BOOTSTRAP MODEL AVERAGING UNIT ROOT INFERENCE

BRUCE E. HANSEN

Department of Economics, Social Science Building, University of Wisconsin, Madison, WI USA 53706-1396, behansen@wisc.edu.

JEFFREY S. RACINE

Department of Economics and Graduate Program in Statistics, McMaster University, racinej@mcmaster.ca; Department of Economics and Finance, La Trobe University; Info-Metrics Institute, American University; Rimini Center for Economic Analysis.

Abstract. Classical unit root tests are known to suffer from potentially crippling size distortions, and a range of procedures have been proposed to attenuate this problem, including the use of bootstrap procedures. It is also known that the estimating equation's functional form can affect the outcome of the test, and various model selection procedures have been proposed to overcome this limitation. In this paper, we adopt a model averaging procedure to deal with model uncertainty at the testing stage. In addition, we leverage an automatic model-free dependent bootstrap procedure where the null is imposed by simple differencing (the block length is automatically determined using recent developments for bootstrapping dependent processes). Monte Carlo simulations indicate that this approach exhibits the lowest size distortions among its peers in settings that confound existing approaches, while it has superior power relative to those peers whose size distortions do not preclude their general use. The proposed approach is fully automatic, and there are no nuisance parameters that have to be set by the user, which ought to appeal to practitioners.

1. Introduction

Though unit root tests were developed over four decades ago, problems with the various approaches that have been proposed persist and, perhaps surprisingly, there remains room for improvement. When testing for a unit root, the null is that the series contains a unit root, with rejection of the null in one direction indicating that a series is stationary, and rejection in the other

Date: August 11, 2018.

Key words and phrases. Inference, model selection, size distortion, time series.

Racine would like to gratefully acknowledge support from the Natural Sciences and Engineering Research Council of Canada (NSERC:www.nserc.ca), the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca). Hansen thanks the National Science Foundation and the Phipps Chair for research support. We would like to thank but not implicate James MacKinnon, Rob Hyndman and In Choi for their insight and encouragement.

direction indicating explosiveness. If these tests exhibit large upwards size distortions, practitioners may wrongly conclude that a time series is stationary (or explosive) when in fact it is not, which can render subsequent inference invalid. Size distortions surface surprisingly often in this setting since practitioners must select from among a range of candidate estimating equations when testing for the presence of a unit root, and estimating equation mis-specification leads to bias in estimated parameters. To deal with the size distortions arising from mis-specification, bootstrap procedures have been proposed (Park (2003), Palm, Smeekes & Urbain (2008)). However, for many of these tests, size distortions and sub-optimal power concerns persist in part because they rely on a model specification which must be selected by the practitioner from amount a set of mis-specified candidate models. In order to attenuate the size distortions arising from the choice of a mis-specified estimating equation, the use of model selection criteria such as the Bayes Information Criterion (BIC) (Schwarz 1978) has been advocated (Ng & Perron 2001). Ng & Perron (2001) assert that the BIC selects overly parsimonious models and propose a Modified Information Criteria (MIC). For Ng & Perron's (2001) approach, size distortions are attenuated as the dimension of the selected model increases, however this is not without cost as power falls as the dimension increases. Furthermore, the maximum lag that must be set by the practitioner affects the outcome of the test and ad-hoc rules are frequently adopted for its selection (Schwert 1989).

As an alternative to model selection, we could instead exploit recent developments in (frequentist) model averaging (Hansen (2007, Mallows Model Averaging (MMA)), Hansen & Racine (2012, Jackknife Model Averaging (JMA)), Hansen (2014)), as it is known that model averaging can overcome limitations associated with the use of model selection methods. In the present context, we propose a unit root statistic that is a weighted average of unit root statistics taken from a set of candidate estimating equations that are the same as those used for Ng & Perron's (2001) or Dickey & Fuller's (1979) approaches. We also adopt an automatic, model-free, time series bootstrap procedure for constructing the null distribution of the proposed statistic, from which nonparametric critical values or nonparametric P-values can be obtained. Most existing bootstrap unit root procedures are model-based, and one problem with model-based resampling is that the data generating process (DGP) is unknown and must be identified from the series at hand. In order to ensure that the bootstrap samples have the same structure as the series at hand, this

identification must be correct. Swensen (2003) considers model-based and model-free bootstrap approaches and demonstrates that a difference-based model-free approach (along the lines of that proposed herein) delivers a bootstrap distribution that approaches the true asymptotic distribution under the null of a unit root (see also Palm et al. (2008, Section 2.4)).

We will see that when model averaging is combined with a model-free, automatic, time series bootstrap procedure (Politis & Romano 1994, Politis & White 2004, Patton, Politis & White 2009), we can obtain a fully automatic data-driven procedure that has superior size *and* power relative to those peers whose size distortions do not preclude their general use, while the sensitivity to the dimension of the model (i.e., the maximum lag that must be set by the practitioner) is attenuated by averaging over a set of candidate estimating equations in a particular manner. Furthermore, unlike its peers, the procedure is very robust to the number of and maximum dimension of the candidate estimating equations over which the averaging is performed (size and power are largely unaffected whether you use augmented Dickey-Fuller models with, e.g., one through two, four, eight, sixteen, or twenty four differenced lags of the time series). To the best of our knowledge, ours is the first to consider model averaging procedures in the unit root setting.

We compare the proposed bootstrap model average approach with the classic Dickey & Fuller (1979) method, Phillips & Perron's (1988) procedure, and with Ng & Perron's (2001) procedure (using the detrending approach of Perron & Qu (2007)) which is a state-of-the-art procedure that appears to be the go-to method for most practitioners. The proposed bootstrap model average approach emerges as the procedure of choice based upon a fairly extensive Monte Carlo comparison with the existing go-to and classical approaches. Ng & Perron's (2001) procedure is based on the same estimating equations as Dickey & Fuller (1979), but they first detrend the series (this reduces size distortions when there is a large negative moving average root in the differenced series) and then use a novel lag selection procedure that chooses a larger lag length than traditional lag selection procedures. The Dickey & Fuller (1979) and Ng & Perron (2001) tests use a parametric autoregression to approximate the ARMA structure of the errors in the test regression, while Phillips & Perron's (1988) procedure instead corrects for any serial correlation and heteroskedasticity in the errors of the estimating equation by directly modifying the test statistic. We direct the reader to the original references for further details.

