Bootstrap Tests in Multiple Structural Change Models

Jamel Jouini* **Mohamed Boutahar†**
GREQAM, Université de la Méditerranée
2 Rue de la Charité, 13002 Marseille, France

July 8, 2003

Abstract

Many time series in diverse fields of application might exhibit the phenomenon of structural change. Some asymptotic tests have been constructed in Bai and Perron (1998) allowing inference to be made about the presence of multiple structural changes. Bai and Perron (2000) carry out some simulations to analyze the adequacy of these tests based on their asymptotic distributions. They find that when allowing for serial correlation and/or different distributions of the data and the errors across segments, the tests show size distortions especially for small minimal number of observations in each segment. In this paper we propose the use of bootstrap methods to yield approximations to the distributions of the test statistics so as to correct the size of the tests and to compare the power of the asymptotic tests to the one of their bootstrap counterparts. We evaluate the performance of two alternative approximations to the finite sample distributions of these test statistics, one based on asymptotics and one based on the bootstrap. We focus on smaller samples and dynamic models and we allow for heterogeneity in the errors across segments. We find that the bootstrap solves the inference problem and appears more accurate than the asymptotic distribution since the error in rejection probability committed by bootstrap tests is very minimal. An other feature that characterizes our results is that the power of bootstrap tests is slightly smaller than the one of the corresponding asymptotic tests since as the intuition suggests, a test statistic with very small size distortions can have lower power.

JEL Classification: C12, C15, C22
Key words: Structural change; Bootstrap; Size; Power

*Corresponding author: Tel.: +33 4 91 14 07 23; fax: +33 4 91 90 02 27; e-mail: jouini@ehess.cnrs-mrs.fr.
†Tel.: +33 4 91 82 90 89; fax: +33 4 91 82 93 56; e-mail: boutahar@lumimath.univ-mrs.fr.

1 Introduction

In diverse fields of application, many time series might exhibit the phenomenon of structural change which is of capital importance in economics and econometrics. The problem consists in testing the null hypothesis of structural stability against the alternative of instability. Under the alternative hypothesis, we can have one or more than one break. The literature addressing the first issue is huge. Indeed, Chow (1960) was the first who considered a classic F test allowing to test for single structural change when the break point is known a priori. Quandt (1960) extends the analysis of Chow (1960) to the case of unknown break date. Indeed, he computes a sequence of Chow statistics for all possible break dates contained in a restricted interval since we cannot consider break dates too close to the boundaries of the sample, as there aren't enough observations to identify all the subsample parameters and the estimated break point is the one that maximizes the Chow test. In the same context, some important tests are the "supremum" tests of Andrews (1993) and the related "average" and "exponential" tests of Andrews and Ploberger (1994). They derive the asymptotic distributions which are useful to provide approximations to the finite sample distributions.

In the analysis of multiple structural change models, the most important contribution is the one of Bai and Perron (1998) who provide a comprehensive treatment of various issues. In particular, they consider tests and tabulate critical values using the derived asymptotic distributions. Bai and Perron (2000) and Jouini and Boutahar (2002) carry out some simulations to analyse the adequacy of these tests based on their asymptotic distributions. They find that the tests show size distortions especially for small values of the trimming when allowing for serial correlation and/or different distributions of the data and the errors across subsamples in the estimated regression models. Thus, the asymptotic distributions may be an unreliable guide to finite sample behavior and as a result the nominal levels of tests based on asymptotic critical values can be very different from the true levels. An alternative approximation is the bootstrap distribution that gives evidence on the adequacy of tests and often provides a tractable way to reduce or eliminate finite sample distortions of the sizes of statistical tests as suggested by Christiano (1992) and Diebold and Chen (1996) for the case of single structural change tests.

In the context of the use of the bootstrap to reduce the size distortions of tests, Rayner (1990) finds that the bootstrap approximation to the finite sample distribution of studentized statistics in an AR(1) model is more performant than the asymptotic approximation for sample sizes as small as $T = 5$ and for degrees of persistence as large as 0.99. Jeong and Maddala (1992) suggest that the careful treatment of the initial value when generating bootstrap samples is one reason for Rayner's success with the bootstrap. He draws the initial value from a stationary distribution instead of assuming it to be known. Diebold and Chen (1996) consider a simulation study to obtain better approximations to finite sample distributions for single structural change tests using the bootstrap procedure. Their analysis presents numerous limitations since they don't allow for multiple breaks and shifts in the

innovation variance across segments. In this paper we generalize their analysis so as to take these limitations into account. Indeed, we extend the analysis to multiple structural change tests allowing for heterogeneity in the errors across segments.

The purpose of this paper is to show the accuracy of the bootstrap methods and their ability to reduce the error in rejection probability committed by the tests and to present some limited results on the power that may be useful in designing an investigation of power. In section 2, we present the basic model and the estimation method permitting to estimate the regression coefficients and the break dates. Section 3 defines multiple structural change tests. In section 4, we introduce the bootstrap technique and the relating basic concepts being useful to carry out our simulation experiments. Monte Carlo evidence on the performance of both asymptotic and bootstrap approximations is given in section 5. This is the heart of the paper in which we execute the test size and power comparison. The results of Monte Carlo experiments show that, in many circumstances, getting the size right and achieving high power are different tasks. Indeed, the tests using the bootstrap approximation have very small size distortions and quite low power for some cases. On the other hand, using the asymptotic critical values, the tests have severe size distortions and much high power. Concluding comments are presented in section 6.

2 The Model and Estimators

Consider the following structural change model with m breaks:

$$y_t = x_t'\beta + z_t'\delta_j + u_t, \qquad t = T_{j-1}+1, \ldots, T_j, \tag{1}$$

for $j = 1, \ldots, m+1$, $T_0 = 0$ and $T_{m+1} = T$. y_t is the observed dependent variable, $x_t \in \Re^p$ and $z_t \in \Re^q$ are the vectors of regressors, β and δ_j are the corresponding vectors of coefficients with $\delta_i \neq \delta_{i+1}$ $(1 \leq i \leq m)$, and u_t is the disturbance that has a distribution $D\left(0, \sigma^2\right)$, with $\sigma^2 < \infty$. The break dates $(T_1, \ldots, T_m)$ are explicitly treated as unknown and for $i = 1, \ldots, m$, we have $\lambda_i = T_i/T$ with $0 < \lambda_1 < \cdots < \lambda_m < 1$. Note that this is a partial structural change model in the sense that β is not subject to shift and is effectively estimated using the entire sample. When $p = 0$, all the coefficients are subject to change and we then obtain a pure structural change model. The above multiple linear regression model may be expressed in matrix form as

$$Y = X\beta + \bar{Z}\delta + U, \tag{2}$$

where $Y = (y_1, \ldots, y_T)'$, $X = (x_1, \ldots, x_T)'$, $\delta = \left(\delta_1', \delta_2', \ldots, \delta_{m+1}'\right)'$, $U = (u_1, \ldots, u_T)'$, and $\bar{Z}$ is the matrix which diagonally partitions Z at the m-partition $(T_1, \ldots, T_m)$, i.e. $\bar{Z} = diag\left(Z_1, \ldots, Z_{m+1}\right)$ with $Z_i = \left(z_{T_{i-1}+1}, \ldots, z_{T_i}\right)'$. Bai and Perron (1998) impose some restrictions on the possible values of the break dates. Indeed, they define the following set for some arbitrary small positive number ε:

$$\Lambda_\varepsilon = \{(\lambda_1, \ldots, \lambda_m) ; |\lambda_{i+1} - \lambda_i| \geq \varepsilon, \lambda_1 \geq \varepsilon, \lambda_m \leq 1 - \varepsilon\}, \tag{3}$$

to restrict each break date to be asymptotically distinct and bounded from the boundaries of the sample.

