Testing Parameter Stability: A Wild Bootstrap Approach

Gerard O'Reilly* and Karl Whelan†
Central Bank and Financial Services Authority of Ireland

December 20, 2005

Abstract

Unknown-breakpoint tests for possible structural change have become standard in recent years, with the most popular being the so-called Sup-F tests, whose asymptotic distribution was derived by Andrews (1993). We highlight two problems that lead to poor performance when testing for structural breaks in dynamic time series models using the Andrews critical values: High persistence of explanatory variables and heteroskedasticity. We propose a so-called "wild bootstrap" approach to generating critical values for the Sup-F statistic and report that this approach performs well across a wide variety of possible data generating processes, including those with large coefficients on lagged dependent variables and heteroskedasticity.

*E-mail: gerard.oreilly@centralbank.ie

†E-mail: karl.whelan@centralbank.ie. The views expressed in this paper are our own, and do not necessarily reflect the views of the Central Bank and Financial Services Authority of Ireland or the ESCB.

1 Introduction

The past decade has seen a substantial shift in the methodologies used to test the hypothesis of parameter stability in time series models. Where previously it was generally assumed that the researcher knew the date of the potential structural change, the modern approach accepts that in practice such knowledge is usually not available. Thus, the standard Chow (1960) test has now largely been replaced by tests such as Quandt's (1960) Sup-F test—based on the maximum of a sequence of Chow statistics—with the critical values coming from the asymptotic theory provided by Donald Andrews (1993a). However, despite the widespread use of procedures such as the Sup-F test, the asymptotic distributions derived by Andrews can, in some cases, provide poor approximations to the relevant finite-sample distributions.

In this paper, we focus on two problems that lead to poor performance when applying asymptotics-based Sup-F tests to dynamic time series models: High persistence of explanatory variables and heteroskedasticity. Specifically, we show that these problems lead to over-sized tests, so that Type I errors falsely indicating structural breaks occur too often. To address these problems, we propose a new bootstrap approach for generating critical values, and show that this approach leads to test procedures with approximately the correct size.

Our paper initially focuses on pure autoregressive models. These models account for a significant fraction of the applications of the unknown-breakpoint tests for structural change, and the problem due to high persistence has been noted before in this context by Frank Diebold and Celia Chen (1996).[1] Their paper reported that increases in the persistence of dependent variable resulted in tests based on asymptotic distributions becoming increasingly oversized when applied to standard sample sizes. In this paper, we confirm these results, but concentrate on the (more realistic) case in which the regression contains an intercept term. We show that in this case, the size distortions at high levels of persistence are extremely large, particularly for tests of a break in the intercept coefficient.

The second problem, due to heteroskedasticity, relates to results reported by Bruce Hansen (2000). Specifically, Hansen shows that structural change in the marginal distribution of a regressor causes a breakdown in the conditions underlying the derivation of

[1] For instance, in illustrating the unknown-breakpoint test procedures to a general economics audience, Hansen (2001) uses an AR(1) model of productivity growth. See also Stock and Watson (1999) for a paper that tests for structural change in AR models for a wide range of macro variables.

the standard asymptotic distributions for unknown-breakpoint tests. In the context of the autoregressive models discussed first in this paper, this problem can occur if there is heteroskedasticity, so that the distribution of the lagged dependent variable changes at some point in the sample.

Both of these previous papers suggested potential solutions to the separate problems with the asymptotics-based tests that they highlighted. Diebold and Chen suggest a "sieve bootstrap" method that simulates the estimated full-sample process for the dependent variable. Hansen suggests a "fixed regressor bootstrap" to adjust for the effect of having a regressor that exhibits structural change in its distribution. In this paper, we document that these methods are relatively successful in dealing with the separate problems they were designed to address. However, we find that the fixed regressor procedure performs poorly when one moves away from the case analyzed by Hansen with a relatively small ($\rho = 0.5$) lagged dependent variable effect. Even with moderate levels of persistence, the fixed regressor bootstrap produces significantly oversized tests, particularly for intercept breaks. We also find that the performance of the Diebold-Chen sieve bootstrap deteriorates when heteroskedasticity is present, with tests again becoming oversized.

In light of these results, and given that both heteroskedasticity and high lagged dependent variable coefficients are common features in empirical applications, we propose a new bootstrap approach to testing parameter stability which attempts to deal with both of these problems. Specifically, we propose a "wild bootstrap" approach, which uses the sieve bootstrap approach to simulating the estimated no-break process while also modelling the heteroskedasticity present in the data.[2] We find that this approach performs well across a wide variety of possible data generating processes, including those with either or both large coefficients on lagged dependent variables and heteroskedasticity.

We also show that problems stemming from persistent regressors also occur in models that incoporate additional explanatory variables: The more persistent such regressors are, the more likely one is to have over-sized tests when using asympototic or fixed-regessor bootstrap techniques. A simple modification to our wild bootstrap procedure, however, gives tests that are also well-sized in this case.

The contents of the paper are as follows. Section 2 sets out the modelling framework, the tests examined, and the design of the Monte Carlo experiments used to assess the ap-

[2]The wild bootstrap has been used recently in other applications unrelated to structural change tests. See, for example, Davidson and Flachaire (2001) and Godfrey and Tremayne (2003).

plication of these tests to univariate time series models. Section 3 reports our main results, while Section 4 discusses the case in which there are additional explanatory variables in addition to lagged dependent variables. Finally, Section 5 reports results from an application of the methodology. We show that the use of the wild bootstrap approach results in different conclusions to that of the asymptotic approach when assessing whether or not there have been changes over time in the aggregate inflation process.

