ELSEVIER

Journal of Econometrics 70 (1996) 221–241

JOURNAL OF
Econometrics

Testing structural stability with endogenous breakpoint A size comparison of analytic and bootstrap procedures

Francis X. Diebold*[a,b], Celia Chen[a]

[a]*Department of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297, USA*
[b]*NBER, Cambridge, MA 02138, USA*

Abstract

We compare the performance of two alternative approximations to the finite-sample distributions of test statistics for structural change, one based on asymptotics and one based on the bootstrap. We focus on tests acknowledging that the breakpoint is selected endogenously – in particular, the 'supremum' tests of Andrews (1993). We explore a variety of issues of interest in applied work, focusing particularly on smaller samples and persistent dynamics. The bootstrap approximation to the finite-sample distribution appears consistently accurate, in contrast to the asymptotic approximation. The results are of interest not only from the perspective of testing for structural change, but also from the broader perspective of compiling evidence on the adequacy of bootstrap approximations to finite-sample distributions in econometrics.

Key words: Structural change; Bootstrap
JEL classification: C12; C15; C22

* Corresponding author.

We are grateful to three referees for their constructive comments. Many colleagues provided helpful comments as well; special thanks go to Don Andrews, Jean-Marie Dufour, Neil Ericsson, Eric Ghysels, Walter Krämer, and Jim Stock, as well as seminar participants at the Montreal/C.R.D.E. Conference on the Econometrics of Structural Change, Penn, SUNY Albany, Indiana, Illinois, Ohio State, the Federal Reserve Bank of Kansas City, and the Federal Reserve Board. We alone are to blame for remaining shortcomings. We gratefully acknowledge the support of the National Science Foundation, the Sloan Foundation, and the University of Pennsylvania Research Foundation.

SSDI 0304407694016902

1. Introduction

Structural change is of paramount importance in economics and econometrics, and the associated literature is huge.[1] An important associated problem is testing the null hypothesis of structural stability against the alternative of a one-time structural break. In standard treatments, the location of the potential break is assumed known *a priori*. The standard approach is often highly unrealistic, however, because of the endogeneity, or sample selection, problem. That is, implicitly or explicitly, data-based procedures are typically used to determine the most likely location of a break, thereby invalidating the distribution theory associated with conventional tests.

Tests formally incorporating selection effects are therefore desirable. Important such tests are the 'supremum' test of Andrews (1993) and the related 'average' and 'exponential' tests of Andrews and Ploberger (1994), Andrews, Lee, and Ploberger (1992), and Hansen (1991, 1992). At issue, however, is how best to approximate their finite-sample distributions under the null hypothesis of structural stability.

One obvious approximation to the finite-sample distribution is the asymptotic distribution obtained by Andrews (1993). The asymptotic distribution is relatively easy to characterize and tabulate, but it may be an unrealiable guide to finite-sample behavior. An alternative approximation is the bootstrap distribution, as suggested by Christiano (1992). Bootstrap procedures play upon the increased sophistication and feasibility of simulation methods in econometrics, together with drastic increases in computing capability. Bootstrap methods are typically employed in hopes of obtaining better approximations to finite-sample distributions, but as of now the theory is far from complete, and the efficacy of the bootstrap in this application remains an open question.

In this paper, we provide a detailed finite-sample evaluation of the size of supremum tests for structural change in a dynamic model, with attention focused on the comparative performance of asymptotic vs. bootstrap procedures. In doing so, we build upon and extend both Andrews' and Christiano's work. The results are of interest not only from the perspective of testing for structural change, but also from the broader perspective of compiling evidence on the adequacy of bootstrap approximations to finite-sample distributions in econometrics. In Section 2, we introduce the basic model and test statistics. The heart of the paper is Section 3, in which we execute the test size comparison and present the results in both tabular and graphical (response surface) form. In Section 4, we extend the analysis to include richer data-generating processes and alternative test statistics. We conclude and sketch directions for future research in Section 5.

[1] Chow (1986) and Hackl and Westlund (1989) provide useful surveys.

2. The basic model and tests

Here we discuss our most basic model and test statistics, which are the centerpiece of the analysis. We work with the Gaussian zero-mean first-order autoregressive process,[2]

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0, 1), \quad t = 1, \ldots, T,$$

$$y_0 \sim \mathrm{N}\left[0, \frac{1}{1-\rho^2}\right], \quad |\rho| < 1.$$

The null hypothesis of structural stability is

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad t = 1, \ldots, T.$$

We call this the restricted, or no-break, model. The alternative hypothesis of a one-time structural break is

$$y_t = \rho y_{t-1} + \varepsilon_t, \qquad t = 1, \ldots, T^*,$$

$$y_t = \rho_2 y_{t-1} + \varepsilon_t, \qquad t = (T^* + 1), \ldots, T.$$

We call these the unrestricted, or subsample, models.

Our basic test statistics are the 'supremum' tests of Andrews (1993),

$$SupW = \max_{\pi} T\left[\frac{\hat{\varepsilon}'\hat{\varepsilon} - \hat{\varepsilon}_1'\hat{\varepsilon}_1 - \hat{\varepsilon}_2'\hat{\varepsilon}_2}{\hat{\varepsilon}_1'\hat{\varepsilon}_1 + \hat{\varepsilon}_2'\hat{\varepsilon}_2}\right],$$

$$SupLM = \max_{\pi} T\left[\frac{\hat{\varepsilon}'\hat{\varepsilon} - \hat{\varepsilon}_1'\hat{\varepsilon}_1 - \hat{\varepsilon}_2'\hat{\varepsilon}_2}{\hat{\varepsilon}'\hat{\varepsilon}}\right],$$

$$SupLR = \max_{\pi} T \log\left[\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\hat{\varepsilon}_1'\hat{\varepsilon}_1 + \hat{\varepsilon}_2'\hat{\varepsilon}_2}\right],$$

where $\hat{\varepsilon}$ is the $T \times 1$ vector of OLS residuals from the restricted model, $\hat{\varepsilon}_1$ is the $T^* \times 1$ vector of OLS residuals from the subsample 1 model, $\hat{\varepsilon}_2$ is the $(T - T^*) \times 1$ vector of OLS residuals from the subsample 2 model, and $\pi = T^*/T$. Following standard procedure, we impose $\pi \in [0.15, 0.85]$.[3]

We consider two ways of approximating the finite-sample distribution of the three statistics (that is, ways of approximating the significance level, or p-value, corresponding to a given value of a test statistic). The first approximation is

[2] We also examined a variety of leptokurtic and skewed innovations, against which all of our results were robust.

[3] We could, of course, trim by other amounts. The boostrap procedure that we introduce can be easily modified to do so. But Andrews tabulated only the asymptotic distribution for $\pi \in [0.15, 0.85]$. So, to main comparability, we follow suit.

Andrews' (1993) asymptotic distribution, which is essentially the distribution of the supremum of a set of χ^2 random variables. This is obviously appropriate asymptotically, although little is known about the reliability of the asymptotic distribution as a guide to finite-sample behavior. Andrews' Monte Carlo analysis indicates that the asymptotic approximation performs quite well; however, he examines only a restrictive static regression model with a single deterministic regressor. In contrast, we focus on dynamics models.

The second approximation is the bootstrap distribution. Bootstrapping was first applied to the endogenous change-point problem by Christiano (1992), who used the supremum of Chow's (1960) F test to assess the likelihood of a structural shift in the dynamics of U.S. aggregate output. Our bootstrap procedures are straightforward generalizations of Christiano's to a richer class of tests.

The bootstrap formalizes a highly intuitive idea. Using observed data, we calculate the test statistic of interest. Then we build up an approximation to its finite-sample null distribution by (1) generating many pseduo-data samples using model parameters estimated under the null hypothesis and pseudo-disturbances drawn with replacement from the estimated model's residuals, and then (2) calculating the statistic for each sample of pseudo-data.[4] We call this the 'bootstrap distribution'. Finally, we compute the approximate p-value as the probability, relative to the bootstrap distribution, of obtaining a larger value of the test statistic than the one actually obtained.

