Tests for an End-of-Sample Bubble in Financial Time Series*

Sam Astill[a], David I. Harvey[b], Stephen J. Leybourne[b] and A.M. Robert Taylor[a]

a. Essex Business School, University of Essex.

b. Granger Centre for Time Series Econometrics and School of Economics, University of Nottingham.

September 28, 2016

Abstract

In this paper we examine the issue of detecting explosive behaviour in economic and financial time series when an explosive episode is both ongoing at the end of the sample, and of finite length. We propose a testing strategy based on the sub-sampling method of Andrews (2003), in which a suitable test statistic is calculated on a finite number of end-of-sample observations, with a critical value obtained using sub-sample test statistics calculated on the remaining observations. This approach also has the practical advantage that, by virtue of how the critical values are obtained, it can deliver tests which are robust to, among other things, conditional heteroskedasticity and serial correlation in the driving shocks. We also explore modifications of the raw statistics to account for unconditional heteroskedasticity using studentisation and a White-type correction. We evaluate the finite sample size and power properties of our proposed procedures, and find that they offer promising levels of power, suggesting the possibility for earlier detection of end-of-sample bubble episodes compared to existing procedures.

Keywords: Rational bubble; Explosive autoregression; Right-tailed unit root testing: Sub-sampling.

JEL Classification: C22; C12; G14.

*We are grateful to the Guest Editors, Peter Phillips and Aman Ullah, and two anonymous referees for their helpful and constructive comments on earlier versions of this paper. Address correspondence to: Robert Taylor, Essex Business School, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK.

1 Introduction

The efficacy of unit root tests for detecting explosive rational asset price bubbles is well documented. In the seminal paper on the presence of explosive rational asset price bubbles in stock prices by Diba and Grossman (1988), the authors note that if the bubble component of a stock price series follows an explosive autoregressive process, the explosive behaviour caused by the bubble component is still manifest in the first difference of the series. This is due to the fact that an explosive autoregressive process cannot be differenced to stationarity. As such, if a series is found to be non-stationary in levels, but stationary in first differences, then the series is not subject to explosive behaviour. Based on this, Diba and Grossman (1988) propose testing the null hypothesis of no explosive behaviour by applying standard left-tailed regression-based unit root tests to a series in both levels and first differences. Recent research on the detection of asset price bubbles, however, has concentrated on applying right-tailed Dickey-Fuller [DF] tests to the levels of a series. The earliest contribution to this approach in the literature was made by Phillips, Wu and Yu (2011) [PWY], who propose a test procedure for detecting explosive rational bubbles in stock markets based on the supremum of a set of forward recursive right-tailed DF test statistics applied to both the price and dividend series in levels. If explosive behaviour is found in the price series, but not in the dividend series, they conclude that the stock price is subject to an explosive rational bubble. PWY apply their test procedure to the NASDAQ composite stock price and dividend index for the period from February 1973 – June 2005 and identify the emergence of the dot-com bubble in the middle of 1995.

Due to the simplicity of the PWY test procedure, and its favourable power properties, this test procedure has been utilised extensively in the both the finance and econometrics literature to detect bubbles in a number of financial series. Gilbert (2010) applies the PWY test procedure to commodities futures prices for the years 2000-2009 and finds evidence of explosive behaviour in the nickel, copper and crude oil series. Homm and Breitung (2012) apply both the PWY test and a Chow-type test to various series including stock prices, commodity prices and house prices, finding evidence of bubbles in a number of the series examined. Bettendorf and Chen (2103) apply the PWY test procedure to the sterling-US dollar exchange rate and find evidence of explosive behaviour in the exchange rate driven by explosive behaviour in the price index for traded goods. In a subsequent paper aimed at dealing with the issue that more than one asset price bubble could potentially be present within a given sample of data, Phillips, Shi and Yu (2015) [PSY] propose a test for at least one bubble based on a supremum of right-tailed DF statistics computed over all possible start and end dates (subject to a minimum sample size). If a rejection is obtained by this test, PSY propose a dating procedure to identify the timing of the bubble episodes, which uses sequential application of a sequence of backward recursive right-tailed DF statistics.

Based on the aforementioned evidence for the presence of explosive asset price bubbles, and the detrimental impact to the economy often caused by the collapse of such bubbles, it is

imperative that bubbles are detected as early as possible. Arguably, the most useful application of tests for asset price bubbles to policy makers is detecting an ongoing asset price bubble as soon as possible. Thus, our focus in this paper is on testing for an explosive asset price bubble of finite length that is ongoing at the end of the sample. Whilst most tests proposed in the literature concentrate on detecting and dating past asset price bubbles, the backward recursive approach of PSY is particularly well-suited to detect an end-of-sample bubble. A potential drawback of the PSY approach to bubble detection, however, is that for its asymptotic validity, it assumes that the length of the bubble regime is some non-vanishing fraction of the total sample size. In the context of detecting end-of-sample bubbles quickly, a more appropriate assumption might be one of a finite length end-of-sample bubble regime, because possibly only a few bubble observations might have been observed at the time when the tests are executed.

The approach we consider in this paper is based on the end-of-sample instability testing approach developed by Andrews (2003) and Andrews and Kim (2006). This involves calculating a test statistic based on a finite sized window of end-of-sample observations, and comparing this with critical values obtained by sub-sampling across the remaining earlier observations. This approach, by design, delivers tests which are robust to serial correlation and conditional heteroskedasticity in the driving shocks, without the need for any correction (parametric or non-parametric) to the test statistics. In this paper we propose such Andrews-type tests, adapted to the case of testing the null of no end-of-sample bubble against the bubble alternative. The statistics we consider for this Andrews-type approach take the form of: (i) a right-tailed DF statistic (notice that because of the robustness to serial correlation mentioned above, no lags are needed in the DF test regression), and (ii) implementations of the Andrews and Kim (2006)-type statistics that are motivated by a first order Taylor series expansion of the first differences of an explosive autoregressive process. We find that all these procedures offer decent finite sample size control, and the Andrews-Kim-type variants in particular offer promising levels of power, suggesting the possibility for earlier detection of end-of-sample bubble episodes compared to extant procedures.

When testing for the possibility of changing autoregressive dynamics in financial series, it can also be important to recognise that the underlying innovation process may be susceptible to changes in unconditional variance. To this end, we also consider variants of the above procedures that are robust to heteroskedasticity. A studentisation of the Andrews and Kim (2006)-type statistics automatically delivers tests which are robust to breaks in the unconditional volatility of the driving shocks occurring before the end-of-sample window of observations used to compute the statistics. To achieve robustness to a wider class of heteroskedastic processes, including volatility changes that occur during the end-of-sample window, we propose a further correction to the statistics based on a White-type heteroskedasticity adjustment. Following Harvey *et al.* (2016) [HLST] who proposed wild bootstrap versions of the PWY test, we also consider a wild bootstrap variant of the backward recursive approach of PSY in order to also render this statistic asymptotically robust to non-stationary volatility.

The remainder of the paper is organised as follows. In section 2 we outline the model and present our proposed test procedures. The finite sample size and power properties of the tests are examined in section 3 using Monte Carlo simulations, where comparisons are made with the PSY approach. In section 4 we consider extensions to the proposed tests, and also to the PSY approach, that are robust to heteroskedasticity, and assess the finite sample and power properties of the different procedures. Section 5 presents an application to the same S&P500 price-dividend ratio data studied by PSY, and section 6 concludes.

2 The Model and Tests for an End-of-Sample Bubble

Consider a time series process $\{y_t\}$ generated according to the following data-generating process [DGP]

$$y_t = \mu + u_t, \quad t = 1, ..., T+m \tag{1}$$

$$u_t = \begin{cases} u_{t-1} + \varepsilon_t, & t = 1, ..., T \\ \phi u_{t-1} + \varepsilon_t, & t = T+1, ..., T+m \end{cases} \tag{2}$$

where $u_0 = O_p(1)$ and where, following Andrews (2003) and Andrews and Kim (2006), we assume the innovation process $\{\varepsilon_t\}$ is mean zero, stationary and ergodic. The series y_t follows a unit root process for the first T observations and is then subject to (potential) explosive behaviour for the final m observations (where m is considered to be small relative to T). The null hypothesis (H_0) of no explosive behaviour corresponds to $\phi = 1$ in (2), so that y_t remains unit root throughout the entire sample of $T+m$ observations, while the alternative (H_1) of an end-of-sample explosive regime occurs when $\phi > 1$ in (2).