2 Model, Tests, and Monte Carlo Design

2.1 Model and Test Statistics

Our initial focus is on tests of parameter stability in the AR(1) model

$$y_t = \alpha + \rho y_{t-1} + \epsilon_t, \tag{1}$$

estimated over the sample $t = 1, 2, ..., T$. In the case where there is a known date, k, for the potential breakpoint, a test of the null hypothesis of parameter stability can be formulated by imposing linear restrictions on an unrestricted regression model. For instance, if one is testing for a break in both the intercept term and the coefficient on the lagged dependent variable at date k, then one can estimate the regression model

$$y_t = \alpha + \rho y_{t-1} + \mu_0 d_k + \mu_1 \left(d_k y_{t-1}\right) + \epsilon_t, \tag{2}$$

where d_k is a dummy variable equalling zero before date k and one thereafter. The null hypothesis of parameter stability can then be formulated through the linear restrictions $H_0 : \mu_0 = \mu_1 = 0$. These linear restrictions can be written in matrix form as $R\beta = 0$, and can be tested with the Wald test statistic

$$W_k = \left(R\hat{\beta}\right)' \left(R\hat{\Omega}R'\right)^{-1} \left(R\hat{\beta}\right), \tag{3}$$

where

$$\Omega = \mathrm{Var}(\hat{\beta}). \tag{4}$$

This statistic has an asymptotic χ^2 distribution when implemented with a consistent estimator of the covariance matrix, Ω. In the case where ϵ_t is assumed to be iid, then the Wald statistic collapses to a multiple of the familiar F statistic based on a percentage difference in sums of squared errors between restricted and unrestricted models. When the errors are

not iid because of heteroskedasticity, then the Wald statistic will only have a χ^2 asymptotic distribution when one uses a heteroskedasticity-consistent covariance matrix estimator.

In the case in which the potential breakpoint is unknown, a popular test statistic, proposed by Quandt (1960), is based on the maximum of a sequence of Wald statistics:

$$SupW = \sup_{\pi} W_k \tag{5}$$

where the supremum is taken over $\pi = (\gamma T, (1-\gamma)T)$. In our calculations, we will follow the usual convention and set the trimming parameter γ equal to 0.15, and will report results for $SupW$ tests based on both the standard Wald statistic (using residual sums of squares) and a heteroskedasticy-consistent version using a White-corrected covariance matrix estimator.

2.2 Methods for Generating Critical Values

We will examine test procedures based on four different methods for generating critical values for the $SupW$ test statistic.

Andrews (1993) Asymptotic Distribution: In an important contribution, Donald Andrews derived the asymptotic distribution for $SupW$-style statistics. These statistics had previously been considered only as an informal diagnostic tool, but the critical values in the Andrews paper are now in widespread usage in formal tests of parameter stability. The critical values used depend on both the trimming parameter, γ, and the number of linear restrictions being tested. In our applications, in which the trimming parameter is 15 per cent, the 10 percent critical value for a test for a break in one parameter is 7.17, while the corresponding critical value for a test for a break in two parameters is 10.01.

Fixed Regressor Bootstrap: In the context of the generic regression model

$$y = \beta x + \epsilon_t, \tag{6}$$

Hansen (2000) has documented that the conditions underlying the derivation of the Andrews asymptotic distributions can break down when there is structural change in the marginal distribution of one or all of the regressors. For example, consider the case in which there is no change in the β coefficients, but there is a mean shift in one of the variables in the x matrix. Hansen shows that the use of asymptotic critical values in this case can lead to oversized $SupW$ tests, with too many Type-I errors incorrectly suggesting structural change

in the β coefficients. In the context of the pure autoregressive model examined here, such a mean change would have to imply a change in the parameters of the model, so this case does not apply here. However, changes over time in the *variance* of a regressor leads to similar problems, and this can apply in the AR(1) model if the residual term displays heteroskedasticity.

In the case where heteroskedastic errors are suspected, Hansen recommended a "fixed regressor bootstrap", based on the following procedure. A random sample $(u_t : t = 1, 2, ...T)$ of $N(0, 1)$ variables is simulated. This series is then scaled by a set of empirical residuals to construct an artificial dependent variable that maintains the pattern of heteroskedasticity seen in the data.[3] A *SupW* test statistic is then constructed based on the regression model relating $u_t\hat{\epsilon}_t$ to x. This process is repeated N times, to generate a bootstrap distribution for the *SupW* statistic, and the α-th percentile of this distribution is used as the $1 - \alpha$ percent critical value for this test procedure.

Though mainly intended as an approach to dealing with structural change in an exogenous x variable, Hansen reports calculations showing that this method can also work when there is a lagged dependent variable, which is also treated as a "fixed regressor" though these calculations are limited to a value of $\rho = 0.5$. And this method has been applied by a number of researchers in the context of AR(1) models.[4]

Sieve Bootstrap: Diebold and Chen (1996) document a separate problem with the Andrews critical values. In the context of the AR(1) model, they demonstrate that despite being asymptotically correct, tests based on asymptotic critical values become increasingly inaccurate in finite samples as the true value of ρ increases. Again, the problem is that the tests produce more Type I errors than their nominal size.

In place of the asymptotic critical values, Diebold and Chen recommend a so-called "sieve bootstrap" method which works as follows. The AR(1) model is estimated via OLS, and the residuals $\hat{\epsilon}_t$ are stored. Then, N different "pseudo-data" time series $(y_t^* : t = 1, 2, ...T)$ are generated in a manner consistent with the estimated (no-break) model:

$$y_t^* = \hat{\alpha} + \hat{\rho} y_{t-1}^* + u_t \tag{7}$$

where $\hat{\alpha}$ and $\hat{\rho}$ are OLS estimates and the u_t pseudo-disturbances are drawn randomly with

[3] We have followed Hansen's suggested approach here and used residuals based on a regression of y on x and the structual break dummies corresponding to the maximum *SupW* breakdate.

[4] See, for instance, Levin and Piger (2003) and Gadzinsky and Orlandi (2004).

replacement from the estimated residuals $\hat{\epsilon}_t$. For each of these N simulated series, a $SupW$ test statistic is calculated, and the α-th percentile of the resulting distribution is used as the $1-\alpha$ percent critical value for this test procedure.[5]

Wild Bootstrap: Bootstrap methods can only be expected to work well when they provide a good approximation to the underlying data generating process. In light of this, one potential weakness of the sieve bootstrap approach is that its approach to constructing the pseudo-disturbances u_t (drawing randomly from the estimated residuals) may provide a poor description of DGPs that exhibit heteroskedasticity. There are a number of possible ways to modify the bootstrap approach to deal with this problem. Here, we consider a version of the so-called "wild bootstrap" approach as discussed by Davidson and Flachaire (2001). Specifically, our version of the wild bootstrap follows the same steps as the sieve bootstrap, but differs in generating the pseudo-disturbances u_t according to the rule

$$u_t = \begin{cases} \hat{\epsilon}_t & \text{with probability 0.5} \\ -\hat{\epsilon}_t & \text{with probability 0.5} \end{cases} \tag{8}$$

For each of these N simulated series, a $SupW$ test statistic is calculated, and the α-th percentile of the resulting distribution is used as the $1-\alpha$ percent critical value.

