same estimator under the null, and to the $\sqrt{n}$ convergence rate of other test statistics under the alternative, like for instance the well-known test of Philips and Perron (1988). We make the case that the underlying reason for the slow rate of convergence of $\widehat{\rho}_n$ under the alternative is that the regressors in equation (1.1) become asymptotically collinear as p increases to infinity. We explain this phenomenon, and show how/why this collinearity problem is responsible for the reduced power of the ADF-test, and explains theoretically the empirically observed fact that the power of this test is a decreasing function of the order p—see e.g. Figure 9.5 of Patterson (2011).

The remaining of the paper is organized as follows. Section 2 states the main assumptions imposed on the underlying process $\{U_t\}$ and derives the asymptotic distribution of the ADF test t_n under the null hypothesis of a unit root. The asymptotic behavior of t_n under the alternative of stationarity is also derived in Section 2, and its consequences for the power properties of the ADF test are discussed. Section 3 showcases a real data example where the reduced power of the ADF test is manifested, and underscores the importance of properly choosing the order p in practice. All technical proofs are deferred to Section 4.

2 Asymptotic Properties of the ADF test

2.1 Assumptions

We first state the conditions we impose on the dependence structure of the underlying second order stationary process $\{U_t\}$ that drives the random walk under the null. Assuming that $\{U_t\}$ is purely non-deterministic, i.e., that it possesses as spectral density whose logarithm is integrable, then the Wold representation yields

$$U_t = \sum_{j=1}^{\infty} \alpha_j \varepsilon_{t-j} + \varepsilon_t \qquad (2.5)$$

where $\sum_{j=1}^{\infty} \alpha_j^2 < \infty$ and ε_t is a zero mean, uncorrelated process with $0 < Var(\varepsilon_t) = \sigma_\varepsilon^2 < \infty$. We slightly restrict the above class of stationary processes to the one satisfying the following assumption.

Assumption 1

(i) The autocovariance function $\gamma_U(h) = Cov(U_t, U_{t+h})$, $h \in \mathbb{Z}$, of $U = \{U_t, t \in \mathbb{Z}\}$ satisfies $\sum_{h\in\mathbb{Z}} |\gamma_U(h)| < \infty$, and the spectral density f_U of U is strictly positive, i.e., $f_U(\lambda) > 0$ for all λ.

(ii) $E(\varepsilon_t^4) < \infty$ and the process $\{\varepsilon_t\}$ satisfies the following weakly dependent condition: $\sum_{n=1}^{\infty} n\|\mathcal{P}_1(\varepsilon_n)\| < \infty$, where $\mathcal{P}_t(Y) = E(Y|\mathcal{F}_t) - E(Y|\mathcal{F}_{t-1})$ and $\mathcal{F}_s = (\ldots, \varepsilon_{s-1}, \varepsilon_s)$.

The Wold representation (2.5) is an MA(∞) equation for U_t with respect to its innovations ε_t. Assumption 1(i) allows for an alternative representation of U_t with respect to the same white noise given in (2.5). In fact, because of the summability of the autocovariance function, the process $\{U_t\}$ has a continuous spectral density $f_U(\lambda) = (2\pi)^{-1}\sum_{h\in\mathbb{Z}} \gamma_U(h)\cos(\lambda h)$. This together with the strict positivity of $f_U(\lambda)$ implies that $\{U_t\}$ possesses a so-called *Wold-type AR-representation*, that is, U_t can be expressed as

$$U_t = \sum_{j=1}^{\infty} b_j U_{t-j} + \varepsilon_t, \tag{2.6}$$

where ε_t is the same white noise innovation process as the one appearing in the Wold representation (2.5). Furthermore, the coefficients b_j are absolutely summable, i.e., $\sum_{j=1}^{\infty} |b_j| < \infty$ and $b(z) = 1 - \sum_{k=1}^{\infty} b_j z^j \neq 0$ for $|z| \leq 1$; see Pourahmadi (2001), Lemma 6.4.

The Wold-type AR-representation (2.6) of U_t with respect to the white noise process ε_t should not be confused with the rather strong assumption of a linear AR(∞) process with respect to i.i.d. innovations. For example, one important difference between the class of process obeying a Wold-type AR-representation and the class of linear AR(∞) processes (1.3) is the linearity of the optimal predictor. In fact, for processes in the class (1.3) with i.i.d. or with martingale difference errors, the optimal k-step ahead predictor, is always the linear predictor. That is, for positive k, the general L_2-optimal predictor of U_{t+k} based on its past U_t, i.e., the conditional expectation $E(U_{t+k}|U_s, s \leq t)$, is for processes (1.3) with i.i.d. or with martingale difference innovations, identical to the best linear predictor $\mathcal{P}_{\mathcal{M}_t}(U_{t+k})$. Here $\mathcal{P}_C(Y)$ denotes orthogonal projection of Y onto the set C and $\mathcal{M}_t = \overline{span}\{U_j : j \leq s\}$, i.e., the closed linear span generated by the random variables $\{U_j : j \leq s\}$. This linearity property of the L_2-optical predictor is not shared by processes having a Wold-type AR-representation with respect to white noise innovations.

It is apparent that the class of processes having a Wold-type AR-representation is very large, and includes basically all linear or nonlinear time series that possess a strictly positive and continuous spectral density. The difference between the Wold-type AR-representation and the linear AR(∞) property (1.3) is further illustrated by means of the following two examples.

Example 1: (Non causal linear processes) Consider the process $U_t = \phi U_{t-1} + e_t$, with $|\phi| > 1$, and e_t a zero mean i.i.d. process with variance σ_e^2. Notice that $\{U_t\}$ is stationary but it does not belong to the linear AR(∞) class (1.3) since it is not causal (the root of $1 - \phi z = 0$ lies outside the unit disc). However, for $\varepsilon_t = \phi^{-2}(e_t - (\phi^2 - 1)\sum_{j\geq 1}\phi^{-j}e_{t+j})$, U_t has the AR-representation $U_t = bU_{t-1} + \varepsilon_t$ with $b = 1/\phi$ and the (causal) Wold representation

$U_t = \sum_{j=1}^{\infty} b^j \varepsilon_{t-j}$ with respect to the white noise process $\{\varepsilon_t\}$, where $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \phi^{-2}\sigma_e^2$.

Example 2: (Non invertible linear processes) Consider the process $U_t = \theta e_{t-1} + e_t$, with $|\theta| > 1$, and e_t a zero mean i.i.d. process with variance σ_e^2. Notice that $\{U_t\}$ is stationary but it does not belong to the linear AR(∞) class (1.3) since it is not invertible (the root of $1-\theta z = 0$ lies outside the unit disc). Now, for $\varepsilon_t = e_t + (\theta^{-1} - \theta)\sum_{j=1}^{\infty} \theta^{-j+1} e_{t-j}$ we get $U_t = \varepsilon_t - \theta^{-1}\varepsilon_{t-1}$ and therefore, U_t has the AR-representation $U_t = \sum_{j=1}^{\infty} bU_{t-j} + \varepsilon_t$, $b_j = -(1/\theta)^j$, with respect to the white noise process $\{\varepsilon_t\}$. Here $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma_e^2(1 + \theta^2 - \theta^4)$.

Assumption 1(ii) is imposed in order to control the dependence structure of the innovation process $\{\varepsilon_t\}$ and consequently of U_t; cf. Wu and Min (2005). It is based on the concept of weak dependence introduced by Wu (2005) and allows together with (2.5) for a very broad class of possible processes. Wu and Min (2005) give many examples of processes belonging to this class, including many well-known processes like, ARCH, GARCH processes, threshold autoregressive processes, bilinear processes and random coefficient autoregressive processes. Notice that instead of the weak dependence assumption above, other measures could be also used to control the dependence structure of the innovations process $\{\varepsilon_t\}$ as well. For instance, the results presented in this paper can be derived also under the alternative assumption that the innovation process $\{\varepsilon_t\}$ in (2.5) is strong mixing with strong mixing coefficient α_m satisfying $\sum_{m=1}^{\infty} \alpha_m^{1/2} < \infty$. In any case, Assumption 1(ii) extends considerably the class of stationary process allowed and goes far beyond linear autoregressive processes with i.i.d. innovations or innovations that form martingale differences.