The estimation method considered is that based on the least-squares principle proposed in Bai and Perron (1998). This method is described as follows. For each m-partition $(T_1, \ldots, T_m)$, denoted $\{T_j\}$, the associated least-squares estimates of β and δ_j are obtained by minimizing the sum of squared residuals $\sum_{i=1}^{m+1} \sum_{t=T_{i-1}+1}^{T_i} (y_t - x_t'\beta - z_t'\delta_i)^2$. Let $\hat{\beta}(\{T_j\})$ and $\hat{\delta}(\{T_j\})$ denote the resulting estimates. Substituting them in the objective function and denoting the resulting sum of squared residuals as $S_T(T_1, \ldots, T_m)$, the estimated break dates $\left(\hat{T}_1, \ldots, \hat{T}_m\right)$ are

$$\left(\hat{T}_1, \ldots, \hat{T}_m\right) = \arg \min_{(T_1, \ldots, T_m)} S_T(T_1, \ldots, T_m), \tag{4}$$

where the minimization is taken over all partitions $(T_1, \ldots, T_m)$ such that $T_i - T_{i-1} \geq [\varepsilon T]$.[1] The break point estimators are thus global minimizers of the objective function. Finally, the estimated regression parameters are the associated least-squares estimates at the estimated m-partition $\left\{\hat{T}_j\right\}$, i.e. $\hat{\beta} = \hat{\beta}\left(\left\{\hat{T}_j\right\}\right)$ and $\hat{\delta} = \hat{\delta}\left(\left\{\hat{T}_j\right\}\right)$. For our Monte Carlo experiments, we use the efficient algorithm developed in Bai and Perron (2003a) based on the principle of dynamic programming which allows global minimizers to be obtained using a number of sums of squared residuals that is of order $O\left(T^2\right)$ for any $m \geq 2$.

3 The Test Statistics

Several tests for structural change have been proposed in the econometrics literature. These tests can be classified in two groups: a) tests for a single structural change;[2] and b) tests for multiple structural breaks. Here we focus on tests for multiple breaks. In this context, Bai and Perron (1998) consider estimating multiple structural changes in a linear model and develop three tests which preclude the presence of trending regressors.

3.1 A test of structural stability versus a fixed number of changes

Bai and Perron (1998) first consider the sup F type test of structural stability against the alternative hypothesis that there is a known number of breaks k:

[1] $[\varepsilon T]$ is then interpreted as the minimal number of observations in each segment. From Bai and Perron (2003a), if the tests are not required and the estimation is the sole concern, the minimal number of observations in each segment can be set to any value greater than q.

[2] Diebold and Chen (1996) compare the performance of two alternative approximations to the finite-sample distributions of test statistics for single structural change, one based on asymptotics and one based on the bootstrap.

$$F_T(\lambda_1,\ldots,\lambda_k;q) = \frac{1}{T}\left(\frac{T-(k+1)q-p}{kq}\right)\frac{\hat{\delta}'R'\left(R\left(\bar{Z}'M_X\bar{Z}\right)^{-1}R'\right)^{-1}R\hat{\delta}}{SSR_k}, \tag{5}$$

where R is the conventional matrix such that $(R\delta)' = \left(\delta_1' - \delta_2', \ldots, \delta_k' - \delta_{k+1}'\right)$, $M_X = I - X\left(X'X\right)^{-1}X'$, and SSR_k is the sum of squared residuals under the alternative hypothesis, which depends on the break dates $(T_1,\ldots,T_k)$. The sup F type test statistic is then defined as

$$\sup F_T(k;q) = \sup_{(\lambda_1,\ldots,\lambda_k)\in\Lambda_\varepsilon} F_T(\lambda_1,\ldots,\lambda_k;q) = F_T\left(\hat{\lambda}_1,\ldots,\hat{\lambda}_k;q\right). \tag{6}$$

where the break fraction estimates $\hat{\lambda}_1,\ldots,\hat{\lambda}_k$ minimize the global sum of squared residuals and are also obtained from the maximization of the following F statistic:[3]

$$F_T(\lambda_1,\ldots,\lambda_k;q) = \frac{1}{T}\left(\frac{T-(k+1)q-p}{kq}\right)\hat{\delta}'R'\left(R\tilde{V}\left(\hat{\delta}\right)R'\right)^{-1}R\hat{\delta},$$

where $\tilde{V}\left(\hat{\delta}\right) = \left(\bar{Z}'M_X\bar{Z}/T\right)^{-1}$ is the covariance matrix of $\hat{\delta}$ assuming spherical errors. Different versions of these tests can be obtained depending on the assumptions made with respect to the distribution of the regressors and the errors across segments (see, e.g., Bai and Perron, 2000 and 2003a).

3.2 A test of structural stability versus an unknown number of breaks

Bai and Perron (1998) also consider tests of no structural change against an unknown number of breaks given some upper bound M for m. The following new class of tests is called double maximum tests and is defined for some fixed weights $\{a_1,\ldots,a_M\}$ as

$$\begin{aligned} D\max F_T(M,q,a_1,\ldots,a_M) &= \max_{1\le m\le M} a_m \sup_{(\lambda_1,\ldots,\lambda_m)\in\Lambda_\varepsilon} F_T(\lambda_1,\ldots,\lambda_m;q) \\ &= \max_{1\le m\le M} a_m\ F_T\left(\hat{\lambda}_1,\ldots,\hat{\lambda}_m;q\right). \end{aligned} \tag{7}$$

The weights $\{a_1,\ldots,a_M\}$ reflect the imposition of some priors on the likelihood of various numbers of structural breaks. Firstly, they set all weights equal to unity, i.e. $a_m = 1$ and label this version of the test as $UD\max F_T(M,q)$. Then, they consider a set of weights such that the marginal p-values are equal across values of m. The weights are then defined as

$$a_m = \begin{cases} 1, & \text{if } m = 1, \\ c(q,\alpha,1)/c(q,\alpha,m), & \text{if } m > 1, \end{cases}$$

where α is the significance level of the test and $c(q,\alpha,m)$ is the asymptotic critical value of the test $\sup_{(\lambda_1,\ldots,\lambda_m)\in\Lambda_\varepsilon} F_T(\lambda_1,\ldots,\lambda_m;q)$. This version of the test is denoted as $WD\max F_T(M,q)$.

[3] The break date estimators are consistent even in the presence of serial correlation.

3.3 A sequential test

The last test developed by Bai and Perron (1998) is a sequential test of l versus $l+1$ structural changes:

$$\begin{aligned} \sup F_T(l+1|l) &= \left\{ S_T\left(\hat{T}_1,\ldots,\hat{T}_l\right) \right. \\ &\quad \left. - \min_{1\leq i\leq l+1} \inf_{\tau\in\Lambda_{i,\eta}} S_T\left(\hat{T}_1,\ldots,\hat{T}_{i-1},\tau,\hat{T}_i,\ldots,\hat{T}_l\right) \right\} / \hat{\sigma}^2, \end{aligned} \tag{8}$$

where

$$\Lambda_{i,\eta} = \left\{ \tau; \hat{T}_{i-1} + \left(\hat{T}_i - \hat{T}_{i-1}\right)\eta \leq \tau \leq \hat{T}_i - \left(\hat{T}_i - \hat{T}_{i-1}\right)\eta \right\},$$

$S_T\left(\hat{T}_1,\ldots,\hat{T}_{i-1},\tau,\hat{T}_i,\ldots,\hat{T}_l\right)$ is the sum of squared residuals resulting from the least-squares estimation from each m-partition $(T_1,\ldots,T_m)$, and $\hat{\sigma}^2$ is a consistent estimator of σ^2 under the null hypothesis. The test amounts to the application of $(l+1)$ tests of the stability null hypothesis against the alternative hypothesis of a single break. It is applied to each segment $\left[\hat{T}_{i-1}+1, \hat{T}_i\right]$ for $i = 1,\ldots,l+1$, and with $\hat{T}_0 = 0$ and $\hat{T}_{l+1} = T$.[4] We reject the null hypothesis and we conclude in favor of a model with $(l+1)$ structural breaks if the sum of squared residuals obtained from the estimated model with l changes is sufficiently larger than the overall minimal value of the sum of squared residuals (over all segments where an additional change is included) and the break point thus selected is the one associated with this overall minimum.[5]

The asymptotic distributions of these three tests are derived in Bai and Perron (1998) and asymptotic critical values are tabulated in Bai and Perron (1998, 2003b) for $\varepsilon = 0.05$ $(M = 9)$, 0.10 $(M = 8)$, 0.15 $(M = 5)$, 0.20 $(M = 3)$, and 0.25 $(M = 2)$. Note that these asymptotic distributions are derived without taking the trending regressors into account. Hence, the asymptotic critical values so found are not valid when we allow for the presence of a trend in the regression. A future investigation that can be carried out consists in using the bootstrap procedure to find approximations to the distributions of test statistics when we allow for the presence of trending regressors. This issue is beyond the scope of the paper.