2.3 Monte Carlo Design

To assess the performance of our various tests procedures, we performed a set of Monte Carlo simulations featuring different values of ρ both with and without heteroskedasticity. Specifically, we considered two different types of data generating process. In the first case, the true model is an AR(1) with spherical errors:

$$y_t = \alpha + \rho y_{t-1} + \epsilon_t, \qquad \epsilon_t \sim N(0,1). \tag{9}$$

In the second, there is a break in the variance half way through the sample. To design an example in which there is the potential for significant problems, we consider a three-fold increase in the standard deviation of the error distribution:

$$y_t = \alpha + \rho y_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \begin{cases} N(0,1) & t \leq \frac{T}{2} \\ 3N(0,1) & t > \frac{T}{2} \end{cases} \tag{10}$$

[5]This bootstrap methodology for generating critical values for $SupW$-style statistics was first suggested by Christiano (1992), but that paper did not present evidence on the size of tests based on this method.

For both models, we considered six different values of ρ, starting with $\rho = 0.5$ and moving up to $\rho = 0.99$.

We report results for $T = 100$. This is approximately an average sample size used in the quarterly regressions reported in empirical macroeconomic studies. For each experiment, the number of Monte Carlo replications is 5000, while for each of the bootstrap methods considered, the number of bootstrap replications is $N = 399$.[6]

3 Results

This section presents the results from the Monte Carlo experiments just described. In each case, we report the actual size obtained for test procedures intended to generate Type-I errors 10 percent of the time. In other words, the tables report the fraction of Monte Carlo replications in which the simulated test statistic based on the null hypothesis of no structural change exceeded the 10 percent critical value implied by the test procedure.[7]

3.1 Asymptotic Critical Values

We start with the results in Table 1, which describe the performance of tests for breaks in the intercept parameter, the ρ parameter, and in both coefficients, when one uses the asymptotic critical values derived by Andrews. The results for the baseline case with white noise errors expand on results previously reported by Diebold and Chen (1996). The detailed results presented in that study related to the case in which there was no intercept in the DGP or regressions, and the paper reported a set of relatively modest size distortions for a variety of cases. For the more realistic case in which an intercept is included, Diebold and Chen noted briefly that size distortions for asymptotic tests "can be extremely large" and this is confirmed in Table 1. Also, while Diebold and Chen reported some results only for the joint test for a simultaneous break in the intercept and ρ coefficients, we report results for all three cases.

The most striking result from the baseline case is the poor performance of tests for a break in the intercept coefficient. Size distortions increase rapidly as the true value of

[6] McKinnon (2002) has pointed out that the sampling errors associated with a low value of N tend to cancel out in Monte Carlo experiments such as this.

[7] The tables do not report standard errors for the estimated sizes of these tests, but using the formula $\sqrt{\frac{p(1-p)}{N}}$, one can calculate that these standard errors are small, ranging from 0.004 for tests that have size close to ten percent, to 0.007 for some of the more inaccurate tests.

ρ increases beyond 0.5, and for high values of ρ the tests falsely indicates breaks in the intercept more than half the time. These results show that the well-known difficulty of distinguishing between a series that has a mean break and a series that is persistent and thus goes through long swings away from the sample mean, applies to this method. This problem for empirical break testing is well known and dates back at least as far as the debate surrounding the results in Perron (1989). Perron argued that evidence for a unit root in GNP was undermined by including dummies for the Great Crash and the oil price shock of the 1970s. Christiano (1992) and others critiqued Perron's results as being based on "pre-test" selection of breakdates, and advocated test procedures based on the maximum of a sequence of break tests. In theory, the Andrews distribution is also supposed to allow for appropriate inference in this case, but our calculations show that, in practice, this is not the case. Indeed, even if one uses the $SupW$ test and the Andrews distribution, it turns out that it is still easy to confuse a persistent time series with no breaks for one that is less persistent but has a break in the intercept.

The test for a break in both coefficients performs nearly as poorly as the intercept break test. However, the test for a break in the ρ coefficient, does not do quite so badly. Unsurprisingly, given that this DGP has spherical errors, the version of the tests using a heteroskedasticity-consistent covariance matrix performs even more poorly than the basic test statistic, sometimes spectacularly so: For $\rho = 0.99$, this test for a break in both coefficients suggests a break over 80 percent of the time.

The results for the variance break DGP confirm Bruce Hansen's (2000) theoretical results that the Andrews asymptotic distribution does not apply when there is a change in the marginal distribution of a dependent variable. Even for the case $\rho = 0.5$ when the asymptotic tests do well if there are spherical errors, the introduction of heteroskedasticity substantially increases the size of the tests. Indeed, heteroskedasticity worsens the performance of the basic version of the test for all values of ρ, with sizes for DGPs exhibiting both high persistence and heteroskedasticity moving into the 70 to 80 percent range. Perhaps surprisingly, the version of the test that uses a robust covariance matrix does not perform better when there is a variance break, and in most cases performs worse, than the tests based on the basic statistic.

3.2 Fixed Regressor Bootstrap

One approach to dealing with the problems caused by heteroskedasticity is to adopt Bruce Hansen's (2000) fixed regressor bootstrap, which has been designed to deal with the problems caused by changes in the marginal distribution of a regressor. Table 2 reports the results for this method. Comparing the right-hand entries of the first line of the table with the corresponding entries on Table 1, we see that the fixed regressor bootstrap, as expected produces better-sized tests for the case with $\rho = 0.5$ and a break in the variance. For instance, the size of the intercept break tests falls from 0.22 to 0.15 and the size of the ρ break test falls from 0.18 to 0.14.

However, once one moves beyond the $\rho = 0.5$ case to higher values of ρ, the performance of this method worsens significantly. This is true for each of the cases examined here, with the worsening performance of this method roughly tracking that of the tests based on asymptotic critical values. While the test sizes reported here are not quite as high as those for the asymptotic method, they are still often very high: For instance, the test for an intercept break with spherical errors and $\rho = 0.99$ has a size of 0.575. The performance of the heteroskedasticity-consistent version of the test mirrors that of corresponding tests using asymptotic critical values. This test again performs worse than the test statistic based on residual sums of squares, even for the case in which there is a variance break.

