GRÉDI

Groupe de Recherche en Économie et Développement International

Cahier de recherche / Working Paper
08-17

Modified Fast Double Sieve Bootstraps for ADF Tests

Patrick Richard

UNIVERSITÉ DE SHERBROOKE

Modified Fast Double Sieve Bootstraps for ADF Tests

Patrick Richard
GREDI and Départment de sciences économiques
Université de Sherbrooke
Sherbrooke, Québec, Canada
J1K 2R1

Abstract

This paper studies the finite sample performance of the sieve bootstrap augmented Dickey-Fuller (ADF) unit root test. It is well known that this test's accuracy in terms of rejection probability under the null depends greatly on the underlying DGP. Through extensive simulations, we find that it also depends on the numbers of lags employed in the bootstrap DGP and in the bootstrap ADF regression. Based on this finding and using some well established theoretical results, we propose a simple modification that significantly improves the test's accuracy. We also introduce different versions of the fast double bootstrap, each modified according to the same theoretical basis. According to our simulations, these new testing procedures have lower error in rejection probability under the null while retaining good power.

This research was supported by a grant from the Fonds Québécois de Recherche sur la Société et la Culture. An earlier version of this paper was presented at the SCSE 2008 Conference at Montebello and CEA annual meeting at UBC, Vancouver. I am grateful to Russell Davidson for helpful comments.

October 2008

1 Introduction

It is a well established fact that Augmented Dickey-Fuller (ADF) tests may severely over-reject the unit root hypothesis when the first difference of the process under scrutiny is a linear model which cannot be written as a finite order AR(p) model, see, among many others, Schwert (1989). The simple bootstrap, which often helps to increase finite sample accuracy of tests statistics, is not appropriate in such cases because it is not generally possible to reduce the problem to i.i.d. resampling. Thus, more complicated and specialised bootstrap methods must be used.

One such method, called the sieve bootstrap, was introduced by Bühlmann (1997) and was applied to unit root testing by Psaradakis (2001), Park (2002), Chang and Park (2003), Palm *et al.* (2007) and Richard (2007). Typically, an autoregression of order p is used to approximate the true model and bootstrap samples are drawn by i.i.d. resampling from its residuals. As shown by Chang and Park (2003) letting p increase at a proper rate of the sample size yields asymptotically valid ADF tests. The choice of an AR(p) approximation is motivated only by ease of estimation considerations. Indeed, Richard (2007) shows that sieve bootstrap ADF tests based on MA and ARMA approximations are asymptotically valid and may sometimes perform better than AR sieve bootstrap tests in finite samples. We will not consider these MA and ARMA sieve bootstraps in the present paper. Thus, the term sieve bootstrap will henceforth be used to designate the AR sieve bootstrap alone.

The theory of the bootstrap for unit root processes is not as thoroughly developed as it is for stationary ones. In particular, in spite of the fact that the ADF test statistics may be shown to be an asymptotic pivot under fairly weak conditions, the existence of bootstrap asymptotic refinements in the sense of Beran (1988) has so far only been proved under somewhat restrictive assumptions.[1] There is therefore no unequivocal theoretical reason to advocate the use of the sieve bootstrap ADF test over that of the usual asymptotic version. However, there is a large body of simulation evidence to the effect that sieve bootstrap ADF tests may be more accurate under the null than asymptotic ones, see for example, Chang and Park (2003), Palm *et al.* (2007) and Richard (2007). According to these papers, the extent to which the sieve bootstrap improves finite sample inference accuracy over asymptotic theory is function of the sample size and of the underlying data

[1] Park (2003) has shown that such refinements do exist for I(1) processes but his proofs require rather strict assumptions such as a first difference process that is AR(p) with p finite and known.

generating process (DGP).

In this paper, we find that the accuracy of the sieve bootstrap also depends greatly on the manner in which the sieve bootstrap model and ADF regressions are set up. In particular, we argue that the difference between the correlation structure of the residuals of the original data ADF regression and that of the sieve bootstrap ADF regression is an important factor. Based on this, we propose a simple modification to the usual sieve bootstrap ADF testing procedure which reduces the error in rejection probability (ERP). In order to obtain a greater accuracy gain under the null, we propose four modified versions of the fast double bootstrap (FDB) introduced by Davidson and MacKinnon (2007). The finite sample properties of these methods are explored through a Monte Carlo study.

The paper is organised as follows: section 2 briefly introduces the ERP problem of the ADF unit root test. Section 3 presents the usual sieve bootstrap test and proposes two slightly modified versions that have, according to simulation evidence, better finite sample properties under the null. In section 4, the fast double bootstrap is introduced along with some modifications that greatly reduce its ERP. The finite sample power of those different versions of the sieve bootstrap ADF test is investigated in section 5. Section 6 concludes.

2 ADF unit root test

We consider an I(1) time series y_t which is generated by a DGP belonging to the model:

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

$$u_t = \pi(L)\varepsilon_t \tag{2}$$

where L is the lag operator and $\pi(L)$ is a lag polynomial such that u_t is an invertible, possibly infinite order MA process. Under proper assumptions on ε_t and $\pi(L)$ such as those made by Chang and Park (2003), this model is general enough to nest all ARIMA$(p, 1, q)$ processes with stationary and invertible first difference. Since we do not derive any theoretical results, we do not state these assumptions here, see Chang and Park (2003) for details. The widely used ADF unit root test is based on testing the hypothesis that $\gamma = 0$ against the alternative $\gamma < 0$ in the regression:

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=1}^{k} \gamma_i \Delta y_{t-i} + e_t \tag{3}$$