It has been established that the classical approach (Dickey & Fuller 1979) which uses tabulated critical values sometimes suffers from crippling size distortions. MacKinnon's (1996) improved tabulated critical values do not attenuate such size distortions, unfortunately. Schwert (1989) conducted extensive Monte Carlo simulations and demonstrated how there can exist considerable bias present in a mis-specified estimating equation for unit root testing in the presence of a moving average error process (i.e., in the presence of a large negative moving average root). In such cases, the critical values depend on the unknown parameters hence tabulated Dickey-Fuller critical values should be avoided, and appropriate bootstrap procedures may be necessary for sound inference in this setting. DeJong, Nankervis, Savin & Whiteman (1992), again via extensive simulations, demonstrate that the augmented Dickey-Fuller tests have low power in the presence of a large autoregressive root. One important practical aspect for augmented Dickey-Fuller unit root tests is the specification of the lag length. If it is too small, then the remaining serial correlation in the errors will bias the test. If it is too large, then the power of the test will suffer. Hansen (1995) demonstrates how large power gains can be achieved by including correlated stationary covariates in the estimating equation (this could be incorporated in our proposed bootstrap model average approach). Ng & Perron (2001) point out that a high order augmented autoregression is often necessary for unit root tests to have good size, but that information criteria such as the BIC tend to select a truncation lag that is small, and propose a Modified Information Criteria (MIC) along with GLS detrended data and demonstrate how this improves size but can lead to a loss in power. Perron & Qu (2007) propose an improved method for detrending for the Ng & Perron (2001) approach. We direct the interested reader to Choi (2015) who presents a state-of-the art treatment of unit root inference.

The rest of this paper proceeds as follows. Section 2 presents some background on unit root inference and presents theoretical underpinnings and the distribution of the weighted average ADF statistic with fixed weights. Section 3 outlines the proposed approach. Section 4 presents results from a Monte Carlo simulation that compares the proposed approach with the classic Dickey & Fuller (1979) procedure, that of Ng & Perron (2001), and a bootstrap BIC model selection procedure. An R (R Core Team 2018) package exists that implements the proposed method (Racine 2018).

2. Testing for a Unit Root

Consider a time series y_t which satisfies an autoregressive equation

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

for $t = 1, \ldots, T$, where u_t is a stationary I(0) process. The latter includes i.i.d. processes, white noise, and mean-reverting stationary processes. When $\rho = 1$ then y_t is I(1) and we say that y_t has a unit root in its autoregressive representation. On the other hand when $|\rho| < 1$ then y_t is I(0), is stationary, and does not have a unit root. When $\rho > 1$ then y_t is explosive. Testing $H_0 \colon \rho = 1$ versus either $H_1 \colon \rho \neq 1$ or $H_1 \colon \rho < 1$ are important practical issues in applied time series modeling.

The most common method for testing the unit root hypothesis is to use the Augmented Dickey-Fuller (ADF) statistic, which is based on the least-squares estimation of an autoregressive (AR) model for y_t. The test can be described as follows. For some lag order k, estimate by least squares the k^{th} order autoregression

$$\Delta y_t = \widehat{\gamma}(k) y_{t-1} + \sum_{j=1}^{k-1} \widehat{a}_j(k) \Delta y_{t-j} + \widehat{\beta}(k) + \widehat{\epsilon}_t(k).$$

Form a t-statistic for the null that $\gamma = 0$, that is

$$ADF(k) = \frac{\widehat{\gamma}(k)}{s\left(\widehat{\gamma}(k)\right)}$$

where $s\left(\widehat{\gamma}(k)\right)$ is a standard error for $\widehat{\gamma}(k)$. The test rejects in favour of a stationary alternative for large negative values of $ADF(k)$ and in favour of non-stationary explosive alternatives for positive (or small negative) values. For stationary trend alternatives a linear time trend is also included in the regression.

The conventional asymptotic distribution theory approximates the null distribution of $ADF(k)$ by either assuming that k is the true autoregressive order (so that the estimated model is correctly specified) or by assuming that $k \to \infty$ as $T \to \infty$ so that the model is approximately correct. We take a different approach and derive the asymptotic distribution without either of these assumptions.

We use the following regularity condition on the fundamental errors u_t defined in (1).

Assumption 2.1. *For some $p > r > 2$, u_t is a strictly stationary, zero mean, strong mixing process of size $-pr/(p-r)$, $E|u_i|^p < \infty$, and $\omega^2 > 0$ where*

$$\omega^2 = \sum_{j=-\infty}^{\infty} E(u_t u_{t-j}).$$

Assumption 2.1 is a mild set of standard mixing conditions which allow for broad I(0) processes, and encompasses standard AR and ARMA processes. The assumption that the long-run variance ω^2 is positive excludes over-differenced processes.

Our representation of the asymptotic distribution of the ADF statistic will be written in terms of the approximating models. For each k define the approximate model

$$u_t = \sum_{j=1}^{k-1} a_j(k) u_{t-j} + \epsilon_t(k)$$

by projection. That is, the coefficients $a_j(k)$ are defined so that $E(u_{t-j}\epsilon_t(k)) = 0$ for $j = 1, \ldots, k-1$. This defines the AR(k) approximate model and error. Given the error $\epsilon_t(k)$, we can define its variance, autocovariance and long-run variance

$$\sigma^2(k) = E\left(\epsilon_t(k)^2\right) \tag{2}$$

$$\lambda(k) = \sum_{j=1}^{\infty} E(\epsilon_t(k)\epsilon_{t-j}(k)) \tag{3}$$

$$\nu^2(k) = \sigma^2(k) + 2\lambda(k). \tag{4}$$

The parameter $\sigma^2(k)$ is the variance of $\epsilon_t(k)$, $\lambda(k)$ is the sum of its autocovariances, and $\nu^2(k)$ is its long-run variance. Under mis-specification the error $\epsilon_t(k)$ has serial correlation so that $\lambda(k) \neq 0$ and $\sigma^2(k) \neq \nu^2(k)$. Under correct specification the error is white noise so $\lambda(k) = 0$ and $\sigma^2(k) = \nu^2(k)$. Thus for small k we expect $\lambda(k) \neq 0$ and $\sigma^2(k) \neq \nu^2(k)$ but for large k we expect $\lambda(k) \simeq 0$ and $\sigma^2(k) \simeq \nu^2(k)$, though there is no reason to expect equality for any finite k.

Theorem 2.1. *Under Assumption 2.1 and H_0: $\rho = 1$, jointly over $k = 1, \ldots, K$, as $T \to \infty$*

$$ADF(k) \xrightarrow{d} \frac{\nu(k)}{\sigma(k)} \frac{\int_0^1 W^* dW}{\left(\int_0^1 W^{*2}\right)^{1/2}} + \frac{\lambda(k)}{\omega\sigma(k)\left(\int_0^1 W^{*2}\right)^{1/2}} \tag{5}$$

where $W(r)$ is a standard Brownian motion, and $W^*(r) = W(r) - \int_0^1 W(r)dr$ (or a detrended Brownian motion if a time trend is included).