3.3 Sieve Bootstrap

The results just reported for the fixed regressor bootstrap support and generalize findings from a recent paper by Todd Clark (2003). Clark finds that the fixed regressor bootstrap produces oversized tests for DGPs based on empirical estimates of inflation processes for various sub-categories of the US CPI. He then recommends use of a sieve bootstrap technique as previously discussed by Diebold and Chen. Table 3 reports the results from our Monte Carlo examination of this method.

The results for the baseline DGP featuring spherical errors essentially confirm the previous conclusions of Diebold and Chen that this methodology works reasonably well across a wide set of values of ρ. Test sizes remain in the 10-13 percent region for values of ρ up to 0.95. The size bias jumps somewhat at very high levels of persistence, but is still well below what was reported in the previous cases. Also unlike the previous cases, the heteroskedasticity-consistent version of the test performs about as well as the basic tests statistic despite the fact that the underlying errors are spherical.

The results for the variance break DGP are less positive. The test based on residual sums of squares exhibits substantial positive size biases for all values of ρ. The heteroskedasticity-consistent version of the test does somewhat better, but still displays quite large biases at high levels of persistence. For instance with $\rho = 0.95$, the sieve bootstrap test for an intercept break has a size of 19 percent.

3.4 Wild Bootstrap

The results in Table 3 show that the sieve bootstrap may not be the most appropriate method for generating critical values when the underlying data generating process exhibits heteroskedasticity. The most likely reason for this is that sieve bootstrap's method for constructing the "pseudo-data" samples, and thus the critical values, relies on drawing randomly from the sample of empirical residuals, $\hat{\epsilon}_t$. As noted in Section 2.2, the wild bootstrap is an alternative methodology that can generate "pseudo-data" samples consistent with the null hypothesis of no structural change, while preserving the pattern of heteroskedasticity seen in the estimated residuals. Thus, this method might be expected to perform somewhat better than the sieve bootstrap in the presence of heteroskedasticity.

Table 4 confirms this conjecture: For the variance break DGP, the wild bootstrap method produces smaller size biases for all of the cases that we considered. These bias reductions are substantial for the basic test statistic, and somewhat smaller for the heteroskedasticity-consistent version of the test. The results for the baseline DGP with white noise errors show that, in this case, there is no reduction in the efficiency when using the wild bootstrap rather than the sieve. Thus, the improved performance of the wild bootstrap in the presence of heteroskedasticity does not come at the cost of a poor performance when errors are white noise.

3.5 A Bias-Adjusted Wild Bootstrap

Although the wild bootstrap produced the best results of the four method compared thus far, with size distortions being very low in most cases, its performance does deteriorate for processes with very high values of ρ. While size biases are negligible for values of ρ less than 0.9, they increase noticeably above this point.

One potential reason for this deterioration in the performance is our use of OLS parameter estimates to generate the simulated "pseudo-data" upon which the critical values are based. It is well-known that OLS estimates of ρ are downward biased in finite samples,

and that this bias becomes larger as ρ approaches one. For this reason, our simulated processes may not mimic the underlying DGP as we would wish. This suggests using a bias-adjusted form of the wild bootstrap. This can be done as follows. First, calculate the OLS estimates $\hat{\rho}$. Second, calculate a median-unbiased estimate of ρ consistent with $\hat{\rho}$ and with the sample size used. Call this $\hat{\rho}^u$. Finally, simulate N different pseudo-data time series $(y_t^* : t = 1, 2, ...T)$ consistent with the model:

$$y_t^* = \hat{\alpha} + \hat{\rho}^u y_{t-1}^* + u_t \tag{11}$$

choosing the u_t in the same manner as in the wild bootstrap method (using equation 8).

Table 5 reports the results obtained from Monte Carlo examination of this method with our two DGPs. We note that the bias-adjustment calculated in these Monte Carlo replications was of a simple variety, based on the tables for finite sample biases reported by Andrews (1993b). However, in practical applications, we recommend using a more sophisticated methodology such as Bruce Hansen's (1999) grid bootstrap method.

The results for this bias-adjusted wild bootstrap are very encouraging. The size biases are, in almost all cases, the smallest obtained for all of the methods examined here, with sizes ranging from about 0.09 to about 0.11 for all values of ρ less than 0.99. The test sizes for the $\rho = 0.99$ case are a little unsatisfactory, (equalling 0.179 for the variance break case), but they are still an improvement on all of the other methods.

3.6 Additional Calculations

In addition to the calculations reported in the tables, we also performed some Monte Carlo exercises for other values of the sample size T, and for other popular test statistics. These findings provided further support for our overall assessment of the various test procedures:

- **Sample Size**: As would be expected, the asymptotics-based tests perform somewhat better than reported here for larger samples, and somewhat poorer for smaller samples. However, for all realistic sample lengths, the empirical sizes for these tests increased substantially for high values of ρ. Similar results also applied for the fixed regressor bootstrap, while the performance of the wild bootstrap remained good across a range of sample sizes.

- **Other Test Statistics**: Andrews and Ploberger (1994) developed alternative procedures to the Sup-W statistic that, under certain conditions, can lead to more powerful

tests. Monte Carlo calculations with these Ave-W and Exp-W statistics revealed finite sample distributions with very similar properties to those reported here for the Sup-W.

4 Allowing for Additional Explanatory Variables

While univariate time series models are commonly used for various applications, it is more common for time series models to mix dynamic lagged dependent variable terms with terms describing the effect of additional explanatory variables. In this section, we assess the performance of the various methods when extended to this case.

Again, we base our assessment on Monte Carlo exercises featuring DGPs with varying levels of persistence for the dependent variable. However, in this case, we also consider the effect of varying the level of persistence of an additional exogenous regressor. The DGP that we consider takes the form

$$y_t = \alpha_y + \rho_y y_{t-1} + \beta x_t + \epsilon_t \tag{12}$$

$$x_t = \alpha_x + \rho_x x_{t-1} + \eta_t \tag{13}$$

where both error terms are iid random variables, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$, $\eta_t \sim N(0, \sigma_\eta^2)$. In relation to the tests, the approach that we take to implementing the sieve and wild bootstraps is to treat the x_t variable as a fixed regressor, taking it as given in each of the bootstrap replications.