4 Bootstrap Tests

The bootstrap technique was invented and introduced by Efron (1979). It is a simple way allowing to find an approximation of the distribution of a test statistic or quantities that are

[4] Given the estimates $\hat{T}_i$, we require that the break fractions $\hat{\lambda}_i = \hat{T}_i/T$ converge to their true values at rate T. Hence, the estimates $\hat{T}_i$ need not be obtained by a global minimization of the sum of squared residuals, we can also use the sequential one-at-a-time estimates which imply break fractions that converge at rate T (see, e.g., Bai, 1997).

[5] In other words, we conclude in favor of a model with $(l+1)$ changes if the overall maximal value of the $\sup F_T(1;q)$ (over all segments where an additional break point is included) is sufficiently large and the break date thus chosen is the one associated with this overall maximum.

very hard, or even impossible to compute analytically. The basic idea is a sort of Monte Carlo experiment in which we create a new sample by drawing the error terms from the empirical distribution function (EDF) of their sample counterparts. This bootstrap sample serves to calculate a bootstrap test statistic $\hat{\tau}^*$ in exactly the same way as the real sample was used to compute $\hat{\tau}$. If we repeat the bootstrap procedure B times, then we can estimate a bootstrap P value by the proportion of bootstrap samples that yield a statistic greater than $\hat{\tau}$. For a one-tailed test with a rejection region in the upper tail, the bootstrap P value is as follows:

$$\hat{p}^*(\hat{\tau}) \equiv \frac{1}{B}\sum_{j=1}^{B} I\left(\hat{\tau}_j^* > \hat{\tau}\right), \tag{9}$$

where $I\left(\hat{\tau}_j^* > \hat{\tau}\right)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise (see, e.g., Davidson and MacKinnon, 1993).

The main reason for using bootstrap tests is that the asymptotic tests may in finite samples be biased in the sense that they have actual sizes that differ from their nominal ones. A main feature of bootstrap tests is that, under certain conditions, their actual sizes will converge to the true ones faster than asymptotic tests and at times converge considerably faster. In other words, if the sample size is T, the error in rejection probability (size distortion) committed by an asymptotic test is, in general, of order $T^{-1/2}$ for a one-tailed test and of order T^{-1} for a two-tailed test. But the use of the bootsrap can reduce this order by a factor of $T^{-1/2}$, T^{-1}, or even more; see Hall (1992) for a very full discussion, based on Edgeworth expansions, of the extent to which asymptotic refinements are available in different contexts.

In the literature, many works study via simulations the behaviour, in finite samples, of different test statistics: the experimental results show that if the statistic is an asymptotic pivot, the bootstrap often reduces the size distortions of tests based on the asymptotic distribution, see Horowitz (1997).

4.1 Parametric Bootstrap

The parametric bootstrap uses a fully specified parametric model, which means that each set of parameter values defines just one data-generating process (DGP). The first step in constructing a parametric bootstrap DGP is to estimate by the ordinary least-squares (OLS) method the model under the null hypothesis yielding the restricted parameter estimates. Note that for the $\sup F_T(k;q)$, $UD\max$ and $WD\max$ tests and under the null hypothesis, the number of breaks m is equal to 0 and consequently the parametric bootstrap samples are generated according to the model

$$y_t^* = x_t'\hat{\beta} + z_t'\hat{\delta}_1 + u_t^*, \qquad t = 1, \ldots, T, \tag{10}$$

where $u_t^* \sim \hat{D}\left(0, \hat{\sigma}^2\right)$, and $\hat{\sigma}^2 = SSR/\left(T - (q+p)\right)$, with SSR is the sum of squared residuals. $\hat{\beta}$, $\hat{\delta}_1$, and $\hat{\sigma}^2$ are consistent least-squares estimates of the parameters β, δ_1, and σ^2. On the other hand, for the sequential test $\sup F_T(l+1|l)$, the null hypothesis is characterized

by the presence of l changes, i.e. $m = l$. In this case, the parametric bootstrap DGP is then given by

$$y_t^* = x_t'\hat{\beta} + z_t'\hat{\delta}_j + u_t^*, \qquad t = \hat{T}_{j-1} + 1, \ldots, \hat{T}_j, \tag{11}$$

for $j = 1, \ldots, l+1$, $\hat{T}_0 = 0$, $\hat{T}_{l+1} = T$, and $u_t^* \sim \hat{D}\left(0, \hat{\sigma}^2\right)$. $\hat{\beta}$, $\hat{\delta}_1, \ldots, \hat{\delta}_{l+1}$, $\hat{\sigma}^2$, and $\hat{T}_1, \ldots \hat{T}_l$ are consistent least-squares estimates of the parameters β, $\delta_1, \ldots, \delta_{l+1}$, σ^2, and $T_1, \ldots T_l$.

4.2 Nonparametric Bootstrap

If we didn't know the true error distribution, we would have to find some way to estimate this distribution. Note that under the null hypothesis, the OLS residual vector for the restricted model is a consistent estimator of the error vector. This means that, if the errors are mutually independent drawings from the error distribution, then so are the residuals, asymptotically. We know that the EDF of the residuals is a consistent estimator of the unknown cumulated distribution function (CDF) of the error distribution. Thus, if we draw bootstrap errors from the EDF of the residuals, we are drawing them from a distribution that tends to the true error distribution as $T \to \infty$. This procedure is called resampling.

We now examine two ways for generating the bootstrap error terms u_t^* by resampling with replacement:

1. The bootstrap error vector u^* is generated by resampling with replacement from the residual vector $\tilde{u} = \{\tilde{u}_t\}_{t=1}^T$, where

$$\tilde{u}_t = \sqrt{\frac{T}{T-(q+p)}}\left(\hat{u}_t - \frac{1}{T}\sum_{s=1}^{T}\hat{u}_s\ \iota\right), \tag{12}$$

   where ι is the unit vector.[6] This argument gives rise to an improved bootstrap DGP since the bootstrap error terms are drawn from a distribution with mean zero and variance $\hat{\sigma}^2$.

2. We now draw the bootstrap error vector u^* by resampling with replacement from the residual vector $\tilde{u} = \{\tilde{u}_t\}_{t=1}^T$, where

$$\tilde{u}_t = \frac{\hat{u}_t}{\sqrt{1-h_t}} - \frac{1}{T}\sum_{s=1}^{T}\frac{\hat{u}_s}{\sqrt{1-h_s}}, \tag{13}$$

   $h_t = V_t\left(V'V\right)^{-1}V_t'$, and $V = \left[X, \bar{Z}\right]$. This solution consists in utilizing the fact that $E\left(\hat{u}_t^2\right) = \sigma^2\left(1-h_t\right)$. This argument also improves the bootstrap DGP since the terms $\tilde{u}_t$ have the same variance, and are recentered.

[6] When the regression contains a constant term, the residuals are by construction centered and we don't need to recenter them.