Theorem 2.1 shows that the ADF t-statistics converge jointly to mis-specified versions of the classic Dickey-Fuller t-distribution. The distortions are due to mis-specified serial correlation. When the autoregression is correctly specified so that the error is white noise, then $\lambda(k) = 0$ in which case the distribution in (5) simplifies to the classical $\int_0^1 W^* dW / \left(\int_0^1 W^{*2}\right)^{1/2}$ found by Dickey & Fuller (1979).

Theorem 2.1 also shows that the sequence of t-statistics (for different autoregressive orders) converge jointly, and are all functions of the same Brownian motion process $W(r)$.

The asymptotic distribution in Theorem 2.1 is generally unknown as it depends on the unknown parameters $\sigma(k)$ and $\lambda(k)$. However, the distribution can be approximated by bootstrap methods since these parameters can be consistently estimated.

By picking a suitably large autoregressive order k the distributional distortions can be minimized. Larger values of k, however, reduce the power of unit root tests in finite samples. Thus it has been viewed as desirable to use an autoregressive order k which is large enough to minimize the size distortions but not so large as to reduce the power of the test. This requires a data-dependent rule $\widehat{k}$ for selection of k. One popular method is BIC selection. However, Ng & Perron (2001) argued that this produces a $\widehat{k}$ which is too small to alleviate the size distortion, and proposed instead a Modified Information Criteria (MIC) designed for the unit root testing problem.

A data-dependent selection rule $\widehat{k}$ leads to a data-dependent ADF test $ADF(\widehat{k})$. The appropriate null distribution for $ADF(\widehat{k})$ is unclear, however, as the use of a selected lag length invalidates the conventional limit theory unless used with an ad hoc assumption that $\widehat{k}$ diverges with T. Bootstrap critical values could be used instead though no formal justification has been provided.

Instead of selection rules, we propose an averaging statistic. For $k = 1, \ldots, K$ let $w(k)$ be a set of non-negative weights which sum to one, and set $w = (w(1), \ldots, w(K))$. Then an averaging ADF statistic is

$$ADF(w) = \sum_{k=1}^{K} w(k) ADF(k).$$

The asymptotic distribution of the averaging ADF statistic can be deduced directly from Theorem 2.1.

Theorem 2.2. *As* $T \to \infty$

$$ADF(w) \xrightarrow{d} \left(\sum_{k=1}^{K} w(k)\frac{\nu(k)}{\sigma(k)}\right)\frac{\int_0^1 W^* dW}{\left(\int_0^1 W^{*2}\right)^{1/2}} + \left(\sum_{k=1}^{K} w(k)\frac{\lambda(k)}{\omega\sigma(k)}\right)\frac{1}{\left(\int_0^1 W^{*2}\right)^{1/2}}.$$

Theorem 2.2 provides the asymptotic distribution of the averaging ADF statistic for fixed weights. Like the distribution in Theorem 2.1, it is a distorted version of the classic Dickey-Fuller t-distribution.

While the distribution in Theorem 2.2 is generally unknown, and dependent on the unknown serial correlation properties of the series Δy_t, it can be approximated by standard bootstrap methods since the serial correlation properties can be consistently estimated.

Proof of Theorem 2.1. For simplicity we omit the deterministic components from the exposition. By the Herrndorf (1984) functional central limit theorem

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} u_t \Rightarrow \omega W(r)$$
$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} \epsilon_t(k) \Rightarrow \nu(k) W(r)$$

where $W(r)$ is standard Brownian motion. This convergence holds jointly over both equations and over k since the $\epsilon_t(k)$ are linear transformations of the errors u_t. Applying Theorem 4.1 of Hansen (1992)

$$\frac{1}{T}\sum_{t=1}^{T} y_{t-1}\epsilon_t(k) \xrightarrow{d} \omega\nu(k)\int_0^1 W dW + \lambda(k)$$

jointly over k.

Let $\widehat{a}(k) = (\widehat{a}_1(k), \ldots, \widehat{a}_{k-1}(k))'$ and $a(k) = (a_1(k), \ldots, a_{k-1}(k))'$. Define $x_t(k) = (y_{t-1}, \ldots, y_{t-k+1})'$, $Q(k) = E\left(x_t(k)x_t(k)'\right)$ and

$$\Omega(k) = \sum_{j=-\infty}^{\infty} E\left(x_t(k)\epsilon_t(k)x_{t-j}(k)'\epsilon_{t-j}(k)\right).$$

By standard manipulations for the asymptotic theory of integrated processes

$$\begin{pmatrix} T\widehat{\gamma}(k) \\ \sqrt{T}\left(\widehat{a}(k) - a(k)\right) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \omega^2 \int_0^1 W^2 & 0 \\ 0 & Q(k) \end{pmatrix}^{-1} \begin{pmatrix} \omega\nu(k) \int_0^1 W dW + \lambda(k) \\ \xi(k) \end{pmatrix}$$

where $\xi(k) \sim N(0, \Omega(k))$. This convergence is joint across k.

Furthermore, the standard errors satisfy

$$T\left(s\left(\widehat{\gamma}(k)\right)\right)^2 \xrightarrow{d} \left(\omega^2 \int_0^1 W^2\right)^{-1} \sigma^2(k).$$

Together, we find

$$ADF(k) \xrightarrow{d} \frac{\omega\nu(k) \int_0^1 W dW + \lambda(k)}{\left(\omega^2 \int_0^1 W^2\right)^{1/2} \sigma(k)} = \frac{\nu(k)}{\sigma(k)} \frac{\int_0^1 W dW}{\left(\int_0^1 W^2\right)^{1/2}} + \frac{\lambda(k)}{\omega\sigma(k) \left(\int_0^1 W^2\right)^{1/2}}$$

as claimed. □

3. A Model Averaging Bootstrap Procedure

3.1. **Model Average Estimators.** The goal in model averaging is to reduce estimation variance while controlling mis-specification bias. The Mallows (Mallows 1973) Criterion for the model average estimator (Hansen 2007) is

$$C_n(w) = w'\hat{E}'\hat{E}w + 2\sigma^2\mathcal{K}'w,$$

where $\hat{E}$ is the $T \times M$ matrix with columns containing the residual vector from the mth candidate estimating equation, $\mathcal{K}$ the $M \times 1$ vector of the number of parameters in each model, and σ^2 the variance from the largest dimensional model.[1] This criterion is used to select the weight vector $\hat{w}$, i.e.,

$$\hat{w} = \arg\min_w C_n(w).$$

Because $\arg\min_w C_n(w)$ has no closed-form solution, the weight vector is found numerically. The solution involves constrained minimization subject to non-negativity and summation constraints, which constitutes a classic quadratic programming problem. This criterion involves nothing more