In the case of the univariate DGPs examined earlier, the only parameter relevant for determining the finite-sample distribution of the tests was the parameter ρ, and thus we were able to provide a comprehensive reporting of the properties of the tests across a wide range of relevant cases. In contrast, in this case, simulating this DGP requires the choice of the parameters such as ρ_y, ρ_x, β, σ_x and σ_y, and experimentation with various values for these parameters have revealed different test sizes for the approaches adopted here.

Our approach has been to calibrate these parameters to an explicit empirical example that we will discuss in the next section. Specifically, we calibrate α_x, α_y, β, σ_y and σ_x based on values obtained for an exercise that treats the Euro-area inflation process as the y_t variable and the Euro-area output gap as the x_t variable. We then consider the performance of the tests for two fixed values of ρ_y (0.80 in Table 6 and 0.95 in Table 7), and for a range of values for ρ_x going from 0.50 to 0.95.

Tables 6 and 7 show that our general assessment of the asymptotic and fixed-regressor methods extends to cases involving additional regressors. The sizes for these tests are commonly above their intended ten percent level, with tests for breaks in α_y and ρ_y having higher sizes in the case $\rho_y = 0.95$ than in the case $\rho_y = 0.80$. The performance of tests for breaks in β, the coefficient on the additional explanatory variable, get worse as the persistence of this variable increases. In contrast, the sieve and wild bootstrap methods both produce tests that are close to their intended size.

The poor performance of the fixed regressor bootstrap in these simulations stands in contrast to Monte Carlo results reported in Hansen (2000) which suggest that the method tends to produce tests with close to the appropriate size. For instance, Table 1 of his paper reports a size of exactly ten percent for an example like this one in which the errors are iid, and the model contains both a lagged dependent variable and an independent x variable. The reason for the difference in our assessment of this procedure stems from the differences in the underlying DGPs used to generate the size statistics. In terms of our example, Hansens's calculations correspond to a DGP with $\rho_y = 0.5$ and $\rho_x = 0.0$. However, our calculations show once one moves to cases in which the dependent and independent variables have more persistence, the performance of this procedure deteriorates.

It is worth emphasizing that the specific numbers reported in Tables 6 and 7 would change if we make different assumptions about the various parameters of the DGP. However, calculations not reported here show that our general assessment of the wild and sieve bootstrap versus asymptotic and fixed-regressor methods are robust to these choices.

5 An Application: Instability in the Inflation Process?

Tables 8 and 9 present some examples of the application of the methods developed in this paper. Specifically, we examine the question of the stability of the inflation processes for both the US and the Euro area. This example is not chosen at random, but rather to provide a good illustration of the types of application that our tests are most likely to be useful for. In particular, macroeconomic time series are likely to be a particularly useful area for the methods developed here. This is for two reasons. First, many macroeconomic time series are highly persistent, with the hypothesis of a unit root often very difficult to reject.[8] Second, following the initial work of McConnell and Peres Quiros (2000), there

[8] See, for instance, Stock (1991).

is now a substantial literature documenting the reduced volatility of macroeconomic time series around the world. Thus, heteroskedasticity seems to be a very common feature of macroeconomic time series.

Macroeconomic series also provide good examples of the importance of structural change tests. Since Lucas (1976), it is well known that the parameters of reduced-form econometric equations for macroeconomic variables can be expected to change over time as policy regimes evolve. Indeed, there has been considerable debate in recent years about whether the parameters of inflation processes have changed due to the shift since the early 1980s towards more aggressive anti-inflationary policies by central banks; this theme has been emphasised by John Taylor (1998), Thomas Sargent (1999) and others.

Table 8 reports the results from applying each of our tests to an AR(4) process for Euro-area and US inflation, as measured by the log-difference in the GDP deflator.[9] White tests suggest that the US regression certainly has heteroskedasticity with a very significant p-value; for the Euro area the p-value for the White (1980) test is 0.15, pointing to a strong likelihood of underlying heteroskedasticity.

For both the US and the Euro area, our preferred tests give substantially different answers to the asyptotic and fixed regressor methods when testing for a break in the intercept term (in other words, testing for a break in the mean of the inflation process). For the Euro area, the asymptotic test using the standard Wald statistic reports a p-value of 0.013 for a break in the intercept term. Similarly, the fixed regressor bootstrap points to a break that is significant at the 10 percent level. In contrast, the sieve, wild, and bias-adjusted wild bootstrap all point against the idea of a statistically significant break. Given the evidence of potential heteroskedasticity in this series, the version of the test based on the robust covariance matrix may be more relevant. In this case, the asymptotic test suggests the likelihood of a break in the ρ coefficient, and also perhaps a joint break in both the intercept and persistence parameters. In contrast, the sieve, wild, and bias-adjusted wild bootstrap tests all indicate that there has not been a significant break. Table 9 shows that these results are generally robust to adding an output gap (defined by HP-filtering real GDP) to the specification.

For the US, the results show that the evidence for a break in the intercept is also

[9]The US data were downloaded from the Bureau of Economic Analysis website, www.bea.gov, while the Euro-area data come from the Area-Wide Model database, as documented in Fagan, Henry, and Mestre (2001).

considerably weaker when one uses our preferred procedures. Perhaps more interesting, though, are the results for the tests for a break in the ρ parameter. One potential concern about the bootstrap tests suggested here is that these procedures may only have obtained correctly-sized tests at the expense of having very poor power. In other words, it may be that even when there is a break in the inflation process, our preferred tests may not be able to detect it. The tests for a break in the ρ coefficient in the US inflation process suggest, however, that these procedures *are* capable of detecting real breaks.

Table 8 shows that the wild bootstrap tests based on the robust covariance matrix produce a p-value of 0.009 for the null hypothesis of no break in ρ for the US. Again, the general nature of the results are not changed by re-doing the results for a model incoporating the output gap. In this case, the robust version of the wild bootstrap test gives a p-value of 0.018. The maximum Sup-W statistic for this test occurs at 1981:Q2, and the point estimates of the pre- and post-break values of ρ are 0.95 and 0.77. While not insignificant, these results show that coefficient changes do not have to be extremely large for our bootstrap procedures to be able to detect them.

6 Conclusion

This paper has made three principal contributions.