Note that the residual vector $\hat{u}$ is obtained under the null hypothesis. If we adopt these resampling procedures, the nonparametric bootstrap DGP is the same as the parametric bootstrap one with the modification that in this case the bootstrap errors are drawn by resampling with replacement from the above residual vectors, i.e. $u_t^* \sim \text{EDF}(\tilde{u}_t)$, where $\text{EDF}(\tilde{u}_t)$ denotes the distribution that assigns probability $1/T$ to each of the elements of the vector $\tilde{u}$.

To keep the precision gain of Davidson and MacKinnon (1999) in the framework of non-parametric bootstrap, the test statistic must asymptotically be independent of the parameters of the model and the EDF of the residuals. The authors show that this condition is satisfied when we use the EDF of the residuals obtained under the null or the alternative hypothesis. Only the Monte Carlo experiments show that the use of the residuals of the restricted model provides a very sensible precision gain (see, e.g., Van Giersbergen and Kiviet, 1994; Li and Maddala, 1993; and Nankervis and Savin, 1994).

4.3 The Number of Bootstrap Samples

Suppose that we wish to compute a bootstrap P value. Then we need to draw B samples from the bootstrap DGP. If the bootstrap test is at level α, then the number of bootstrap simulations should be chosen to satisfy the condition that $\alpha\,(B+1)$ is an integer (see, e.g., Davidson and MacKinnon (2000) for more details). This choice of B deletes all eventual bias of the bootstrap estimation of a P value. Davidson and MacKinnon (2000) propose an alternative approach to choose B. It is a bit more complicated but can save a lot of computer time and allows to minimize the power loss.

It is important that B be sufficiently large because the power of a bootstrap test depends on the number of bootstrap samples. Indeed, the ability of a bootstrap test to reject a false null hypothesis declines as B becomes smaller. A reasonable rule of thumb is that power loss will very rarely be a problem when $B = 999$, and that it will never be a problem when $B = 9999$, see Davidson and MacKinnon (2003).

In the same context of choice of B, when the null hypothesis is true, B can safely be small, because we are not concerned about power at all. Similarly, when the null hypothesis is false and test power is extremely high, B does not need to be large, because power loss is not a serious issue. However, when the null hypothesis is false and test power is moderately high, B needs to be large in order to avoid power loss. Note that the procedure proposed by Davidson and MacKinnon (2000) tends to make B small when it can safely be small and large when it needs to be large.

4.4 The Bootstrap when the Null Hypothesis is False

We know that to obtain an exact finite sample distribution, it is preferred to construct the bootstrap DGP under the null hypothesis.[7] When this hypothesis is false, a reasonable choice is the pseudo-true null, which is the DGP that satisfies the null hypothesis using pseudo-true values for the unknown parameters obtained under the incorrect assumption that the null hypothesis is true. Consequently, bootstrap sampling when the null hypothesis is false is equivalent to sampling from the null hypothesis model with pseudo-true parameter values; for more details, see Horowitz (1994, 1997).

4.5 The Bootstrap Procedure

The bootstrap P value can be approximated according to the following algorithm:

1. We compute the test statistic (see Section 3), say $\hat{\tau}$, in the usual way.

2. We estimate the model under the null hypothesis by the least-squares principle and we construct a bootstrap DGP as in sections 4.1 or 4.2 to generate a data vector y^*. Based on these data, we compute a bootstrap test statistic $\hat{\tau}^*$ in exactly the same way as the real sample was used to compute $\hat{\tau}$.

3. We draw B bootstrap samples from the DGP constructed in the second step so as to obtain B bootstrap statistics $\left\{\hat{\tau}_j^*\right\}_{j=1}^B$. The EDF constructed using these B realizations is an approximation of the bootstrap distribution: $\hat{G}^*(x) = B^{-1}\sum_{j=1}^B I\left(\hat{\tau}_j^* \leq x\right)$.

4. The bootstrap P value is then estimated by the proportion of bootstrap samples that yield a statistic $\hat{\tau}_j^*$ greater than $\hat{\tau}$ as in (9) since we have one-tailed tests with a rejection region in the upper tail.

4.6 Bootstrap Refinements

Davidson and MacKinnon (1999) show that an extra refinement which is of order $T^{-1/2}$ in most cases, is quite generally available if the test statistic is asymptotically independent of the bootstrap DGP. The parametric bootstrap and in many cases the nonparametric bootstrap satisfy this condition if the parameters are estimated under the null hypothesis. Hence this result, together with existing results, suggests that the bootstrap tests can be more accurate than asymptotic tests by a full order of T^{-1} in many circumstances.

Note that the main result of the power analysis found in Davidson and MacKinnon (1996) is that the bootstrap tests with correct sizes can also often be shown to have the same power properties as their asymptotic counterparts.

[7] The idea of constructing the bootstrap DGP under the null hypothesis is proposed by Beran and Srivastava (1985) and Beran (1986).

5 Monte Carlo Analysis

Bai and Perron (2000) and Jouini and Boutahar (2002) investigate via simulations the performance of the above-mentioned tests considering various data-generating processes that allow for general conditions on the data and the errors including differences across segments. The results show that if serial correlation and/or heterogeneity in the data and the errors across segments are not allowed in the estimated regression model and not present in the DGP, the use of any value of the trimming ε asymptotically leads to tests with adequate sizes. However, if such features are allowed in the estimated regression, a higher value of the trimming is needed. They find that correcting for heterogeneity in the distribution of the data or the errors and for serial correlation, and a large magnitude of change improve the power of the tests. Diebold and Chen (1996) compare the performance of two approximations to the finite sample distributions of test statistics for single structural change developed in Andrews (1993), one based on asymptotics and the other based on the bootstrap. They find that the bootstrap works well for AR(1) processes. In contrast, this paper focusses on multiple structural change tests.

In this section we report some Monte Carlo experiments to investigate the size and the power of the tests for an AR(1) process. We consider three ways of approximating the finite sample distributions of the statistics. The first approximation is Bai and Perron' (1998) asymptotic distribution, the second approximation is the parametric bootstrap distribution and the last one is the nonparametric bootstrap distribution. The data-generating processes used to generate the bootstrap samples are defined in sections 4.1 and 4.2.

For each procedure, we explore the relationship between actual and nominal test size ($\alpha = 5\%$) and we are also interested in computing the true power of the tests. Note that the asymptotic standard error of the empirical size $\hat{\alpha}$ is $(\alpha(1-\alpha)/N)^{1/2}$ which is estimated by $(\hat{\alpha}(1-\hat{\alpha})/N)^{1/2}$ that is decreasing in N, the number of Monte Carlo replications. Finally, in the estimated regression model we allow for heterogeneity in the errors across segments.

The Monte Carlo experiments were carried out using a program written in GAUSS with GAUSS pseudo-random number generators.

5.1 Test Size

We are interested in estimating the probability of rejecting the null hypothesis when it is true for each of the test procedures outlined above, and to see how those probabilities vary with the sample size T, the persistence as measured by the coefficient on the lagged dependant variable, the trimming and the distribution of the errors across segments.

The theoretical results of Davidson and MacKinnon (1999) show that the bootstrap tests should be performant. But in many circumstances, in small samples, these tests can show size distortions. Here we show that the bootstap procedures correct the size of the tests and quasi-perfectly solve the inference problem since the error in rejection probability committed by bootstrap tests is very minimal.

The number of Monte Carlo replications is set at $N = 500$ and we choose $B = 199$, a smaller value than we would recommend using in practice, in order to reduce the costs of doing the Monte Carlo experiments. This choice of B does not materially affect the results of the experiments because experimental errors tend to cancel out across replications. An other argument that justifies our choice of B is that we are interested in doing inference under the null hypothesis. Increasing the number of bootstrap samples beyond 199 had little effect on the results of the experiments. It follows from Hall (1986) that the error in rejection probability made by a test using bootstrap-based critical values is $O\left(T^{-(j+1)/2}\right)$ (for some integer $j \geq 1$), regardless of the number of bootstrap samples used to estimate the bootstrap critical values. Consequently, the bootstrap methods provide an asymptotic refinement even with small numbers of bootstrap samples.