The results of Gené and Zinn (1990) and Stinchcombe and White (1993) show that the bootstrap procedures we use in this paper are asymptotically valid. Recent research shows that, roughly speaking, the bootstrap is similar to a two-term empirical Edgeworth expansion, although in some important cases it can be shown to be even better (e.g., Bhattacharya and Qumsiyeh, 1989; Hall, 1992). At the very least, the likely impressive finite-sample performance of the bootstrap has emerged as an empirical 'folk theorem', supported by various Monte Carlo studies (see, for example, the insightful survey of Jeong and Maddala, 1992).

Having sketched the basic model and test statistics, we now proceed to the finite-sample comparisons of test size.

3. Monte Carlo analysis of test size

First let us define some notation. *ChiSupW*, *ChiSupLR*, and *ChiSupLM* refer to tests against the $\chi^2(1)$ distribution. *AsySupW*, *AsySupLR*, and *AsySupLM*

[4] Our procedure is a 'nonparametric bootstrap' by virtue of the fact that no distributional assumptions are made. If distributional information is available, one may of course draw the pseudo-disturbances from the appropriate parametric distribution fitted to the model residuals, resulting in a 'parametric bootstrap'. In practice, the nonparametric bootstrap is more popular, as one is usually unsure of the underlying density.

refer to tests against Andrew's (1993) asymptotic distribution. *BootSupW*, *BootSupLR*, and *BootSupLM* refer to tests against the bootstrap distribution.

For each procedure, we explore the relationship between empirical and nominal test size, varying the sample size (T), persistence as measured by the autoregressive coefficient (ρ), and nominal test size (α). We explore sample sizes of $T = 10, 50, 100, 500, 1000$, persistence parameters of $\rho = 0.01, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.83, 0.85, 0.87, 0.90, 0.92, 0.95, 0.96, 0.97, 0.98, 0.99$, and nominal test sizes of $\alpha = 1\%, 2.5\%, 5\%, 10\%$. We perform $RM = 1000$ Monte Carlo replications. Given the choices of T, ρ, and α, we use a full factorial experimental design for examining the properties of the *AsySup* tests, and a subset of the full factorial design for examining the *ChiSup* and *BootSup* tests.

We seek to estimate the probability of rejecting the hypothesis of structural stability when it is true (for a fixed nominal test size α), for each of the test procedures outlined above, and to see how those probabilities vary with T, ρ, and α. Let the true, or empiricial, test size be denoted by α_0. Following standard practice, we use the unbiased and asymptotically (in RM) normal estimator $\hat{\alpha}_0 = S/RM$, where S is the total number of times the procedure rejects. The asymptotic (in RM) standard error of $\hat{\alpha}_0$ is $(\alpha_0(1 - \alpha_0)/RM)^{1/2}$ which we estimate by $(\hat{\alpha}_0(1 - \hat{\alpha}_0)/RM)^{1/2}$.

In addition to presenting the actual Monte Carlo test-size results for each approximation to the finite-sample distribution, we summarize the results succinctly by estimating response surfaces, as advocated by Hendry (1984) and Ericsson (1991). To fit a response surface, we use data from all relevant Monte Carlo experiments to determine how the size distortion, $\hat{\alpha}_0 - \alpha$, depends on T, ρ, and α. Typically, we start with a second-order expansion in powers of $T^{-1/2}$, ρ, and α, as well as terms capturing interaction of ρ and α with T. We then obtain the final response-surface specifications by dropping variables with insignificant t-ratios. Because the variance of $\hat{\alpha}_0$ varies with the value of $\hat{\alpha}_0$, the t-ratios are constructed using White's (1980) heteroskedasticity-consistent standard errors. The final specifications of the response surfaces are given in Table 1.[5]

All computations were done using 386-MATLAB version 3.5k on an IBM-compatible 486 machine running at 33 MHz. All N(0, 1) deviates were generated using the Forsythe et al. (1977) algorithm, as implemented in MATLAB, and all U(0, 1) deviates were generated using the Park–Miller (1988) algorithm, also as implemented in MATLAB.

[5] The response surfaces were also estimated separately for each nominal size. The R^2's for those regressions were of course a bit higher than for the surfaces reported here. The actual graphs of the separately estimated surfaces, however, are virtually identical to those of the surfaces reported here, all of which still have very high R^2.

Table 1
Response surface specifications

	AsySup						
ChiSupW	*W*	*LR, LM*	*BootSup*	*AsySupW* intercept	*BootSup* intercept	*AsyExpW* intercept	*BootExpW* intercept
$T^{1/2}$	$T^{-1/2}$	$T^{-1/2}$	$T^{-1/2}$	$T^{-1/2}$	$T^{-1/2}$	$T^{-1/2}$	$T^{-1/2}$
T	T^{-1}	T^{-1}	T^{-1}	T^{-1}	T^{-1}	T^{-1}	T^{-1}
T^2	$T^{-3/2}$	$T^{-3/2}$		$T^{-3/2}$		$T^{-3/2}$	
α	T^{-2}	T^{-2}				T^{-2}	
α^2	$T^{-1/2}\alpha$	$T^{-1/2}\alpha$					$T^{-1/2}\alpha$
$\rho^{1/2}$	$T^{-1/2}\rho$	$T^{-1/2}\rho$	$T^{-1/2}\rho$	$T^{-1/2}\rho$	$T^{-1/2}\rho$	$T^{-1/2}\rho$	$T^{-1/2}\rho$
ρ	$T^{-1}\alpha$	$T^{-1}\alpha$			$T^{-1}\alpha$		
ρ^2	$T^{-1}\rho$	$T^{-1}\rho$	$T^{-1}\rho$	$T^{-1}\rho$	$T^{-1}\rho$	$T^{-1}\rho$	$T^{-1}\rho$
	$T^{-1/2}\alpha^2$				$T^{-1/2}\alpha^2$		
	$T^{-1/2}\rho^2$	$T^{-1/2}\rho^2$	$T^{-1/2}\rho^2$	$T^{-1/2}\rho^2$	$T^{-1/2}\rho^2$	$T^{-1/2}\rho^2$	$T^{-1/2}\rho^2$
					$T^{-1}\alpha^2$		$T^{-1}\alpha^2$
	$T^{-1}\rho^2$	$T^{-1}\rho^2$	$T^{-1}\rho^2$	$T^{-1}\rho^2$	$T^{-1}\rho^2$	$T^{-1}\rho^2$	$T^{-1}\rho^2$
	$T^{-1/2}\alpha\rho$	$T^{-1/2}\alpha\rho$		$T^{-1/2}\alpha\rho$	$T^{-1/2}\alpha\rho$	$T^{-1/2}\alpha\rho$	$T^{-1/2}\alpha\rho$
	$T^{-1}\alpha\rho$	$T^{-1}\alpha\rho$		$T^{-1}\alpha\rho$	$T^{-1}\alpha\rho$	$T^{-1}\alpha\rho$	$T^{-1}\alpha\rho$

The columns give the regressors for each response surface. T is sample size, α is nominal test size, and ρ is the AR(1) parameter. All test names are of the form '*xyz*', where *x* denotes the type of approximation to the null distribution made ('*Chi*' for chi-squared, '*Asy*' for asymptotic, and '*Boot*' for bootstrap), *y* denotes the type of test ('*Sup*' for supremum, '*Exp*' for exponential, and '*Avg*' for average), and *z* denotes the testing principle employed ('*W*' for Wald, '*LR*' for likelihood ratio, and '*LM*' for Lagrange multiplier).

3.1. *Results for ChiSupW, ChiSupLR, ChiSupLM*

As is now well-known, the χ^2 distribution does not obtain, even asymptotically, because the location of the break point is selected with the aid of the data. Nevertheless, the use of such approximations remains commonplace in applied work. Examination of the finite-sample properties of the *ChiSup* tests illustrates the severity of the size distortions and provides a key benchmark against which to compare the results for other tests.