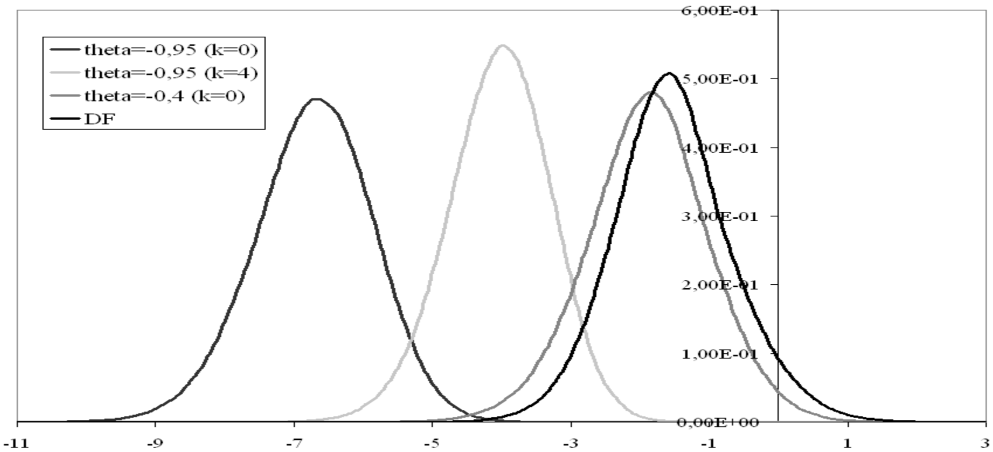
where k has to be chosen by the investigator and a constant and a deterministic trend may be added. In all that follows, we will refer to this equation as the ADF regression. The ADF unit root test may be conducted by computing the usual t-statistic for $\gamma = 0$, which we will refer to as τ.

Econometric theory provides results about the asymptotic distribution of τ when k tends to infinity at a proper rate of the sample size. Most notably, Phillips (1987) and Chang and Park (2002) show that the asymptotic distribution of τ is the now well known DF distribution (which we will denote by F_{DF}). In practice however, one has a finite number of observations and must choose one value of k to carry out the test. Techniques for the choice of k in finite samples are discussed in several papers see, among others, Ng and Perron (2001).

Galbraith and Zinde-Walsh (1999) derive the asymptotic distribution of τ when k is fixed, even as n tends to infinity, and u_t is an invertible MA(q) process. It turns out that this distribution is a function of F_{DF}, but it is shifted and somewhat deformed by factors that depend on k and the sign of the root of the MA(q) process that is closest to unity.

In particular, when u_t is a MA(1) process with a sole parameter θ, Galbraith and Zinde-Walsh (1999) show that the asymptotic distribution of the ADF test, $F_\infty(\tau)$, shifts to the left when θ is negative and that the extent of this shift increases as $\theta \rightarrow -1$ and is inversely related to k. Intuitively, this shift is due to the near cancellation of the unit root with the root of the MA polynomial. As θ approaches -1, this near cancellation becomes more complete, so that y_t increasingly resembles a stationary process. Increasing k allows one to capture more of the infinite autoregressive structure of u_t and thus remove the near-cancellation problem. An illustration of this point is found in figure 1, which shows the position of $F(\tau)$, the true finite sample distribution of τ under the null, with respect to F_{DF}. This figure was generated using 500 000 Monte Carlo samples of an I(1) process such as the one described in equations (1) and (2) where u_t was a simple MA(1) process with a single parameter θ and $k = 0$ and $k = 4$.

Figure 1. ADF test distribution, $n = 100$.



When u_t is a MA(1) process with a positive parameter, no near cancellation of the roots occur. Galbraith and Zinde-Walsh (1999) show that $F_\infty(\tau)$ is then not much different from F_{DF} even for small values of k. Thus, their results provide a theoretical explanation to the often observed fact that the ADF test based on F_{DF} over-rejects when a large negative MA(1) component is present but not when the MA(1) parameter is positive.

3 Sieve bootstrap unit root tests

The sieve bootstrap consists of approximating (2) with a finite order AR(p) model

$$u_t = \alpha_1 u_{t-1} + \alpha_2 u_{t-2} + ... + \alpha_p u_{t-p} + \varepsilon_{p,t} \tag{4}$$

where $\varepsilon_{p,t}$ is uncorrelated with u_{t-i}, $i = 1, 2, ..., p$. The sieve bootstrap DGP is then

$$y_t^\star = y_{t-1}^\star + u_t^\star$$

where

$$u_t^\star = \sum_{i=1}^{p} \hat{\alpha}_i u_{t-i}^\star + \varepsilon_t^\star. \tag{5}$$

where $\varepsilon_t^\star$ is drawn from the empirical density function (EDF) of the residuals of the AR(p) model estimated on Δy_t (thus imposing the null hypothesis) and $\hat{\alpha}_i$ is a consistent estimator of α_i.[2] The sieve bootstrap test statistics ($\tau_j^\star$) are obtained by testing the hypothesis that $\gamma = 0$ against the alternative $\gamma < 0$ in the regression:

$$\Delta y_t^\star = \gamma y_{t-1}^\star + \sum_{i=1}^{k'} \gamma_i \Delta y_{t-i}^\star + v_t. \tag{6}$$

Let $\hat{F}_B^\star(\tau)$ denote the EDF of the $\tau_j^\star$, $j = 1, 2, ..., B$ and $F^\star(\tau)$ be its limit as $B \to \infty$, where B is the number of bootstrap samples generated. Under the assumption that p and k' go to infinity at a proper rate of n, Chang and Park (2003) show that this sieve bootstrap ADF test is asymptotically valid, that is, that $F^\star(\tau) \overset{a}{=} F_{DF} + o(1)$.

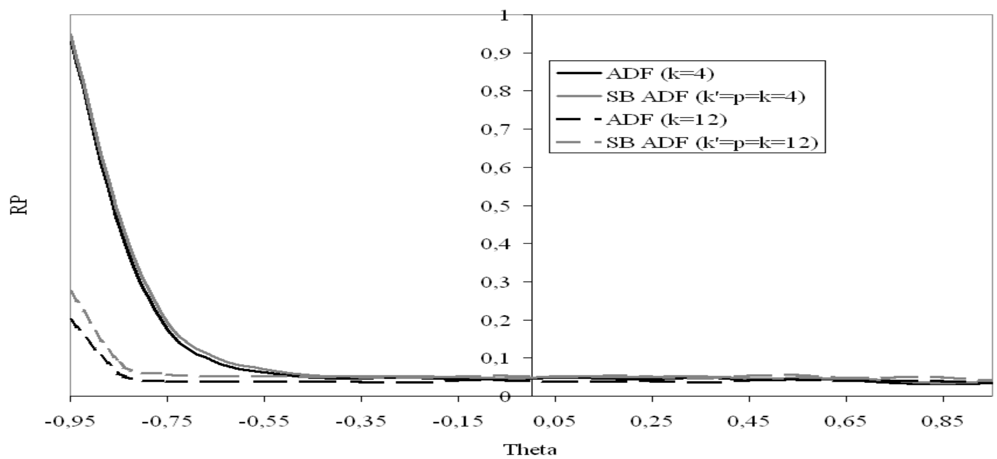
If u_t is a MA(1) process with a negative parameter (or any other MA(q) DGP that results in over-rejection), then, according to the results of Galbraith and Zinde-Walsh (1999), in order to reduce the ERP for some fixed k, a sieve bootstrap test procedure must provide a null distribution $\hat{F}_B^\star(\tau)$ that is shifted to the left with respect to F_{DF}. As we will now see, whether of not this goal is achieved greatly depends on how the sieve bootstrap DGP is built and on the specification of the sieve bootstrap ADF regression (equation 6, hereafter, SB ADF regression).