Based on Assumption 1 regarding the class of stationary process $\{U_t\}$, Assumption 2 below specifies the generation mechanism of the underlying and observable process $\mathbf{X} = \{X_t, t \geq 0\}$.

Assumption 2 The process $\mathbf{X}$ satisfies one (and only one) of the following two conditions:

(i) (Unit root case:) $X_0 = 0$ and $X_t = X_{t-1} + U_t$ for $t = 1, 2, \ldots$.

(ii) (Stationary case:) $X_t = U_t$, for $t = 0, 1, 2, \ldots$,

where $\{U_t\}$ is the second order stationary process specified in Assumption 1.

The assumption $X_0 = 0$ simplifies notation and does not affect the asymptotic results of this paper for the unit root case. It can be replaced by other assumptions concerning the starting value X_0 provided this random variable remains bounded in probability. Assumption 2 simply states that X_t is either a stationary process satisfying Assumption 1 or it is obtained by integrating such a stationary process.

Notice that if Assumption 2 is true, then X_t obeys also the useful representation

$$X_t = \rho X_{t-1} + \sum_{j=1}^{\infty} a_j \Delta X_{t-j} + \varepsilon_t, \tag{2.7}$$

with ε_t the white noise process discussed in Assumption 1. To see this, notice that (2.7) is obviously true if Assumption 2(i) is satisfied with the choices $\rho = 1$ and $\Delta X_{t-1} = X_t - X_{t-1} = U_t$. Furthermore, if Assumption 2(ii) is true then it is easily verified that $X_t =$

$(\sum_{j=1}^{\infty} b_j)X_{t-1} - \sum_{j=1}^{\infty}(\sum_{s=j+1}^{\infty} b_s)\Delta X_{t-j} + \varepsilon_t$, which implies that also in this case (2.7) is true with

$$\rho = \sum_{j=1}^{\infty} b_j, \quad \text{and} \quad a_j = -\sum_{s=j+1}^{\infty} b_s, \quad j = 1, 2, \ldots.$$

Now, let

$$\rho_{min} = \inf \Big\{ \rho = \sum_{j=1}^{\infty} b_j : \; b_j, j = 1, 2, \ldots \; \text{ and } \; b(z) = 1 - \sum_{j=1}^{\infty} b_j z^j \neq 0 \quad \text{for} \quad |z| \leq 1 \Big\}.$$

The null and alternative hypothesis of interest can then be stated as

$$H_0: \; \rho = 1, \qquad H_1: \; \rho \in (\rho_{min}, 1). \tag{2.8}$$

Notice that H_0 is equivalent to Assumption 2(i) while H_1 to Assumption 2(ii). The range of values of ρ under the alternative H_1 is an interval since $B = \{b_j, j = 1, 2, \ldots : b(z) \neq 0 \quad \text{for} \quad |z| \leq 1\}$ is a convex set and the mapping $g : B \rightarrow \mathbb{R}$ with $g(b_1, b_2, \ldots) = \sum_{j=1}^{\infty} b_j \equiv \rho$ is continuous. Notice that $\rho_{min} < -1$ is also possible, for instance if $U_t = \varepsilon_t - \theta\varepsilon_{t-1}$ with $\theta \in (0.5, 1)$.

Remark 2.1 It is common in the econometric literature to state Assumption 2 in the following different form:

$$X_t = aX_{t-1} + U_t, \tag{2.9}$$

where $\{U_t\}$ is some zero mean, second order stationary process satisfying certain conditions. In this formulation, the case $a = 1$ is associated with the null hypothesis of unit root, while the case $|a| < 1$ with the alternative; see among others Ng and Perron (1995) and Chang and Park (2002). It is easily seen that the above formulation is a restatement of Assumption 2 in the sense that in both cases the same conditions are imposed on the underlying process $\{X_t\}$. If $a = 1$ this is obviously true while for $|a| < 1$, using the backshift operator $L^s X_t = X_{t-s}$ we have that $X_t = (1 - aL)^{-1}U_t = \sum_{j=0}^{\infty} a^j U_{t-j}$ and $\{X_t\}$ is stationary. However, if (2.9) is considered as a model for X_t, then identifiability and interpretability problems occur for the parameter a unless, of course, $a = 1$. To see why, let X_t be a stationary series, and define the new stationary series $V_t = X_t - bX_{t-1}$ where b is arbitrary; i.e., in the stationary case, eq. (2.9) holds true for *any value* of the parameter a as long as it is not one. To make the parameter a identifiable in the stationary case, an additional condition must be imposed, e.g., that the series U_t is the innovation series of X_t. Our Assumption 2 avoids these difficulties.

2.2 Limiting Distribution under the Null

The following theorem establishes the limiting distribution of the test statistic t_n under H_0 in (2.8). It shows that this limiting distribution is identical to that obtained under the AR(∞) linearity or weak linearity assumption for $\{U_t\}$; cf. Dickey and Fuller (1981) and Chang and Park (2002).

Theorem 2.1 *Let Assumption 1 and Assumption 2(i) be satisfied and suppose that* $p_n \to \infty$ *as* $n \to \infty$ *such that* $p_n/\sqrt{n} \to 0$. *Then*

$$t_n \;\Rightarrow \int_0^1 W(t)dW(t)\Big/\Big(\int_0^1 W^2(t)dt\Big)^{1/2},$$

where $\{W(t), t \in [0,1]\}$ *is the standard Wiener process on* $[0,1]$.

By the above theorem, an asymptotic α-level test of the null hypothesis of a unit root is given by rejecting H_0 whenever t_n is smaller than C_α, where C_α is the lower α-percentage point of the distribution of $\int_0^1 W(t)dW(t)\Big/\sqrt{\int_0^1 W^2(t)dt}$. Notice that since the class of stationary processes satisfying Assumption 1 is very rich and contains as special case many linear and nonlinear processes including the commonly used linear AR(∞) process driven by i.i.d. innovations or by martingale differences, Theorem 2.1 generalizes considerably previous results regarding the limiting distribution of the ADF-test under the null hypothesis of a unit root.

Remark 2.2 Theorem 2.1 can be also extended to cover the case of a deterministic trend. In particular, if the process under H_0 is generated by the equation $Y_t = X_t + a + bt$, where (X_t) satisfies Assumption 2(i) and the regression equation

$$Y_t = \rho Y_{t-1} + a + bt + \sum_{j=1}^{p} a_{j,p}\Delta X_{t-j} + e_{t,p},$$

is fitted to the observed time series $Y_1, Y_2, \ldots, Y_n$, then the distribution of the least squares estimator $\widehat{\rho}_n$ of ρ is the same as the one given in Thorem 2.1 with the standard Brownian motion $W(t)$ replaced by

$$\widetilde{W}(t) = W(t) + (6t-4)\int_0^1 W(s)ds - (12t-6)\int_0^1 sW(s)ds.$$

2.3 Behavior Under the Alternative

Before stating the distribution of the least squares estimator $\widehat{\rho}_n$ under the alternative of stationarity, we discuss an asymptotic collinearity problem that occurs when regression equation (1.1) is fitted to a time series stemming from a stationary process. This collinearity problem is essential for understanding the effects of choosing the truncation parameter p on the power behavior of the test. The following proposition summarizes this behavior and is of interest on its own.