Under the null hypothesis of structural stability, our model is a Gaussian zero-mean first-order process:

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

where $u_t \sim NID\left(0, \sigma^2 = 1\right)$. We call this the no-break model. The bootstrap samples must be constructed recursively because of the presence of a lagged dependent variable. This is necessary because y_t^* must depend on y_{t-1}^* and not on y_{t-1} from the observed data. The recursive rule for generating a bootstrap sample is

$$y_t^* = \hat{\rho} y_{t-1}^* + u_t^*, \qquad t = 1, \ldots, T, \tag{14}$$

where $y_0^* = y_0$ and $\hat{\rho}$ is a consistent least-squares estimate of ρ. Note that for the parametric bootstrap $u_t^* \sim NID\left(0, \hat{\sigma}^2\right)$, where $\hat{\sigma}^2$ is a consistent estimator of σ^2 whereas for the nonparametric one the error vector u^* is generated by resampling with replacement from the residual vector given in Eq. 12. As we see every bootstrap sample is conditional on the observed value of y_0. This initialization is certainly the most convenient and it may, in many circumstances, be the only method that is feasible. The bootstrap P values are computed using the four-step procedure described in section 4.5.

5.1.1 The asymptotic approximation

The results are reported in tables 1-5. The examination of the finite sample properties of the asymptotic approximation illustrates that the size distortions of the tests are very severe and huge even often for some selected parameter values for which the asymptotic tests may not encounter problems. Indeed, for $T = 250$, $\varepsilon = 0.20$ and $\rho = 0.75$, the maximum difference between the actual and nominal sizes is 0.078 for the 0.05-level test when we allow for no heterogeneity in the errors across segments. As the intuition suggests, the tests tend to overreject especially when we allow for heterogeneity in the errors across segments. The tendency of the tests to overreject decreases as the sample size and/or the trimming increases. The serial correlation also affects the actual size of the tests. Indeed, a nearly

unit-root greatly affects the empirical size even for large samples and when we allow for or not different variances of the errors across segments. For example, the UD max and WD max tests almost reject 90% of the time the stability null hypothesis for $\rho = 0.95$ and $T = 250$. Note that allowing for different variances of the errors induces substantial size distortions unless the trimming is large. When the degree of persistence is small and for all values of the trimming, the asymptotic tests slightly underreject for any sample size and when there is not heterogeneity in the errors across subsamples since the actual test size tends to be pushed downward relative to nominal size. When the degree of persistence is small or equal to 0.5, the distribution of the errors greatly affects the empirical size of the tests whereas when serial correlation is high, this distribution has a slight effect on the size performance of the tests. Asymptotically, the tests almost have the right size for $\rho = 0.5$, $\varepsilon \geq 0.10$ and when the errors have the same variance across segments.

Since the finite sample performance of the asymptotic tests is not good, it is judicious to find an other approximation to the finite sample distribution so as to correct the deficiencies of the asymptotic distribution. This leads us to the bootstrap methods. This choice is justified by the fact that the bootstrap approximations to finite sample distributions in econometrics are adequate. This is shown below.

5.1.2 The bootstrap procedures

The results are presented in tables 6 and 7. The bootstrap techniques consistently outperform the asymptotic distribution in approximating the finite sample distribution since they dramatically reduces the differences between the actual and nominal sizes even for some chosen parameter values for which the bootstrap tests may encounter problems. Indeed, with the nonparametric bootstrap approximation and $\varepsilon = 0.05$, the maximum difference between the empirical and nominal sizes is 0.016 for the 0.05-level test when the errors are not heterogenous across segments. Thus, for our examined model, the bootstrap eliminates the problem of excessive finite sample size of the tests. The bootstrap performance is nearly perfect even when we allow for heterogeneity in the errors across segments regardless of the value of the nuisance parameter ρ with slight superiority to the nonparametric bootstrap. Thus we are now able to choose large values of the degree of persistence to do correct inferences. Consequently, there is a marked contrast between the performance of asymptotic tests and their bootstrap counterparts. Indeed, the use of tests based on the conventional asymptotic approximation is associated with severe size distortions which heavily depend on ρ. For $T = 50$, $\rho \leq 0.50$, $\varepsilon \geq 0.10$ and in the case of no heterogeneity in the errors, the sizes obtained with asymptotic critical values are quite close to the nominal size but the bootstrap-based size is closer to the nominal size than is the size based on asymptotic critical values.

The results show that the bootstrap methods improve on asymptotic approximations and quasi-perfectly solve the inference problem since the error in rejection probability is very minimal even when we allow for shifts in the innovation variance across subsamples. Thus, the bootstrap distribution is a good approximation to the finite sample distribution when

the sample size is not large, even when the parametric serial correlation is high and the distribution of the errors differs across segments. For these procedures, there is no point in exploring large values of the sample size T. Since the bootstrap is very accurate (in the sense that the error in rejection probability made by the bootstrap tests is very minimal) with samples of 50 in the cases that are investigated here, there is little to be gained by adding experiments with larger samples. Moreover, the sample sizes were limited by the very long computing times that are involved. Note that the samples used in empirical applications often are much larger than those used in our experiments.

5.2 Test Power[8]

In this case we are interested in estimating the frequency of rejecting the null hypothesis when it is false for all the above-mentioned test procedures. We attempt to see how these frequencies vary with the trimming, the break size and the distribution of the errors across segments. We show that what is important to have good power properties is not the sample size but the magnitude of change.

Here the null hypothesis is false, then we must choose a large number of bootstrap samples so as to avoid the power loss which can be generated by the tests in many circumstances. But, the power of the tests is high (see, e.g., Bai and Perron, 2000; and Jouini and Boutahar, 2002) and hence B does not need to be large, because power loss is not a serious issue (see, Davidson and MacKinnon, 2000). Thus, we choose $B = 199$, and the number of Monte Carlo replications is set at $N = 500$, a smaller value than we would recommend using in practice because Monte Carlo experiments with bootstrapping are much more time-consuming than experiments that do not involve bootstrapping. Then long computing time presents a potentially serious barrier to investigating the power of structural break tests using the bootstrap methods. As we will show below, this choice of N does not materially affect the results of the experiments.

5.2.1 The Case of One Break

We now look into the simulation results when there is effectively one break in the DGP which is as follows:

$$y_t = \begin{cases} \rho_1 y_{t-1} + u_t, & 1 \leq t \leq T_1, \\ \rho_2 y_{t-1} + u_t, & T_1 < t \leq T, \end{cases}$$

where $u_t \sim NID\left(0, \sigma^2\right)$ and the break date is $T_1 = T/2$. We call this the one-break model. In our simulation experiments, σ^2 takes value 1 and the sample size is set at $T = 100$. We consider small and large break sizes. For the small one, $\rho_1 = 0.50$ and $\rho_2 = 0.95$ and for the large one, $\rho_1 = 0.05$ and $\rho_2 = 0.95$. For the $\sup F_T(k)$, $UD\max$ and $WD\max$ tests,

[8] Davidson and MacKinnon (1996) find that the bootstrap tests with correct sizes can also often be shown to have the same power properties as their asymptotic counterparts.