Under the null hypothesis that $\gamma = 0$, equation (3) becomes an AR(k) model. Assuming invertibility, Δy_t has an AR(∞) form, so that this AR(k) model is under-specified. Thus, for any finite k, the errors of (3) are not i.i.d. On the other hand, $\Delta y_t^\star$ is by construction an AR(p) process and $y_t^\star$ is I(1). Thus, the errors of (6) are i.i.d. whenever $k' \geq p$. In such cases, standard results imply that $\hat{F}_B^\star(\tau)$ is close to F_{DF}. Hence, if (3) is largely under-specified and if $k' \geq p$, then the sieve bootstrap should not be expected to work any better than the asymptotic test. Specifically, for relatively low values of p and $k' \geq p$, the SB ADF test should be expected to over-reject as much as the ADF test when θ is close to -1 because the bootstrap does not properly replicate the near-cancellation of the unit root and the errors of (6) are not correlated in a manner similar to those of (3). This point is illustrated by the simulations reported in figure 2, which are based on 10 000 Monte Carlo repetitions and $B = 499$.[3]

[2] Paparoditis and Politis (2005) suggest resampling from $\tilde{\varepsilon}_t = y_t - \tilde{\rho} y_{t-1} - \sum_{i=1}^{p} \tilde{\alpha}_i \Delta y_{t-i}$. They show that this may result in increased power. According to some simulations based on this alternative procedure, it results in tests with, on average, slightly higher ERP than tests based on resampling the residuals of (4). Further, the modifications proposed below do not work as well for this alternative resampling scheme.

[3] Unless otherwise specified, all ADF tests are performed with a constant and no deterministic trend.

Figure 2. Rejection probability (RP) at nominal level 5%, $n = 100$.



Nevertheless, according to the simulations of Chang and Park (2003), the sieve bootstrap is able to provide more precise inference than the test based on asymptotic theory. This results from the fact that they do not impose that $k' = p$ but rather select the two independently using the AIC. Figure 3 shows the results of a set of simulations using samples of 100 observations where p, k and k' are chosen by the AIC with a maximum of 12 lags. It is obvious that relaxing the constraint $k' = p$ results in the sieve bootstrap providing more accurate ADF tests under the null.

Figure 4 shows the mean values of p, k and k' chosen by the AIC. The three of them are approximately the same for all $\theta \geq -0.35$, which correspond to the area in figure 3 where the two tests perform similarly. On the other hand, there are important differences between them for $\theta < -0.35$. Precisely, the AIC selects lower average orders for the ADF regressions estimated on the original and bootstrap data (k and k') than for the sieve bootstrap DGP (p). Since this difference increases as $\theta \rightarrow -1$, it is not unreasonable to postulate that this results from the root near cancellation known to occur between the MA(1) part and the unit root. Both figures are based on 10 000 Monte Carlo samples with $B = 499$.

It is easy to see that the difference between k' and p is responsible for the accuracy gain of the SB ADF test. Indeed, under the null, whenever $k' < p$ equation (6) is an AR(k') process that underspecifies the AR(p) sieve bootstrap DGP. Furthermore, as noted before, the ADF regression (3) is an AR(k) that also underspecifies the true AR(∞) DGP.

Figure 3. RP at nominal level 5%, $n = 100$.

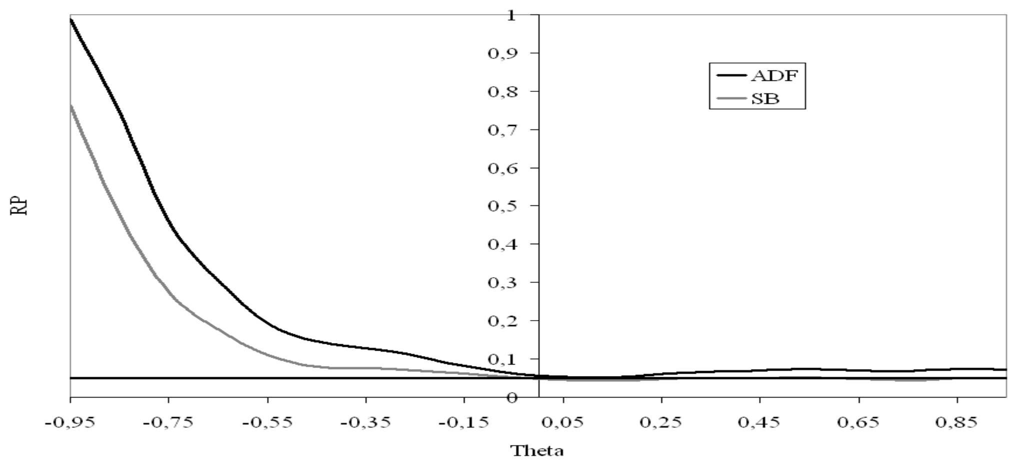
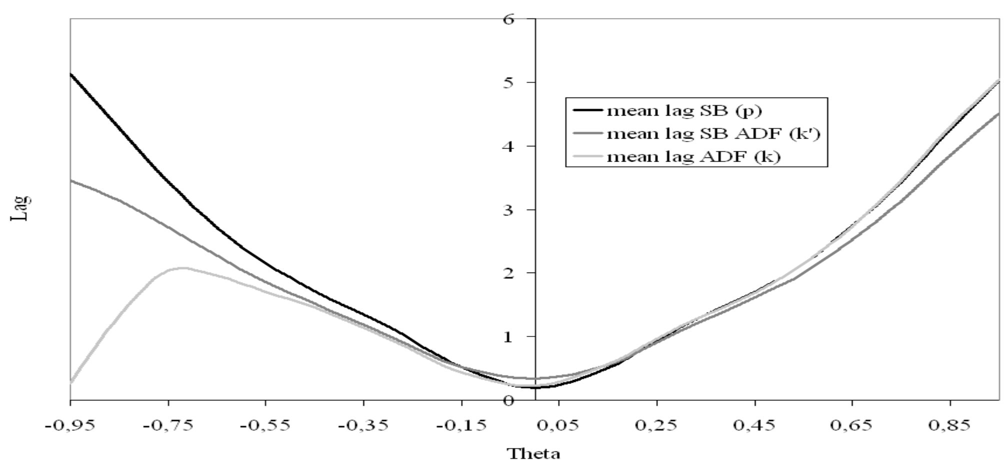