First, we provided more extensive documentation than previous papers of how persistent regressors and heteroskedasticity affect the performance of tests for structural change in time series models based on the asymptotic distributions documented in Andrews (1993). In particular, we document how these features have different effects on tests for breaks in the intercept, breaks in the lagged dependent variable parameter, and joint breaks in both of these parameters.

Second, we described some of the limitations of the fixed regressor bootstrap methodology introduced by Bruce Hansen (2000). Specifically, we show that this procedure results in substantially over-sized tests when the dependent variable displays moderate or high levels of persistence.

Finally, we introduced an alternative "wild bootstrap" procedure for generating critical values for structural change tests for time series models. We show that this procedure gives tests that have approximately the correct size even when there are persistent regressors and heteroskedasticity. Because high persistence and heteroskedasticity are common features of

time series studied in fields such as macroeconomics and finance, we hope that our proposed method will prove useful for a large number of future applications.

References

[1] Andrews, Donald (1993a). "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821-856.

[2] Andrews, Donald (1993b). "Exactly Median-Unbiased Estimation of First-Order Autoregressive/Unit Root Models," *Econometrica*, 61, 139-165.

[3] Andrews, Donald and Werner Ploberger (1994). "Optimal Tests When a Nuisance Parameter is Present Only Under the Alternative," *Econometrica*, 61, 1383-1414.

[4] Chow, Gregory (1960). "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28, 591-605.

[5] Christiano, Lawrence (1992). "Searching for a Break in GNP," *Journal of Business and Economic Statistics*, 10, 237-249.

[6] Clark, Todd (2003). "Disaggregate Evidence on the Persistence of Consumer Price Inflation," *Journal of Applied Econometrics*, forthcoming.

[7] Cogley, Timothy and Thomas Sargent (2001). "Evolving Post-Word War II Inflation Dynamics" in *NBER Macroeconomics Annual*, Vol. 16.

[8] Davidson, Russell and Emmanuel Flachaire (2001). The Wild Bootstrap, Tamed at Last, working paper, McGill University.

[9] Diebold, Frank and Celia Chen (1996). "Testing Structural Stability With Endogeneous Break Point: A Size Comparison of Analytic and Bootstrap Procedures," *Journal of Econometrics*, 70, 221-241.

[10] Fagan, Gabriel, Jerome Henry and Ricardo Mestre (2001). An Area-Wide Model (AWM) for the Euro Area, ECB Working Paper No. 42.

[11] Gadzinsky, Gregory and Fabrice Orlandi (2004). "Inflation Persistence in the European Union, the Euro Area, and the United States," ECB Working Paper No. 414.

[12] Godfrey, Leslie and Andy Tremayne (2003). Using the Wild Bootstrap to Implement Heteroskedasticity-Robust Tests for Serial Correlation in Dynamic Regression Models, working paper, University of York.

[13] Hansen, Bruce (1999). "The Grid Bootstrap and the Autoregressive Model," *Review of Economics and Statistics*, 81, 594-607.

[14] Hansen, Bruce (2000). "Testing for Structural Change in Conditional Models," *Journal of Econometrics*, 97, 93-115.

[15] Hansen, Bruce (2001). "The New Econometrics of Structural Change: Dating Breaks in US Labor Productivity," *Journal of Economic Perspectives*, 15(4), 117-128.

[16] Levin, Andrew and Jeremy Piger (2003). Is Inflation Persistence Intrinsic in Industrial Economies?, working paper, Federal Reserve Bank of St. Louis.

[17] Lucas, Robert (1976). "Econometric Policy Evaluation: A Critique," *Carnegie-Rochester Series on Public Policy*, 1, 19-46.

[18] McConnell, Margaret and Gabriel Peres Quiroz (2000). "Output Fluctuations in the United States: What has Changed Since the Early 1980s?" *American Economic Review*, 90, 1464-1476.

[19] McKinnon, James (2002). "Bootstrap Inference in Econometrics," *Canadian Journal of Economics*, 35, 615-645.

[20] Perron, Pierre (1989). "The Great Crash, the Oil-Price Shock, and the Unit-Root Hypothesis," *Econometrica*, 57, 1361-1401.

[21] Pivetta, Frederic, and Ricardo Reis (2003). "The Persistence of Inflation in the United States," mimeo, Princeton University.

[22] Quandt, Richard (1960). "Tests of the Hypothesis that a Linear Regression Obeys Two Separate Regimes," *Journal of the American Statistical Association*, 55, 324-330.

[23] Sargent, Thomas (1999). *The Conquest of American Inflation*, Princeton: Princeton University Press.

[24] Stock, James (1991). "Confidence Intervals for the Largest Autoregressive Root in US Macroeconomic Time Series," *Journal of Monetary Economics*, 28, 435-459.

[25] Stock, James and Mark Watson (1999). Business Cycle Fluctuations in U.S. Macroeconomic Time Series, in *Handbook of Macroeconomics*, edited by John Taylor and Michael Woodford, North Holland, Vol. 1a, pages 3-64.

[26] Taylor, John B. (1998). "Monetary Policy Guidelines for Unemployment and Inflation Stability," in John Taylor and Robert Solow (eds.) *Inflation, Unemployment, and Monetary Policy*, Cambridge: MIT Press.

[27] White, Halbert (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

Table 1: Size of 10% Test ($T = 100$): Andrews (1993a) Asymptotics

	Baseline			*Variance Break*		
	Intercept	ρ	All	Intercept	ρ	All
$\rho = 0.50$	0.129	0.109	0.111	0.222	0.179	0.268
$\rho = 0.70$	0.171	0.125	0.145	0.262	0.198	0.335
$\rho = 0.80$	0.204	0.143	0.189	0.297	0.224	0.402
$\rho = 0.90$	0.320	0.173	0.301	0.443	0.260	0.562
$\rho = 0.95$	0.434	0.204	0.414	0.571	0.297	0.667
$\rho = 0.99$	0.578	0.361	0.546	0.714	0.398	0.786
	Heteroskedasticity Robust Version of Test					
$\rho = 0.50$	0.165	0.185	0.296	0.165	0.177	0.306
$\rho = 0.70$	0.210	0.233	0.375	0.201	0.207	0.378
$\rho = 0.80$	0.249	0.256	0.456	0.237	0.228	0.457
$\rho = 0.90$	0.376	0.316	0.606	0.373	0.283	0.607
$\rho = 0.95$	0.499	0.344	0.719	0.508	0.326	0.700
$\rho = 0.99$	0.649	0.480	0.804	0.665	0.436	0.801

Note: Results relate to the fraction of Type I errors recorded in 5000 Monte Carlo replications of two DGPs: The baseline DGP of equation (9) and the variance break DGP of equation (10).