[1]Note that the residual vectors will be of different lengths when the model incorporates lags, so some care must be exercised when populating $\hat{E}$, i.e., the first $k-1$ elements from the residual vector for the estimating equation models not containing lags must be discarded.

than computing the residuals for each candidate estimating equation, obtaining the rank of each candidate estimating equation, and solving a simple quadratic program. The MMA $C_n(w)$ criterion provides an estimate of the average squared error from the model average fit, and has been shown to be asymptotically optimal in the sense of achieving the lowest possible squared error in a class of model average estimators. See Hansen (2007) for further details. See also Hansen (2014) who explores the use of the Mallows criterion in a time series autoregression setting and notes that averaging estimators have reduced risk relative to unconstrained estimation when the covariates are grouped in sets of four or larger so that a Stein shrinkage effect holds, and suggests that averaging estimators be restricted to models in which the regressors have been grouped in this manner. See also the related work of Hansen (2010) who uses a Mallows criterion for combining forecasts local to a unit root. In our procedure we use Schwert's (1989) rule for determining the dimension of the largest ADF model that is averaged over, and group candidate models per Hansen (2014).

Hansen & Racine (2012) propose an alternative jackknife model averaging (JMA) criterion for the model average estimator given by

$$CV_n(w) = \frac{1}{n}(y - \tilde{X}w)'(y - \tilde{X}w),$$

where $\tilde{X}$ is the $T \times M$ matrix with columns containing the jackknife estimator from the m^{th} candidate estimating equation formed by deleting the t observation when constructing the t^{th} prediction. Like its Mallows counterpart, this involves solving a quadratic program where we minimize $(y - \tilde{X}w)'(y - \tilde{X}w) = y'y + w'\tilde{X}'\tilde{X}w - 2y'\tilde{X}w$ and the first term is ignorable. In the presence of homoskedastic errors, JMA and MMA are nearly equivalent, but when the errors are heteroskedastic, JMA has significantly lower MSE.

To obtain a model average test statistic, we take the $ADF(k)$ statistic from each of the K candidate estimating equations and average them using the weight vector $\hat{w}$, and call this averaged statistic $ADF(w) = \sum_{k=1}^{K} \hat{w}(k)ADF(k)$. In order to obtain the null distribution of this statistic, we use a time series bootstrap with automatic choice of the expected block length.

3.2. **A Unit Root Model Average Bootstrap Procedure.** We consider a first-difference-based bootstrap procedure for obtaining the sampling distribution of $ADF(w)$ under the null of a unit root

along the lines of Swensen (2003), who proves the consistency of the standard (non-averaged) test without deterministic components based on the stationary bootstrap (deterministic components can be added in the same manner as in Psaradakis (2001); see Palm et al. (2008, page 382)). The bootstrap procedure is as follows:

(1) Take the first difference of the series at hand $\epsilon_t = \Delta y_t$, $t = 1, \ldots, T$ (ϵ_t could be, e.g., an ARMA process)
(2) Apply a time series bootstrap with automated block length choice to ϵ_t, $t = 1, \ldots, T$ (l is the expected block length obtained for the geometric bootstrap; see Patton et al. (2009), Politis & White (2004), Politis & Romano (1994))
(3) Take the cumulative sum of this bootstrap residual ϵ_t^*, $t = 1, \ldots, T$ initializing the sum to the first realization of the series, which will generate a bootstrap series containing a unit root, i.e., $y_t^* = y_1 + \sum_{i=2}^{t} \epsilon_i^* = y_{t-1}^* + \epsilon_t^*$
(4) Next, take the bootstrap $ADF(k)^*$ statistics from each of the K candidate estimating equations and average them using the weight vector for the original series $\hat{w}$, which delivers a bootstrap model average statistic $ADF(w)^*$ generated under the null
(5) Repeat this process B times to obtain the B bootstrap statistics $ADF(w)^*_{1,K}, \ldots, ADF(w)^*_{B,K}$.
(6) For the one-sided stationary alternative $H_1 \colon \rho < 1$ compute the α empirical quantile q_α from the bootstrap statistics. Reject $H_0 \colon \rho = 1$ in favor of $H_1 \colon \rho < 1$ if $ADF(w) < q_\alpha$.
(7) For the two-sided alternative $H_1 \colon \rho \neq 1$ compute the $\alpha/2$ and $1 - \alpha/2$ empirical quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ from the bootstrap statistics. Reject $H_0 \colon \rho = 1$ in favor of $H_1 : \rho \neq 1$ if $ADF(w) < q_{\alpha/2}$ or $ADF(w) > q_{1-\alpha/2}$.

4. Monte Carlo Simulation for Bootstrap Unit Root Test

In order to assess the finite-sample performance of the proposed approach relative to its peers, we conduct a series of Monte Carlo simulation experiments. In particular, we construct power curves for a handful of procedures based upon three DGPs, one a simple AR(1), one an ARMA(1,1), and one an ARMA(1,2). The ARMA(1,1) DGP was used in simulations appearing in Palm et al. (2008) ($y_t = \rho y_{t-1} + \epsilon_t - 0.8\epsilon_{t-1}$) and is known to confound existing tests. It is noteworthy that Phillips & Perron (1988, p. 344) point out a limitation of their approach writing that their

tests "have significant size distortions and are too liberal to be useful for $\theta = -0.5, -0.8$" (here $\theta = -0.8$ is the MA coefficient in the ARMA(1,1) model). For the ARMA(1,2) DGP we use $y_t = \rho y_{t-1} + \epsilon_t + 0.3\epsilon_{t-1} - 0.2\epsilon_{t-1}$, while for the AR(1) DGP we use $y_t = \rho y_{t-1} + \epsilon_t$.

We conduct $B = 399$ bootstrap replications and $M = 2500$ Monte Carlo replications for sample sizes $T = (50, 100, 200, 400)$. We then conduct the two-sided bootstrap test described above using the MMA and JMA weighting scheme and report the empirical rejection frequency of it and the one-sided tests of Ng & Perron (2001), Phillips & Perron (1988) and Dickey & Fuller (1979), where each test is conducted using a 5% nominal level. When $\rho = 1$ we can assess each test's empirical size. We examine power against stationary alternatives $\rho < 1$, and we consider $\rho \in [0.75, 1]$ using a grid of 15 equally spaced values for ρ when constructing each power curve.