the parametric and nonparametric bootstrap samples must be constructed under the null hypothesis of stability according to the recursive rule given in Eq. 14 using the pseudo-true parameter values as defined in section 4.4. For the $\sup F_T(l+1|l)$ tests, we only compute the test statistics for any $l = 1, 2$ for time-computing reason. For $l = 1$, the bootstrap DGP is constructed under the null of one break[9] according to the following recursive rule:

$$y_t^* = \begin{cases} \hat{\rho}_1 y_{t-1}^* + u_t^*, & 1 \leq t \leq \hat{T}_1, \\ \hat{\rho}_2 y_{t-1}^* + u_t^*, & \hat{T}_1 < t \leq T, \end{cases} \tag{15}$$

where $\hat{\rho}_1$ and $\hat{\rho}_2$ are the estimators of the regression coefficients ρ_1 and ρ_2, and $\hat{T}_1$ is the estimator of the break date T_1. These estimates are obtained under the correct assumption that the null hypothesis is true. For $l = 2$, the bootstrap samples are constructed under the null of two breaks using the pseudo-true parameter values as defined in section 4.4 according to the following recursive rule:

$$y_t^* = \begin{cases} \hat{\rho}_1 y_{t-1}^* + u_t^*, & 1 \leq t \leq \hat{T}_1, \\ \hat{\rho}_2 y_{t-1}^* + u_t^*, & \hat{T}_1 < t \leq \hat{T}_2, \\ \hat{\rho}_3 y_{t-1}^* + u_t^*, & \hat{T}_2 < t \leq T, \end{cases} \tag{16}$$

where $\hat{\rho}_1$, $\hat{\rho}_2$ and $\hat{\rho}_3$ are the estimators of the regression coefficients ρ_1, ρ_2 and ρ_3, and $\hat{T}_1$ and $\hat{T}_2$ are the estimators of the break dates T_1 and T_2. These estimators are then obtained under the uncorrect assumption that the null hypothesis is true.

For the two rules $y_0^* = y_0$ and for the parametric bootstrap $u_t^* \sim NID\left(0, \hat{\sigma}^2\right)$, where $\hat{\sigma}^2$ is a consistent estimator of σ^2 whereas for the nonparametric one the error vector u^* is generated by resampling with replacement from the residual vector given in Eq. 12.

The asymptotic approximation The results are presented in tables 8.a and 8.b. Table 8.a presents a case with a small change in the persistence as measured by the coefficient on the lagged dependent variable whereas table 8.b reports results with a large break size. The magnitude of change has an effect on the results of the tests which have high power for large magnitude. The heterogeneity in the errors across segments has a slight positive effect on the power of the tests for small magnitude of change. The power of the $\sup F_T(k)$ test is almost invariant as k increases. However, both UD max and WD max tests have power as high as the case with $k = 1$ which gives the highest power for large magnitude of change. When the break size is small, both the UD max and WD max tests have the highest power. The value of the trimming ε has a slight effect on the power of the tests especially for a small break size. We have carried out some other experiments in which we have increased the sample size for the same small break size.[10] We have remarked that this induces a negligible increase in power compared to the case of large break size. Hence, we can conclude that what is important

[9] For this case, the null hypothesis is satisfied since the DGP effectively contains one break. Hence, the construction of the bootstrap DGP is easy using the estimators of the unknown parameters.

[10] The results are not reported here and are available upon request from the authors.

to have good power properties is not the size of the sample but the magnitude of change. The power of the $\sup F_T(l+1|l)$ tests is very low since the true DGP is a one-break model and consequently the tests have not a tendency to reject the null hypothesis. To conclude it may seem unsurprising that the $\sup F_T(k)$, $UD\max$ and $WD\max$ tests have high power especially for large break size since they have severe size distortions for some cases.

The bootstrap procedures The results are reported in tables 9.a and 10.a for the small break size and in tables 9.b and 10.b for the large one. The break size greatly affects the power of the tests. For the large magnitude of change, the powers are different from the ones obtained with asymptotic critical values for small trimming. The powers increase as the trimming increases and are affected by the distribution of the errors across segments for small values of the trimming. As the intuition suggests, the bootstrap-based powers are smaller than those obtained with asymptotic critical values especially for small trimming. This result may seem unsurprising since a test statistic with very small size distortions can have lower power. A future research that may be investigated consists in focussing on the emphasis on getting the size right without paying much attention to power for small values of the trimming ε.

6 Conclusion

This paper has presented bootstrap methods and illustrated their ability to overcome the problem of incorrect finite sample size of multiple structural change tests for dynamic models. The paper also provides some limited results on the power of the tests. These results show that getting the finite sample size right and obtaining high power are different objectives even for large break size especially when the errors are heterogenous across subsamples and the trimming is small. Hence achieving one objective does not insure that the other is also achieved. The results of the tests are affected by the change of some factors and the nonparametric bootstrap appears to be slightly more performant than the parametric one for structural change tests applied to AR(1) processes. These conclusions are of course subject to the framework specified here.

Our Monte Carlo simulations were limited because experiments with bootstrapping are much more time-consuming than the ones that do not involve bootstrapping. Hence the required computations cannot be carried out quickly. The results are of interest not only from the perspective of using recent tests to test for multiple structural changes, but also from the perspective of providing evidence on the accuracy and adequacy of bootstrap methods to eliminate the problem of size distortions and of designing future investigations of power of the tests.

References

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

[2] Andrews, D. W. K. and W. Ploberger (1994), "Optimal Tests when a Nuisance Parameter is Present Only Under the Alternative," *Econometrica*, **62**, 1383-1414.

[3] Bai, J. (1997), "Estimating Multiple Breaks one at a Time," *Econometric Theory*, **13**, 315-352.

[4] Bai, J. and P. Perron (1998), Estimating and Testing Linear Models with Multiple Structural Changes, *Econometrica*, **66**, 47-78.

[5] Bai, J. and P. Perron (2000), Multiple Structural Change Models: A Simulation Analysis, unpublished manuscript, Department of Economics, Boston University.

[6] Bai, J. and P. Perron (2003a), Computation and Analysis of Multiple Structural Change Models, *Journal of Applied Econometrics*, **18**, 1-22.

[7] Bai, J. and P. Perron (2003b), Critical Values for Multiple Structural Change Tests, *Econometrics Journal*, **1**, 1-7.

[8] Beran, R. (1986), "Simulating Power Functions," *Annals of Statistics*, **14**, 151-173.

[9] Beran, R. and Srivastava (1985), "Bootstrap Tests and Confidence Regions for Functions of a Covariance Matrix," *Annals of Statistics*, **13**, 95-115.

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

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

[12] Davidson, R. and J. MacKinnon (1993), *Estimation and Inference in Economics*. Oxford University Press, New York.

[13] Davidson, R. and J. MacKinnon (1996), "The Power of Bootstrap and Asymptotic Tests," *Queen's University Institute for Economic Research, Discussion Paper 937.*

[14] Davidson, R. and J. MacKinnon (1999), "The Size Distortion of Bootstrap Tests," *Econometric Theory*, **15**, 361-376.

[15] Davidson, R. and J. MacKinnon (2000), "Bootstrap Tests: How many Bootstraps," *Econometric Reviews*, **19,** 55-68**.**

[16] Davidson, R. and J. MacKinnon (2003), *Econometric Theory and Methods***.** Oxford University Press, New York.

[17] Diebold, F. X. and C. Chen (1996), "Testing Structural Stability with Endogenous Break Point: A Size Comparison of Analytic and Bootstrap Procedures," unpublished manuscript, Department of Economics, University of Pennsylvania.

[18] Efron, B. (1979), "Bootstrap Methods; Another Look at the Jacknife," *Annals of Statistics*, **7**, 1-26.

[19] Hall, P. (1986), "On the Number of Bootstrap Simulations Required to Construct a Confidence Interval," *Annals of Statistics*, **14**, 1453-1462.

[20] Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics, New York: Springer Verlag.

[21] Horowitz, J. L. (1994), "Bootstrap-Based Critical Values for the Information Matrix Test," *Journal of Econometrics*, **61**, 395-411.

[22] Horowitz, J. L. (1997), "Bootstrap Methods in Econometrics: Theory and Numerical Performance," *In Advances in Economics and Econometrics: Theory and Applications*, Volume **3**, pp. 188-222. David M. Kreps and Kenneth F. Wallis (eds), Cambridge, Cambridge University Press.