Table 2: Size of 10% Test ($T = 100$): Fixed Regressor Bootstrap

	Baseline			*Variance Break*		
	Intercept	ρ	All	Intercept	ρ	All
$\rho = 0.50$	0.163	0.155	0.155	0.145	0.142	0.135
$\rho = 0.70$	0.191	0.172	0.177	0.196	0.173	0.177
$\rho = 0.80$	0.204	0.149	0.193	0.234	0.182	0.218
$\rho = 0.90$	0.340	0.210	0.287	0.353	0.216	0.308
$\rho = 0.95$	0.462	0.236	0.388	0.459	0.237	0.392
$rho = 0.99$	0.575	0.364	0.485	0.573	0.366	0.480
	Heteroskedasticity Robust Version of Test					
$\rho = 0.50$	0.185	0.194	0.258	0.168	0.191	0.239
$\rho = 0.70$	0.215	0.230	0.326	0.218	0.236	0.320
$\rho = 0.80$	0.253	0.266	0.459	0.262	0.256	0.392
$\rho = 0.90$	0.375	0.290	0.520	0.387	0.303	0.534
$\rho = 0.95$	0.502	0.326	0.627	0.510	0.328	0.627
$rho = 0.99$	0.630	0.452	0.706	0.628	0.451	0.712

Note: Results relate to the fraction of Type I errors recorded in 5000 Monte Carlo replications of two DGPs: The baseline DGP of equation (9) and the variance break DGP of equation (10).

Table 3: Size of 10% Test ($T = 100$): Sieve Bootstrap

	Baseline			*Variance Break*		
	Intercept	ρ	All	Intercept	ρ	All
$\rho = 0.50$	0.104	0.102	0.104	0.194	0.223	0.267
$\rho = 0.70$	0.106	0.101	0.100	0.192	0.224	0.279
$\rho = 0.80$	0.104	0.105	0.106	0.189	0.228	0.287
$\rho = 0.90$	0.114	0.099	0.114	0.234	0.232	0.338
$\rho = 0.95$	0.128	0.104	0.128	0.265	0.221	0.357
$\rho = 0.99$	0.163	0.129	0.155	0.332	0.222	0.406
	Heteroskedasticity Robust Version of Test					
$\rho = 0.50$	0.106	0.105	0.103	0.117	0.117	0.146
$\rho = 0.70$	0.106	0.108	0.106	0.122	0.134	0.156
$\rho = 0.80$	0.106	0.111	0.112	0.126	0.129	0.158
$\rho = 0.90$	0.116	0.107	0.120	0.161	0.132	0.180
$\rho = 0.95$	0.127	0.113	0.129	0.190	0.133	0.176
$\rho = 0.99$	0.167	0.147	0.154	0.246	0.151	0.197

Note: Results relate to the fraction of Type I errors recorded in 5000 Monte Carlo replications of two DGPs: The baseline DGP of equation (9) and the variance break DGP of equation (10).

Table 4: Size of the 10% Test ($T = 100$): Wild Bootstrap

	Baseline			*Variance Break*		
	Intercept	ρ	All	Intercept	ρ	All
$\rho = 0.50$	0.103	0.103	0.107	0.113	0.107	0.120
$\rho = 0.70$	0.107	0.101	0.101	0.109	0.112	0.117
$\rho = 0.80$	0.102	0.102	0.105	0.103	0.113	0.118
$\rho = 0.90$	0.117	0.104	0.115	0.130	0.110	0.137
$\rho = 0.95$	0.128	0.105	0.128	0.155	0.118	0.142
$\rho = 0.99$	0.166	0.126	0.155	0.213	0.137	0.173
	Heteroskedasticity Robust Version of Test					
$\rho = 0.50$	0.107	0.106	0.110	0.110	0.108	0.112
$\rho = 0.70$	0.107	0.108	0.114	0.104	0.110	0.118
$\rho = 0.80$	0.103	0.115	0.117	0.102	0.104	0.118
$\rho = 0.90$	0.118	0.105	0.124	0.122	0.107	0.131
$\rho = 0.95$	0.130	0.114	0.134	0.137	0.109	0.142
$\rho = 0.99$	0.170	0.149	0.160	0.197	0.126	0.161

Note: Results relate to the fraction of Type I errors recorded in 5000 Monte Carlo replications of two DGPs: The baseline DGP of equation (9) and the variance break DGP of equation (10).

Table 5: Size of the 10% Test (T=100): Bias-Adjusted Wild Bootstrap

	Baseline			*Variance Break*		
	Intercept	ρ	All	Intercept	ρ	All
$\rho = 0.50$	0.096	0.096	0.101	0.108	0.104	0.110
$\rho = 0.70$	0.109	0.102	0.108	0.109	0.109	0.108
$\rho = 0.80$	0.098	0.095	0.101	0.105	0.102	0.106
$\rho = 0.90$	0.095	0.085	0.101	0.107	0.096	0.110
$\rho = 0.95$	0.105	0.075	0.103	0.125	0.101	0.110
$\rho = 0.99$	0.153	0.112	0.137	0.201	0.135	0.145
	Heteroskedasticity Robust Version of Test					
$\rho = 0.50$	0.101	0.102	0.108	0.104	0.105	0.104
$\rho = 0.70$	0.112	0.106	0.110	0.107	0.105	0.106
$\rho = 0.80$	0.099	0.108	0.111	0.101	0.098	0.101
$\rho = 0.90$	0.094	0.101	0.107	0.103	0.097	0.113
$\rho = 0.95$	0.104	0.103	0.119	0.118	0.095	0.112
$\rho = 0.99$	0.153	0.149	0.132	0.179	0.129	0.138

Note: Results relate to the fraction of Type I errors recorded from 5000 Monte Carlo replications of the two DGP's, Baseline DGP in equation (1) and the Variance break DGP in equation (2).