The Monte Carlo experiments using the χ^2 distribution use the T and α values described above, and persistence parameters of $\rho = 0.01, 0.50, 0.70, 0.80, 0.83, 0.85, 0.87, 0.90, 0.92, 0.95, 0.96, 0.97, 0.98, 0.99$. To save space, we present in Table 2 the results from only a small but representative subset of the parameter space points actually explored. As intuition suggests, the *ChiSup* tests tend to overreject. Moreover, the size distortions are *huge*. For example, the response surface for *ChiSupW* (Fig. 1), which plots the size distortion, $\hat{\alpha}_0 - \alpha$, as a function of sample size and persistence when nominal test size is fixed at 10%, highlights their poor performance: empirical test size typically dwarfs nominal

Table 2
Empirical size of *ChiSup* tests, normal innovations, AR(1)

		$\rho = 0.01$			$\rho = 0.50$			$\rho = 0.80$			$\rho = 0.99$		
T	*Sup* test	1.0	5.0	10.0	1.0	5.0	10.0	1.0	5.0	10.0	1.0	5.0	10.0
10	*W*	8.8	23.7	38.1	12.0	27.6	41.4	16.8	35.7	49.1	24.7	42.3	56.3
		(0.9)	(1.3)	(1.5)	(1.0)	(1.4)	(1.6)	(1.2)	(1.5)	(1.6)	(1.4)	(1.6)	(1.6)
	LR	4.5	16.9	32.1	4.9	20.9	36.2	9.6	27.2	44.6	14.2	35.5	51.5
		(0.7)	(1.2)	(1.5)	(0.7)	(1.3)	(1.5)	(0.9)	(1.4)	(1.6)	(1.1)	(1.5)	(1.6)
	LM	0.4	10.2	24.9	0.7	13.8	29.4	1.6	18.1	37.3	3.0	27.0	43.0
		(0.2)	(1.0)	(1.4)	(0.3)	(1.1)	(1.4)	(0.4)	(1.2)	(1.5)	(0.5)	(1.4)	(1.6)
50	*W*	6.2	26.9	46.6	7.5	29.0	46.9	9.4	33.7	51.8	15.1	41.8	61.1
		(0.8)	(1.4)	(1.6)	(0.8)	(1.4)	(1.6)	(0.9)	(1.5)	(1.6)	(1.1)	(1.6)	(1.5)
	LR	4.5	25.0	44.9	5.8	27.4	44.6	8.2	31.4	50.3	13.0	39.5	60.4
		(0.7)	(1.4)	(1.6)	(0.7)	(1.4)	(1.6)	(0.9)	(1.5)	(1.6)	(1.1)	(1.5)	(1.5)
	LM	3.4	22.9	42.9	4.5	25.6	43.1	6.7	28.5	48.6	10.6	36.8	58.8
		(0.6)	(1.3)	(1.6)	(0.7)	(1.4)	(1.6)	(0.8)	(1.4)	(1.6)	(1.0)	(1.5)	(1.6)
100	*W*	7.5	29.8	52.5	8.9	30.9	54.8	10.3	35.1	54.7	18.1	46.1	64.5
		(0.8)	(1.4)	(1.6)	(0.9)	(1.5)	(1.6)	(1.0)	(1.5)	(1.6)	(1.2)	(1.6)	(1.5)
	LR	6.6	28.9	51.9	7.6	29.9	53.4	9.4	33.8	53.7	15.8	44.9	63.9
		(0.8)	(1.4)	(1.6)	(0.8)	(1.4)	(1.6)	(0.9)	(1.5)	(1.6)	(1.2)	(1.6)	(1.5)
	LM	5.9	27.4	50.2	6.7	29.0	53.0	8.5	32.7	53.2	14.2	44.0	63.0
		(0.7)	(1.4)	(1.6)	(0.8)	(1.4)	(1.6)	(0.9)	(1.5)	(1.6)	(1.1)	(1.6)	(1.5)
500	*W*	13.0	38.2	60.9	12.1	37.4	58.0	13.4	38.3	59.5	20.6	49.3	67.6
		(1.1)	(1.5)	(1.5)	(1.0)	(1.5)	(1.6)	(1.1)	(1.5)	(1.6)	(1.3)	(1.6)	(1.5)
	LR	12.8	37.6	60.7	11.9	37.2	57.9	13.3	38.2	59.3	20.1	48.9	67.3
		(1.1)	(1.5)	(1.5)	(1.0)	(1.5)	(1.6)	(1.1)	(1.5)	(1.6)	(1.3)	(1.6)	(1.5)
	LM	12.8	37.4	60.5	11.5	37.0	57.9	13.2	38.0	59.3	19.5	48.4	67.0
		(1.1)	(1.5)	(1.5)	(1.0)	(1.5)	(1.6)	(1.1)	(1.5)	(1.6)	(1.3)	(1.6)	(1.5)
1000	*W*	11.3	40.3	60.9	12.4	38.3	61.5	12.7	39.4	60.3	17.0	47.9	64.4
		(1.0)	(1.6)	(1.5)	(1.0)	(1.5)	(1.5)	(1.1)	(1.5)	(1.5)	(1.2)	(1.6)	(1.5)
	LR	11.3	40.1	60.8	12.4	38.3	61.3	12.5	39.3	60.2	17.0	47.9	64.4
		(1.0)	(1.5)	(1.5)	(1.0)	(1.5)	(1.5)	(1.0)	(1.5)	(1.5)	(1.2)	(1.6)	(1.5)
	LM	11.2	40.1	60.7	12.3	38.2	61.3	12.1	39.1	60.2	16.9	47.9	64.4
		(1.0)	(1.5)	(1.5)	(1.0)	(1.5)	(1.5)	(1.0)	(1.5)	(1.5)	(1.2)	(1.6)	(1.5)

T is a sample size and ρ is the AR(1) parameter. For each (T, ρ) pair, three nominal test sizes (1%, 5%, 10%) are explored. See footnote to Table 1 for test naming conventions.

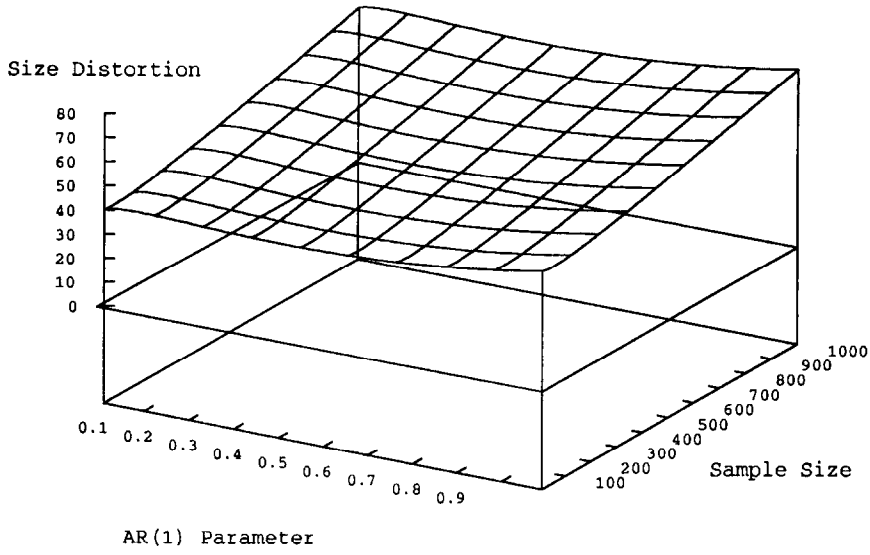

Fig. 1. Response surface for *ChiSupW*, 10% nominal size, AR(1). See footnote to Table 1 for test naming conventions.

test size.[6,7] Furthermore, the tendency for the *ChiSup* tests to overreject *worsens* as sample size increases, due to the increased severity of data mining.[8] Serial correlation also affects the empirical size of the *ChiSup* tests. As the degree of serial correlation increases, the *ChiSup* tests tend to reject more often. The extent to which the tests overreject from using the incorrect asymptotic distribution, however, far overpowers this effect.

3.2. *Results for AsySupW, AsySupLR, AsySupLM*

Having seen how poorly the χ^2 distribution performs as an approximation to the finite-sample distribution, we now turn to the asymptotic approximation. The full factorial experimental design is used here, although again we present only a subset of the results, this time in Table 3 and Figs. 2–4.[9] The basic

[6] The response surfaces for *ChiSupLR* and *ChiSupLM* are similar to that of *ChiSupW* and therefore are not included here.

[7] We superimpose graphs of the x-y plane on most of our graphs of response surfaces. A response surface that lies near the x-y plane indicates that the distribution approximates the finite sample distribution well. Clearly, this is not the case for the *ChiSup* tests.

[8] Hence T appears in positive rather than negative powers in the response surface.

[9] Note the scale change of the z axis relative to Fig. 1.