Figure 4. Mean values of p, k and k', $n = 100$.



Since the sieve bootstrap DGP (5) consistently estimates the true AR(∞) DGP as $n \to \infty$ (Park, 2002), it follows that the errors of (6) have a correlation structure similar to that of the errors of (3). Thus, $F^{\star}(\tau)$, and consequently $\hat{F}^{\star}_{B}(\tau)$, also shift to the left. Hence, the critical value of $\hat{F}^{\star}_{B}(\tau)$ is larger in absolute value than that of F_{DF}, which reduces the rejection frequency. In fact, if $k' = k$ is fixed and finite as $n \to \infty$ while $p \to \infty$, the results of Galbraith and Zinde-Walsh (1999) applied to the bootstrap DGP imply that, asymptotically, $F^{\star}(\tau)$ shifts to the left in exactly the same fashion as $F(\tau)$, so that the sieve bootstrap test is asymptotically valid.[4] This is illustrated in table 1, which reports the rejection probability of the sieve bootstrap test obtained from a set of Monte Carlo simulations in which $k' = k = 4$ for different values of n and with p increasing as a function of n. We used 4 000 Monte Carlo samples with $\theta = -0.85$ and 499 bootstrap samples each. As can be seen, the sieve bootstrap test's RP converges to 0.05 as $n \to \infty$ even though k and k' remain fixed.

Table 1. Rejection probability, $\theta = -0.85$.

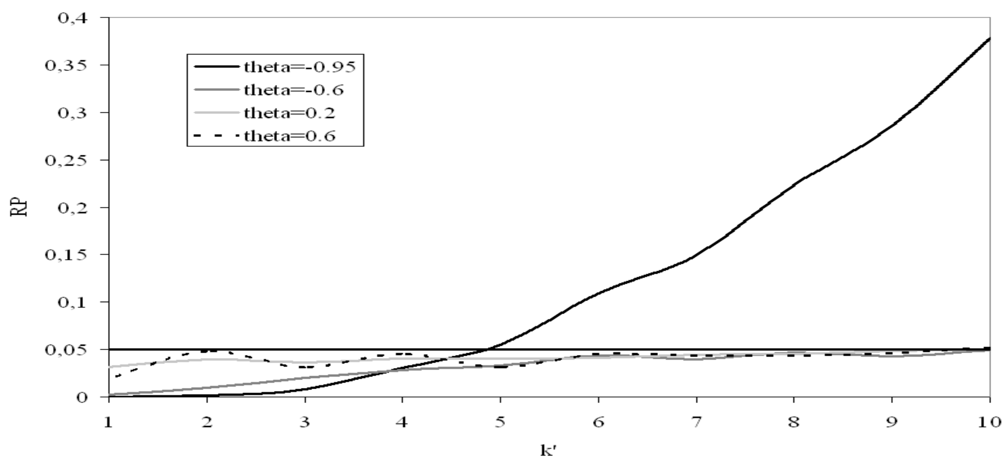
n	$p \approx n^{2/5}$	RP
100	6	0.2345
500	12	0.1195
1000	15	0.09975
5000	30	0.04425
10000	39	0.04875

To investigate this point further, we ran simulations for four MA(1) DGPs with $\theta = -0.95$, -0.6, 0.2 and 0.6 with p and k fixed at 10 and $k' = p - \ell$, with ℓ going from 0 to 9. The sample size was again 100, $B = 499$ and 10 000 Monte Carlo repetitions were used. The results are presented in figure 5. The sensitivity of the RP to the difference between p and k' is evidently greatly dependent on the underlying DGP. Of particular interest is the fact that, for all values of θ, there is at least one value of ℓ such that the SB ADF test has no or very little ERP.

The curve for $\theta = -0.6$ provides some interesting insight on the effect of ℓ. Since $k = 10$, the errors of the ADF regression (3) are almost uncorrelated. However, as ℓ increases, the correlation captured by the $p - \ell$ last lags of the sieve model moves into the errors of the SB ADF regression. For low ℓ, these last $p - \ell$ lags do not represent much dependence, so that the SB ADF regression's errors are only weakly correlated and the ERP remains low. As $p - \ell$ increases however, more correlation gets transferred to the errors of the SB ADF regression. As this happens, $\hat{F}_B^\star(\tau)$ shifts to the left and its critical value becomes larger in absolute value, which eventually causes under-rejection. As the difference between p and k' increases, the severity of the under-rejection becomes greater. Eventually, a point is reached where the bootstrap distribution is so far to the left that rejection does not occur anymore.

[4] This asymptotic validity is actually shown by Park (2002) for the simple DF test ($k' = k = 0$).