Standard Dickey-Fuller-type approaches to testing H_0 against H_1 have been developed (see, among others, PWY, Homm and Breitung, 2012, and PSY), and rely on large sample theory to establish properties of their test procedures, implicitly treating the length of the bubble regime to be of order T. Given that our focus is on developing tests to detect end-of-sample bubble behaviour in y_t when there are only a few observations from the bubble regime in the sample, an alternative approach to consider is that of the end-of-sample instability tests that follow the work of Andrews (2003) and Andrews and Kim (2006), where the asymptotics rely on $T \to \infty$ while, importantly, m is allowed to remain finite. Initially treating m as known, the general Andrews-type approach involves calculating a test statistic based on a finite length window of m end-of-sample observations, and comparing this with critical values obtained by sub-sampling using the first T observations. Specifically, $T-m+1$ analogous test statistics are computed, each using a rolling window of m observations, from $t = 1, ..., m$ to $t = T-m+1, ..., T$; the α-level critical value is then equal to the $1-\alpha$ empirical quantile of these $T-m+1$ sub-sample statistics. In this paper we consider a number of suitably designed tests within this Andrews-type framework, where the intention is to distinguish between $H_0 : \phi = 1$ and $H_1 : \phi > 1$ by comparing a statistic that detects explosivity based on $t = T+1, ..., T+m$, with a critical value

obtained from this same statistic applied to the $T - m + 1$ prior sub-samples.

A natural candidate statistic to use in the Andrews-type approach is the Dickey-Fuller t-ratio (DF_m) associated with the OLS estimator of ρ in the regression

$$\Delta y_t = \mu^* + \rho y_{t-1} + \varepsilon_t, \quad t = j+1, ..., j+m$$

defined for the sub-samples $j \in [1, ..., T-m]$ (for critical value calculation) and $j = T$ (for the test statistic), where a rejection of H_0 in favour of H_1 is signalled by an upper-tail rejection. Notice that lagged difference augmentation is not required because any dependence in ε_t is common to all sub-samples. Under our assumptions, the Andrews-type approach applied to DF_m will result in a correctly sized test under H_0 for large T. A potential problem for this implementation of the Andrews-type approach is that the estimator of the autoregressive parameter ρ is likely to be inaccurate for the small values we envisage using for m, which may have a detrimental effect on power.

A simple alternative statistic to employ in the Andrews-type framework can be motivated by considering the properties of the first differences of y_t. Under H_0, it is clear that $\Delta y_t = \varepsilon_t$ throughout the full sample period, while under H_1, $\Delta y_t = \varepsilon_t$ up to time $t = T$, at which point the bubble regime commences and $\Delta y_t = (\phi - 1)u_{t-1} + \varepsilon_t$. Defining the explosive offset $\delta > 0$ as $\delta := \phi - 1$, we can write, for $t = T+1, ..., T+m$,

$$\begin{aligned} u_t &= (1+\delta)^{t-T} u_T + \sum_{j=0}^{t-T-1} (1+\delta)^j \varepsilon_{t-j} \\ \Delta y_t &= \Delta u_t \\ &= \delta(1+\delta)^{t-T-1} u_T + \sum_{j=0}^{t-T-1} (1+\delta)^j \Delta\varepsilon_{t-j}. \end{aligned} \tag{3}$$

Notice that the stochastic behaviour of Δy_t is dominated by the first term on the right hand side of (3), with, for finite m, $\delta(1+\delta)^{t-T-1} u_T = O_p(T^{1/2})$ and $\sum_{j=0}^{t-T-1}(1+\delta)^j \Delta\varepsilon_{t-j} = O_p(1)$. Next consider approximating $(1+\delta)^{t-T-1}$ using a first order Taylor series expansion around $\delta = 0$. We find

$$(1+\delta)^{t-T-1} \approx 1 + (t-T-1)\delta$$

giving the approximation

$$\Delta y_t = \delta(1-\delta)u_T + \delta^2 u_T(t-T) + e_t \tag{4}$$

where e_t contains the higher order terms in the Taylor series expansion and the $O_p(1)$ term from (3).

Using the approximation in (4), an obvious candidate statistic for an Andrews-type instability test would be the F-statistic for joint significance of the estimated coefficients in a regression of Δy_t on u_T and $u_T(t-T)$, which is identical to calculating the F-statistic in a regression of Δy_t on a constant and linear trend. However, such a statistic is inherently two-sided and

does not take account of the fact that the form of explosive behaviour we are trying to detect imposes positivity constraints on both the constant and linear trend terms in (4). A natural one-sided possibility would be to focus on just the trend term, and simply test for an upward trend in a regression of Δy_t on a constant and linear trend using a t-statistic. The drawback of this approach is that it involves estimating the constant term in all of the rolling sub-sample regressions from which the critical value is obtained (as well as the end-of-sample regression). This constitutes an inefficient approach to testing because it does not make use of the fact that the rolling sub-samples up to time T contain a zero intercept in population terms.

An alternative is instead to simply test for an upward trend in a regression of Δy_t on a linear trend alone using a standard t-statistic. This is correctly specified for the rolling sub-sample statistics (relevant for the critical value); the end-of-sample statistic is then based on an under-specified model, but will retain power against H_1 nonetheless, and unreported simulations confirm that this restricted testing approach yields higher power than either the F-statistic approach or the t-statistic approach that fits a constant. The restricted regression t-statistic in question is a studentised version of the trend coefficient estimator

$$\frac{\sum_{t=j+1}^{j+m}(t-j)\Delta y_t}{\sum_{t=j+1}^{j+m}(t-j)^2}$$

defined for the sub-sample $t = j+1, ..., j+m$, with $j \in [1, ..., T-m]$ (for critical value calculation) and $j = T$ (for the test statistic). In fact, given the nature of the Andrews-type methodology, we can simply consider the numerator

$$S_m := \sum_{t=j+1}^{j+m}(t-j)\Delta y_t$$

because the denominator is numerically identical across j, and neither is a studentisation required under the assumption that ε_t is stationary and ergodic. Under our assumptions, the Andrews-type approach applied to S_m will result in a correctly sized test under H_0 for large T.

It is interesting to note the relation of S_m to the R statistic of Andrews and Kim (2006) for testing an end-of-sample change from I(0) to I(1) behaviour. In the present context, Δy_t is I(0) up to time T, and, given that an explosive autoregressive process retains explosive behaviour when first differenced, it would be expected that a test for a change to I(1) behaviour in Δy_t will also reject in the presence of a change to explosivity. The Andrews-Kim R statistic in this context would be

$$R_m := \sum_{t=j+1}^{j+m}\left(\sum_{s=t}^{j+m}\Delta y_s\right)^2.$$

Note that our statistic S_m can equivalently be expressed as

$$S_m = \sum_{t=j+1}^{j+m}\sum_{s=t}^{j+m}\Delta y_s \tag{5}$$

so in a sense, R_m could be interpreted as a two-sided variant of S_m.

An attractive feature of the Andrews-type approach is that the asymptotic (in T) size of S (and R) will be unaffected by the presence of a finite number of bubbles, each of finite length, occurring earlier in the sample period. This arises because the Δy_t from these bubble regimes affect only an asymptotically negligible number of the sub-sample statistics used for computing the critical value.

In practice, the true value of the putative bubble regime length m is of course unknown. In the remainder of the paper we use m' to denote the sub-sample window width used when construction the tests, denoting the procedures hereafter by $S_{m'}$, $R_{m'}$ and $DF_{m'}$, retaining m as the DGP parameter in (1)-(2).

3 Finite Sample Simulations

In this section we perform a set of Monte Carlo simulation exercises to examine the finite sample (empirical) size and power properties of the end-of-sample $S_{m'}$, $R_{m'}$ and $DF_{m'}$ tests proposed in the previous section. Because in practice the true value of m (the length of the end-of-sample bubble period) will be unknown, we will consider the properties of the tests with different window width settings, m'.

The performance of these tests is assessed in relation to a recursive Dickey-Fuller-based approach following the work of PWY and PSY. Of the procedures proposed by PWY and PSY, the most suitable for testing for the presence of a bubble that occurs at the end of the sample period is an implementation of the $BSADF$ test of PSY. Specifically, we consider the statistic

$$BSADF := \sup_{r\in[0,1-r_0]} ADF_r^1$$

where ADF_r^1 denotes the standard augmented Dickey-Fuller statistic based on fitting the following regression

$$\Delta y_t = \mu + \rho y_{t-1} + \sum_{i=1}^{k} \gamma_k \Delta y_{t-k} + \text{error}_t$$

over the sub-sample period $t = \lfloor rT^* \rfloor + 1, ..., T^*$ (where $\lfloor . \rfloor$ denotes the integer part of its argument), with T^* the sample size (i.e. $T^* := T + m$). The $BSADF$ statistic is therefore a supremum of a sequence of backward recursive unit root statistics running to the end of the sample period. The minimum sub-sample length is given by $\lfloor r_0 T^* \rfloor$, with r_0 chosen to ensure that the sub-samples exceed an appropriate minimum length; we follow PSY and set $r_0 = 0.01 + 1.8/\sqrt{T^*}$. PSY recommend using a small fixed lag length in the Dickey-Fuller regressions, so in the simulations that follow the $BSADF$ test statistic is calculated with $k = 1$. The limiting null distribution of the $BSADF$ statistic is obtained from the result in equation (5) of PSY, on fixing $r_2 = 1$. We employ asymptotic critical values for this test, and we obtained these by simulating the limiting null distribution.