Table 3
Empirical size of *AsySup* tests, normal innovations, AR(1)

		$\rho = 0.01$			$\rho = 0.50$			$\rho = 0.80$			$\rho = 0.99$		
T	*Sup* test	1.0	5.0	10.0	1.0	5.0	10.0	1.0	5.0	10.0	1.0	5.0	10.0
10	*W*	1.9	5.3	7.4	2.7	6.6	10.7	4.9	11.5	15.3	8.2	15.8	21.2
		(0.4)	(0.7)	(0.8)	(0.5)	(0.8)	(1.0)	(0.7)	(1.0)	(1.1)	(0.9)	(1.2)	(1.3)
	LR	0.1	1.4	3.8	0.2	2.3	4.0	1.0	3.8	7.4	2.1	6.4	11.4
		(0.1)	(0.4)	(0.6)	(0.1)	(0.5)	(0.6)	(0.3)	(0.6)	(0.8)	(0.5)	(0.8)	(1.0)
	LM	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	1.0	0.0	0.1	1.7
		(0.0)	(0.0)	(0.1)	(0.0)	(0.0)	(0.1)	(0.0)	(0.0)	(0.3)	(0.0)	(0.1)	(0.4)
50	*W*	0.2	1.8	4.3	0.2	3.0	5.6	0.8	4.1	8.1	1.8	7.4	12.4
		(0.1)	(0.4)	(0.6)	(0.1)	(0.5)	(0.7)	(0.3)	(0.6)	(0.9)	(0.4)	(0.8)	(1.0)
	LR	0.0	1.0	3.1	0.0	1.9	4.5	0.5	3.3	6.3	1.3	5.2	10.3
		(0.0)	(0.3)	(0.5)	(0.0)	(0.4)	(0.7)	(0.2)	(0.6)	(0.8)	(0.4)	(0.7)	(1.0)
	LM	0.0	0.4	2.3	0.0	0.9	3.5	0.2	2.2	4.7	0.7	3.9	8.0
		(0.0)	(0.2)	(0.5)	(0.0)	(0.3)	(0.6)	(0.1)	(0.5)	(0.7)	(0.3)	(0.6)	(0.9)
100	*W*	0.5	3.3	5.5	0.6	3.7	6.3	0.8	5.1	8.2	1.6	7.4	13.5
		(0.2)	(0.6)	(0.7)	(0.2)	(0.6)	(0.8)	(0.3)	(0.7)	(0.9)	(0.4)	(0.8)	(1.1)
	LR	0.4	2.8	4.8	0.5	3.2	5.4	0.4	3.9	7.2	1.2	6.3	12.2
		(0.2)	(0.5)	(0.7)	(0.2)	(0.6)	(0.7)	(0.2)	(0.6)	(0.8)	(0.3)	(0.8)	(1.0)
	LM	0.3	2.2	4.2	0.2	2.8	4.8	0.4	3.3	6.5	0.6	5.3	10.4
		(0.2)	(0.5)	(0.6)	(0.1)	(0.5)	(0.7)	(0.2)	(0.6)	(0.8)	(0.2)	(0.7)	(1.0)
500	*W*	0.9	5.0	10.5	0.9	5.1	9.1	1.2	5.6	11.0	3.1	10.7	16.3
		(0.3)	(0.7)	(1.0)	(0.3)	(0.7)	(0.9)	(0.3)	(0.7)	(1.0)	(0.5)	(1.0)	(1.2)
	LR	0.9	4.9	10.3	0.9	5.1	8.8	1.2	5.4	10.7	3.0	10.3	15.9
		(0.3)	(0.7)	(1.0)	(0.3)	(0.7)	(0.9)	(0.3)	(0.7)	(1.0)	(0.5)	(1.0)	(1.2)
	LM	0.9	4.3	10.3	0.9	5.0	8.5	0.8	5.1	10.4	2.9	10.3	15.8
		(0.3)	(0.7)	(1.0)	(0.3)	(0.7)	(0.9)	(0.3)	(0.7)	(1.0)	(0.5)	(1.0)	(1.2)
1000	*W*	0.9	4.3	8.2	1.2	5.5	10.1	0.8	4.8	9.8	1.5	8.1	13.8
		(0.3)	(0.6)	(0.9)	(0.3)	(0.7)	(1.0)	(0.3)	(0.7)	(0.9)	(0.4)	(0.9)	(1.1)
	LR	0.9	4.1	8.1	1.2	5.2	9.9	0.7	4.8	9.7	1.4	8.0	13.8
		(0.3)	(0.6)	(0.9)	(0.3)	(0.7)	(0.9)	(0.3)	(0.7)	(0.9)	(0.4)	(0.9)	(1.1)
	LM	0.9	4.0	8.1	1.2	5.2	9.9	0.7	4.5	9.7	1.3	7.7	13.7
		(0.3)	(0.6)	(0.9)	(0.3)	(0.7)	(0.9)	(0.3)	(0.7)	(0.9)	(0.4)	(0.8)	(1.1)

T is the sample size and ρ is the AR(1) parameter. For each (T, ρ) pair, three nomianl test sizes (1%, 5%, 10%) are explored. See footnote to Table 1 for test naming conventions.

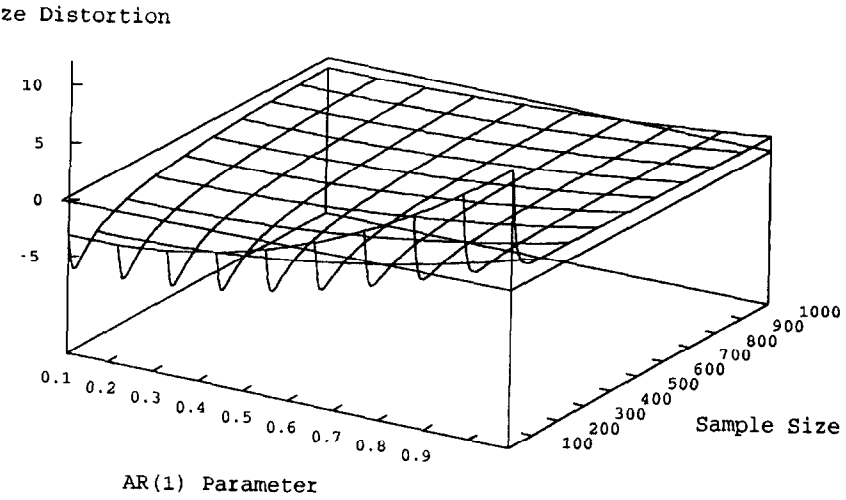

Fig. 2. Response surface for *AsySupW*, 10% nominal size, AR(1). See footnote to Table 1 for test naming conventions.

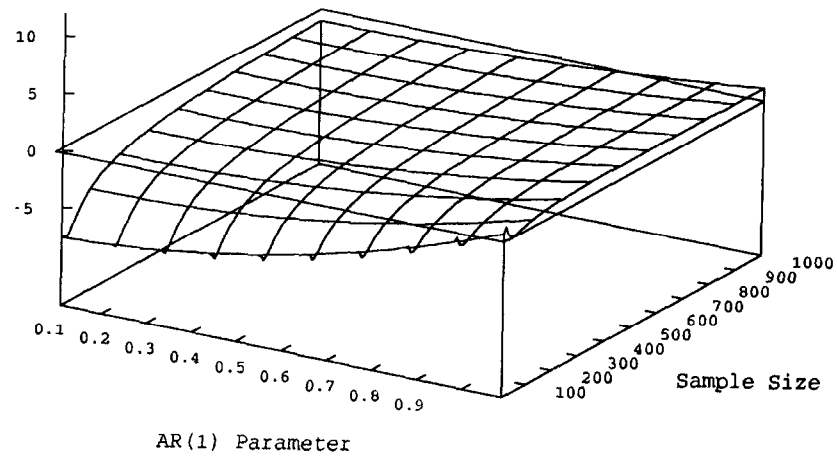

Fig. 3. Response surface for *AsySupLR*, 10% nominal size, AR(1). See footnote to Table 1 for test naming conventions.

observations are:

(i) Use of the asymptotic distribution yields large reductions in size distortion relative to use of the χ^2 distribution. Significant distortions remain, however.