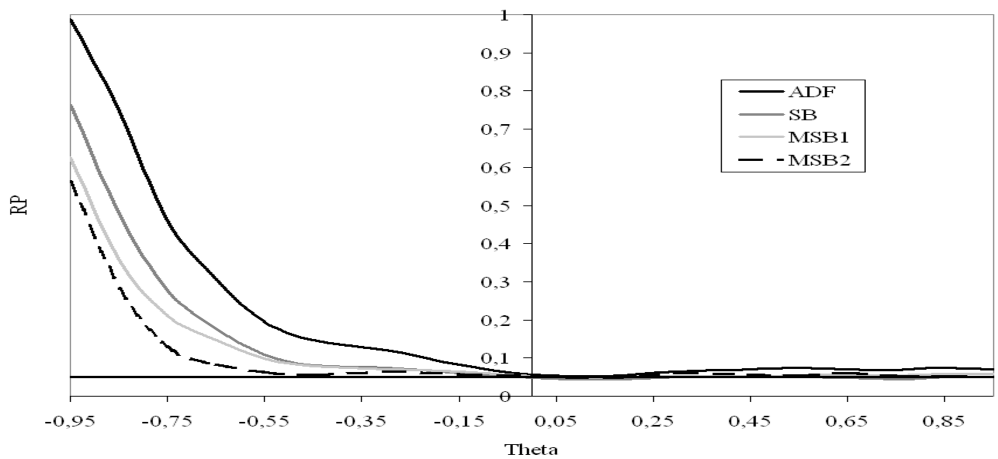
Figure 5. RP at nominal level 5% as a function of k', $n = 100$.



Of course, ℓ also decreases the RP of the SB ADF test under the alternative. Thus, in order to use these findings to reduce the SB ADF test's ERP with a minimum chance of under-rejection and power loss, it is necessary to use a rather large ℓ when θ is close to -1 and a small one otherwise. Figure 4 suggests that a very simple way to accomplish this is to have k and p selected by AIC and to set $k' = k$. This idea was put to the test in yet another set of simulations, the results of which are presented in figure 6. The setting was exactly the same as in figures 3 and 4, except that the restriction $k' = k$ was imposed. We will call this version of the sieve bootstrap the modified sieve bootstrap of the first type (MSB1). Obviously, imposing this restriction reduces the SB ADF test's ERP.

Nonlinear restrictions can also be imposed on k'. A simple one is $k' = \max\{k'_0 - (k'_0 - k)^i, 0\}$, where k'_0 is the lag order selected by AIC for the SB ADF regression. For $i = 1$, this rule simplifies to $k' = k$. Simulations were run with $i = 2$ and are reported in figure 6 under the label MSB2. As can be seen this MSB2 ADF test has even lower ERP than the MSB1 ADF test. Additional simulations have shown that these restrictions have similar effects for samples of 50 and 200 observations, although the MSB2 ADF test under-rejects a little for θ around -0.7 with 200 observations. The results are also robust to the inclusion of a deterministic time trend in the ADF regressions and hold for tests at nominal level 1% and 10%. The restrictions' effect on power is investigated in section 5.

Figure 6. RP at nominal level 5%, $n = 100$.



4 The fast double bootstrap

The previous section has shown that the ERP of the ADF test can be reduced by various versions of the sieve bootstrap but that substantial ERP remains. The fast double bootstrap (FDB) introduced by Davidson and MacKinnon (2007) is a method that is often useful when normal bootstrap procedures fail to reduce ERP in a satisfactory manner. The FDB is inspired by the double bootstrap proposed by Beran (1988). Let $G(x)$ denote the CDF of the bootstrap test's P value. Then, in the ideal case where $F(\tau) = F^{\star}(\tau)$ and $B = \infty$, $G(x)$ simply is the uniform distribution on the unit interval. In reality, B is finite and there may be differences between $F(\tau)$ and $\hat{F}^{\star}_{B}(\tau)$. Thus, $G(x)$ is almost certainly not $U(0,1)$.

The double bootstrap attempts to estimate $G(x)$ by generating B' second level bootstrap samples for, and based on, each first level bootstrap sample. Thus, every first level bootstrap test statistics $\tau^{\star}_{j}$ is accompanied by B' second level bootstrap statistics, $\tau^{\star\star}_{j,i}$. This allows to compute a set of B second level bootstrap P values $\hat{p}^{\star\star}_{j}$ which are used to obtain an estimate of $G(x)$. The double bootstrap P value is then calculated as:

$$\hat{p}^{\star\star}(\hat{\tau}) = \hat{G}_{B'}(\hat{p}^{\star}(\hat{\tau})) = \frac{1}{B}\sum_{j=1}^{B} I\left(\hat{p}^{\star\star}_{j} \leq \hat{p}^{\star}(\hat{\tau})\right) \tag{7}$$

where $\hat{p}^{\star}(\hat{\tau})$ is the first level bootstrap P value and $\hat{\tau}$ denotes the value of τ

calculated on the given sample. If $G(x)$ is $U(0,1)$ and if B' and B are infinite, then $\hat{p}^{\star\star}(\hat{\tau}) = \hat{p}^{\star}(\hat{\tau})$. On the other hand, suppose that the first level bootstrap test tends to over-reject. This means that $\hat{F}^{\star}_{B}(\tau)$ generates too few extreme values of the test statistic compared to $F(\tau)$. Hence, $\hat{p}^{\star}(\hat{\tau})$ tends to be too low. If $\hat{F}^{\star\star}_{j\,B'}(\tau)$, the distribution of the second level bootstrap statistics, also generates too few extreme test statistics with respect to $\hat{F}^{\star}_{B}(\tau)$, then $\hat{p}^{\star\star}_{j}$ will tend to be too low as well compared to $\hat{p}^{\star}(\hat{\tau})$. Thus, by definition (7), the double bootstrap P value will tend to be higher than $\hat{p}^{\star}(\hat{\tau})$ and, consequently, the double bootstrap will not over-reject as much as the bootstrap. See Davidson and MacKinnon (2007) and MacKinnon (2006) for a more detailed exposition.

Although the idea of the double bootstrap is quite compelling, it has one major drawback in that, in order to achieve any acceptable level of accuracy, it requires that both B and B' be large, which often translates in intolerably long computing times. The FDB is designed to achieve the same objective as the double bootstrap at a much lower computational cost.