All Monte Carlo simulations that follow were conducted in Gauss 9.0. The size simulations were computed using 50,000 replications, while powers were evaluated with 5,000 replications,

all tests being performed at the nominal 0.05-level. We generate data according to the DGP (1)-(2), setting $\mu = 0$ without loss of generality, and with an initial value $u_0 = 100$, chosen such that under the alternative hypothesis, the bubbles generated are generally *upwardly* explosive (note that the tests are invariant to u_0 under the null).

3.1 Empirical Size

To examine the size of the test procedures discussed in this paper, we set $\phi = 1$, and let ε_t be generated according to the moving average process $\varepsilon_t = v_t + \theta v_{t-1}$ with $v_t \sim IIDN(0,1)$ for $\theta = \{0, \pm 0.3, \pm 0.5\}$. Table 1 reports the empirical size of the test procedures for the total sample sizes $T^* := T + m = \{100, 200\}$, where the $S_{m'}$, $R_{m'}$ and $DF_{m'}$ tests are implemented using $m' = 5$ and $m' = 10$.

The overall picture from Table 1 is that all procedures control size fairly well, particularly for the larger sample size, and with the exception of $BSADF$, the tests are largely unaffected by the presence of serially correlated innovations. Concentrating on the IID case $\theta = 0$, we see that for the smaller sample size of $T^* = 100$ the newly proposed $S_{m'}$, $R_{m'}$ and $DF_{m'}$ tests exhibit some modest oversize; the size of these tests is also increasing in the window width, m', used in their construction. The $BSADF$ test also exhibits mild oversize in this scenario, with maximum size similar to S_5, R_5 and somewhat lower than S_{10}, R_{10}. As we increase the sample size the degree of oversize exhibited by the tests is generally decreasing. The reduction in oversize for $S_{m'}$, $R_{m'}$ and $DF_{m'}$ is due to the fact that, as the sample size increases, we are able to calculate more sub-sample test statistics, allowing more accurate calculations of the critical values of the tests. For non-zero values of θ we see a distinction between the $BSADF$ test and our proposed $S_{m'}$, $R_{m'}$ and $DF_{m'}$ tests. While the latter three are little affected by the value of θ for any sample size, the $BSADF$ test can suffer from undersize for the negative values of θ considered. The relative robustness of the $S_{m'}$, $R_{m'}$ and $DF_{m'}$ tests to moving average components is explained by the fact that the same serial correlation properties present in observations used by the end-of-sample statistic are also present in all of the sub-sample statistics used for critical value computation, thereby rendering the size of the test relatively unaffected.

Given that the undersize observed for $BSADF$ could be attributable to the fact that k in the Dickey-Fuller regressions is fixed at $k = 1$ rather than being data-dependent, we also investigated the properties of $BSADF$ where k is chosen according to the Bayes Information Criterion (BIC). Table 1 also reports results for this variant, which we denote by $BSADF_B$, and where the maximum value of k is set to 6. We find that the undersize is indeed generally removed, but at the expense of substantial oversize, particularly for $T = 100$ and also for $\theta > 0$ for the larger sample sizes. As a result, we do not consider the $BSADF_B$ procedure further in this paper.

3.2 Empirical Power

We now examine the power of the tests to detect an end-of-sample bubble. To do so, we generate $\varepsilon_t \sim IIDN(0,1)$ innovations, and consider bubble lengths of $m = \{2, 5, 10\}$ in sample sizes of $T^* = \{100, 200\}$. Figure 1 reports power curves across $\phi \in [1, \phi_{\max}]$ using a grid of 50 steps, with $\phi_{\max} = 1.05$ for $m = 2$, and $\phi_{\max} = 1.02$ for $m = 5$ and 10, respectively (reflecting the fact that, for a given value of ϕ, a bubble of longer duration is easier to detect).

For the shortest bubble length, $m = 2$, there is a fairly clear ranking of the tests in terms of power, with the best overall performance given by the $S_{m'}$ tests, followed by the $R_{m'}$ tests, with the results qualitatively similar across $T^* = 100$ and $T^* = 200$. The $DF_{m'}$ tests exhibit substantially lower power, and the $BSADF$ test has the poorest power performance of all. It is clear, then, that the $S_{m'}$ and $R_{m'}$ tests are well suited to detect end-of-sample bubbles of very short duration, unlike the $DF_{m'}$ and $BSADF$ tests. It is also interesting to note that the choice of the end of sample window width, m', used in the $S_{m'}$ and $R_{m'}$ tests has an impact on their respective power levels, with the shorter window settings (i.e. S_5 and R_5) delivering relatively higher power for this short duration end-of-sample bubble than the longer window widths (i.e. S_{10} and R_{10}). This ranking is reversed for the $DF_{m'}$ tests, where DF_5 has lower power than DF_{10}.

Moving to the case of $m = 5$, we observe a broadly similar power ranking among the tests. In particular, the $S_{m'}$ tests continue to display the best overall power profiles, with both window width variants S_5 and S_{10} now emerging as unambiguously the most powerful procedures. The power of the $R_{m'}$ tests again lie between the power curves for the $S_{m'}$ and $BSADF$ tests, but interestingly, the power of the $DF_{m'}$ tests is now very sensitive to the choice of m', with DF_5 displaying very low power levels, below that of $BSADF$.

For the case of $m = 10$, all the $DF_{m'}$ tests have poor power performance, while $BSADF$ now has a more competitive power profile for this bubble of longer duration. The power of the S_5 and R_5 tests is lower for this case where the bubble duration is considerably longer than the window width used in the tests; this arises because bubble observations are now being included in the sample period from which the critical values are derived. However, the S_{10} test retains its position as the best performing test.

From a real-time monitoring perspective, it is interesting to investigate how quickly a bubble is likely to be detected by the different procedures. One way of measuring this speed of detection is to examine the powers of the tests when the sample contains just a single bubble observation at the end, then when the last two observations correspond to the bubble regime, then the last three, and so on. Other things being equal, a good test for real-time monitoring purposes will be one that has high power for a low number of bubble observations, so that in practice the null would be expected to be rejected in favour of a bubble relatively early into the bubble regime. We now consider such power comparisons, restricting our attention to the best-performing procedure $S_{m'}$ and the comparator test $BSADF$, by simulating processes with $T^* = 200$ and

a single end-of-sample bubble of length $m = 20$. We initially apply the tests to the simulated series using only the observations $t = 1, ..., 160$, thereby evaluating the rejection frequencies of the tests as if we were at the point in time 160. We then repeat the simulation exercise using the observations $t = 1, ..., 161$, again calculating the rejection frequencies of the tests (now associated with time period 161), and continue in this manner until we are simulating the rejection frequencies based on the full sample $t = 1, ..., 200$. Of course, for the simulation experiments up to and including time period 180, no bubble is present in the data so we expect to see rejection frequencies close to the nominal size for these cases. After this point, a bubble is present of increasing duration, and we can evaluate the powers of the tests to detect it, giving an indication of the relative performance of the procedures to provide an early warning of a bubble in an evolving real-time situation.

Denoting the end-date of the sample to which the tests are applied by E, Figure 2 reports the rejection frequencies of the $S_{m'}$ and $BSADF$ tests for $E = \{160, 161, ..., 200\}$ for $\phi = 1.01$ and $\phi = 1.02$.[1] We observe that the rejection frequencies for all the tests are approximately equal to their nominal size up to $E = 180$, after which time the bubble enters the samples and the rejection frequencies begin to rise. It can be seen that the powers reinforce the results from our earlier power simulations, with the $S_{m'}$ test most likely to reject early into the bubble than $BSADF$. As we move further into the bubble, the rejection frequency of some of the $S_{m'}$ tests begins to plateau or decrease; this feature arises because when $E > 180 + m'$, the critical values start to increase due to contamination by bubble observations. The $BSADF$ test is not subject to these power decreases due to its construction, and power continues to rise with increasing numbers of bubble observations. However, the results demonstrate that it is the $S_{m'}$ procedure that is particularly well-suited to early detection of an end-of-sample bubble.[2]

4 Accounting for Heteroskedasticity

The tests considered thus far implicitly assume that the unconditional variance of the innovation process $\{\varepsilon_t\}$ is constant throughout the sample period. However, when dealing with financial time series, it is important to recognise that the underlying innovations may be susceptible to

[1] Note that the $BSADF$ results depend on ϕ but do not of course change across the settings for m'; the results are simply repeated for ease of comparison. Figure 2 also reports results for tests that will be introduced and discussed later in the paper.