Proposition 2.1 *Let* $\{W_t, t \in \mathbb{Z}\}$ *be a zero mean, second order stationary process with autocovariance function* $\gamma_W(h) = E(W_t W_{t+h})$ *and spectral density* f_W *satisfying* $f_W(0) > 0$. *Denote by* $\mathcal{M}_{t,t-p} = \overline{sp}\{\Delta W_t,\ \Delta W_{t-1},\ \cdots \Delta W_{t-p}\}$ *the closed linear span generated by the differences* ΔW_{t-j}, $j = 0, 1, \ldots, p$ *and by* $\mathcal{P}_A(Y)$ *the orthogonal projection of* Y *onto the closed set* A.

(i) *If* $\gamma_W(h) \to 0$ *as* $h \to \infty$ *then* $E(W_t - \mathcal{P}_{M_{t,t-p}}(W_t))^2 \;\to\; 0$ *as* $p \to \infty$.

(ii) *If* $\sum_{h=-\infty}^{\infty} |\gamma_W(h)| < \infty$ *then* $p \cdot E(W_t - \mathcal{P}_{M_{t,t-p}}(W_t))^2 \;\to\; 2\pi f_W(0)$ *as* $p \to \infty$ *where* $f_W(\cdot)$ *denotes the spectral density of* $\{W_t\}$.

What the above proposition essentially says is that if W_t is a second order stationary process, then W_{t-1} can be expressed as a linear combination of it own differences ΔW_{t-j}, $j = 1, 2, \ldots$. Now, this proposition has serious consequences for the power behavior of the ADF-test under the alternative H_1. In particular, it implies a severe asymptotic collinearity problem that shows up when regression (1.1) is fitted to a stationary time series $X_1, X_2, \ldots, X_n$. To elaborate, under the alternative of stationarity, the random variables X_{t-1} and ΔX_{t-j}, $j = 1, 2, \ldots$ appearing on the right hand side of (2.7) are perfectly collinear. Consequently, in fitting (the truncated) equation (1.1), the random variables X_{t-1} and ΔX_{t-j}, $j = 1, 2, \ldots, p$ which appear as regressors, become asymptotically collinear as the truncation parameter p increases to infinity. Furthermore, the corresponding mean square prediction error $E(X_{t-1} - \mathcal{P}_{M_{t-1,t-p}}(X_{t-1}))^2$ converges at the rate $1/p$ as $p \to \infty$. This could be a severe problem even for small values of p, as in our data example, especially when the covariance structure of the process is significant only at small lags. This asymptotic collinearity problem occurs even if the underlying process $\{U_t\}$ is a finite, p-th order stationary autoregressive process and equation (1.1) is fitted to the observed time series using an truncation order p_n which is allowed to increase (to infinity) as the sample size n increases.

The next theorem establishes the limiting distribution of the least squares estimator $\widehat{\rho}_n$ under the alternative H_1 in (2.8).

Theorem 2.2 *Let Assumption 1 and Assumption 2(ii) be satisfied and suppose that* $p = p_n \to \infty$ *as* $n \to \infty$ *such that* $p_n^4/n^{1/2} \to 0$ *and* $\sqrt{n}\sum_{j=p+1}^{\infty} |a_j| \to 0$. *Then, as* $n \to \infty$,

(i) $\frac{n}{p} Var(\widehat{\rho}_n) \;\to\; (1-\rho)^2$, *in probability, and*

(ii) $\sqrt{\frac{n}{p}}(\widehat{\rho}_n - \rho) \;\Rightarrow\; N(0, (1-\rho)^2)$,

where $\rho = \sum_{j=1}^{\infty} b_j$.

Remark 2.3 Notice that because in regression (1.1) we are interested in estimating the parameter ρ only, we would expect, under the alternative of stationarity, that the estimator $\widehat{\rho}_n$ will be $\sqrt{n}$-consistent. However, the lower $\sqrt{n/p}$ convergence rate of $\hat{\rho}_n$ is due to the fact that estimating ρ is tantamount to estimating the spectral density of $\{X_t\}$ at frequency zero. In fact, using $2\pi f_X(\lambda) = \sigma_\varepsilon^2/|1 - \sum_{j=0}^{\infty} b_j \exp\{i\lambda j\}|^2$, we get that $\rho \;=\; 1 \;-\; \sigma_\varepsilon/\sqrt{2\pi f_X(0)}$. This makes it clear that although ρ appears to be a single parameter in the regression equation (1.1), estimating ρ is essentially a nonparametric estimation problem. This behavior of the estimator $\widehat{\rho}_n$ is regression (1.1) is different compared to the least squares estimator $\widehat{a}_n$ in regression (2.9) considered by Phillips and Perron (1988) that is $\sqrt{n}$-consistent under H_1. The reason for the different convergence rates of the two estimators under H_1 lies in the fact that $\widehat{\rho}_n$ in regression (1.1) estimates a function of the spectral density of $\{X_t\}$ at frequency zero, while $\widehat{a}_n$ in regression (2.9) estimates the first order autocorrelation, see also Remark 2.1. Note, however, that the Phillips and Perron (1988) test suffers from a difficulty of its own in that its estimated critical value is typically not $\sqrt{n}$–consistent as it is itself a function of the underlying spectral density.

Remark 2.4 For the alternative of stationary ARMA processes, Lopez (1997) claimed that he has derived the asymptotic distribution of $n^{1/2}p^{-1/2}(\widehat{\rho}_n - \rho)$ under the weaker condition that $p^3/n \to 0$ as $n \to \infty$. Apart from the more restrictive ARMA process set-up, this statement seems to be wrong. To derive the asymptotic distribution of $n^{1/2}p^{-1/2}(\widehat{\rho}_n - \rho)$ for this class of alternatives, Lopez (1997) proceeds by first replacing the ADF regression equation (1.1) by an autoregression containing only the levels of the X_t's, that is he considers instead of equation (1.1), the autoregression equation $X_t = \phi_{1,p}X_{t-1} + \phi_{2,p}X_{t-2} + \ldots + \phi_{p,p}X_{t-p} + v_{t,p}$; see equation (9) in Lopez (1997). Instead of the estimator $\widehat{\rho}_n$ he then investigates the estimator $\widehat{\phi}_n = \sum_{i=1}^{p} \widehat{\phi}_{i,p}$, where $\widehat{\phi}_{i,p}$ is the least squares estimator of $\phi_{i,p}$ in the aforementioned autoregression containing only levels. Using results obtained by Berk (1974), the limiting distribution of $\widehat{\phi}_n$ is then easily established allowing for the turncation lag p to increase to infinity such that $p^3/n \to 0$. However, the important step missing in this proof is the theoretical justification for the validity of this replacement in the regression problem considered. In fact, what one needs to show is that under the assumptions made,

$$\sqrt{\frac{n}{p}}(\widehat{\phi}_n - \widehat{\rho}_n) \to 0, \quad \text{in probability.}$$

We strongly doubt the validity of such a statement, since the regression equation using only levels of the X_t's does not suffer from the collinearity problems that are present in the regression equation (1.1) which contains also differences, see Proposition 2.1. The collinearity problems under the alternative, are those that force the truncation lag p in regression (1.1) to increase to infinity much slower compared to n in order to obtain consistency and asymptotic normality of the estimator $\widehat{\rho}_n$; compare Theorem 2.2.