The FDB consists of drawing one second level bootstrap sample from each first level bootstrap sample and calculating the relevant test statistic from each of these samples. What results is a set of B first level bootstrap statistics ($\tau^{\star}_{j}$) and a set of B second level bootstrap statistics, which we call $\tau^{\star\star}_{j}$. Then, for a one-tailed test that rejects to the left, the FDB P value is calculated as follows:

$$\hat{p}^{\star\star}_{F}(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} I\left(\tau^{\star}_{j} < \hat{Q}^{\star\star}_{B}(\hat{p}^{\star}(\hat{\tau}))\right), \tag{8}$$

where $\hat{Q}^{\star\star}_{B}(\hat{p}^{\star}(\hat{\tau}))$ is the $\hat{p}^{\star}(\hat{\tau})$ quantile of the distribution of the $\tau^{\star\star}_{j}$ and is defined by the equation:

$$\frac{1}{B}\sum_{j=1}^{B} I\left(\tau^{\star\star}_{j} < \hat{Q}^{\star\star}_{B}(\hat{p}^{\star}(\hat{\tau}))\right) = \hat{p}^{\star}(\hat{\tau}). \tag{9}$$

MacKinnon (2006) shows that if the distribution of $\tau^{\star\star}_{j,i}$ does not depend on $\tau^{\star}_{j}$, then $\hat{p}^{\star\star}_{F}(\hat{\tau})$ is equivalent to $\hat{p}^{\star\star}(\hat{\tau})$ when B and B' tend to infinity. It is well known that F_{DF} does not depend on any of the parameters of the DGP. Further, the results of Chang and Park (2003) imply that the asymptotic distribution of the $\tau^{\star\star}_{j,i}$ also is F_{DF} (indeed, the bootstrap DGP (5) meets all the conditions necessary for the results of Park, 2002 and Chang and Park, 2003 to apply). We may thus expect the FDB to perform asymptotically as well as the double bootstrap. The extent to which this is true in finite samples depends on the degree of finite sample independence between $\tau^{\star\star}_{j,i}$ and $\tau^{\star}_{j}$, which itself depends on several factors such as the DGP and the sample size. Simulation results comparing the double bootstrap with the FDB are provided at the end of the next subsection.

The accuracy of the FDB therefore depends on that of the double bootstrap. In turn, this depends on how well $\hat{G}_{B'}$ estimates G. For this estimation to be precise, it must be the case that $\hat{F}^{\star\star}_{j\,B'}(\tau)$ generates too few extreme test statistics with respect to $\hat{F}^{\star}_{B}(\tau)$ in a proportion similar to that of $\hat{F}^{\star}_{B}(\tau)$ relative to $F(\tau)$. For this to be so, it is necessary that a similar relation exists between the second level bootstrap DGPs and the first level bootstrap DGP than between the first level bootstrap DGP and the original DGP. In subsection 4.2, we will introduce some modifications to the FDB that make use of this fact to obtain more precise ADF tests.

4.1 A fast double sieve bootstrap

We propose to use a fast double sieve bootstrap (FDSB) ADF test which is carried out through the following steps.

1. Estimate the ADF regression (3) with k lags and calculate the ADF statistic $\hat{\tau}$. Fit an AR(p) model to Δy_t. Call the residuals $\hat{\varepsilon}_{t,p}$. It may be wise to center the $\hat{\varepsilon}_{t,p}$ to make sure that their average is 0. Also, if OLS is used, the centered residuals should be rescaled to correct their variance. See Davidson and MacKinnon (2004, chap. 4).

2. Draw bootstrap errors from the EDF of the centered and rescaled $\hat{\varepsilon}_{t,p}$ and build the bootstrap pseudo-series $u^{\star}_t$ using the parameters estimated in step 1.

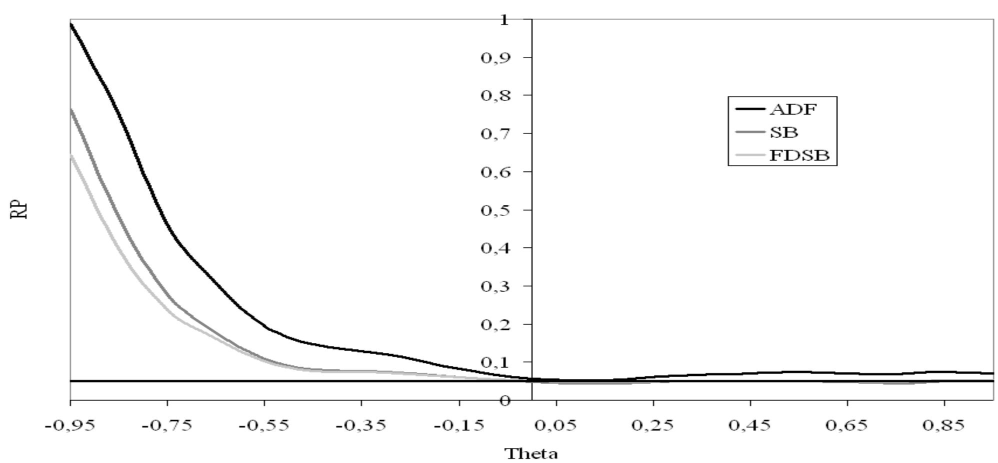
3. For each of the bootstrap samples, calculate $\hat{\tau}^{\star}_j$ using the ADF regression (6) with k' lags. Also, fit an AR(p') model to $u^{\star}_t$ and save its rescaled and centered residuals $\hat{\varepsilon}^{\star}_{t,p'}$.

4. Draw bootstrap errors from the EDF of $\hat{\varepsilon}^{\star}_{t,p'}$ and build the second level bootstrap pseudo-series $u^{\star\star}_t$ and $y^{\star\star}_t$ using the parameters estimated in step 3. For each of the bootstrap samples, calculate $\hat{\tau}^{\star\star}_j$ using the ADF regression (6) estimated with $y^{\star\star}_t$ and k'' lags.

5. Repeat steps 2 to 4 B times and calculate the FDSB ADF test P value as defined by equation (8).

In order to evaluate the accuracy of the FDSB ADF test, we have conducted an experiment based on 5 000 simulated samples of 100 realisations of the ARIMA(0,1,1) process described above. In our experiment, we have set $B = 599$ and used the AIC to select p, k, k', p' and k'' separately. Once more, a maximum of 12 lags was imposed.[5] The results are presented in figure 7. The FDSB has an ERP comparable to that of the simple sieve bootstrap for most of the parameter space considered except when θ is large and negative, where the FDSB has a smaller ERP than the usual sieve bootstrap.