[2] In a companion discussion paper version of this paper (Astill *et al.*, 2016), we also considered DGPs where a bubble abruptly collapses after a number of periods, and also examined the impact of a previously collapsed bubble on the power of the tests to detect an end-of-sample bubble. Overall, we find similar power patterns to those in Figure 2, apart from when a previously collapsed bubble is relatively close to the end-of-sample bubble. In this latter case, the $S_{m'}$ tests recover their properties from the collapse of the first bubble much more rapidly than the $BSADF$ test, which has relatively poor power to detect the second bubble. Of course, a prior bubble of long duration can adversely affect the powers of the $S_{m'}$ tests, since a large proportion of sub-sample statistics used for computing the critical values are affected by the earlier bubble, hence caution should be exercised if a long bubble is present in the sample period used to obtain critical values.

variance changes. To this end, we now consider variants of the better-performing $S_{m'}$ tests that account for unconditional heteroskedasticity (note that conditional heteroskedasticity is already permitted under the conditions on ε_t; see Andrews, 2003).

A first step in this direction would be to consider a simple studentised version of $S_{m'}$, taking the form

$$S^*_{m'} := \frac{S_{m'}}{\sqrt{\sum_{t=j+1}^{j+m'} (\Delta y_t)^2}}. \tag{6}$$

Such a modification imbues the $S^*_{m'}$ tests with robustness to a finite number of volatility shifts that occur over the period $t = 1, ..., T^* - m'$. This arises because, for all sub-samples which do not contain a variance break, the statistics are correctly studentised, while only a finite number of sub-sample statistics will have a studentisation that is contaminated by the variance change. Given that only an asymptotically negligible number of the sub-sample statistics used for computing the critical value are affected, $S^*_{m'}$ will be asymptotically correctly size. While $S^*_{m'}$ would not deliver size control if a volatility shift occurred in the final m' observations of the series, in some circumstances it may be deemed that the greater concern is robustness to volatility shifts that arise over the much longer sample period used to obtain the critical values.

A further modification that would produce a test robust to unconditional heteroskedasticity of more general form across the full sample period (including the final m' observations) is to adopt a White-type correction in the studentisation, i.e.:

$$S^{*w}_{m'} := \frac{S_{m'}}{\sqrt{\sum_{t=j+1}^{j+m'} \{(t-j)\Delta y_t\}^2}}. \tag{7}$$

In what follows we assess the relative size and power performance of $S^*_{m'}$ and $S^{*w}_{m'}$, under both homoskedastic and heteroskedastic DGPs, and also consider their properties in relation to the unmodified $S_{m'}$ tests.

In a recent paper, HLST developed wild bootstrap variants of the PWY test that deliver asymptotic robustness to non-stationary volatility. In the current setting, it is natural to consider a similar wild bootstrap approach applied to the $BSADF$ statistic outlined above, which will also serve as a useful comparator for the $S^*_{m'}$ and $S^{*w}_{m'}$ tests. HLST propose two bootstrap algorithms, one based on the wild bootstrap applied to the first differences of the series, the other based on the wild bootstrap applied to residuals from a fitted model, which makes use of the BIC-based approach of Harvey, Leybourne and Sollis (2016). We also consider the equivalent two wild bootstrap methods here, although in the latter case, because we are purely interested in modelling a putative bubble that occurs at the end of the sample, we restrict attention to Model 1 of that paper in the model fitting stage, thereby fitting a unit root to bubble model with the change-point date identified by minimising the sum of squares residuals.[3] Asymptotic results similar to those of HLST would apply to such bootstrap tests, ensuring the

[3] The HLST dating methodology only identifies valid bubble dates where the end of bubble date (denoted y_{T^*} here) exceeds the start of bubble date (denoted y_T here). In cases where this is not satisfied for any T, we revert to using Δy_t for the model-based residuals.

asymptotic validity of these procedures under heteroskedasticity of the form considered here. In the sequel, we denote the first difference-based wild bootstrap approach by $BSADF_b^1$, and the model-based variant by $BSADF_b^2$.

We now evaluate the finite sample size and power of the heteroskedasticity-adjusted tests using a similar set of Monte Carlo simulations to those of the previous section. Table 2 reports the sizes of $S_{m'}^*$, $S_{m'}^{*w}$, $BSADF_b^1$ and $BSADF_b^2$, along with the original $S_{m'}$ and $BSADF$ tests for comparison. Here, we introduce a single shift in the variance of the innovations, with $\varepsilon_t \sim IIDN(0,1)$ for $t = 1, ..., T_\sigma$ and $\varepsilon_t \sim IIDN(0,\sigma^2)$ for $t = T_\sigma + 1, ..., T^*$, for $\sigma^2 = \{1/10, 1/5, 1, 5, 10\}$. We consider two cases: (i) $T_\sigma = T/2$, allowing for a mid sample shift, and (ii) $T_\sigma = T^* - 5$, where the shift occurs five observations from the end of the sample, commensurate with our focus on changes occurring late in the sample period. First, in the homoskedastic case ($\sigma^2 = 1$), we find that the S_5^*, S_5^{*w} and S_{10}^*, S_{10}^{*w} tests display very similar sizes to their uncorrected counterparts S_5 and S_{10}, respectively. Similarly, the size of $BSADF_b^2$ is similar to that of $BSADF$, with the size of $BSADF_b^1$ a little lower.

When the innovation variance changes, the impact on the tests is dependent on both the timing and the direction of the change. The unadjusted tests S_5, S_{10} and $BSADF$ lack robustness to σ^2, and, relative to the homoskedastic case, size decreases when there is a downward variance shift, and size increases when the shift is upwards. The extent of the size distortions is relatively modest in the case of a mid sample variance change, but is more exaggerated when the change occurs late. Indeed, quite large oversize is seen in all these tests when a late upward change arises. For the S_5^* and S_{10}^*, as would be expected, size robustness is seen when the volatility change occurs mid sample, although when the volatility change is only present in the last five observations, the $S_{m'}^*$ approach does not generally deliver robustness. This is seen in the size distortions manifest in the S_{10}^* test, with undersize associated with an increase in variance, and oversize with a decrease in variance. Note that here, S_5^* is numerically invariant to σ^2 as the window width coincides with the number of observations in the final variance regime for this particular case. The $S_{m'}^{*w}$ tests achieve good size control across σ^2 when $T^* = 200$, demonstrating the robustness of this approach to heteroskedasticity. When $T^* = 100$, some upward size distortions are present, but these are modest in nature compared to the unadjusted tests. The asymptotically heteroskedasticity-robust $BSADF_b^1$ and $BSADF_b^2$ tests improve finite sample size relative to $BSADF$, although the $BSADF_b^2$ variant can still have size in excess of 0.10 for late upward volatility shifts, even when $T^* = 200$.

Figure 3 reports finite sample power results for $S_{m'}^*$, $S_{m'}^{*w}$, $BSADF_b^1$ and $BSADF_b^2$ for the same homoskedastic DGPs as were considered in Figure 1. The original $S_{m'}$ and $BSADF$ power curves are also super-imposed for comparison purposes. Consider first $m = 2$ where the bubble begins very close to the sample end. It is evident that the heteroskedasticity corrections applied to $S_{m'}^*$ and $S_{m'}^{*w}$ have a cost in terms of power, with the power ranking being, for a given m', $S_{m'}$ followed by $S_{m'}^*$ and then $S_{m'}^{*w}$. As with the $S_{m'}$ tests, S_5^* outperforms S_{10}^* here, although interestingly, S_{10}^{*w} displays greater power than S_5^{*w}. The $BSADF_b^2$ test displays similar levels

of power to $BSADF$, as might be expected given the results of HLST who show that this wild bootstrap approach involves no loss in (size-adjusted) power. It is noticeable that $BSADF_b^1$ does not achieve the same power as $BSADF$, in contrast to HLST's findings for this test when applied to bubbles of longer duration. On comparing the $S_{m'}^{*w}$ and $BSADF_b^2$ tests, $S_{m'}^{*w}$ offers higher power for the smaller values of ϕ, while the ranking is reversed for larger ϕ, suggesting a possible role for $BSADF_b^2$ in the early detection of large bubbles. Turning to $m = 5$, we see that $S_{m'}^{*}$ and $S_{m'}^{*w}$ have levels of power closer to each other, and also closer to the uncorrected $S_{m'}$ tests. Here, the $S_{m'}^{*}$ and $S_{m'}^{*w}$ tests have superior power to $BSADF_b^2$ (and $BSADF_b^1$) across ϕ for both values of m'. For $m = 10$, the S_{10}^{*w} becomes the best performing of all the corrected tests, dominating S_{10}^{*} and $BSADF_b^2$, as well as S_5^{*} and S_5^{*w}. On the basis of these results, our recommendation would be for the S_{10} test in the absence of heteroskedasticity concerns, and the S_{10}^{*w} variant if full robustness to heteroskedasticity is desired.