It might seem odd that we compare our proposed two-sided tests with existing one-sided tests. We select our two-sided test as we believe it is important to be agnostic about the alternative. We use existing one-sided tests as these are the common implementation. If we replace the latter tests by two-sided versions this substantially decreases their power, so this is a fair comparison regarding power.

The legends in the figures that follow use the following abbreviations; MMA (proposed bootstrap model average ADF test with Hansen (2007) weight selection); JMA (proposed test with Hansen & Racine (2012) weight selection); N-P (Ng & Perron (2001) with MIC model selection using Perron & Qu (2007) detrending based on asymptotic critical values); P-P (Phillips & Perron (1988) based on asymptotic critical values); ADF (Dickey & Fuller (1979) with BIC model selection based on MacKinnon's (1996) asymptotic critical values). The N-P and ADF tests are from the R package `CADFtest` (Lupi 2009) while the P-P test is from the R package `tseries` (Trapletti & Hornik 2018). The proposed test is from the R package `hr` (Racine 2018).

Among all tests considered, the preferred test would have a power curve that would exhibit correct size (i.e., when $\rho = 1$ would have an empirical rejection frequency that is approximately 5% or otherwise exhibit the lowest size distortions), and otherwise would have uniformly higher power than its peers (i.e., when $\rho < 1$ its power curve would lie above those of its peers). Any test procedure exhibiting large upwards size distortions in standard settings cannot in good conscience be recommended to practitioners since it could lead to the rejection of the null at levels far in excess

of the nominal level of the test when the null is in fact correct. Results for each DGP and a brief discussion follow.

4.1. **AR(1) DGP.** Figure 1 presents power curves generated under the AR(1) DGP. For the proposed test we present results using both the MMA or JMA weight selection schemes outlined in Section 3 based upon the bootstrap procedure outlined in Section 3. For the remaining tests we use asymptotic critical values as outlined above.

Figure 1 reveals that this is a textbook case where all tests are approximately correctly sized, though it is evident that the proposed approach dominates in terms of power. As expected, for a given value of ρ (< 1), power increases as the sample size increases for all tests. For the proposed test, Hansen's (2007) MMA weight selection scheme delivers a test with more power than that based on Hansen & Racine's (2012) JMA approach for this DGP. Based on this set of power curves, for this DGP the proposed approach with MMA weight selection dominates its peers.

Next, we address the question of the fact that there are two comparisons being made, one between model selection and model averaging, and one between asymptotic and bootstrap inference. To assess these interplay of these issues we report both the asymptotic and bootstrapped versions of the N-P approach in Figure 2. We report this simply because the reader might reasonably wonder whether the power gains associated with the proposed procedure arise from the use of the bootstrap procedure rather than from the use of model averaging.

By way of illustration, compare the performance of the asymptotic and bootstrapped MIC model selected Ng & Perron (2001) test (N-P and N-P*, respectively) with the bootstrapped model averaged ADF test (MMA or JMA) in Figure 2. A comparison of the asymptotic and bootstrapped versions of Ng & Perron's (2001) MIC model selected test reveals no consequential power gains associated with bootstrapping the process using the identical bootstrap procedure that underpins the proposed approach (we present results for $T = 50$ and $T = 100$ only as results do not differ qualitatively for $T = 200$ and $T = 400$). We can therefore safely conclude that any power gains associated with the proposed approach arise from the use of model averaging versus model selection and cannot be attributed to the use of a bootstrap procedure.

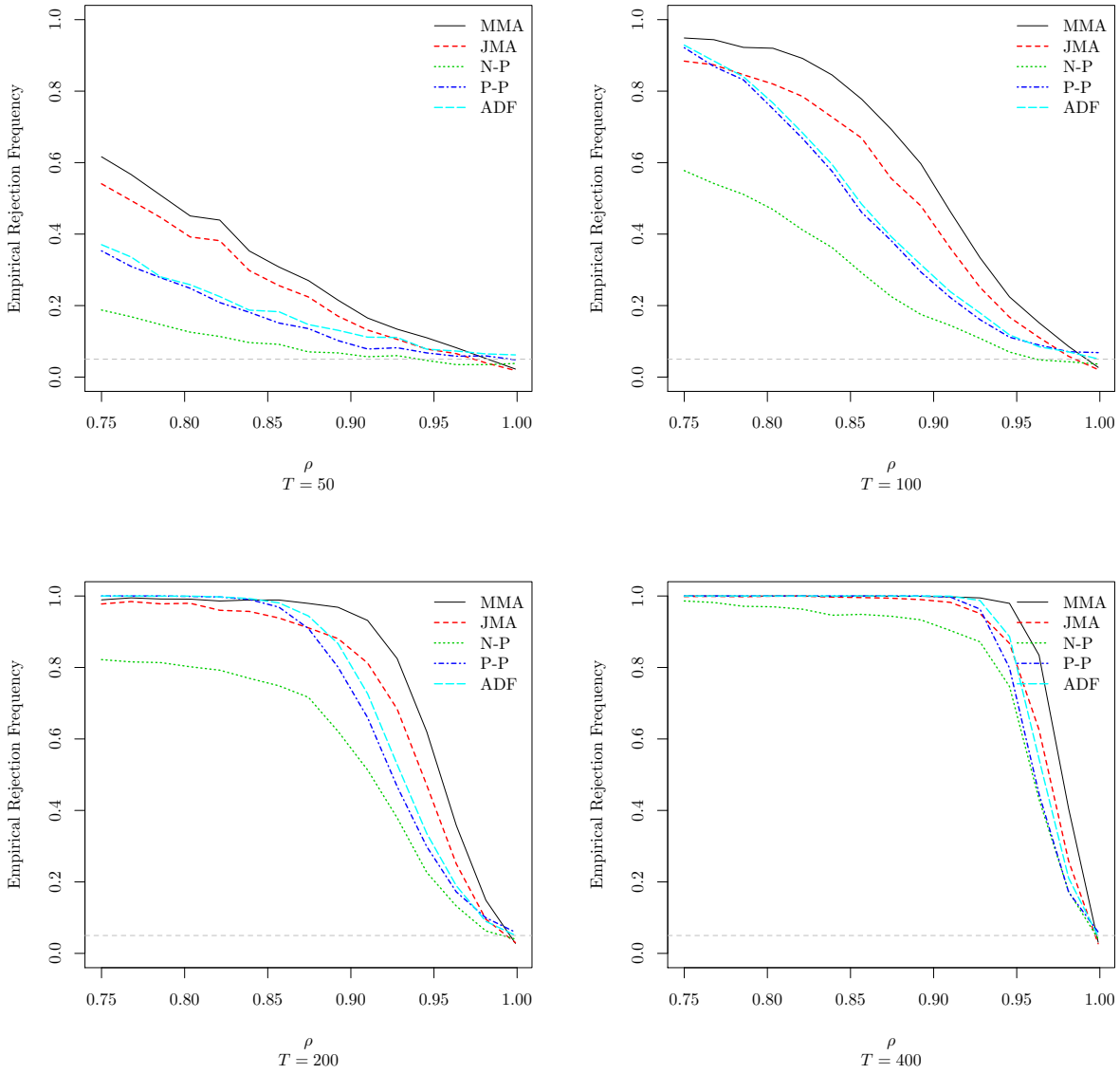

FIGURE 1. Power curves for the AR(1) DGP $y_t = \rho y_{t-1} + \varepsilon_t$ for unit root tests using a $\alpha = 0.05$ level of significance. When $\rho = 1$ a unit root is present (empirical size is the height of the power curve at $\rho = 1$, i.e., at the right of each figure).