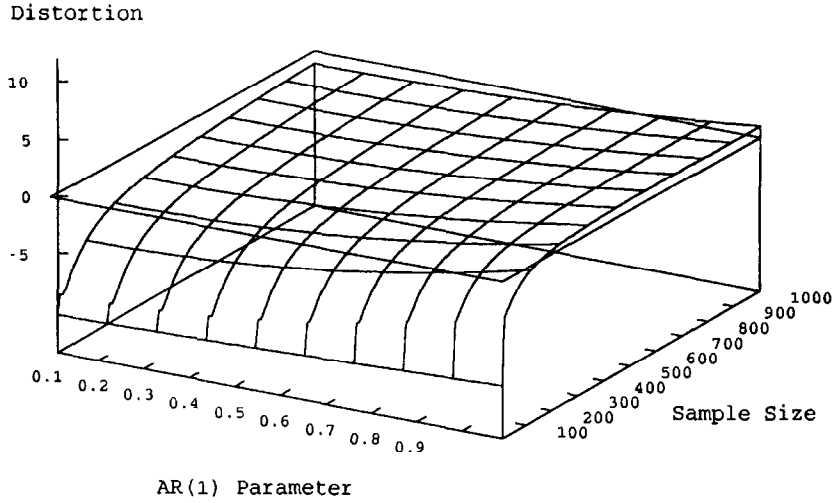

Fig. 4. Response surface for *AsySupLM*, 10% nominal size, AR (1). See footnote to Table 1 for test naming conventions.

(ii) Clearly, and as expected, nominal and empirical test size converge as T grows.

(iii) Table 3 confirms that the expected relationship, *SupW* $\geqslant$ *SupLR* $\geqslant$ *SupLM*, holds throughout. (This must hold at any break point, and therefore for the supremum.) In line with this, for small samples, *AsySupW* tends to overreject, *AsySupLR* to underreject, and *AsySupLM* to underreject drastically.

(iv) The response surfaces in Figs. 2–4 show that the size distortion, $\hat{\alpha}_0 - \alpha$, is generally large (and negative) for small T, and decreases with T. The convergence of empirical to nominal test size, however, can be slow and nonmonotone, particularly with the amounts of serial correlation typically present in economic data.

(v) Serial correlation tends to inflate empirical test size. The stronger the serial correlation, the greater the empirical size inflation. Additionally, for *AsySupW* and *AsySupLR* the sensitivity to serial correlation is greater the smaller the sample size. This can be seen most clearly for the case of *AsySupW* where the values for $T = 10$ are larger than for $T = 100$.

(vi) The size distortions associated with *AsySupLM* procedure in small samples are particularly severe.

Let us elaborate upon points (iv) and (v). It is important to note that there are effectively two sources of test size distortion. First, in the absence of serial correlation, finite-sample empirical test size tends to be pushed downward

relative to nominal size. Serial correlation works in the opposite direction, pushing empirical test size upward. Thus, the occasional good test size performance is merely an artifact of a happenstance cancellation of competing biases, and should never be relied upon to work in practice.

The upshot of this subsection is that, although the finite-sample performance of the *AsySup* statistics *is* much better than that of the *ChiSup* statistics, it is *not* as good as one would hope. Whether there exists an even better approximation to the finite-sample distribution – one that corrects the deficiencies of the asymptotic distribution – remains an open question. This leads us to the bootstrap procedure, to which we now turn.

3.3. *Results for BootSupW, BootSupLR, BootSupLM*

We provide a flow chart illustrating our Monte Carlo procedure for evaluating the bootstrap in Fig. 5. The details of the procedure, and our Monte Carlo examination of its properties, are as follows. Let $i = 1, \ldots, RM$ index Monte Carlo replications, and let $j = 1, \ldots, RB$ index the bootstrap replications inside each Monte Carlo replication.[10] Nominal test size is α. Then:

(1) Draw the vector of innovations $\{\varepsilon_1^i, \varepsilon_2^i, \ldots, \varepsilon_T^i\} \sim \mathrm{N}(0, 1)$. Then generate a vector of 'true' data $\{y_1^i, y_2^i, \ldots, y_T^i\}$ from

$$y_t^i = \rho y_{t-1}^i + \varepsilon_t^i, \quad t = 1, \ldots, T,$$

$$y_0 \sim \mathrm{N}(0, 1/(1-\rho^2)),$$

and compute the OLS estimator for ρ, $\hat{\rho}^i$, and the associated de-meaned residuals $\{\hat{\varepsilon}_1^i, \hat{\varepsilon}_2^i, \ldots, \hat{\varepsilon}_T^i\}$. Finally, compute $SUP^i = \{SupW^i, SupLR^i, SupLM^i\}$.

(2a) Draw $\{e_1^j, e_2^j, \ldots, e_T^j\}$ by sampling with replacement from $\{\hat{\varepsilon}_1^i, \hat{\varepsilon}_2^i, \ldots, \hat{\varepsilon}_T^i\}$. Then generate the pseudo-data $\{y_1^{ij}, y_2^{ij}, \ldots, y_T^{ij}\}$ via $y_t^{ij} = \hat{\rho}^i y_{t-1}^{ij} + e_t^j$. Choose the startup value, y_0^{ij}, randomly from the stationary distribution, as proxied by the vector of 'true' data $\{y_1^i, y_2^i, \ldots, y_T^i\}$. Finally, compute SUP^{ij}.

(2b) Repeat step (2a) RB times, yielding a $(RB \times 1)$ vector of SUP^{ij} values. This vector constitutes the bootstrap distribution for Monte Carlo replication i. The 10% critical value of the bootstrap distribution, for example, is estimated as the 900th element in the vector, after sorting from smallest to largest.[11]

[10] Throughout, we set $RB = 1000$.

[11] This estimator is consistent (in RB), but other estimators may be available with even better properties. We leave this to future research.

(2c) Compare the SUP^i value from step (1) to the $\alpha\%$ bootstrap critical value from (2b), and determine whether the critical value is exceeded.

(3) Repeat steps (1)–(2) RM times.

(4) Compute the percentage of times a rejection occurs in Step (2c). If nominal and empirical test size are equal, rejection should occur $\alpha\%$ of the time (up to Monte Carlo error).

We present the results in Table 4. The bootstrap consistently outperforms the asymptotic distribution in approximating the finite-sample distribution. In fact, as we proceeded with the experiments, it became apparent that the bootstrap's performance was nearly perfect. Thus, we were able to choose rather widely-spaced values of ρ, relative to those explored for the *AsySup* tests. Similarly, there was no point in exploring values of T greater than 50, because the bootstrap approximation was nearly perfect even for sample sizes *smaller* than 50. In the end, we explored $\rho = 0.01, 0.50, 0.80, 0.90, 0.99$ and $T = 10, 25, 50$.

Note that the three bootstrapped tests have identical empirical size. This occurs because, as Engle (1984) shows, the Lagrange multiplier and likelihood ratio tests for linear restrictions have the following relationships with the Wald test in the linear regression model:

$$LR = T\log\left(1+\frac{W}{T}\right) = T\log\left[1+\frac{\hat{\varepsilon}'\hat{\varepsilon}-\hat{\varepsilon}_1'\hat{\varepsilon}_1-\hat{\varepsilon}_2'\hat{\varepsilon}_2}{\hat{\varepsilon}_1'\hat{\varepsilon}_1+\hat{\varepsilon}_2'\hat{\varepsilon}_2}\right] = T\log\left[\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\hat{\varepsilon}_1'\hat{\varepsilon}_1+\hat{\varepsilon}_2'\hat{\varepsilon}_2}\right],$$

$$LM = \frac{W}{1+\frac{W}{T}} = \frac{T\dfrac{\hat{\varepsilon}'\hat{\varepsilon}-\hat{\varepsilon}_1'\hat{\varepsilon}_1-\hat{\varepsilon}_2'\hat{\varepsilon}_2}{\hat{\varepsilon}_1'\hat{\varepsilon}_1+\hat{\varepsilon}_2'\hat{\varepsilon}_2}}{\dfrac{\hat{\varepsilon}'\hat{\varepsilon}}{\hat{\varepsilon}_1'\hat{\varepsilon}_1+\hat{\varepsilon}_2'\hat{\varepsilon}_2}} = T\frac{\hat{\varepsilon}'\hat{\varepsilon}-\hat{\varepsilon}_1'\hat{\varepsilon}_1-\hat{\varepsilon}_2'\hat{\varepsilon}_2}{\hat{\varepsilon}'\hat{\varepsilon}}.$$

Because the first derivatives,

$$\frac{\partial LR}{\partial W} = \frac{1}{1+W/T} \quad \text{and} \quad \frac{\partial LM}{\partial W} = \frac{1}{(1+W/T)^2},$$

are greater than zero (so long as $W > 0$), LR and LM are monotone increasing functions of W. This means that the breakpoint that gives the *SupW* statistic will also be the breakpoint that gives the *SupLR* and *SupLM* statistics. If the 'true' $SupW^i$ is equal to the 900th largest $SupW^{ij}$ of the bootstrap $SupW^i$ distribution, then the 'true' $SupLR^i$ is equal to the 900th largest $SupLR^{ij}$ in the bootstrapped finite-sample distribution for the *SupLR* test statistic, and similarly for the *SupLM* test. Thus, the bootstrap distributions adjust for the differences among the *SupW*, *LR*, and *LM* statistics.