Figure 7. RP at nominal level 5%, $n = 100$.



We have mentioned that all conditions are satisfied for the fast double sieve bootstrap to be asymptotically equivalent to the double bootstrap. This does not mean that they should have similar characteristics in small samples. Hence, we have compared the two methods with a small simulation experiment. Because of the very high computational time of the double bootstrap, we have restricted ourselves to 1 000 simulated samples and have set $B = 399$ and $B' = 299$. For the same reason, samples of only 50 observations were used. The results, which are reported in table 2, indicate that the FDSB ADF test has similar properties than its double bootstrap counter-part.

[5]Similar results were obtained with a maximum of 16 lags.

Table 2. RP, double bootstrap and FDSB ADF tests.

θ	-0.85	-0.55	0.55	0.85
FDSB ADF	0.420	0.144	0.046	0.039
DSB ADF	0.402	0.128	0.049	0.043

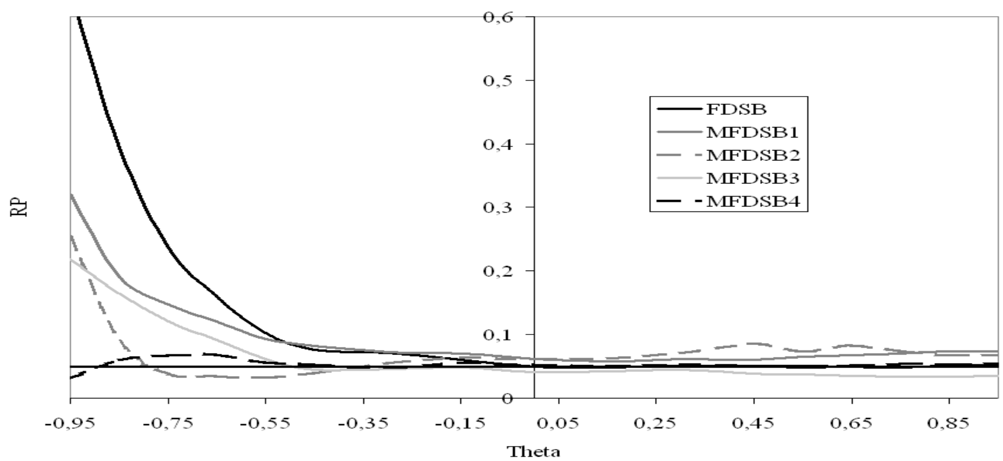
4.2 Modified fast double sieve bootstraps

We saw in section 3 that basing the choice of k' on our knowledge of the cause of the sieve bootstrap ADF test's over-rejection allowed us to obtain a more precise test. Conducting a similar exercise here seems in order. Simulations based on 5 000 samples and $B = 599$ and imposing $k' = k$ in step 3 of the FDSB algorithm are reported in figure 8. We call this testing procedure the modified fast double sieve bootstrap of the first type (MFDSB1). It can be seen that the resulting MFDSB1 ADF test has significantly lower ERP for DGPs with θ close to -1 than any procedure considered thus far. On the other hand, it slightly over-rejects for some positive values of θ. In addition, figure 8 shows that using a procedure which we label MFDSB2, by which the nonlinear restriction that leads to the MSB2 is imposed rather than $k' = k$, decreases the ERP even further for values of θ close to -1.

It is also possible to reduce ERP by imposing a restriction on the lag order of the second level bootstrap ADF regression. For DGPs where the SB ADF test over-rejects, $F(\tau)$ is shifted to the left with respect to $\hat{F}_B^\star(\tau)$. It is therefore necessary that $\hat{F}_B^\star(\tau)$ be to the left of $\hat{F}_B^{\star\star}(\tau)$, the distribution of the second level bootstrap statistics. Meanwhile, any steps we take to achieve this should not cause the FDSB ADF test to under-reject with DGPs where it already performs well. A very simple way to achieve this is to set $k'' = p'$ in step 4 of the FDSB algorithm, so that the second level bootstrap ADF regression has errors that are uncorrelated. It then follows that the second level test statistics have a distribution close to F_{DF}, which is to the right of $\hat{F}_B^\star(\tau)$ when θ is close to -1 but approximately coincides with it when θ is far from -1. We call the FDSB imposing this restriction the MFDSB3. Figure 8 shows that this provides a test with somewhat lower ERP than the MFDSB1 for values of θ close to -1. This MFDSB3 test does, however, under-reject a bit for some positive θs. Finally, we investigate the accuracy of a procedure that combines this restriction with $k' = k$. We call it MFDSB4. The simulations reported in figure 8 indicate that it has extremely low ERP. [6]

[6]The features of figure 8 are robust to the inclusion of a deterministic time trend, different

Figure 8. RP at nominal level 5%, $n = 100$.



5 Power Considerations

Before advocating the use of any of the modified sieve bootstrap procedures proposed above in practical applications, it is necessary to study their power. Let $u_t = \theta\varepsilon_{t-1} + \varepsilon_t$ and $y_t = \rho y_{t-1} + u_t$, where $\rho = 1 - c/n$, $c = 5$ or 10 and ε_t is as before. We look at two cases, one in which none of the proposed modifications improve the size properties of the ordinary sieve bootstrap test because it is already very precise ($\theta = 0.8$) and one in which they all do ($\theta = -0.8$). All these simulations are based on 5 000 samples with $B = 599$. Table 3 reports ERP-adjusted power when the only deterministic component in the ADF regression is a constant.[7]

The numbers in table 3 indicate that the power of the MSB1 and MSB2 ADF tests is similar to, though usually slightly lower than, that of the SB ADF test. Thus, their lower ERP when θ is close to -1 apparently comes at a very small power cost. It therefore seems appropriate to say that one should always impose restrictions such as $k' = \max\{k'_0 - (k'_0 - k)^i, 0\}$ when carrying out a sieve bootstrap

sample sizes and hold for all conventional nominal levels.

[7]DGPs with $\theta = -0.4$, 0 and 0.4 were also considered. The results were similar to those for $\theta = 0.8$. Simulations with $n = 50$ were also carried out and yielded results similar to those reported in table 3. Finally, the inclusion of a time trend does not significantly alter the results.