4.2. **ARMA(1,1) DGP.** Figure 3 presents power curves generated under the ARMA(1,1) DGP containing a coefficient of -0.8 on the lagged MA component that is known to confound existing tests.

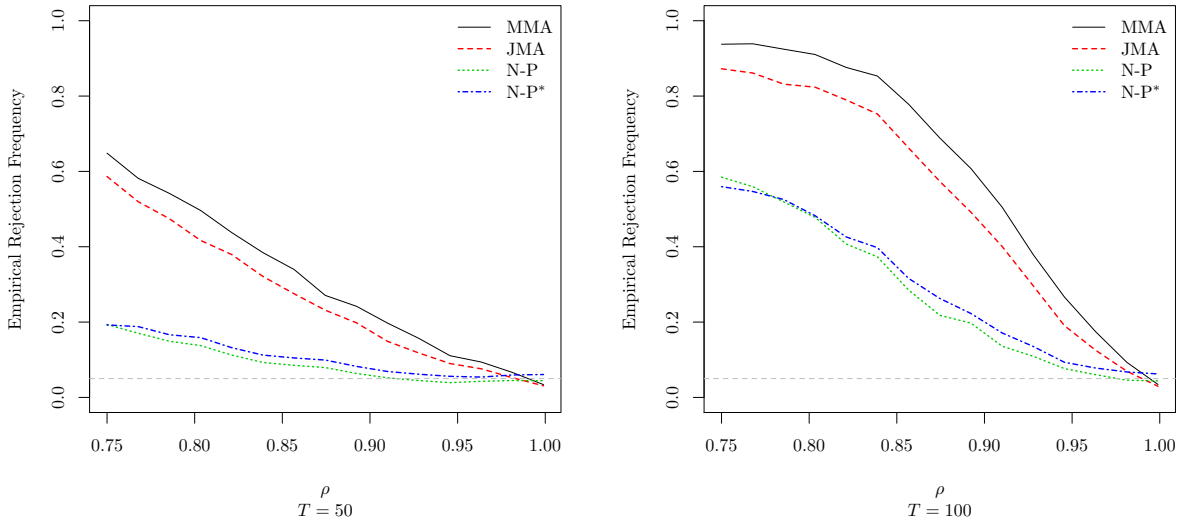

FIGURE 2. Power curves for the AR(1) DGP $y_t = \rho y_{t-1} + \varepsilon_t$ for unit root tests using a $\alpha = 0.05$ level of significance. When $\rho = 1$ a unit root is present (empirical size is the height of the power curve at $\rho = 1$, i.e., at the right of each figure). The power curve N-P is based on asymptotic critical values, while N-P* is based on bootstrapped critical values.

It is evident from Figure 3 that the Phillips & Perron (1988) and Dickey & Fuller (1979) approaches completely fail for this DGP as was pointed out by Phillips & Perron (1988) and Palm et al. (2008). In particular, for this DGP the Phillips & Perron (1988) test has empirical size equal to 1.00 for any T when $\alpha = 0.05$ thereby rejecting 100% of the time when in fact the null is true, while the Dickey & Fuller (1979) approach displays similarly crippling upward size distortions that very slowly approach nominal size as T increases but otherwise remain unacceptably high.

In light of the extreme size distortions that surface for this (and similar) DGPs when using the Phillips & Perron (1988) and Dickey & Fuller (1979) tests, it is difficult to recommend either test to practitioners, therefore the choice of tests therefore comes down to either Ng & Perron's (2001) procedure or the proposed bootstrap model averaging approach.

For this DGP, the proposed approach is slightly over-sized for $T \leq 100$ (for the MMA weighting scheme the empirical rejection frequencies are approximately 8% for $T = 50$ and 6% for $T = 100$, respectively; for the JMA weighting scheme the empirical rejection frequencies are approximately 7% for $T = 50$ and 5% for $T = 100$, respectively) while Ng & Perron's (2001) approach exhibits