Remark 2.5 Theorem 2.2 allows for the following approximative expression for the power function of the ADF-test for fixed alternatives,

$$P_{H_1}(t_n < C_\alpha) \approx \Phi\Big(\frac{\sqrt{n}\,\widehat{Std}(\widehat{\rho}_n)}{\sqrt{p}(1-\rho)}C_\alpha + \sqrt{\frac{n}{p}}\Big) \approx \Phi\Big(C_\alpha + \sqrt{\frac{n}{p}}\Big), \tag{2.10}$$

where C_α denotes the upper α-percentage point of the limiting distribution given in Theorem 2.1 and the second approximation follows since under H_1, $\widehat{Std}(\widehat{\rho}_n) = \sqrt{p}(1-\rho)/\sqrt{n} + o_P(\sqrt{p/n})$. Therefore, and since $n^{1/2}/p^{1/2} \to \infty$, the test is consistent but with a rate which is smaller than the parametric rate $n^{1/2}$. Furthermore, as it is seen from (2.10), asymptotically the power of the test is not affected by the distance between ρ and its value under the null hypothesis ($\rho = 1$) and the dominating $\sqrt{n/p}$-term is a decreasing function of p. This last property of the power function explains the empirically observed fact that increasing the truncation parameter p in (1.1) leads to a drop of power of the ADF-test; see e.g. Figures 9.2 and 9.5 of Patterson (2011).

3 A real data example and some practical issues

We now turn to a real data example that shows the power issues associated with the ADF test in practice. To this end, consider the dataset of Figure 1 that is extensively discussed in Example 1.2 of the well-known textbook by Shumway and Stoffer (2010). The data represent yearly average global temperatures deviations from 1880 to 2009 where the deviations are measured in degrees Celcius from the 1951-1980 average.

A familiar question is whether the data are trending and/or is temperature taking a 'random walk', i.e., does the temperature dataset have a unit root? Indeed, a linear trend can be readily noticed in the temperature dataset, and can help explain (at least in part) the strong autocorrelation characterizing the data pictured in Figure 2. However, we do not wish to focus on the Global Warming hypothesis here that would amount to checking the statistical significance of the linear trend. Rather, we want to test if there is a unit root process superimposed on the estimated trend whether the latter is negligible or not.

Using the `tseries` package in the R language, the P-values of the two aforementioned unit root tests were computed via fitting an equation that includes estimating a linear trend as in Remark 2.2. The P-value of the Phillips and Perron (1988) test was 0.01 while the P-value of the ADF test was 0.70. Obviously, this tremendous difference in the P-values raises serious concerns.

From the documentation of the `tseries` package it is made apparent that the `adf.test` function uses a default value for the order p given by the formula $p = \lfloor (n-1)^{1/3} \rfloor$; this implies a choice of $p = 5$ for our dataset where $n = 130$. We first note that this formula gives an acceptable rate for p under H_0 (where it is just needed that $p/\sqrt{n} \to 0$) but the rate is *not* acceptable under H_1 where it is needed that $p^4/\sqrt{n} \to 0$. In other words, the default formula gives a value for p that is too large for the asymptotics to work in the stationary case; indeed, we would need $p << n^{1/8} \approx 1.8$ when $n = 130$.

The detrended data are shown in Figure 3, and their correlogram of Figure 4 does not show particularly strong dependence. Indeed, the estimated lag-1 autocorrelation is about 0.6 which does not give strong evidence for a unit root. Furthermore, the partial autocorrelation of the detrended data shown in Figure 5 suggests that a stationary AR(1) model might be quite appropriate—at least if one is ready to treat the lag-4 value as negligible for the purposes of parsimony. Running the ADF regression with the choice $p = 1$ actually results in a P-value just slightly under 0.01 that is in close agreement with the P-value of the Phillips and Perron (1988) test.

The above discussion helps underscore both the claimed loss of power associated with even a moderately large value of p in the ADF test, as well as the need to scrutinize the choice of p in practice as the implications can be quite severe.

4 Proofs

Proof of Theorem 2.1: Using the notation $\Delta X_{t,p} = (\Delta X_t, \Delta X_{t-1}, \ldots, \Delta X_{t-p+1})'$, $\varepsilon_{t,p} = X_t - \rho X_{t-1} - \sum_{j=1}^{p} a_{j,p} \Delta X_{t-j}$ and $\varepsilon_{t,p} = X_t - \widehat{\rho}_n X_{t-1} - \sum_{j=1}^{p} \widehat{a}_{j,p} \Delta X_{t-j}$ with least squares

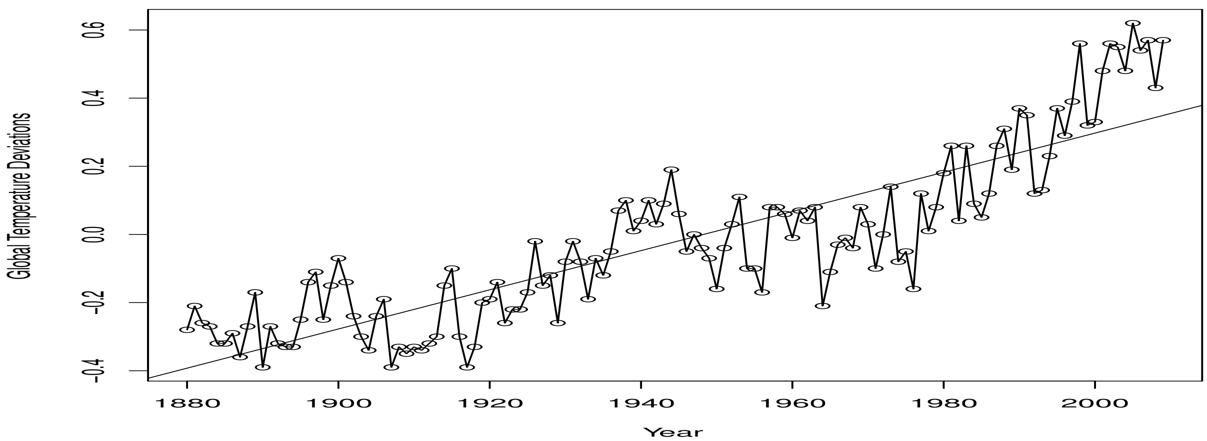

Figure 1: Yearly average global temperatures deviations with superimposed fitted linear trend; sample size $n = 130$.

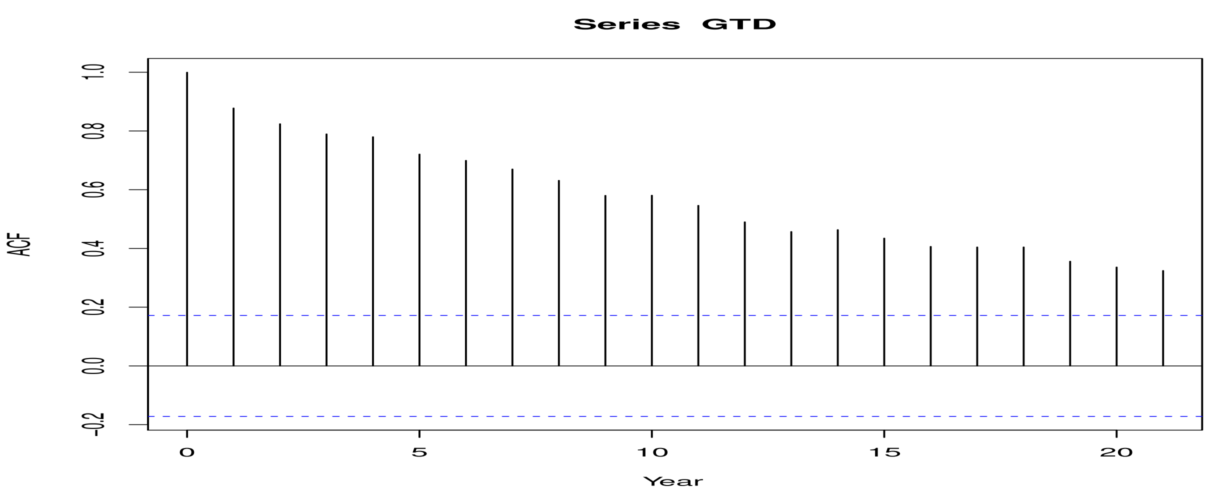

Figure 2: Correlogram of yearly average global temperatures deviations.