The rejection frequency simulations across sample end-dates reported in Figure 2 also contain results for the $S_{m'}^{*}$, $S_{m'}^{*w}$, $BSADF_b^1$ and $BSADF_b^2$ tests. In line with the results of Figure 3, we observe that the $S_{m'}^{*}$ and $S_{m'}^{*w}$ follow the same broad rejection patterns as $S_{m'}$, but with reduced power levels. It can be seen that across all the figures, $S_{m'}^{*}$ is more likely to reject early into the bubble regime compared with $S_{m'}^{*w}$, but then the $S_{m'}^{*}$ tests achieve a greater rejection frequency when further into the bubble. The $BSADF_b^2$ test displays a similar rejection pattern to $BSADF$ (again in line with Figure 3), with the $BSADF_b^1$ powers somewhat lower. As was the case with $BSADF$, $BSADF_b^1$ and $BSADF_b^2$ have power that always rises with increasing numbers of bubble observations, while the $S_{m'}^{*}$ and $S_{m'}^{*w}$ powers eventually plateau and decrease. As before, however, the Andrews-based approaches deliver greater early rejection frequencies than the $BSADF$ approach and its bootstrap variants.

Finally, in Figure 4 we consider powers when heteroskedasticity is present in the innovations. We restrict attention to $T^* = 200$ and $m = 5$, and simulate the powers of the $S_{m'}^{*}$, $S_{m'}^{*w}$, $BSADF_b^1$ and $BSADF_b^2$ tests for four cases, covering both a mid sample increase and decrease in volatility ($\sigma^2 = 1/5$ and $\sigma^2 = 5$), and volatility shifts of the same magnitude that occur in the last five observations. Consider first the results for the mid sample volatility shifts. Here, all tests are asymptotically robust to the heteroskedasticity, as is reflected in the $\phi = 1$ power curve intercepts. Compared to the corresponding DGP without any variance shift (i.e. Figure 3(d)), the powers of the tests are increased for $\sigma^2 = 1/5$ and decreased for $\sigma^2 = 5$. However, the relative rankings of the procedures are broadly unaffected by the presence of heteroskedasticity, with the most noticeable feature being the dominance of $S_{m'}^{*}$ and $S_{m'}^{*w}$ over $BSADF_b^1$ and $BSADF_b^2$. When the variance change applies only to the last five observations, S_{10}^{*} is no longer robust, and is subject to undersize when $\sigma^2 = 1/5$ and oversize when $\sigma^2 = 5$. For the volatility decrease, the $S_{m'}^{*w}$ tests substantially outperform $BSADF_b^1$ and $BSADF_b^2$, while the ranking is less clear with respect to $BSADF_b^2$ when the volatility increases, partly because $BSADF_b^2$ displays some finite sample oversize in this case. Overall, our recommendation remains for the S_{10}^{*w} test in the presence of possible heteroskedasticity.

5 An Empirical Application

We now examine the ability of our test procedures to detect bubbles in an empirical data series. PSY apply their real-time dating strategy (based on sequential application of a sequence of backward recursive right-tailed DF statistics) to the S&P500 price-dividend ratio, using monthly data over the period 1871M01-2010M12. They identify five primary bubble episodes: the post long-depression period (1879M10-1880M04), the Great Crash episode (1928M11-1929M10), the postwar boom (1955M01-1956M04), Black Monday in October 1987 (1986M06-1987M09) and the dot-com bubble (1995M11-2001M08). Focusing on these episodes, we apply the $S_{m'}$, $S^*_{m'}$ and $S^{*w}_{m'}$ tests in a pseudo-real-time manner to the same dataset, beginning the testing with the first 100 observations (1871M01-1879M4), to examine whether these new procedures could have detected the onset of these bubble episodes sooner than using PSY's approach. Table 3 reports, for each bubble episode, the first date for which each test rejects in favour of explosive behaviour (the first of the PSY bubble regime dates is also listed for comparison in each case).

For the post long-depression, there is little to choose between the $S_{m'}$ and $S^*_{m'}$ tests, with all of these tests first rejecting in either 1879M10 or 1879M11, broadly in line with the PSY date of 1879M10. The $S^{*w}_{m'}$ tests do not detect this episode, possibly due to the reduced power of this test when allowing for heteroskedasticity in the final m' observations. For the Great Crash episode, the $S_{m'}$ tests reject in exactly the same period identified by PSY. The $S^*_{m'}$ and $S^{*w}_{m'}$ tests reject well before this, generally in late 1925 (S^*_{10} first rejects in 1927M08, and S^{*w}_5 also rejects at this point in time as well as in 1925M10), potentially indicating an early detection of explosive behaviour in the run-up to the Great Crash episode. In the case of the postwar boom, the $S^*_{m'}$ and $S^{*w}_{m'}$ tests reject several months before the initial bubble date identified by PSY. The first rejection is in 1954M02 for S^*_5 and S^{*w}_5, while rejections are first found in 1954M05 and 1954M06 for S^*_{10} and S^{*w}_{10}, respectively; these are to be compared with the date of 1955M01 for PSY, demonstrating clear evidence of earlier detection of this bubble episode (in contrast, however, the $S_{m'}$ tests only show a rejection six months after the date identified by PSY). Turning to the Black Monday period, S_5 and S_{10} reject three to four months sooner than PSY, S^*_{10} rejects two months earlier, and S^{*w}_{10} rejects at the same time as PSY. The S^*_5 and S^{*w}_5 tests fail to reject for this episode, reinforcing our overall preference for the $m' = 10$ tests. Finally, for all of our proposed tests, the first rejections seen for the dot-com bubble are well ahead of the date identified by PSY, with S_5, S^*_5 and S^{*w}_5 rejecting six months before the PSY date of 1995M11, and S_{10}, S^*_{10} and S^{*w}_{10} rejecting four to five months ahead of PSY's dates.

In addition to the exuberance periods focused on above, the $S_{m'}$, $S^*_{m'}$ and $S^{*w}_{m'}$ tests also reject for a number of other sequential dates across the sample period, suggesting that there may well have been additional periods of explosive autoregressive behaviour in this series that the newly proposed tests detect. For example, our proposed tests find evidence of explosive behaviour in both the years leading up to and during the Second World War, and also find evidence of a period of explosivity following the end of the First World War. In summary then,

the $S_{m'}$, $S^*_{m'}$ and $S^{*w}_{m'}$ tests would in many cases have detected well-documented periods of exuberance before the PSY approach, and also find evidence of some periods of explosive behaviour not identified by PSY. This suggests a worthwhile role for the new tests, in complement to existing procedures such as, in particular, that of PSY.

6 Conclusions

In this paper we have proposed test procedures for the detection of an end-of-sample asset price bubble of finite length. These involve calculating the test statistic of interest on a small number of end-of-sample observations, with a critical value obtained by sub-sampling using the same statistic calculated on the remaining observations. Simulation evidence highlights the size robustness properties of our tests in finite samples, and also their potential power advantages when compared to existing approaches, particularly in terms of the possibility for early detection of an ongoing end-of-sample bubble. A (pseudo) real-time monitoring exercise using the S&P500 price dividend ratio was performed, and it was found that our testing approach detected a number of past bubble episodes a number of months in advance of the dates suggested by PSY. As such we believe the $S_{m'}$, $S^*_{m'}$ and $S^{*w}_{m'}$ tests developed in this paper are a valuable addition to the suite of recently developed bubble detection procedures when the focus is on early bubble detection in a real-time setting.

References

Andrews, D.W.K. (2003). End-of-sample instability tests. Econometrica 71, 1661-1694.

Andrews, D.W.K. and Kim, J.-Y. (2006). Tests for cointegration breakdown over a short time period. Journal of Business and Economic Statistics 24, 379-394.

Astill, S., Harvey, D.I., Leybourne, S.J. and Taylor, A.M.R. (2016). Tests for an end-of-sample bubble in financial time series. Granger Centre Discussion Paper No. 16/02, School of Economics, University of Nottingham.

Bettendorf. T. and Chen, W. (2013). Are there bubbles in the sterling-dollar exchange rate? New evidence from sequential ADF tests. Economics Letters 120, 350-353.

Diba, B.T. and Grossman, H.I. (1988). Explosive rational bubbles in stock prices? American Economic Review 78, 520-530.

Gilbert, C.L. (2010). Speculative influence on commodity prices 2006-08. Discussion Paper 197, United Nations Conference on Trade and Development (UNCTAD), Geneva.

Harvey, D.I., Leybourne, S.J. and Sollis, R. (2016). Improving the accuracy of asset price bubble start and end date estimators. Discussion Paper, School of Economics, University of Nottingham.

Harvey, D.I., Leybourne, S.J., Sollis, R. and Taylor, A.M.R. (2016). Tests for explosive financial bubbles in the presence of non-stationary volatility. Journal of Empirical Finance 38, 548-574.

Homm, U. and Breitung, J. (2012). Testing for speculative bubbles in stock markets: a comparison of alternative methods. Journal of Financial Econometrics 10, 198-231.