[23] Jeong, J. and G. S. Maddala (1992), "A Perspective on Applications of Bootstrap Methods in Econometrics," *In G. S. Maddala, ed.*, Volume II, Handbook of Statistics: Econometrics.

[24] Jouini, J. and M. Boutahar (2002), "L'étude des modèles avec changements structurels," *Document de Travail n°* 02C01, *GREQAM*, *Université de la Méditerranée*, Marseille.

[25] Li, H. and G. S. Maddala (1993), "Bootstrapping Cointegrating Regressions,". Paper presented at the Fourth Meeting of the European Conference Series in Quantitative Economics and Econometrics.

[26] Nankervis, J. C. and N. E. Savin (1994), "The Level and Power of the Bootstrap-t Test in the AR(1) Model With Trend," unpublished manuscript, Department of Economics, University of Surrey and University of Lowa.

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

[28] 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.

[29] Van Giersbergen, N. P. A. and J. F. Kiviet (1994), "How to Implement Bootstrap Hypothesis Testing in Static and Dynamic Regression Model," Discussion Paper TI 94-130, Amsterdam: Tinbergen Institute. Paper presented at *ESEM*'94 and $\mathbf{EC^2}$'93.

Table 1. Empirical size of the asymptotic tests for T = 50

$T = 50$		No heterogeneity in the errors					Heterogeneity in the errors				
		ρ					ρ				
ε	Tests	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
0.05	$\sup F_T(1)$	12.4	14.4	19.4	28.8	40.2	24.6	26.8	30.6	36.4	45.2
	$\sup F_T(2)$	22.8	21.8	23.8	35.2	51.8	45.0	46.8	48.6	54.8	69.2
	$\sup F_T(3)$	27.8	28.8	36.0	45.6	66.4	60.0	59.8	62.6	71.2	79.4
	$\sup F_T(4)$	32.6	33.4	38.6	52.0	76.0	69.4	68.0	72.2	79.8	90.0
	$\sup F_T(5)$	39.6	39.2	48.6	60.0	82.4	77.4	74.6	80.8	86.4	95.2
	$UD\max$	39.8	39.2	46.2	58.0	79.6	79.4	78.6	80.0	87.6	95.0
	$WD\max$	42.8	43.0	50.6	62.8	85.4	82.2	81.0	85.2	90.2	96.6
0.10	$\sup F_T(1)$	4.2	4.8	9.6	17.6	28.8	9.8	11.0	17.4	22.2	30.4
	$\sup F_T(2)$	3.4	4.0	7.6	21.8	42.6	12.8	16.8	23.6	33.4	51.6
	$\sup F_T(3)$	3.8	5.2	9.8	28.2	51.6	16.4	18.8	25.6	41.6	60.6
	$\sup F_T(4)$	4.2	4.8	10.6	28.6	59.0	18.4	22.8	32.4	45.4	69.0
	$\sup F_T(5)$	4.2	5.8	12.2	29.6	62.4	22.0	25.2	35.0	50.0	74.2
	$UD\max$	4.4	5.6	12.4	29.2	56.8	20.8	22.6	33.2	48.0	68.4
	$WD\max$	5.6	6.8	14.0	34.4	64.2	26.0	29.2	39.8	56.4	76.2
0.15	$\sup F_T(1)$	3.6	4.4	10.4	16.6	26.2	7.6	9.2	15.0	20.6	27.0
	$\sup F_T(2)$	3.8	3.6	8.2	19.8	40.0	9.6	11.4	16.6	28.6	45.2
	$\sup F_T(3)$	3.4	4.2	8.2	23.4	50.2	12.0	14.2	20.2	33.6	56.0
	$\sup F_T(4)$	3.0	4.8	9.0	24.4	55.2	15.4	17.8	22.8	37.8	62.0
	$\sup F_T(5)$	3.8	4.8	10.8	28.2	60.0	19.0	22.8	27.6	44.8	67.2
	$UD\max$	4.4	4.6	10.6	23.8	49.0	11.0	14.2	21.8	36.6	57.8
	$WD\max$	4.6	5.2	11.2	30.6	59.6	17.0	21.6	28.0	46.2	68.4
0.20	$\sup F_T(1)$	3.4	4.4	9.6	14.6	22.8	6.0	7.4	12.2	16.6	22.0
	$\sup F_T(2)$	3.0	3.0	5.6	16.2	32.4	6.2	8.8	12.2	21.6	36.4
	$\sup F_T(3)$	1.6	2.4	5.0	15.2	36.2	7.2	7.8	12.2	23.6	43.2
	$UD\max$	3.0	3.8	10.0	18.4	34.6	7.0	8.8	15.0	23.8	39.6
	$WD\max$	3.4	3.6	7.8	20.6	38.4	8.8	10.0	16.2	26.2	44.4
0.25	$\sup F_T(1)$	3.2	4.2	8.0	13.4	20.6	4.6	6.0	9.6	15.2	20.0
	$\sup F_T(2)$	3.0	3.0	5.2	14.8	28.8	6.6	7.0	9.8	18.6	31.6
	$UD\max$	3.2	4.2	8.4	15.2	27.2	6.0	6.6	11.2	18.4	29.4
	$WD\max$	3.6	3.2	8.4	15.4	29.4	7.0	6.2	12.6	20.0	32.0

Table 2. Empirical size of the asymptotic tests for T = 100

$T = 100$		No heterogeneity in the errors					Heterogeneity in the errors				
		ρ					ρ				
ε	Tests	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
0.05	$\sup F_T(1)$	3.2	3.2	6.4	15.8	35.6	9.0	10.6	13.6	20.6	39.6
	$\sup F_T(2)$	3.0	4.4	8.8	26.2	54.8	15.2	17.6	26.8	39.0	62.0
	$\sup F_T(3)$	3.2	3.2	8.2	27.6	61.6	18.2	20.6	28.2	45.4	69.2
	$\sup F_T(4)$	3.2	3.4	9.2	30.4	69.2	21.2	22.6	32.6	51.8	76.8
	$\sup F_T(5)$	3.4	3.6	9.0	31.4	73.8	22.2	25.4	36.6	55.6	82.4
	$UD\max$	4.6	4.4	11.6	34.0	74.6	24.6	28.2	38.4	57.4	83.2
	$WD\max$	4.2	5.2	14.0	39.2	78.6	28.0	32.6	44.4	63.4	88.2
0.10	$\sup F_T(1)$	3.0	3.4	5.6	13.0	28.4	5.4	8.6	9.2	15.4	30.6
	$\sup F_T(2)$	1.4	2.4	8.4	20.6	50.4	9.8	10.8	19.2	29.8	55.4
	$\sup F_T(3)$	1.8	2.2	6.2	20.6	54.8	9.4	10.6	19.6	30.0	62.0
	$\sup F_T(4)$	1.4	1.8	5.4	23.0	60.2	9.2	11.6	21.0	38.0	65.8
	$\sup F_T(5)$	1.6	2.6	6.0	24.4	63.2	11.2	12.2	20.0	38.8	69.8
	$UD\max$	3.0	3.4	7.6	23.4	61.8	9.4	12.8	21.4	35.4	69.2
	$WD\max$	2.8	3.0	9.2	29.4	68.6	14.4	17.8	27.2	42.8	73.8
0.15	$\sup F_T(1)$	2.8	3.6	5.2	10.6	25.4	6.4	6.6	8.0	12.2	26.6
	$\sup F_T(2)$	1.8	2.4	6.6	15.2	44.8	5.2	6.6	12.4	23.6	48.4
	$\sup F_T(3)$	1.2	1.8	4.6	15.0	47.4	6.4	6.2	10.8	22.6	54.2
	$\sup F_T(4)$	0.6	1.2	5.0	14.0	49.6	6.0	5.8	11.8	24.0	53.6
	$\sup F_T(5)$	0.6	0.8	3.8	11.2	46.2	6.0	6.6	10.8	22.8	51.8
	$UD\max$	2.6	3.2	6.6	16.2	46.6	6.2	8.4	12.0	23.0	53.8
	$WD\max$	2.2	2.2	5.8	18.6	54.4	7.6	9.4	15.0	28.4	60.8
0.20	$\sup F_T(1)$	3.8	3.2	3.8	9.0	22.8	6.0	5.6	4.4	10.4	22.6
	$\sup F_T(2)$	2.2	3.0	5.4	11.6	37.2	4.2	5.8	8.4	15.2	39.6
	$\sup F_T(3)$	1.8	2.0	4.2	10.6	37.6	4.4	5.0	8.0	15.0	40.0
	$UD\max$	3.2	2.6	4.8	11.6	37.6	6.0	6.2	7.0	15.8	40.6
	$WD\max$	2.6	2.8	4.8	12.8	40.8	5.0	6.8	8.8	17.6	43.4
0.25	$\sup F_T(1)$	3.8	3.2	4.2	7.8	18.8	4.2	3.8	4.2	7.6	19.4
	$\sup F_T(2)$	2.8	3.6	4.8	8.6	27.8	4.6	6.2	7.0	11.4	27.6
	$UD\max$	3.8	3.0	4.6	9.2	28.2	4.6	4.4	5.6	10.8	29.0
	$WD\max$	3.0	3.0	4.8	10.4	29.0	4.8	5.6	7.4	12.2	30.2