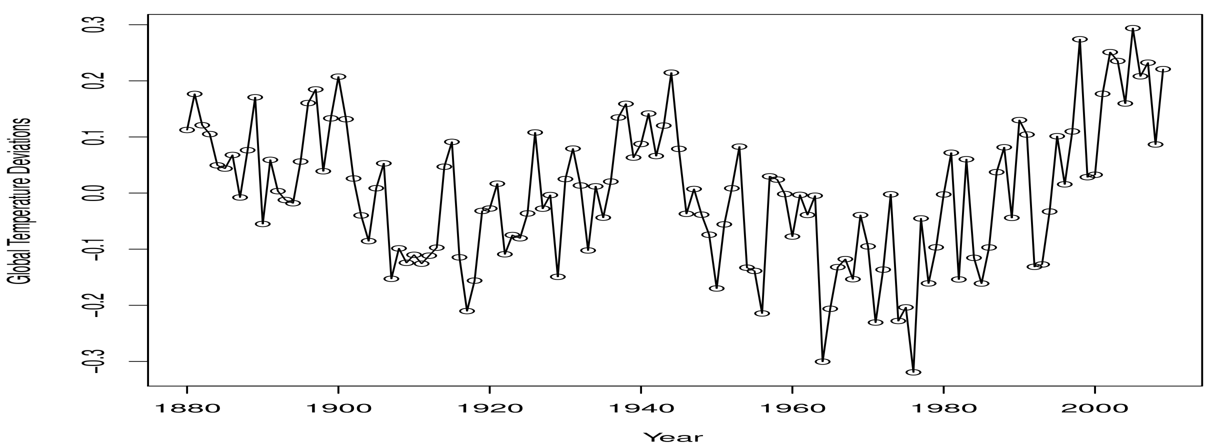

Figure 3: The dataset of yearly average global temperatures deviations after removal of a linear trend.

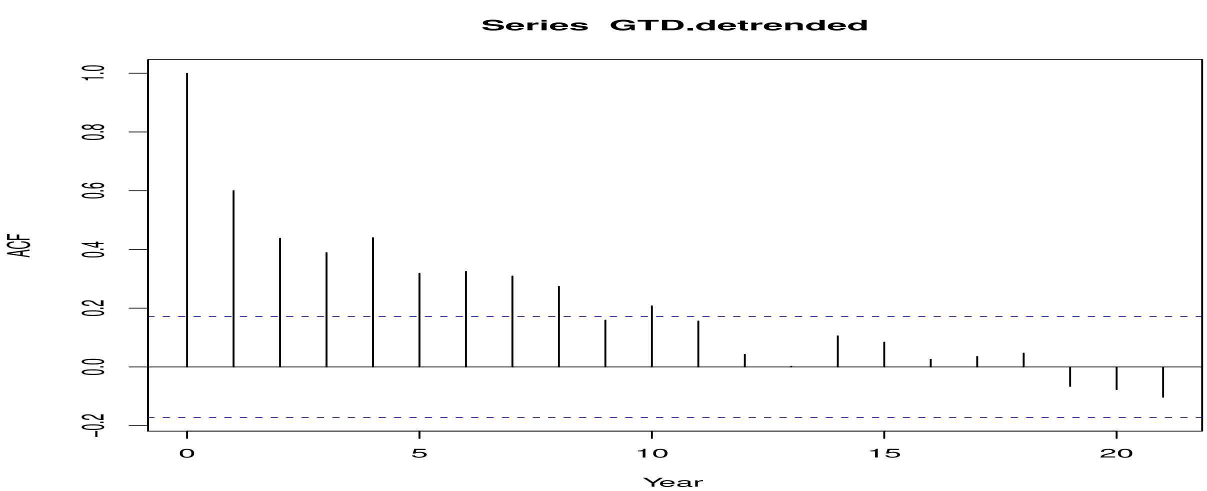

Figure 4: Correlogram of the detrended dataset of yearly average global temperatures deviations.

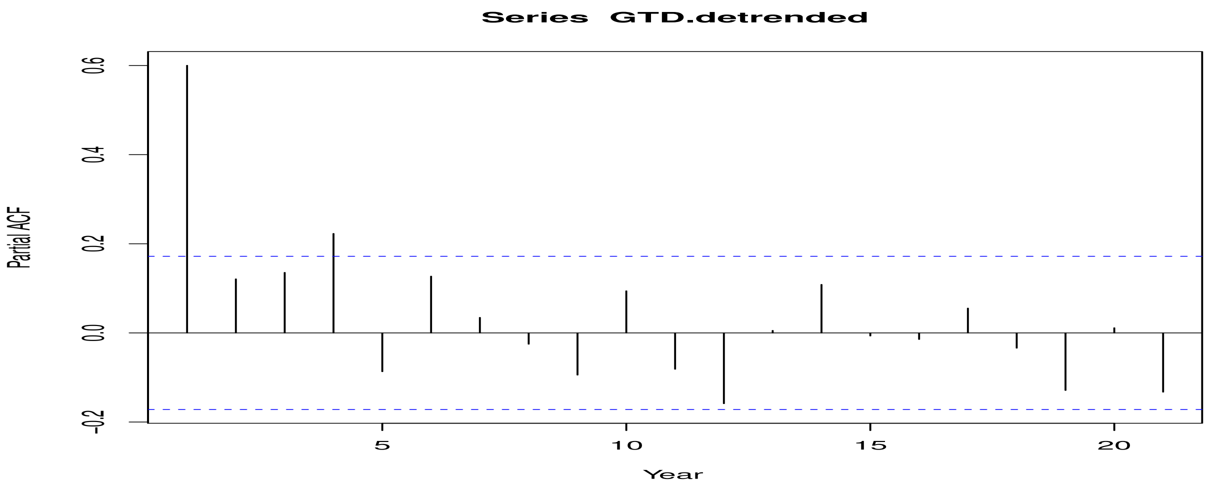

Figure 5: Partial autocorrelation of the detrended dataset of yearly average global temperatures deviations.

estimators $\widehat{\rho}_n$ and $\widehat{a}_{j,p}$, $j = 1, 2, \ldots, p$, it is easily verified that

$$t_n = (\widehat{\rho}_n - 1)/\widehat{Std}(\widehat{\rho}_n) = L_n R_n^{-1}/(\widehat{\sigma}_n^2 R_n^{-1})^{1/2},$$

where

$$L_n = \sum_{t=p+1}^{n} X_{t-1}\varepsilon_{t,p} - (\sum_{t=p+1}^{n} X_{t-1}\Delta X'_{t-1,p})(\sum_{t=p+1}^{n} \Delta X_{t-1,p}\Delta X'_{t-1,p})^{-1}(\sum_{t=p+1}^{n} \Delta X_{t,p}\varepsilon_{t,p}), \tag{4.11}$$

$$R_n = \sum_{t=p+1}^{n} X_{t-1}^2 - (\sum_{t=p+1}^{n} X_{t-1}\Delta X'_{t-1,p})(\sum_{t=p+1}^{n} \Delta X_{t-1,p}\Delta X'_{t-1,p})^{-1}(\sum_{t=p+1}^{n} \Delta X_{t,p}X_{t-1}) \tag{4.12}$$

and $\widehat{\sigma}_n^2 = (n-p)^{-1}\sum_{t=p+1}^{n} \widehat{\varepsilon}_{t,p}^2$ is the error variance estimator. Now, for $i, j \in \{1, 2, \ldots, p\}$ we have that, as $n \to \infty$, $n^{-1}\sum_{t=p+1}^{n} \Delta X_{t-i}\Delta X_{t-j} \to \gamma_U(i-j)$ in probability, and that, by the same arguments as in Berk (1974), p.493,