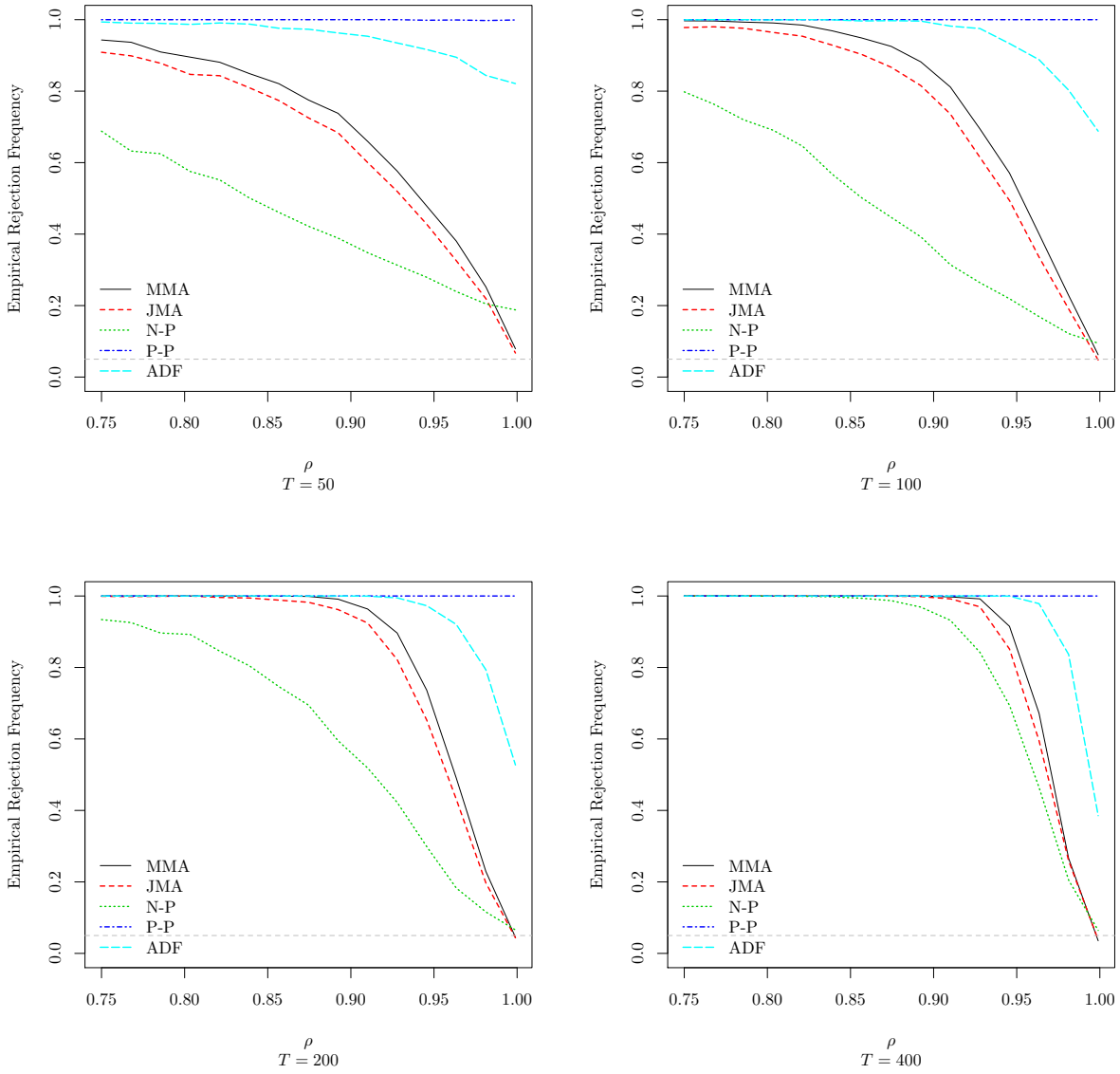

FIGURE 3. Power curves for the ARMA(1,1) DGP $y_t = \rho y_{t-1} + \varepsilon_t - 0.8\varepsilon_{t-1}$ for unit root tests using a $\alpha = 0.05$ level of significance. When $\rho = 1$ a unit root is present (empirical size is the height of the power curve at $\rho = 1$, i.e., at the right of each figure).

substantially larger size distortions for samples of size $T \leq 100$ (approximately 18% and 10% for $T = 50$ and $T = 100$, respectively), while both tests are approximately correctly sized for $T \geq 200$. In addition, the proposed approach has substantially higher power than Ng & Perron's (2001) approach for this DGP, size distortions notwithstanding, regardless of whether the MMA or JMA

weighting schemes are adopted. For this DGP, the proposed approach is first among its peers when judged by its power curve.

4.3. **ARMA(1,2) DGP.** Figure 4 presents power curves generated under the ARMA(1,2) DGP. Given that neither the Phillips & Perron (1988) nor Dickey & Fuller (1979) tests can be recommended to practitioners given their failure for the ARMA(1,1) DGP above and for a range of similar DGPs, we drop them in the power analysis that follows.

Figure 4 reveals that, for this DGP, all three procedures are approximately correctly sized for all T while the proposed approach that employs the MMA weight scheme dominates in terms of power.

4.4. **Discussion.** On the basis of our simulation results, we feel confident recommending the proposed procedure for testing for the presence of a unit root (extensive simulations that examine the effect of the largest dimension model that is averaged over, the distribution of the weight vector and so forth are available upon request). Our approach attenuates the large size distortions that can arise when using the classical approach (Dickey & Fuller 1979) that relies on tabulated critical values, distortions that can also arise when using in Ng & Perron's (2001) approach. Furthermore, the proposed approach exhibits higher power, does not require specification of the model from which the bootstrap resamples are drawn, and uses and an automatic block length selection procedure for the dependent bootstrap method. Though the MMA procedure has higher power than the JMA procedure for the DGPs considered above, extensive simulations not reported here indicate that the JMA procedure exhibits lower size distortions, when present, than the MMA approach, particularly when a trend is included in the model. A conservative approach would therefore be to use the JMA weight selection scheme for this reason. This procedure ought to appeal to practitioners as there are no unknown parameters that must be specified by the user, an R implementation exists, and the procedure is not computationally demanding.

5. Conclusion

We propose a bootstrap model averaging procedure capable of attenuating large upward size distortions that can arise when testing for the presence of a unit root while possessing power that dominates its peers. We adopt a model-free bootstrap procedure where the null is imposed