Phillips, P.C.B., Wu, Y. and Yu, J. (2011). Explosive behavior in the 1990s Nasdaq: when did exuberance escalate stock values? International Economic Review 52, 201-226.

Phillips, P.C.B., Shi, S.-P. and Yu, J. (2015). Testing for multiple bubbles: historical episodes of exuberance and collapse in the S&P 500. International Economic Review 56, 1043-1078.

Table 1. Finite sample size - serial correlation

$T^* = 100$								
θ	S_5	S_{10}	R_5	R_{10}	DF_5	DF_{10}	$BSADF$	$BSADF_B$
-0.5	0.064	0.075	0.064	0.072	0.060	0.071	0.011	0.106
-0.3	0.067	0.081	0.066	0.076	0.060	0.070	0.040	0.127
0.0	0.069	0.086	0.067	0.081	0.061	0.068	0.073	0.148
0.3	0.071	0.088	0.069	0.083	0.061	0.067	0.066	0.193
0.5	0.072	0.088	0.069	0.083	0.060	0.067	0.055	0.208
$T^* = 200$								
θ	S_5	S_{10}	R_5	R_{10}	DF_5	DF_{10}	$BSADF$	$BSADF_B$
-0.5	0.057	0.062	0.056	0.061	0.057	0.059	0.007	0.055
-0.3	0.058	0.064	0.058	0.062	0.055	0.059	0.035	0.068
0.0	0.059	0.066	0.059	0.064	0.057	0.058	0.065	0.082
0.3	0.059	0.066	0.059	0.064	0.057	0.058	0.056	0.121
0.5	0.059	0.067	0.059	0.064	0.056	0.058	0.044	0.129

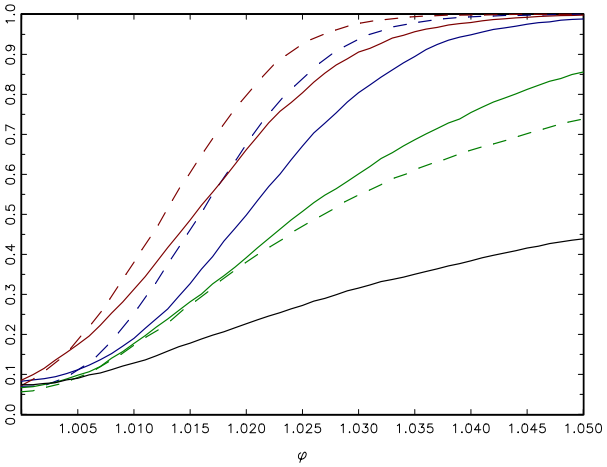
Table 2. Finite sample size - variance shift

$T^* = 100$																		
	Mid sample									Last 5 observations								
σ^2	S_5	S_{10}	S_5^*	S_{10}^*	S_5^{*w}	S_{10}^{*w}	$BSADF$	$BSADF_b^1$	$BSADF_b^2$	S_5	S_{10}	S_5^*	S_{10}^*	S_5^{*w}	S_{10}^{*w}	$BSADF$	$BSADF_b^1$	$BSADF_b^2$
1/10	0.004	0.013	0.071	0.089	0.070	0.086	0.063	0.067	0.049	0.000	0.009	0.069	0.027	0.070	0.085	0.021	0.057	0.062
1/5	0.012	0.024	0.070	0.088	0.070	0.086	0.063	0.066	0.057	0.002	0.016	0.069	0.040	0.070	0.086	0.029	0.058	0.062
1	0.069	0.086	0.069	0.084	0.070	0.085	0.073	0.063	0.073	0.069	0.086	0.069	0.084	0.070	0.085	0.073	0.063	0.073
5	0.126	0.147	0.067	0.080	0.071	0.085	0.098	0.061	0.073	0.247	0.245	0.069	0.128	0.070	0.086	0.164	0.066	0.111
10	0.135	0.159	0.067	0.079	0.071	0.085	0.107	0.060	0.072	0.314	0.309	0.069	0.140	0.070	0.086	0.208	0.065	0.120
$T^* = 200$																		
	Mid sample									Last 5 observations								
σ^2	S_5	S_{10}	S_5^*	S_{10}^*	S_5^{*w}	S_{10}^{*w}	$BSADF$	$BSADF_b^1$	$BSADF_b^2$	S_5	S_{10}	S_5^*	S_{10}^*	S_5^{*w}	S_{10}^{*w}	$BSADF$	$BSADF_b^1$	$BSADF_b^2$
1/10	0.001	0.003	0.059	0.069	0.059	0.068	0.055	0.062	0.050	0.000	0.003	0.058	0.013	0.059	0.066	0.029	0.057	0.062
1/5	0.007	0.011	0.059	0.068	0.059	0.068	0.055	0.061	0.056	0.001	0.007	0.058	0.024	0.059	0.067	0.034	0.057	0.062
1	0.059	0.066	0.058	0.067	0.059	0.067	0.065	0.060	0.068	0.059	0.066	0.058	0.067	0.059	0.067	0.065	0.060	0.068
5	0.114	0.122	0.057	0.064	0.060	0.067	0.091	0.058	0.068	0.239	0.230	0.058	0.111	0.059	0.067	0.142	0.063	0.099
10	0.119	0.130	0.057	0.064	0.060	0.068	0.100	0.058	0.068	0.308	0.298	0.058	0.122	0.059	0.067	0.187	0.062	0.112

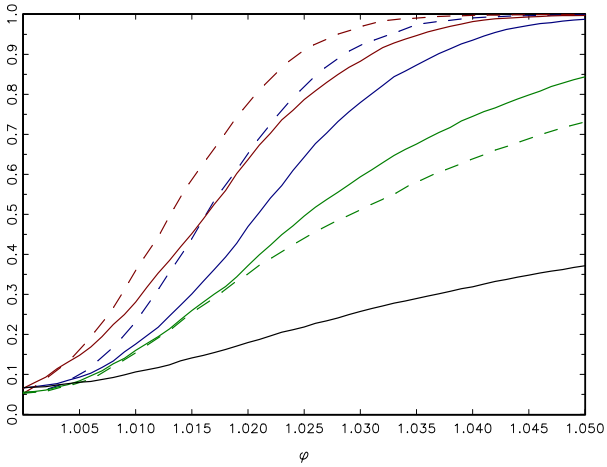
Table 3. Identified exuberance period start dates

Exuberance period	PSY	S_5	S_{10}	S_5^*	S_{10}^*	S_5^{*w}	S_{10}^{*w}
Post long-depression	1879M10	1879M10	1879M10	1879M10	1879M11	-	-
Great Crash	1928M11	1928M11	1928M11	1925M10	1927M08	1925M09	1925M12
Postwar boom	1955M01	1955M07	1955M07	1954M02	1954M05	1954M02	1954M06
Black Monday	1986M06	1986M02	1986M03	-	1986M04	-	1986M06
Dot-com bubble	1995M11	1995M05	1995M06	1995M05	1995M06	1995M05	1995M07

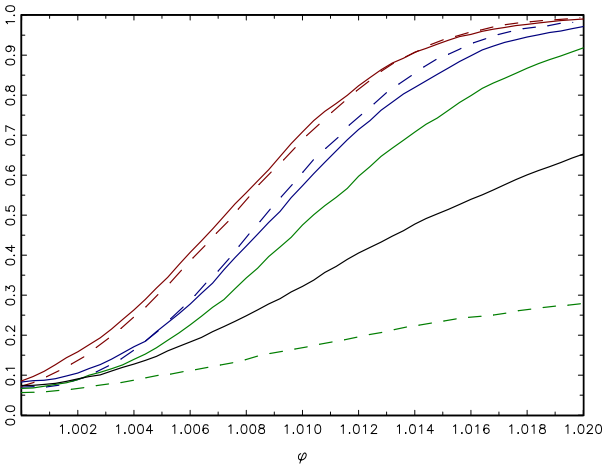
Notes: The column headed PSY records the start dates of the exuberance periods identified by Phillips, Shi and Yu (2015). The remaining columns record the first date for which a particular test rejects in favour of a bubble. For a given window width $m' = \{5, 10\}$, $S_{m'}$ denotes our proposed Andrews-type statistic given in equation (5), $S_{m'}^*$ denotes the studentized version given in (6) which robustifies the procedure to volatility shifts that occur prior to the testing window, and $S_{m'}^{*w}$ denotes the White-type corrected variant given in (7) which delivers robustness to heteroskedasticity across the full sample period.