$$\|n^{-1}\sum_{t=p+1}^{n} \Delta X_{t-1,p}\Delta X'_{t-1,p}\| = O_P(1), \tag{4.13}$$

where for a matrix C, the norm $\|C\| = \sup_{\|x\|\leq 1}\|Cx\|$ is used and $\|x\|$ denotes the Euclidean norm of the vector x. Furthermore,

$$\begin{aligned} \frac{1}{np^{1/2}}\|\sum_{t=p+1}^{n} \Delta X_{t-1,p}X_{t-1}\| &= \Big(p^{-1}\sum_{j=1}^{p}(n^{-1}\sum_{t=p+1}^{n} \Delta X_{t-j}X_{t-1})^2\Big)^{1/2} \\ &= O_P(1) \end{aligned} \tag{4.14}$$

since $n^{-1}\sum_{t=p+1}^{n}\Delta X_{t-j}X_{t-1} = n^{-1}\sum_{t=p+1}^{n}\sum_{l=1}^{t-1}U_{t-j}U_l = O_P(1)$. Finally, since $n^{-1/2}\sum_{t=p+1}^{n}\Delta X_{t-j}\varepsilon_{t,p} = n^{-1/2}\sum_{t=p+1}^{n}U_{t-j}\varepsilon_{t,p} = O_P(1)$ we get that

$$\begin{aligned}\sqrt{p}\, n^{-1}\|\sum_{t=p+1}^{n}\Delta X_{t-1,p}\varepsilon_{t,p}\| &= \frac{p}{\sqrt{n}}\Big(p^{-1}\sum_{j=1}^{p}(\frac{1}{\sqrt{n}}\sum_{t=p+1}^{n}\Delta X_{t-j}\varepsilon_{t,p})^2\Big)^{1/2} \\ &= \frac{p}{\sqrt{n}}O_P(1)\to 0 \qquad (4.15)\end{aligned}$$

as $n\to\infty$. Now, equations (4.13) to (4.15) implies that $n^{-1}L_n = n^{-1}\sum_{t=p+1}^{n}X_{t-1}\varepsilon_{t,p}+o_P(1)$. Furthermore, because

$$n^{-1}\sum_{t=p+1}^{n}X_{t-1}(\varepsilon_{t,p}-\varepsilon_t) = \sum_{j=1}^{p}(a_{j,p}-a_j)n^{-1}\sum_{t=p+1}^{n}X_{t-1}U_{t-j} + \sum_{j=p+1}^{\infty}a_j n^{-1}\sum_{t=p+1}^{n}X_{t-1}U_{t-j},$$

if follows using $n^{-1}\sum_{t=p+1}^{n}X_{t-1}U_{t-j} = O_P(1)$ and Baxter's inequality, cf. Lemma 2.2 of Kreiss et al. (2011), that

$$|n^{-1}\sum_{t=p+1}^{n}X_{t-1}(\varepsilon_{t,p}-\varepsilon_t)| \leq O_P(\sum_{j=p+1}^{\infty}|a_j|)\to 0,$$

as $p\to\infty$. Thus,

$$n^{-1}L_n = n^{-1}\sum_{t=p+1}^{n}X_{t-1}\varepsilon_t + o_P(1). \qquad (4.16)$$

Similarly, using (4.13) and (4.14) we obtain that

$$n^{-2}R_n = n^{-2}\sum_{t=p+1}^{n}X_{t-1}^2 + o_P(1). \qquad (4.17)$$

Now, as in the proof of Theorem 3.1 in Phillips (1987) and using the invariance principle for the partial sum process $S_{[nr]} = n^{-1/2}\sum_{j=1}^{[nr]}\varepsilon_j$ of zero mean weakly dependent random variables satisfying Assumption 1(ii), established in Theorem 1 of Wu and Min (2005), we get that

$$n^{-1}L_n \Rightarrow \sigma_\varepsilon^2\int_0^1 W(t)dW(t), \quad \text{and} \quad n^{-2}R_n \Rightarrow \sigma_\varepsilon^2\int_0^1 W^2(t)dt.$$

□

Proof of Proposition 2.1: Let $\delta_{j,p}$, $j=0,1,\ldots,p$ be the coefficients of ΔX_{t-j} in the best linear prediction of X_t based on ΔX_{t-j} and define $l_{j,p} = (1-j/p)$, $j=0,1,\ldots,p$. We have

$$\begin{aligned}E(X_t-\mathcal{P}_{M_{t,t-p}}X_t)^2 &= E(X_t-\sum_{j=0}^{p}l_{j,p}\Delta X_{t-j})^2 + E(\sum_{j=0}^{p}(l_{j,p}-\delta_{j,p})\Delta X_{t-j})^2 \\ &\quad -2E(\sum_{j=0}^{p}(l_{j,p}-\delta_{j,p})\Delta X_{t-j})(X_t-\sum_{j=0}^{p}l_{j,p}\Delta X_{t-j}).\end{aligned}$$

Note first that $E(X_t - \sum_{j=0}^p l_{j,p}\Delta X_{t-j})^2 = p^{-1}[\gamma(0) + 2\sum_{s=1}^{p-1}(1-s/p)\gamma(s)]$, which converges to zero if $\gamma(h) \to 0$. Furthermore, if $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$ then $p\cdot E(X_t - \sum_{j=0}^p l_{j,p}\Delta X_{t-j})^2 \to \sum_{h=-\infty}^{\infty} \gamma(h) = 2\pi f_{X_t}(0)$ by the dominate convergence theorem. Now, let $X_t(p) = (X_t, X_{t-1}, \ldots, X_{t-p})'$, $\Gamma_{p+1} = E(X_t(p)X_t(p)')$ and define the $(p+1)$-dimensional vectors $\widetilde{\delta}(p) = ((1-\delta_{0,p}), (\delta_{0,p} - \delta_{1,p}), \ldots, (\delta_{p-1,p} - \delta_{p,p}), \delta_{p,p})$ and $\widetilde{l}(p) = (0, 1/p, 1/p, \ldots, 1/p, 0)'$. Then the following upper bound is valid,

$$\begin{aligned} E(\sum_{j=0}^p (l_{j,p} - \delta_{j,p})\Delta X_{t-j})^2 &= (\widetilde{l}(p) - \widetilde{\delta}(p))'\Gamma_{p+1}(\widetilde{l}(p) - \widetilde{\delta}(p)) \\ &\leq \max_{\lambda\in[0,\pi]} f_{X_t}(\lambda)\|\widetilde{l}(p) - \widetilde{\delta}(p)\|^2 \\ &\leq \max_{\lambda\in[0,\pi]} f_{X_t}(\lambda)\Big(2\|\widetilde{l}(p)\|^2 + 2\|\widetilde{\delta}(p)\|^2\Big). \end{aligned}$$

It is easily seen that $\|\widetilde{l}(p)\|^2 = O(p^{-1}) \to 0$. Furthermore, using the following lower bound for the mean square prediction error

$$E(X_t - \mathcal{P}_{M_{t,t-p}}X_t)^2 = \int_{-\pi}^{\pi} |\sum_{j=0}^p \widetilde{\delta}_{j,p}e^{-ij\lambda}|^2 f_{X_t}(\lambda)d\lambda \geq \inf_{\lambda\in[0,\pi]} f_{X_t}(\lambda)\|\widetilde{\delta}(p)\|^2,$$

we get that $\|\widetilde{\delta}(p)\|^2 \to 0$ as $p \to \infty$ from which it follows that $E(\sum_{j=0}^p(l_{j,p} - \delta_{j,p})\Delta X_{t-j})^2 \to 0$ as $p \to 0$. Finally, by the above results and Cauchy-Schwarz's inequality, it follows that $|E(\sum_{j=0}^p(l_{j,p} - \delta_{j,p})\Delta X_{t-j})(X_t - \sum_{j=0}^p l_{j,p}\Delta X_{t-j})| \to 0$ which concludes the proof. □