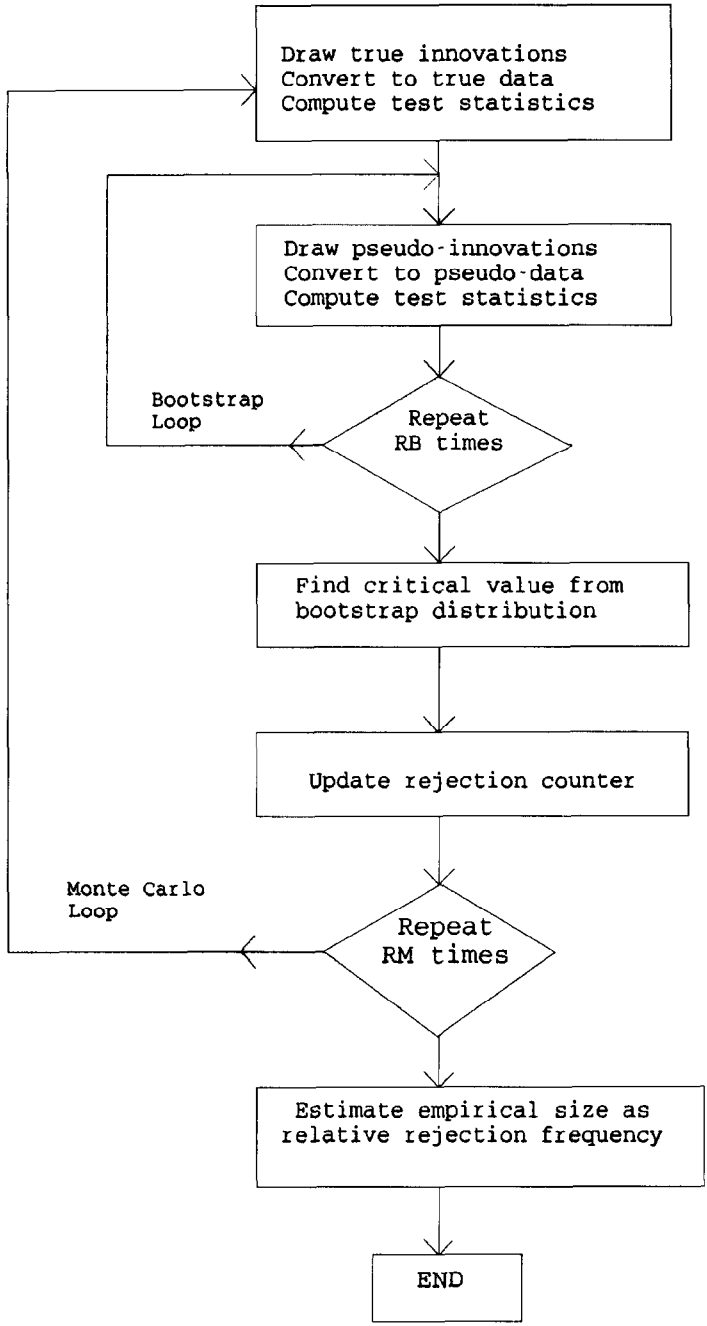

Fig. 5. Flow chart for bootstrap Monte Carlo.

Table 4
Empirical size of *BootSup* tests, normal innovations, AR(1)

	$\rho = 0.01$			$\rho = 0.50$			$\rho = 0.80$			$\rho = 0.90$			$\rho = 0.99$		
T	1.0	5.0	10.0	1.0	5.0	10.0	1.0	5.0	10.0	1.0	5.0	10.0	1.0	5.0	10.0
10	0.8	5.0	8.8	0.8	4.5	9.5	0.8	5.5	10.4	0.9	5.3	9.8	0.6	5.2	10.4
	(0.3)	(0.7)	(0.9)	(0.3)	(0.7)	(0.9)	(0.3)	(0.7)	(1.0)	(0.3)	(0.7)	(0.9)	(0.2)	(0.7)	(1.0)
25	1.3	4.9	10.3	0.7	5.0	10.4	1.3	5.3	9.3	1.1	5.3	10.0	1.5	4.5	9.6
	(0.3)	(0.7)	(1.0)	(0.3)	(0.7)	(1.0)	(0.4)	(0.7)	(0.9)	(0.3)	(0.7)	(0.9)	(0.4)	(0.7)	99.0
50	1.5	6.6	11.4	0.8	5.2	10.2	1.1	4.2	10.3	1.3	5.7	11.4	1.4	6.1	12.0
	(0.4)	(0.8)	(1.0)	(0.3)	(0.7)	(1.0)	(0.3)	(0.6)	(1.0)	(0.4)	(0.7)	(1.0)	(0.4)	(0.8)	(1.0)

The *Wald*, *LR*, *LM* results are identical, for reasons discussed in the text, so no distinction is made in the table. T is sample size and ρ is the AR(1) parameter. For each (T, ρ) pair, three nomianl test sizes (1%, 5%, 10%) are explored. See footnote to Table 1 for test naming conventions.

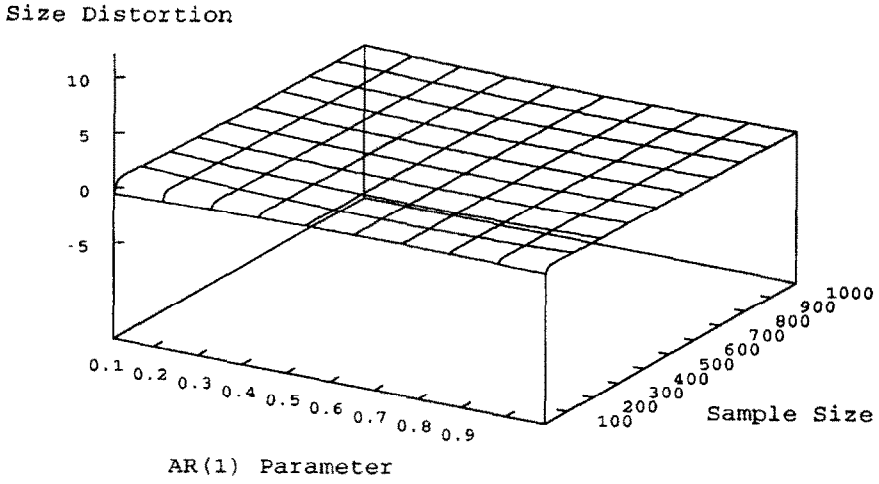

Fig. 6. Response surface for *BootSup*, 10% nominal size, AR(1). See footnote to Table 1 for test naming conventions.

We also present the estimated response surface for the bootstrap in Fig. 6. The performance of the bootstrap is visually striking. As expected, it has little curvature; effectively, it is just the plane containing the x and y axes.[12]

The results demonstrate that the bootstrap distribution can be a much better approximation to the finite-sample distribution than the asymptotic distribu-

[12] For that reason, the x-y plane is not graphed in Fig. 6.

tion when sample size is not large and/or persistence is high. Note, in particular, that the finite-sample size distortion associated with the use of bootstrapped critical values is minimal, *regardless of the value of the nuisance parameter* ρ. This stands in sharp contrast to the finite-sample size distortion associated with the use of asymptotic critical values, which depends heavily on ρ, in spite of the fact that dependence on ρ vanishes in the limit.