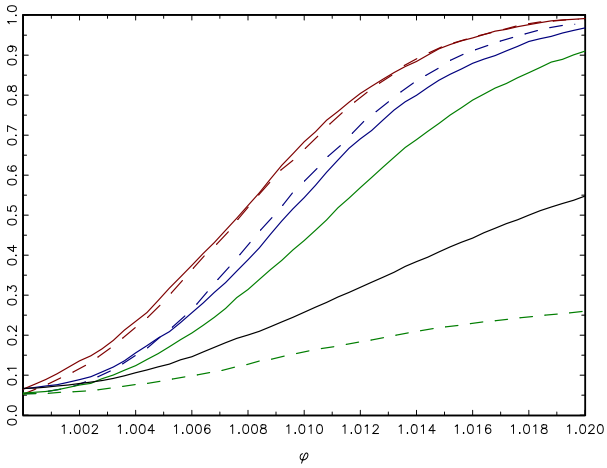
(a) $T^* = 100$, $m = 2$



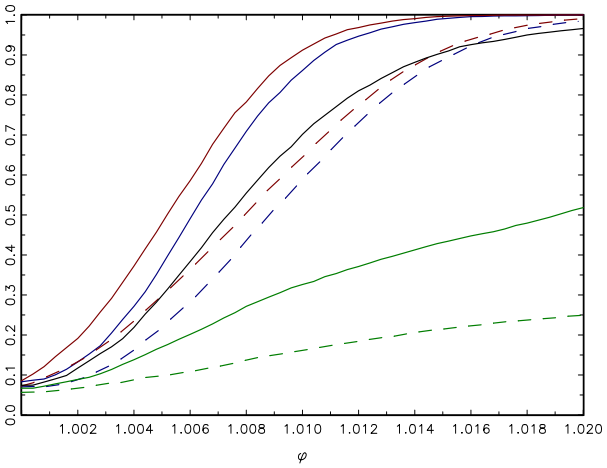
(b) $T^* = 200$, $m = 2$



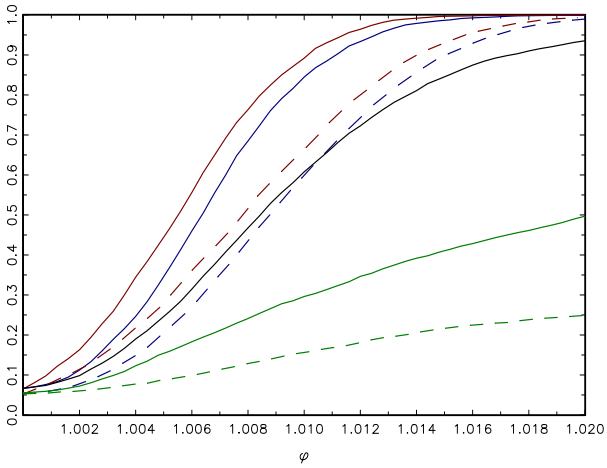
(c) $T^* = 100$, $m = 5$



(d) $T^* = 200$, $m = 5$

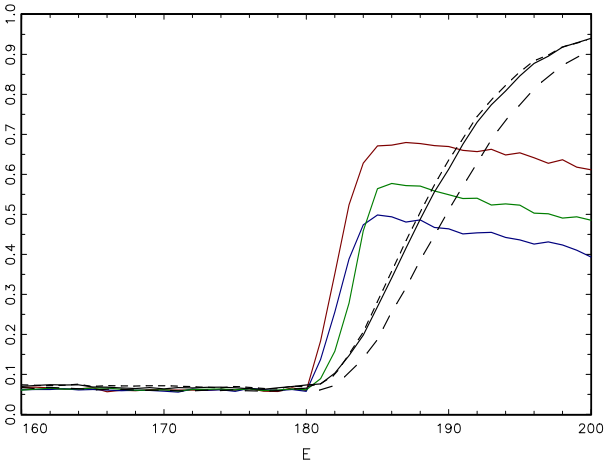


(e) $T^* = 100$, $m = 10$

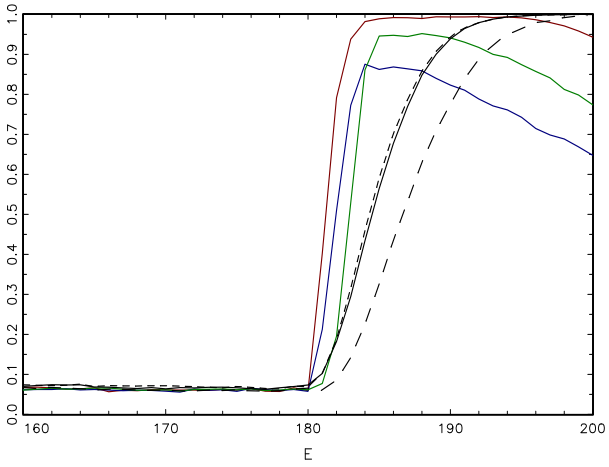


(f) $T^* = 200$, $m = 10$

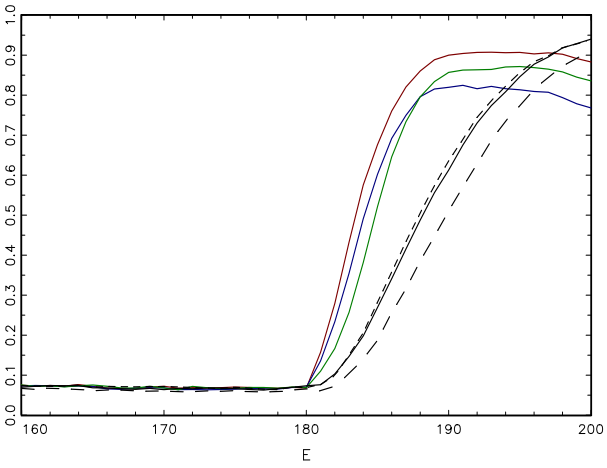
Figure 1. Finite sample power of nominal 0.05-level tests: i.i.d. innovations:
S_5: – –, S_{10}: ——, R_5: – –, R_{10}: ——, DF_5: – –, DF_{10}: ——, $BSADF$: ——



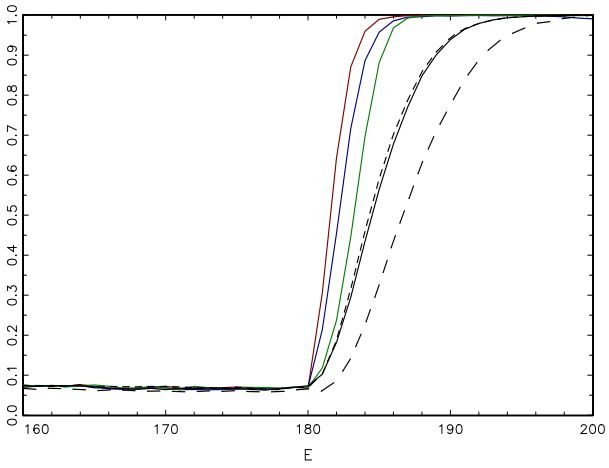
(a) $\phi = 1.01$, $m' = 5$



(b) $\phi = 1.02$, $m' = 5$

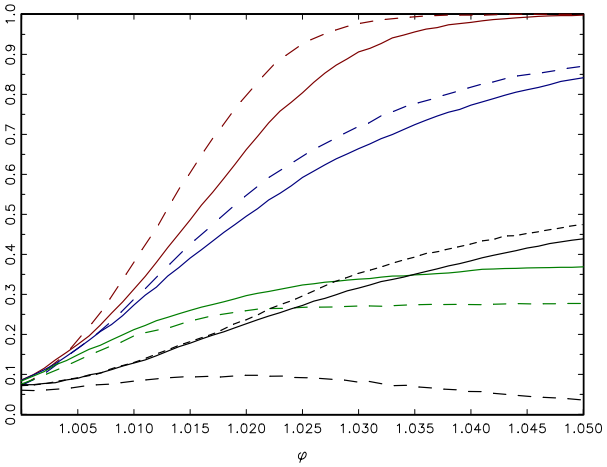


(c) $\phi = 1.01$, $m' = 10$

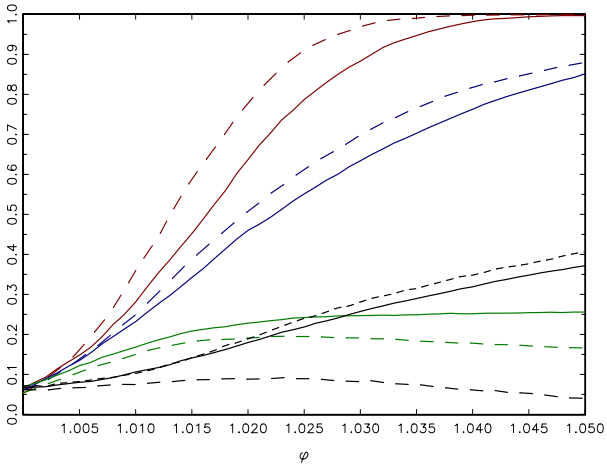


(d) $\phi = 1.02$, $m' = 10$

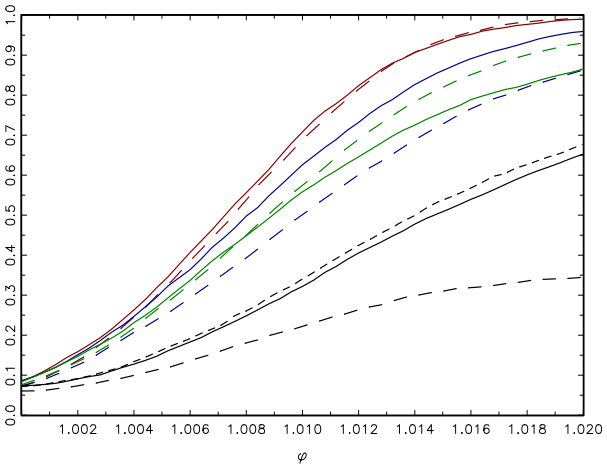
Figure 2. Rejection frequencies of nominal 0.05-level tests: single end-of-sample bubble: $S_{m'}$: ——, $S^*_{m'}$: ——, $S^{*w}_{m'}$: ——, $BSADF$: ——, $BSADF^1_b$: – –, $BSADF^2_b$: - - -