Proof of Theorem 2.2: Note that $\sqrt{n/p}\,(\widehat{\rho}_n - \rho) = L_n R_n^{-1}$ where L_n and R_n are defined in (4.11) and (4.12). Let $\hat{\gamma}_0 = (n-p)^{-1}\sum_{t=p+1}^n X_{t-1}^2$,

$$\hat{d}_p = \Big(\frac{1}{n-p}\sum_{t=p+1}^n \Delta X_{t-i}X_{t-1},\ i = 1, 2, \ldots, p\Big)', \quad \text{and} \quad \hat{C}_p = \Big(\frac{1}{n-p}\sum_{t=p+1}^n \Delta X_{t-i}\Delta X_{t-j}\Big)_{i,j=1,2,\ldots,p}.$$

We have that

$$\begin{aligned} n^{-1}R_n &= \widehat{\gamma}_0 - \widehat{d}_p'\widehat{C}_p^{-1}\widehat{d}_p \\ &= \gamma_0 - d_p'C_p^{-1}d_p + O_P(p^3/n^{1/2}), \end{aligned} \tag{4.18}$$

where $d_p' = (E(X_{t-1}\Delta X_{t-j}), j = 1, 2, \ldots, p)$ and $C_p = E(\Delta X_{t-1,p}\Delta X_{t-1,p}')$. Notice that the $O_P(p^3/n^{1/2})$ term in (4.18) appears because using the notation $\tau_p^2 = \gamma_0 - d_p'C_p^{-1}d_p$ and $\widehat{\tau}_p^2 = \widehat{\gamma}_0 - \widehat{d}_p'\widehat{C}_p^{-1}\widehat{d}_p$, we have that

$$|\widehat{\tau}_p^2 - \tau_p^2| \leq |\widehat{\gamma}_0 - \gamma_0| + \|\widehat{\delta}_p - \delta_p\|\|d_p\| + \|\widehat{d}_p - d_p\|\|\widehat{\delta}_p\|$$

where $\widehat{\delta}_p = \widehat{C}_p^{-1}\widehat{d}_p$ and $\delta_p = C_p^{-1}d_p$. Now, $\|\widehat{d}_p - d_p\| = O_P(p^{1/2}/n^{1/2})$ and

$$p^{-1/2}\|\widehat{d}_p\| = \{p^{-1}\sum_{j=1}^p((n-p)^{-1}\sum_{t=p+1}^n \Delta X_{t-j}X_{t-1})^2\}^{1/2} = O_P(1).$$

Furthermore,

$$\|\widehat{\delta}_p - \delta_p\| = O_P(p^{5/2}/n^{1/2}), \tag{4.19}$$

and

$$p^{-1}\|\widehat{\delta}_p\| = o_P(1), \tag{4.20}$$

which implies that $|\widehat{\tau}_p^2 - \tau_p^2| = O_P(p^3/n^{1/2})$. We show that (4.19) and (4.20) are true.

To see (4.19) notice first that

$$\widehat{\delta}_p - \delta_p = \widehat{C}_p^{-1}((n-p)^{-1}\sum_{t=p+1}^{n} \Delta X_{t-1,p} u_{t-1,p}),$$

where $u_{t-1,p} = X_{t-1} - \sum_{j=1}^{p} \delta_{j,p}\Delta X_{t-j}$ and $\delta_p = (\delta_{1,p}, \delta_{2,p}, \ldots, \delta_{p,p})$ are the coefficients of the best linear predictor of X_{t-1} based on ΔX_{t-j}, $j = 1, 2, \ldots, p$. Now,

$$\|\widehat{\delta}_p - \delta_p\| \leq \|\widehat{C}_p^{-1}\|\|(n-p)^{-1}\sum_{t=p+1}^{n} \Delta X_{t-1,p} u_{t-1,p}\| = O_P(p^{5/2}/n^{1/2}),$$

since $\|(n-p)^{-1}\sum_{t=p+1}^{n} \Delta X_{t-1,p}u_{t-1,p}\| = O_P(p^{1/2}/n^{1/2})$, and

$$\|\widehat{C}_p^{-1}\| = O_P(p^2). \tag{4.21}$$

To see why (4.21) is true notice that for every $p \in \mathbb{N}$ the matrix C_p is positive definite, $\|C_p^{-1}\|$ is the reciprocal of the minimal eigenvalue of C_p. Notice that the spectral density $f_{\Delta X_t}(\lambda)$ of the differenced process $\{\Delta X_t, t \in \mathbb{Z}\}$ satisfies $f_{\Delta X_t}(\lambda) = |1 - e^{i\lambda}|^2 f_{X_t}(\lambda)$ and that for the minimal eigenvalue of C_p we have

$$\begin{aligned}
\inf_{\|x\|=1}\sum_{j=1}^{p}\sum_{k=1}^{p} x_j \mathrm{Cov}(\Delta X_{t-j}, \Delta X_{t-k}) x_k &= \inf_{\|x\|=1}\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big|\sum_{j=1}^{p} x_j e^{ij\lambda}\Big|^2 f_{X_t}(\lambda)|1-e^{i\lambda}|^2 d\lambda \\
&\geq \inf_{\lambda\in[0,\pi]} f_{X_t}(\lambda) \inf_{\|x\|=1}\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big|\sum_{j=1}^{p} x_j e^{ij\lambda}\Big|^2 |1-e^{i\lambda}|^2 d\lambda \\
&= \inf_{\lambda\in[0,\pi]} f_{X_t}(\lambda)\widetilde{\lambda}_{\min},
\end{aligned}$$

where $\widetilde{\lambda}_{\min}$ denotes the minimal eigenvalue of the $p \times p$ covariance matrix of the process with spectral density $(2\pi)^{-1}|1-e^{i\lambda}|^2 = (2\pi)^{-1}2(1-\cos(\lambda))$, i.e. of the noninvertible MA(1) process $Y_t = \varepsilon_t - \varepsilon_{t-1}$. For this process the eigenvalues of the p-dimensional correlation matrix are given by $\widetilde{\lambda}_k = 2(1-\cos((k\pi)/(p+1)))$, $k = 1, 2, \ldots, p$. Thus, $\|C_p^{-1}\| \leq 1/K(1-\cos(\pi/(p+1))) = O(p^2)$, where the last equality follows because $[1-\cos(\pi/(p+1))] \sim p^{-2}$. Note that $\|C_p^{-1}\| \to \infty$ as $p \to \infty$ since the minimal eigenvalue of C_p approaches zero as $p \to \infty$. Furthermore, it is easily seen that $\|\widehat{C}_p - C_p\| = O_P(p/\sqrt{n})$ from which we get using $\|C_p^{-1}\| = O(p^2)$ that

$$\|\widehat{C}_p^{-1} - C_p^{-1}\| \leq \frac{\|C_p^{-1}\|^2\|\widehat{C}_p - C_p\|}{1 - \|\widehat{C}_p - C_p\|\|C_p^{-1}\|} = O_P(p^{5/2}/n^{1/2}).$$

Thus,

$$\|\widehat{C}_p^{-1}\| \leq \|C_p^{-1}\| + \|\widehat{C}_p^{-1} - C_p^{-1}\| = O_P(p^2 + p^{5/2}/n^{1/2}).$$

To see (4.20) notice that

$$\|\widehat{\delta}_p\| \leq \|\delta_p\| + \|\widehat{\delta}_p - \delta_p\|.$$

Now, $\|\widehat{\delta}_p - \delta_p\| = O_P(p^{5/2}/n^{1/2})$. Furthermore, for $l_p = (l_{1,p}, l_{2,p}, \ldots, l_{p,p})'$, $l_{j,p} = (1 - j/p)$, $j = 1, 2, \ldots, p$ we have $\|\delta_p\| \leq \|l_p\| + \|\delta_p - l_p\|$. Now, $\|l_p\| = O(\sqrt{p})$ while $\|l_p - \delta_p\| = o(p)$ which follows because

$$\begin{aligned}
(l_p - d_p)' C_p (l_p - d_p) &\geq \lambda_{\min} \|l_p - \delta_p\|^2 \\
&\geq \inf_{\lambda \in [0,\pi]} f_{X_t}(\lambda) 2(1 - \cos(\pi/(p+1)) \|l_p - d_p\|^2 \\
&\sim K p^{-2} \|l_p - \delta_p\|^2 \ \geq \ 0,
\end{aligned}$$

and as the proof of Proposition 2.1 shows $(l_p - d_p)' C_p (l_p - d_p) = E(\sum_{j=0}^{p} (l_{j,p} - \delta_{j,p}) \Delta X_{t-j})^2 \rightarrow 0$, as $p \rightarrow \infty$.