Table 6: Baseline simulation with $\rho_y = 0.8$. 1000 Monte Carlo simulations with 399 bootstraps

	Asymptotics			Fixed Regressor Bootstrap			Sieve Bootstrap			Wild Bootstrap		
	α_y	ρ_y	β	α_y	ρ_y	β	α_y	ρ_y	β	α_y	ρ_y	β
$rho_\pi = 0.50$	0.19	0.12	0.09	0.24	0.16	0.14	0.09	0.11	0.09	0.09	0.10	0.10
$rho_\pi = 0.70$	0.22	0.10	0.11	0.22	0.15	0.17	0.12	0.09	0.09	0.12	0.09	0.10
$rho_\pi = 0.80$	0.21	0.13	0.12	0.24	0.18	0.15	0.10	0.10	0.09	0.11	0.10	0.09
$rho_\pi = 0.90$	0.20	0.14	0.17	0.25	0.19	0.19	0.11	0.08	0.11	0.11	0.09	0.11
$rho_\pi = 0.95$	0.22	0.17	0.20	0.22	0.20	0.20	0.11	0.10	0.12	0.11	0.10	0.12

Note: The underlying DGP is given by

$$\begin{aligned} y_t &= \alpha_y + \rho_y y_{t-1} + \beta x_t + \epsilon_t \\ x_t &= \alpha_x + \rho_x x_{t-1} + \eta_t \\ \epsilon_t \sim N(0, \sigma_\epsilon^2) &\qquad \eta_t \sim N(0, \sigma_\eta^2) \end{aligned}$$

with parameter values calibrated based on Euro area inflation and these take the following values $\alpha_x = 0.1$, $\alpha_y = 0.15$, $\beta = 0.5$ and variances $\sigma_\eta^2 = 1.8$, $\sigma_\epsilon^2 = 0.24$.

Table 7: Baseline simulation with $\rho_y = 0.95$. 1000 Monte Carlo simulations with 399 bootstraps

	Asymptotics			Fixed Regressor Bootstrap			Sieve Bootstrap			Wild Bootstrap		
	α_y	ρ_y	β	α_y	ρ_y	β	α_y	ρ_y	β	α_y	ρ_y	β
$rho_\pi = 0.50$	0.40	0.27	0.14	0.43	0.28	0.15	0.13	0.13	0.14	0.13	0.12	0.14
$rho_\pi = 0.70$	0.41	0.27	0.12	0.39	0.26	0.17	0.11	0.11	0.09	0.13	0.11	0.09
$rho_\pi = 0.80$	0.35	0.28	0.15	0.38	0.31	0.19	0.13	0.11	0.11	0.12	0.12	0.11
$rho_\pi = 0.90$	0.34	0.29	0.18	0.30	0.27	0.21	0.11	0.11	0.11	0.12	0.12	0.11
$rho_\pi = 0.95$	0.31	0.27	0.22	0.28	0.27	0.23	0.11	0.11	0.11	0.11	0.11	0.11

Note: The underlying DGP is given by

$$\begin{aligned} y_t &= \alpha_y + \rho_y y_{t-1} + \beta x_t + \epsilon_t \\ x_t &= \alpha_x + \rho_x x_{t-1} + \eta_t \\ \epsilon_t \sim N(0, \sigma_\epsilon^2) &\qquad \eta_t \sim N(0, \sigma_\eta^2) \end{aligned}$$

with parameter values calibrated based on Euro area inflation and these take the following values $\alpha_x = 0.1$, $\alpha_y = 0.15$, $\beta = 0.5$ and variances $\sigma_\eta^2 = 1.8$, $\sigma_\epsilon^2 = 0.24$.

Table 8: P-values for *SupW* Tests for Euro Area and US inflation

	Euro Area			US		
	Intercept	ρ	All	Intercept	ρ	All
Andrews Asymptotics	0.013	0.094	0.032	0.056	0.000	0.000
Fixed Regressor Bootstrap	0.070	0.195	0.113	0.030	0.014	0.022
Sieve Bootstrap	0.186	0.327	0.303	0.062	0.001	0.002
Wild Boostrap	0.325	0.426	0.434	0.261	0.014	0.030
Bias Adjusted Wild Bootstrap	0.272	0.476	0.390	0.535	0.029	0.074
	Heteroskedasticity Robust Version of Test					
Andrews Asymptotics	0.044	0.014	0.037	0.024	0.000	0.000
Fixed Regressor Bootstrap	0.068	0.026	0.074	0.025	0.000	0.001
Sieve Bootstrap	0.364	0.201	0.434	0.208	0.005	0.023
Wild Boostrap	0.304	0.190	0.380	0.203	0.009	0.024
Bias Adjusted Wild Bootstrap	0.266	0.138	0.326	0.338	0.009	0.043

Notes: Results refer to AR(4) regressions for GDP price inflation. Euro Area sample is 1971:2 to 2003:4, US sample period is 1960:1 to 2004:2. Bootstrap tests use $N = 5000$ replications.

Table 9: P Values for SupW Tests for Euro Area and US inflation when include an output gap

	Euro Area			US		
	α	ρ	β	α	ρ	β
Asymptotics Asymptotics	0.030	0.085	0.007	0.071	0.000	0.697
Fixed Regressor Bootstrap	0.088	0.173	0.010	0.044	0.023	0.647
Sieve Bootstrap	0.175	0.216	0.006	0.320	0.007	0.703
Wild Bootstrap	0.307	0.316	0.007	0.301	0.027	0.735
Bias Adjusted Bootstrap	0.207	0.315	0.008	0.584	0.047	0.723
	Heteroskedasticity Robust Version of Test					
Asymptotics Asymptotics	0.045	0.030	0.007	0.035	0.000	0.658
Fixed Regressor Bootstrap	0.073	0.059	0.004	0.042	0.001	0.771
Sieve Bootstrap	0.236	0.207	0.013	0.235	0.015	0.750
Wild Bootstrap	0.184	0.181	0.041	0.225	0.016	0.820
Bias Adjusted Bootstrap	0.166	0.143	0.015	0.336	0.018	0.824

Notes: Results refer to following regression

$$\pi_t = \alpha + \rho\pi_{t-1} + \sum_{i=1}^{3} \Delta\gamma_i \pi_{t-i} + \beta gap_t$$

for GDP price inflatio. Euro Area sample is 1971:2to 2003:4, US sample period is 1960:1 to 2004:2. Bootstrap tests use N = 5000 replications.