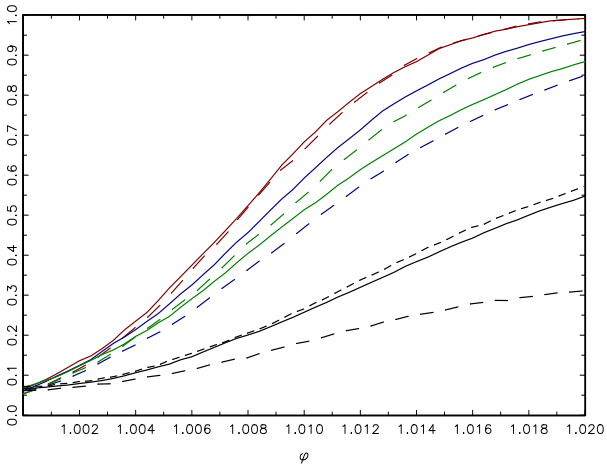
(a) $T^* = 100$, $m = 2$



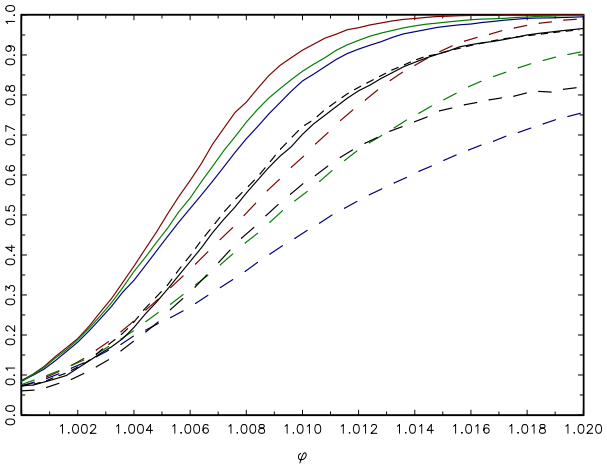
(b) $T^* = 200$, $m = 2$



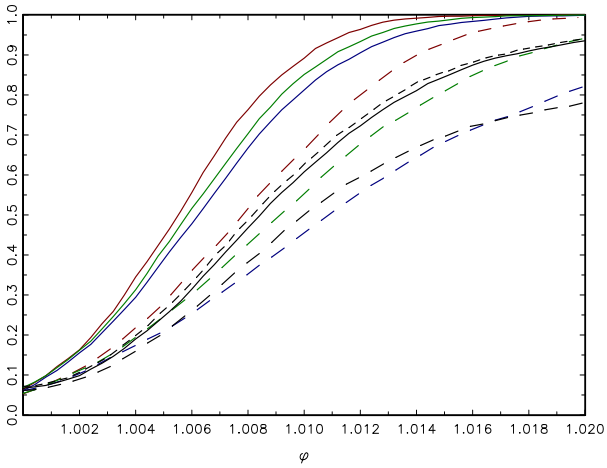
(c) $T^* = 100$, $m = 5$



(d) $T^* = 200$, $m = 5$

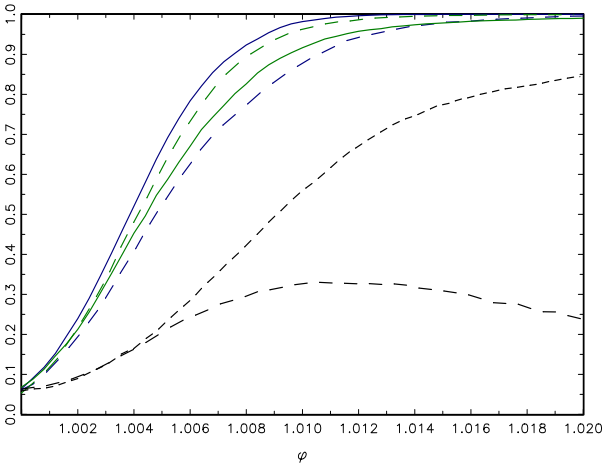


(e) $T^* = 100$, $m = 10$

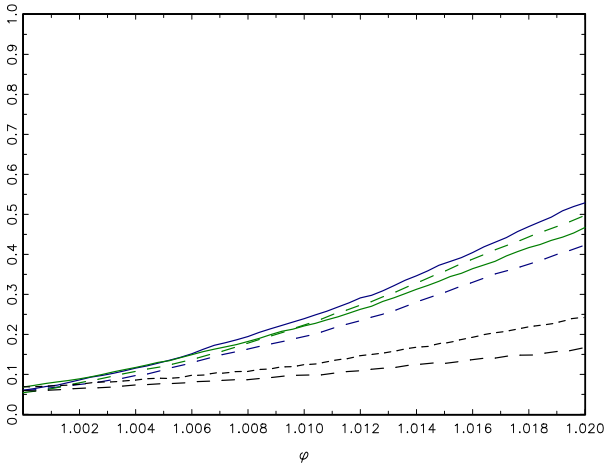


(f) $T^* = 200$, $m = 10$

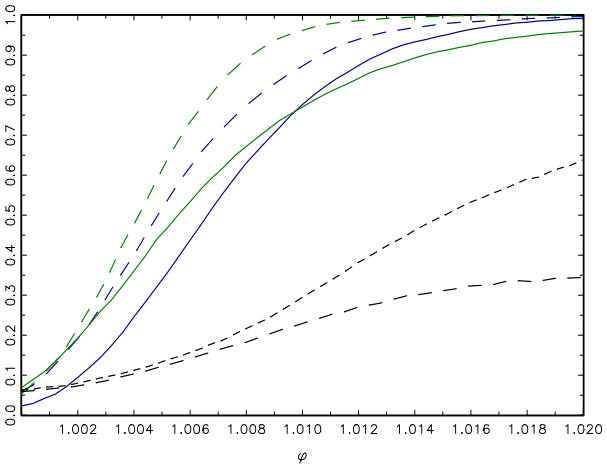
Figure 3. Finite sample power of nominal 0.05-level tests: i.i.d. innovations:
S_5: – –, S_{10}: ——, S_5^*: – –, S_{10}^*: ——, S_5^{*w}: – –, S_{10}^{*w}: ——, $BSADF$: ——, $BSADF_b^1$: – –, $BSADF_b^2$: - - -



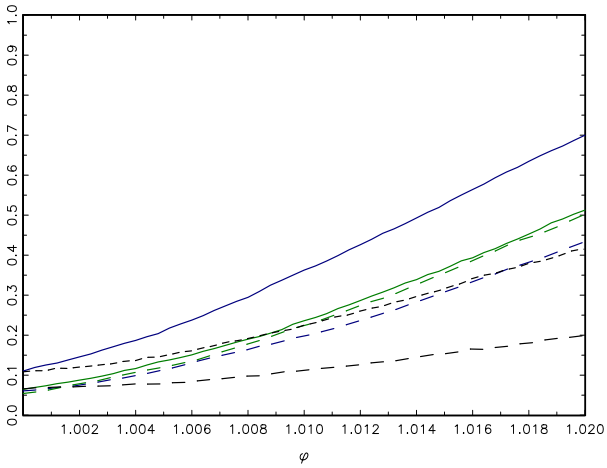
(a) Mid sample, $\sigma^2 = 1/5$



(b) Mid sample, $\sigma^2 = 5$



(c) Last 5 observations, $\sigma^2 = 1/5$



(d) Last 5 observations, $\sigma^2 = 5$

Figure 4. Finite sample power of nominal 0.05-level tests: shift in volatility, $T = 200$, $m = 5$: S_5^*: – –, S_{10}^*: ——, S_5^{*w}: – –, S_{10}^{*w}: ——, $BSADF_b^1$: – –, $BSADF_b^2$: - - -