ADF test.

Also, according to the tables, the ERP-adjusted power of versions 1 and 3 of the MFDSB ADF tests are very similar. They both have power comparable to, yet smaller than, that of the FDSB when $\theta = 0.8$ or -0.8. Thus, the cost of their important ERP improvement, though more substantial than in the MSB case, still is acceptable. The MFDSB2 ADF test usualy has lower power than the MFDSB1 and 3. Finally, the MFDSB4 ADF test does not have good power at all.

Table 3. ERP-adjusted power, no time trend.

		SB	MSB1	MSB2	FDSB	MFDSB1	MFDSB2	MFDSB3	MFDSB4
				$n = 100$					
$\theta = 0.8$	$\rho = 1$	-0.0062	0.0046	0.0042	0.0008	0.0178	0.0243	-0.0174	0.0014
	$\rho = 0.95$	0.1152	0.1020	0.0929	0.1051	0.1056	0.0778	0.1072	0.0986
	$\rho = 0.90$	0.2076	0.1867	0.1432	0.2032	0.1733	0.1191	0.1947	0.1691
				$n = 200$					
	$\rho = 1$	-0.0036	0.0012	0.0004	0.0000	0.0072	0.0066	-0.0056	-0.0004
	$\rho = 0.975$	0.1383	0.1343	0.1135	0.1348	0.1218	0.1098	0.1313	0.1193
	$\rho = 0.95$	0.2531	0.2278	0.2229	0.2372	0.2168	0.1929	0.2254	0.2110
				$n = 100$					
$\theta = -0.8$	$\rho = 1$	0.2882	0.2092	0.1306	0.2388	0.1026	0.0005	0.0828	0.0126
	$\rho = 0.95$	0.1204	0.1265	0.1212	0.1103	0.0998	0.1185	0.0981	0.0777
	$\rho = 0.90$	0.1906	0.1968	0.2186	0.1743	0.1287	0.2196	0.1167	0.0589
				$n = 200$					
	$\rho = 1$	0.2176	0.1808	0.0044	0.1864	0.1100	-0.0402	0.1116	0.0556
	$\rho = 0.975$	0.1303	0.1369	0.1160	0.1463	0.1166	0.09776	0.1174	0.0984
	$\rho = 0.95$	0.2743	0.2651	0.2812	0.2349	0.1849	0.1355	0.1940	0.1220

6 Conclusion

After studying the impact of lag selection on the finite sample performances of the sieve bootstrap ADF test, we introduce some modifications to the sieve bootstrap and fast double sieve bootstrap specifically designed to reduce the test's ERP under the null. These modifications are based on the theory of Galbraith and Zinde-Walsh (1999). They consist of restricting the numbers of lags to be included in the sieve bootstrap DGP and in the ADF regressions in ways that allow the null distribution of the bootstrap test statistics to be closer to the true distribution.

In the presence of a MA(1) component, our simulations indicate that the resulting MSB ADF tests have lower ERP than the usual sieve bootstrap test in the root near-cancellation case and comparable power.[8] This unequivocally makes it a desirable alternative to the usual sieve bootstrap. Further, two versions of the modified fast double sieve bootstrap tests have much smaller ERPs than the fast double sieve bootstrap test while retaining accpetable, yet lower, power. Arguably, their very low ERP, even in extreme cases where the ordinary ADF tests almost always rejects the true null, more than compensates their lower power.

References

Beran, R. (1988). "Prepivoting test statistics: A bootstrap view of asymptotic refinements," *Journal of the American Statistical Association*, **83**, 687-97.

Bühlmann, P. (1997). "Sieve bootstrap for time series," *Bernoulli* **3**, 123-48.

Chang, Y. and J. Y. Park (2002). "On the asymptotics of ADF tests for unit roots," *Econometric Reviews* **21**, 431-47.

Chang, Y. and J. Y. Park (2003). "A sieve bootstrap for the test of a unit root," *Journal of Time Series Analysis* **24**, 379-400.

Davidson, R. (2008). "Size distortion of bootstrap unit root tests," Working paper, Department of Economics, McGill University.

Davidson, R. and J. G. MacKinnon (2007). "Improving the reliability of bootstrap tests with the fast double bootstrap," *Computational Statistics and Data Analysis*, **51**, 3259-81.

[8]See Davidson (2008) for a different approach to bootstrap unit root tests with MA(1) first difference processes.

Davidson, R. and J. G. MacKinnon (2004). *Econometric Theory and Methods*, Oxford, Oxford University Press.

Galbraith, J. W. and V. Zinde-Walsh (1999). "On the distributions of Augmented Dickey-Fuller statistics in processes with moving average components," *Journal of Econometrics*, **93**, 25-47.

MacKinnon, J. G. (2006). "Applications of the fast double bootstrap," Working paper 1023, Queen's University, Department of economics.

Ng. S. and P. Perron (2001). "Lag length selection and the construction of unit root tests with good size and power," *Econometrica* **69**, 1519-54.

Palm, F. C., S. Smeekes and J. P. Urbain (2006). "Bootstrap unit root tests: comparison and extensions." *Journal of Time Series Analysis*, **29**, 371-401.

Paparoditis, E. and D. N. Politis (2005). "Bootstrapping unit root tests for autoregressive time series," *Journal of the American Statistical Association*, **100**, 545-53.

Park, J. Y. (2002). "An invariance principle for sieve bootstrap in time series," *Econometric Theory* **18**, 469-90.

Park, J. Y. (2003). "Bootstrap Unit root tests," *Econometrica*, **71**, 1845-95.

Phillips, P. C. B. (1987). "Time series regression with a unit root," *Econometrica*, **55**, 277-301.

Psaradakis, Z. (2001). "Bootstrap tests for a autoregressive unit root in the presence of weakly dependent errors." *Journal of Time Series Analysis* **22**, 577-94.

Richard, P. (2007). "ARMA sieve bootstrap unit root tests," GREDI working paper no. 07-05, Université de Sherbrooke.

Schwert, G. W. (1989). "Testing for unit roots: A Monte Carlo investigation," *Journal of Business and Economic Statistics* **7**, 147-59.