Moreover, our results are entirely in line with those of Rayner (1990), who finds that the bootstrap approximation to the finite-sample distribution of studentized statistics in an AR(1) model outperforms the asymptotic approximation for sample sizes as small as $T = 5$ and for degrees of persistence as large as $\rho = 0.99$. Jeong and Maddala (1992) suggest that one reason for Rayner's success with the bootstrap is his careful treatment of the initial value when generating bootstrap samples. Instead of assuming the initial value to be known, he draws it from its stationary distribution. We of course followed the same approach.

We conclude this subsection with some discussion of the computational requirements of the bootstrap. Although it is true that a Monte Carlo evaluation of the bootstrap is a formidable computational task, actual use of the bootstrap on a real dataset is a simple matter. For example, performing the *BootSupLR* test for a sample of size 50 requires 1,8398,849 floating point operations, which takes about ten minutes on a 33 MHz 486 machine.

4. Extensions to other models and other tests

4.1. Other models

We now extend the analysis to richer models, by allowing for inclusion of an intercept. For each *AsySup* test statistic, the experiments are over $T = 10, 50, 100, 500$ and $\rho = 0.01, 0.50, 0.80, 0.99$, while for the *BootSup* tests, the experiments are over $T = 10, 25, 50$ and $\rho = 0.01, 0.50, 0.80, 0.90, 0.99$. The size distortions of the *AsySup* tests can be extremely large when allowance for intercept is made. For example, in Fig. 7 we graph the response surface for the *AsySupW* test when an intercept is included in the estimation. Comparing this against Fig. 2, the zero-mean case, the basic shape of the response surfaces remains the same. But when a mean is included, the size distortions become much worse.[13,14] In contrast, the bootstrap (Fig. 8) appears to continue to

[13] Again note the change in the scale of the z axis.

[14] Of course, as sample size grows and for moderate levels of serial correlation, the size distortion almost disappears, as is consistent with the asymptotic behavior of the tests.

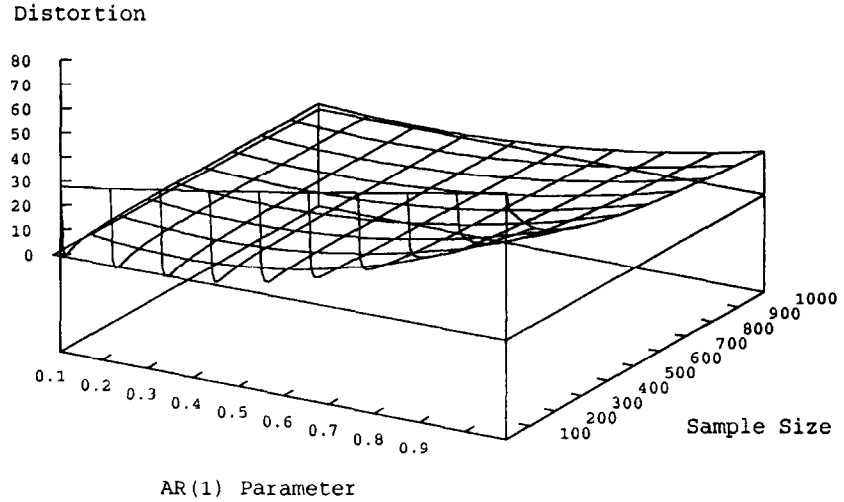

Fig. 7. Response surface for *AsySupW*, 10% nominal size, AR(1) with intercept. See footnote to Table 1 for test naming conventions.

deliver outstanding approximations to the null distributions of the test statistics, with the response surface lying almost completely on the x-y plane.

4.2. Other tests

Now we consider additional tests. Andrews and Ploberger (1994), for example, propose an asymptotically optimal class of tests that can be applied to the change-point problem. Andrews, Lee, and Ploberger (1992) derive the finite-sample distribution for these tests, but the derivation applies only under a model with fixed regressors and normally distributed errors with known variance. We explore the finite-sample performance of six of these tests under much more general conditions. The tests are the exponential Wald (*ExpW*), exponential LR (*ExpLR*), exponential LM (*ExpLM*), average Wald (*AvgW*), average LR (*AvgLR*), and average LM (*AvgLM*), given by

$$ExpK = \log\left[\frac{1}{0.85T - 0.015T + 1}\sum_{\pi}\exp(\tfrac{1}{2}K(\pi))\right],$$

$$AvgK = \frac{1}{0.85T - 0.15T + 1}\sum_{\pi}K(\pi),$$

where $K = W$, LR, or LM.

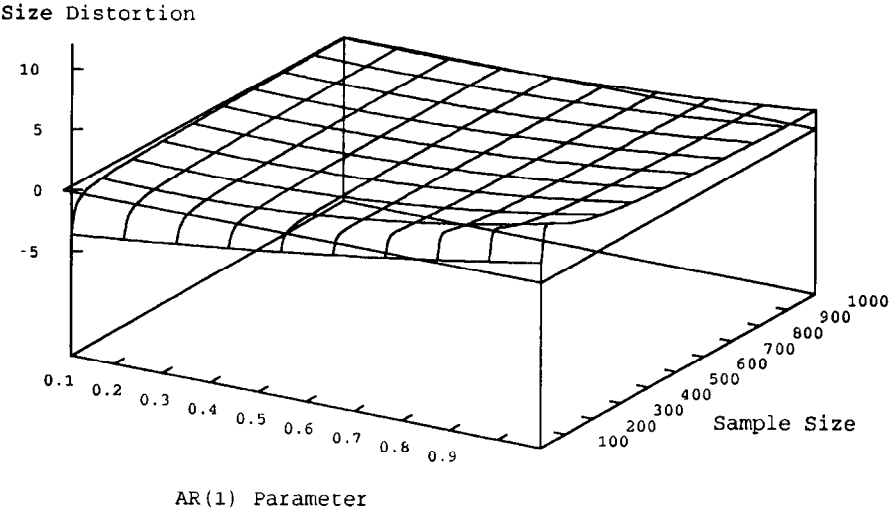

Fig. 8. Response surface for *BootSup*, 10% nominal size, AR(1) with intercept. See footnote to Table 1 for test naming conventions.

We replicated all the experiments for $\alpha = 1\%, 5\%, 10\%$, performed on the *Sup* tests using the *Exp* and *Avg* tests. The size behavior of these tests is similar to that of the *Sup* tests, which is expected because they are all based on the same principles. In each case the *AsyExp* tests display large size distortions, while the *BootExp* tests do not.[15] Compare, for example, the response surface for *AsyExpW* with intercept (Fig. 9) to the response surface for *BootExpW* with intercept (Fig. 10). Qualitatively identical results hold for the *AsyAvg* and *BootAvg* tests as well.[16] In fact, all the results reported earlier for *Sup* tests carry over completely to the *Exp* and *Avg* tests.

5. Conclusions and directions for future research

Our results sound a cautionary note regarding the use of the various asymptotic test procedures in applied time-series econometric analyses of structural change, because of the deviations between nominal and empirical test size that can arise in dynamic models. However, our rather pessimistic conclusion

[15] To save space we report only the results for Wald tests here, but we note that the earlier-discussed equivalence of *BootSupW*, *BootSupLR*, and *BootSupLM* does not carry over to the *BootExp* or *BootAvg* cases.

[16] To save space, we do not report those response surfaces.

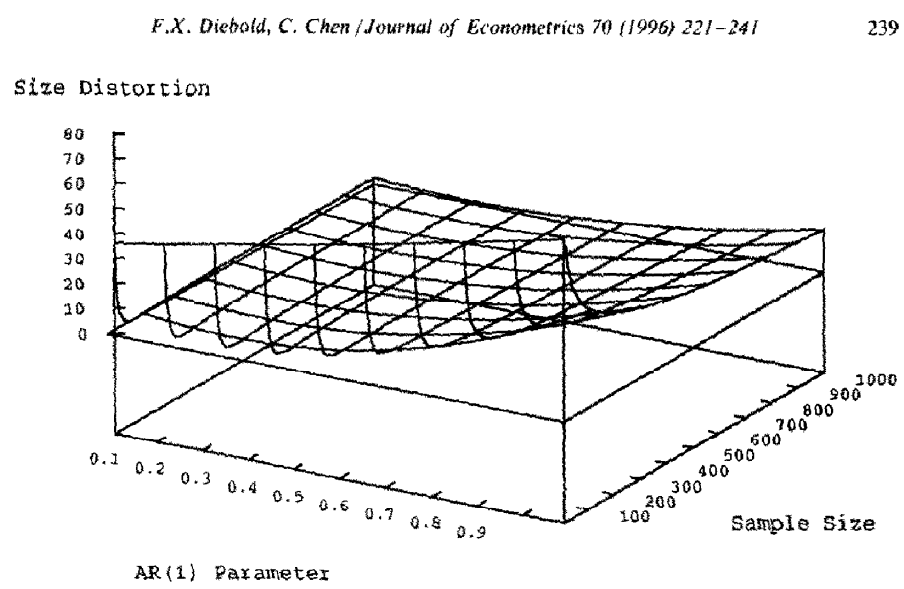

Fig. 9. Response surface for *AsyExpW*, 10% nominal size, AR(1) with intercept. See footnote to Table 1 for test naming conventions.