This concludes the proof of assertion (4.18).

Now,

$$p\tau_p^2 \rightarrow 2\pi f_{X_t}(0) = \sigma_\varepsilon^2 (1 - \rho)^{-2}, \tag{4.22}$$

by Proposition 2.1. Thus

$$\frac{p}{n} R_n = \sigma_\varepsilon^2 (1 - \rho)^{-2} + O_P(p^4/n^{1/2}). \tag{4.23}$$

Let $\widehat{V}_{t-1,p} = \sqrt{p}(X_{t-1} - \widehat{d}_p' \widehat{C}_p^{-1} \Delta X_{t-1,p})$ and $V_{t-1,p} = \sqrt{p}(X_{t-1} - d_p' C_p^{-1} \Delta X_{t-1,p})$. Then,

$$\begin{aligned}
\sqrt{\frac{p}{n}} L_n &= \frac{1}{\sqrt{n}} \sum_{t=p+1}^{n} \widehat{V}_{t-1,p} \varepsilon_{t,p} \\
&= \frac{1}{\sqrt{n}} \sum_{t=p+1}^{n} V_{t-1,p} \varepsilon_t + \frac{1}{\sqrt{n}} \sum_{t=p+1}^{n} (\widehat{V}_{t-1,p} - V_{t-1,p}) \varepsilon_t \\
&\quad + \frac{1}{\sqrt{n}} \sum_{t=p+1}^{n} \widehat{V}_{t-1,p} (\varepsilon_{t,p} - \varepsilon_t) \\
&= \frac{1}{\sqrt{n}} \sum_{t=p+1}^{n} V_{t-1,p} \varepsilon_t + L_{1,n} + L_{2,n},
\end{aligned}$$

with an obvious notation for the remainder terms $L_{1,n}$ and $L_{2,n}$. For these terms we have

$$\begin{aligned}
L_{1,n} &\leq \sqrt{p} \|\widehat{\delta} - \delta\| \|n^{-1/2} \sum_{t=p+1}^{n} \Delta X_{t-1,p} \varepsilon_t\| \\
&= O_P(p^{7/2}/n^{1/2}) \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty.
\end{aligned}$$

and

$$\begin{aligned} |L_{2,n}| &\leq (n^{-1}\sum_{t=p+1}^{n}\widehat{V}_{t-1,p}^2)^{1/2}(\sum_{t=p+1}^{n}(\varepsilon_{t,p}-\varepsilon_t)^2)^{1/2} \\ &= O_P(\sqrt{n}\sum_{j=p+1}^{n}|a_j|). \end{aligned}$$

Notice that the last equality above follows since under the assumptions made,

$$n^{-1}\sum_{t=p+1}^{n}\widehat{V}_{t-1,p}^2 = n^{-1}\sum_{t=p+1}^{n}p(X_{t-1}-\delta_p^{'}\Delta X_{t-1,p})^2 + o_P(1) = O_P(1),$$

and by Baxter's inequality, see Kreiss et al. (2011), Lemma 2.2,

$$E(\sum_{t=p+1}^{n}(\varepsilon_{t,p}-\varepsilon_t)^2) = O(\sqrt{n}\sum_{j=p+1}^{\infty}|a_j|).$$

The proof of the theorem is then concluded since under the assumptions made and by a central limit theorem for martingale differences, see Theorem 1 of Brown(1971),

$$\frac{1}{\sqrt{n}}\sum_{t=p+1}^{n}V_{t-1,p}\varepsilon_t \Rightarrow N(0,\sigma_\varepsilon^4(1-\rho)^{-2}). \tag{4.24}$$

□

Acknowledgment. The authors are grateful to David Stoffer and Adrian Trapletti for helpful discussions. Many thanks are also due to Dr. Li Pan and to the students of the MATH181E class at UCSD in Fall 2012 for their valuable input on the data example of Section 3. The research of the second author was partially supported by NSF grants DMS 12-23137 and DMS 13-08319.

References

[1] Abadir, K. M. (1993), On the Asymptotic Power of Unit Root Tests. *Econometric Theory*, 9, 189-221.

[2] Berk, K. (1974), Consistent Autoregressive Spectral Estimates. *Annals of Statistics*, 2, 489-502.

[3] Brown, B. M. (1971), Martingale Central Limit Theorems. *Annals of Mathematical Statistics*, 42, 59-66.

[4] Chang, Y. and J. Y. Park (2002), On the Asymptotics of ADF Tests for Unit Roots. *Econometric Reviews*, 21, 431-447.

[5] Dickey, D. A. and Fuller, W. A. (1979), Distribution of Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, 74, 427-431.

[6] Dickey, D. A. and Fuller. W. A. (1981), Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica*, 49, 1057-1072.

[7] Hamilton, J.D. (1994). *Time Series Analysis*, Princeton University Press.

[8] Kreiss, J. P., E. Paparoditis and D. N. Politis (2011), On the Range of Validity of the Autoregressive Sieve Bootstrap. *Annals of Statistics*, 39, 2103-2130.

[9] Lopez, J. H. (1997), The Power of the ADF Test. *Economics Letters*, 57, 5-10.

[10] Nabeya, S. and Tanaka, K. (1990), A General Approaxh to the Limiting Distribution for Estimators in Time Series Regression with Nonstable Autoregressive Errors. *Econometrica*, 58, 145-163.

[11] Ng, S. and P. Perron (1995), Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag. *Journal of the American Statistical Association*, 90, 268-281.

[12] Patterson, K. (2011), *Unit Root Tests in Time Series, Vol. I*, Palgrave Texts in Econometrics, Palgrave Macmillan, Hampshire, U.K.

[13] Pourahmadi, M. (2001), *Foundations of Time Series Analysis and Prediction Theory,* John Wiley, New York.

[14] Perron, P. (1991), A Continous Time Approximation to the Unstable First-Order Autoregressive Process: The Case Without an Intercept. *Econometrica*, 59, 211-236.

[15] Phillips, P. C. B. (1987), Time Series Regression with a Unit Root. *Econometrica*, 2, 277-301.

[16] Phillips, P. C. B. and P. Perron (1988), Testing for a Unit Root in Time Series Regression. *Biometrika*, 75, 335-346.

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

[18] Shumway, R.H. and Stoffer, D.S. (2010). *Time Series Analysis and Its Applications: With R Examples*, 3rd Ed., Springer, New York.

[19] Wu, B. W. (2005), Nonlinear System Theory: Another look at Dependence. *Proc. Nat. Journal Acad. Sci. USA*, 102, 14150-14154.

[20] Wu, B. W. and Min, W. (2005), On Linear Processes with Dependent Innovations. *Stochastic Processes and their Applications*, 115, 939-958.