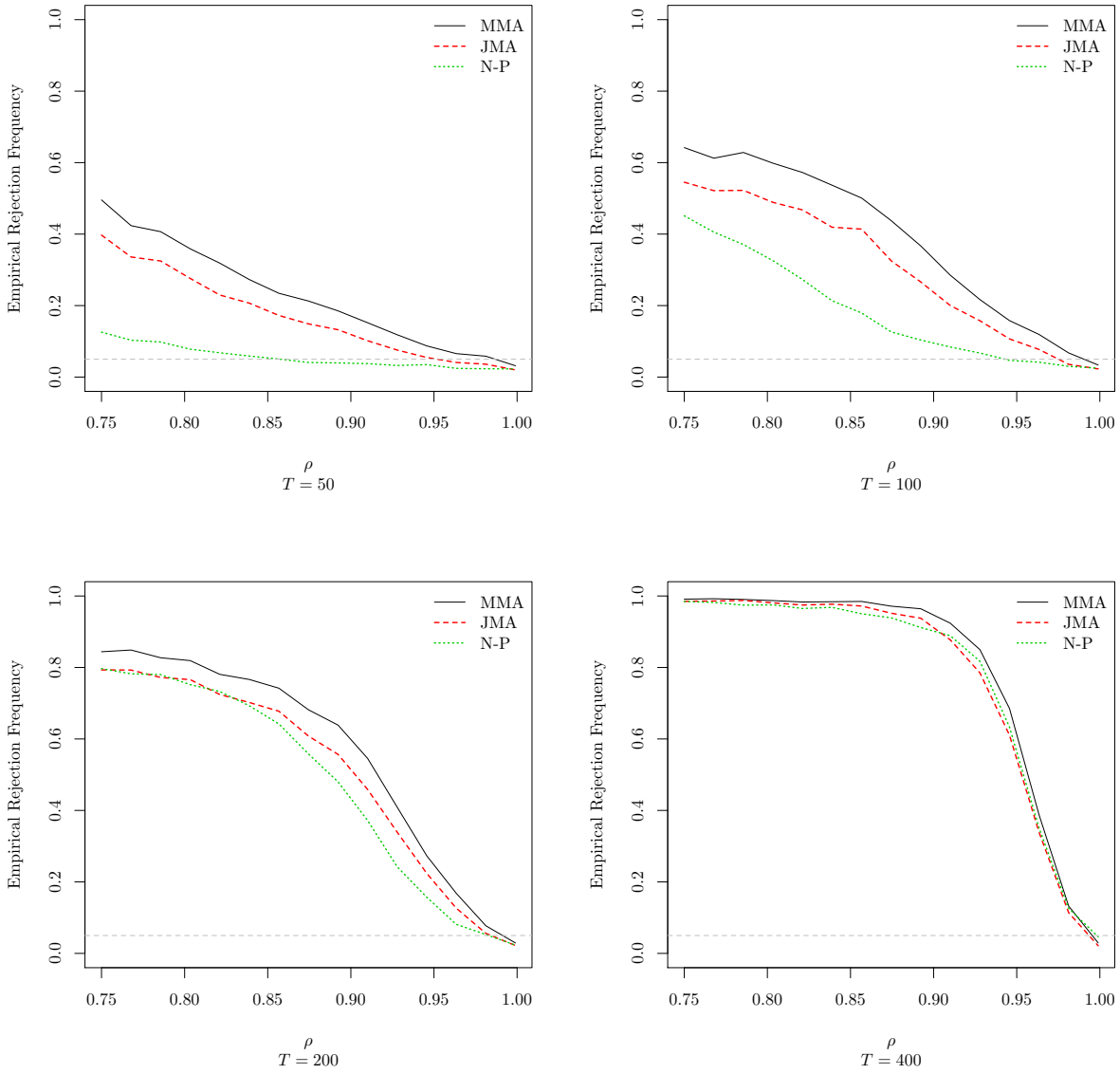

FIGURE 4. Power curves for the DGP $y_t = \rho y_{t-1} + \varepsilon_t + 0.3\varepsilon_{t-1} - 0.2\varepsilon_{t-2}$ for unit root tests using a $\alpha = 0.05$ level of significance. When $\rho = 1$ a unit root is present (empirical size is the height of the power curve at $\rho = 1$, i.e., at the right of each figure).

by simple differencing, exploit recent developments in automatic block length selection for the geometric bootstrap procedure invoked, and adopt a novel model averaging procedure to address model uncertainty. Theoretical support is provided, and a set of simulation exercises underscore its advantages relative to its peers. An R (R Core Team 2018) package exists that implements the

proposed method (Racine 2018). Since there are no nuisance parameters to be set by the user, and in light of its performance in a range of simulated scenarios, we are optimistic that the proposed approach will appeal to practitioners.

References

Choi, I. (2015), *Almost All about Unit Roots: Foundations, Developments, and Applications*, Cambridge University Press.

DeJong, D. N., Nankervis, J. C., Savin, N. & Whiteman, C. H. (1992), 'The power problems of unit root test in time series with autoregressive errors', *Journal of Econometrics* **53**(1), 323–343.

Dickey, D. A. & Fuller, W. A. (1979), 'Distribution of the estimators for autoregressive time series with a unit root', *Journal of the American Statistical Association* **74**(366), 427–431.

Hansen, B. E. (1992), 'Convergence to stochastic integrals for dependent heterogeneous processes', *Econometric Theory* **8**, 489–500.

Hansen, B. E. (1995), 'Rethinking the univariate approach to unit root testing: Using covariates to increase power', *Econometric Theory* **11**(5), 1148–1171.

Hansen, B. E. (2007), 'Least squares model averaging', *Econometrica* **75**, 1175–1189.

Hansen, B. E. (2010), 'Averaging estimators for autoregressions with a near unit root', *Journal of Econometrics* **158**(1), 142–155.

Hansen, B. E. (2014), 'Model averaging, asymptotic risk, and regressor groups', *Quantitative Economics* **5**(3), 495–530.

Hansen, B. E. & Racine, J. S. (2012), 'Jackknife model averaging', *Journal of Econometrics* **167**(1), 38–46.

Herrndorf, N. (1984), 'A functional central limit theorem for weakly dependent sequences of random variables', *The Annals of Probability* **12**, 141–153.

Lupi, C. (2009), 'Unit root cadf testing with R', *Journal of Statistical Software* **32**(2), 20.

MacKinnon, J. (1996), 'Numerical distribution functions for unit root and cointegration tests', *Journal of Applied Econometrics* **11**, 601–618.

Mallows, C. L. (1973), 'Some comments on c_p', *Technometrics* **15**, 661–675.

Ng, S. & Perron, P. (2001), 'Lag length selection and the construction of unit root tests with good size and power', *Econometrica* **69**(6), 1519–1554.

Palm, F. C., Smeekes, S. & Urbain, J.-P. (2008), 'Bootstrap unit-root tests: Comparison and extensions', *Journal of Time Series Analysis* **29**(2), 371–401.

Park, J. Y. (2003), 'Bootstrap unit root tests', *Econometrica* **71**(6), 1845–1895.

Patton, A., Politis, D. N. & White, H. (2009), 'CORRECTION TO "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White', *Econometric Reviews* **28**(4), 372–375.

Perron, P. & Qu, Z. (2007), 'A simple modification to improve the finite sample properties of ng and perron's unit root tests', *Economics Letters* **94**(1), 12–19.

Phillips, P. C. B. & Perron, P. (1988), 'Testing for a unit root in time series regression', *Biometrika* **75**(2), 335–346.

Politis, D. N. & Romano, J. P. (1994), 'Limit theorems for weakly dependent Hilbert space valued random variables with applications to the stationary bootstrap', *Statistica Sinica* **4**, 461–476.

Politis, D. N. & White, H. (2004), 'Automatic block-length selection for the dependent bootstrap', *Econometric Reviews* **23**, 53–70.

Psaradakis, Z. (2001), 'Bootstrap tests for an autoregressive unit root in the presence of weakly dependent errors', *Journal of Time Series Analysis* **22**(5), 577–594.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Racine, J. S. (2018), *hr: Bootstrap Model Average Unit Root Test*. R package version 1.0-2.
URL: *https://github.com/JeffreyRacine/R-Package-hr*

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**, 461–464.

Schwert, G. W. (1989), 'Tests for unit roots: A monte carlo investigation', *Journal of Business & Economic Statistics* **7**(2), 147–159.

Swensen, A. R. (2003), 'Bootstrapping unit root tests for integrated processes', *Journal of Time Series Analysis* **24**(1), 99–126.

Trapletti, A. & Hornik, K. (2018), *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-43.
URL: *https://CRAN.R-project.org/package=tseries*