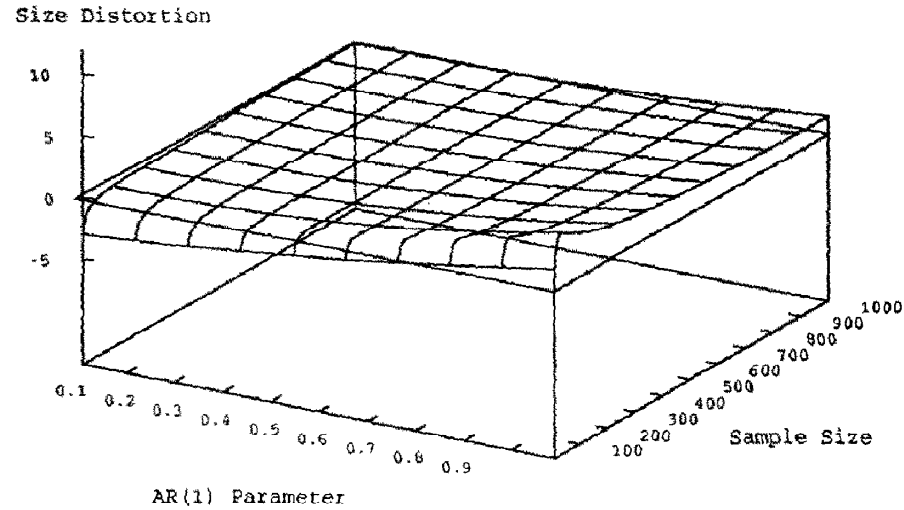

Fig. 10. Response surface for *BootExpW*, 10% nominal size, AR(1) with intercept. See footnote to Table 1 for test naming conventions.

regarding the asymptotic distribution is offset by our optimistic results for the bootstrap distribution. The bootstrap procedures perform very well, even in small samples with high serial correlation.

There are, of course, numerous limitations to our analysis. For example, we don't consider cases of gradual structural change, we don't allow for multiple

breaks, we don't explicitly treat multivariate models, we don't allow for shifts in the innovation variance across subsamples, and so forth. But although we have not addressed such issues here, it's easy to accommodate them with the bootstrap in actual applications. In fact, flexibility and adaptability are two of the bootstrap's greatest virtues. This contrasts with analytic derivation and tabulation of asymptotic distributions.

Given that the bootstrap procedures appear to maintain correct test size, while the asymptotic procedures do not, it will be of obvious interest in future work to compare the *power* of the bootstrapped versions of the *Sup*, *Exp*, and *Avg* tests in dynamic models with small samples, with particular attention paid to location of the break and 'distance' of the alternative from the null. Although, for example, the *Exp* and *Avg* tests enjoy certain optimality properties in large samples, little is known about their comparative power properties under the conditions maintained here. It will also be of interest to examine the power gains accruing to an explicitly multivariate framework.[17]

Preliminary results for the zero-mean AR(1) model show:

(1) The bootstrapped *AsySup* tests have excellent power for large breaks, with power diminishing as the break size diminishes and as the location of the break moves toward the edge of the sample.
(2) The *Avg* and *Exp* tests never perform much better than their *Sup* counterparts in our experiments run thus far, and they sometimes perform worse.
(3) The degree of persistence has a relatively small effect on power.
(4) The power of the standard Chow test depends critically on whether the assumed breakpoint and the true breakpoint coincide. When they do, the Chow test has good power; when they don't, the Chow test can have very poor power.
(5) The *BootSup* tests, which do not require *a priori* specification of the breakpoint, consistently display power performance almost as good as the standard Chow test with *correctly* specified breakpoint. Thus, it appears that there is potentially much to be gained by using the *BootSup* tests, and little to be lost.

References

Andrews, D.W.K., 1993, Tests for parameter instability and structural change with unknown change point, Econometrica 61, 821–856.

Andrews, D.W.K. and W. Ploberger, 1994, Optimal tests when the nuisance parameter is present only under the alternative, Econometrica 62, 1383–1414.

[17] Working with different tests, Bai, Lumsdaine, and Stock (1991) find impressive power gains in a multivariate framework.

Andrews, D.W.K., I. Lee, and W. Ploberger, 1992, Optimal changepoint tests for normal linear regression, Manuscript (Department of Economics, Yale University, New Haven, CT).

Bai, J., R.L. Lumsdaine, and J.H. Stock, 1991, Testing for and dating breaks in integrated and cointegrated time series, Manuscript (Department of Economics, Harvard University, Cambridge, MA).

Bhattacharya, R. and M. Qumsiyeh, 1989, Second order and L^p-comparisons between the bootstrap and empirical Edgeworth expansion methodologies, Annals of Statistics 17, 160–169.

Chow, G.C., 1960, Tests of equality between sets of coefficients in two linear regressions, Econometrica 28, 591–603.

Chow, G.C., 1986, Random and changing coefficient models, in: Z. Griliches and M.D. Intriligator, eds., Handbook of econometrics, Vol. II (North-Holland, Amsterdam) 1213–1245.

Christiano, L.J., 1992, Searching for a break in GNP, Journal of Business and Economic Statstics 10, 237–249.

Engle, R.F., 1984, Wald, likelihood ratio and Lagrange multiplier tests in econometrics, in: Z. Griliches and M.D. Intriligator, eds., Handbook of econometrics, Vol. II (North-Holland, Amsterdam) 775–826.

Ericsson, N., 1991, Monte Carlo methodology and the finite-sample properties of instrumental variables statistics for testing nested and non-nested hypotheses, Econometrica 59, 1249–1277.

Forsythe, G.E., M.A. Malcom, and C.B. Moler, 1977, Computer methods for mathematical computations (Prentice-Hall, Englewood Cliffs, NJ).

Freedman, D.A., 1981, Bootstrapping regression models, Annals of Statistics 9, 1218–1228.

Gené, E. and J. Zinn, 1990, Bootstrapping general empirical measures, Annals of Probability 18, 851–869.

Hackl, P. and A.H. Westlund, 1989, Statistical analysis of structural change: An annotated bibliography, Empirical Economics 14, 167–192.

Hall, P., 1992, The bootstrap and Edgeworth expansion (Springer-Verlag, New York, NY).

Hansen, B., 1991, A comparison of tests for parameter instability: An examination of asymptotic local power, Manuscript (Department of Economics, University of Rochester, Rochester, NY).

Hansen, B., 1992, Testing for parameter instability in linear models, Journal of Policy Modeling 14, 517–533.

Hendry, D., 1984, Monte Carlo experimentation in econometrics, in: Z. Griliches and M.D. Intriligator, eds., Handbook of econometrics, Vol. II (North-Holland, Amsterdam) 939–976.

Jeong, J. and G.S. Maddala, 1992, A perspective on applications of bootstrap methods in econometrics, in: G.S. Maddala, ed., Handbook of statistics: Econometrics, Vol. II, forthcoming.

Park, S.K. and K.W. Miller, 1988, Random number generators: Good ones are hard to find, Communications of the ACM 32, 1192–1201.

Rayner, R.K., 1990, Bootstrapping p values and power in the first-order autoregression: A Monte Carlo investigation, Journal of Business and Economic Statistics 8, 251–263.

Stinchcombe, M.B. and H. White, 1993, Consistent specification testing with unidentified nuisance parameters using duality and Banach space limit theory, Discussion paper 93-14 (Department of Economics, University of California, San Diego, CA).

White, H., 1980, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, Econometrica 48, 817–838.