Table 3. Empirical size of the asymptotic tests for T = 150

$T = 150$		No heterogeneity in the errors					Heterogeneity in the errors				
		ρ					ρ				
ε	Tests	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
0.05	$\sup F_T(1)$	4.4	5.6	9.0	18.0	36.8	10.0	11.0	13.6	21.4	38.8
	$\sup F_T(2)$	2.4	4.0	9.8	31.4	62.8	12.2	15.0	24.8	39.2	65.2
	$\sup F_T(3)$	1.2	3.0	9.0	30.2	67.4	14.2	15.6	25.6	42.2	73.6
	$\sup F_T(4)$	2.8	3.4	7.8	33.4	73.6	14.2	17.8	27.2	54.2	79.2
	$\sup F_T(5)$	1.8	3.2	7.6	34.6	80.6	16.2	17.8	28.2	55.4	84.8
	$UD\max$	3.4	6.2	12.0	37.0	80.0	18.0	19.6	31.8	54.8	85.8
	$WD\max$	3.6	5.8	12.4	42.6	84.0	21.6	22.2	38.2	60.6	90.8
0.10	$\sup F_T(1)$	4.6	5.4	8.0	12.6	32.0	6.6	7.6	9.8	14.2	33.2
	$\sup F_T(2)$	2.0	3.4	8.8	23.0	55.6	6.8	7.2	16.6	26.0	59.8
	$\sup F_T(3)$	2.4	3.6	6.8	23.6	56.2	8.2	9.6	14.8	28.2	58.8
	$\sup F_T(4)$	2.0	2.8	6.2	23.4	61.0	6.8	9.4	15.8	30.4	66.0
	$\sup F_T(5)$	1.8	2.2	6.2	21.6	63.6	7.4	8.8	17.0	30.6	68.6
	$UD\max$	4.0	6.0	10.2	23.8	64.2	7.8	10.6	18.6	31.0	68.6
	$WD\max$	2.8	5.2	9.6	28.6	70.2	9.0	11.4	21.6	37.8	75.4
0.15	$\sup F_T(1)$	4.8	5.4	7.0	10.8	27.2	5.8	6.4	8.2	11.2	27.8
	$\sup F_T(2)$	2.2	4.2	6.4	18.0	44.8	6.2	5.2	10.6	20.6	47.6
	$\sup F_T(3)$	2.4	3.2	6.4	16.4	45.4	5.4	7.0	10.2	19.0	49.4
	$\sup F_T(4)$	2.2	2.8	5.5	16.4	47.0	6.0	6.6	10.2	21.6	51.0
	$\sup F_T(5)$	2.4	2.6	5.6	17.2	49.2	7.0	7.4	10.2	22.8	52.6
	$UD\max$	4.2	5.0	8.2	16.2	51.0	7.2	7.2	11.0	21.0	53.2
	$WD\max$	4.0	4.8	8.4	20.6	55.4	8.2	8.2	13.8	24.8	59.8
0.20	$\sup F_T(1)$	3.6	4.2	6.2	9.0	22.2	4.2	5.6	7.6	10.2	23.6
	$\sup F_T(2)$	2.2	3.2	6.2	13.6	35.2	4.2	4.0	7.6	14.6	36.8
	$\sup F_T(3)$	2.4	3.2	5.4	12.4	34.4	4.0	4.8	7.0	14.4	37.6
	$UD\max$	3.4	4.4	6.8	13.6	36.8	5.8	6.6	8.8	16.0	40.0
	$WD\max$	3.6	4.2	6.8	14.8	41.0	6.0	6.4	9.2	16.6	42.2
0.25	$\sup F_T(1)$	3.4	3.8	6.8	8.0	19.2	3.6	4.6	7.0	8.2	19.8
	$\sup F_T(2)$	2.8	3.2	4.4	9.6	25.6	4.8	4.6	6.0	10.4	28.4
	$UD\max$	3.6	4.0	6.2	11.0	27.4	4.2	5.2	6.4	12.0	28.2
	$WD\max$	3.0	3.6	6.4	11.4	28.8	4.4	4.8	6.6	12.0	30.2

Table 4. Empirical size of the asymptotic tests for T = 200

$T = 200$		No heterogeneity in the errors					Heterogeneity in the errors				
		ρ					ρ				
ε	Tests	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
0.05	$\sup F_T(1)$	4.0	4.4	7.0	17.0	32.2	7.4	8.8	12.2	19.6	33.4
	$\sup F_T(2)$	1.6	4.0	11.0	33.2	66.2	9.2	14.6	21.4	40.0	67.8
	$\sup F_T(3)$	2.0	2.6	7.2	30.6	70.6	12.0	12.8	21.2	39.6	74.0
	$\sup F_T(4)$	1.6	2.0	8.8	35.4	76.8	12.0	13.0	25.8	47.6	78.0
	$\sup F_T(5)$	2.0	2.4	7.6	35.2	78.2	14.2	16.4	26.6	48.0	84.4
	$UD\max$	3.4	4.4	11.0	36.4	81.4	13.0	17.6	28.4	52.6	85.4
	$WD\max$	2.6	3.8	11.6	41.8	85.6	16.8	22.0	33.6	59.6	89.8
0.10	$\sup F_T(1)$	4.4	3.8	6.4	12.6	27.4	6.4	6.0	9.2	14.8	28.4
	$\sup F_T(2)$	2.0	3.8	8.4	23.4	56.4	8.0	9.8	15.2	28.6	59.0
	$\sup F_T(3)$	2.2	2.8	6.4	21.6	59.0	7.4	7.8	12.0	26.6	63.4
	$\sup F_T(4)$	1.6	2.6	7.2	20.2	62.2	7.8	10.0	15.2	26.2	65.4
	$\sup F_T(5)$	1.8	1.8	6.2	20.2	63.6	7.2	9.2	14.2	27.6	68.8
	$UD\max$	4.2	4.2	8.4	22.8	65.2	8.2	9.6	15.4	29.2	69.6
	$WD\max$	2.6	4.2	9.2	25.8	70.8	10.0	12.0	17.8	35.8	74.2
0.15	$\sup F_T(1)$	3.8	4.6	6.6	11.6	22.6	4.4	5.6	7.8	12.8	23.0
	$\sup F_T(2)$	2.4	2.4	7.0	16.4	43.8	6.2	6.4	9.8	19.2	45.8
	$\sup F_T(3)$	2.4	3.0	5.4	16.8	46.4	5.4	5.4	8.6	19.8	47.6
	$\sup F_T(4)$	2.6	2.4	5.2	14.6	48.8	5.6	6.0	9.0	17.6	51.8
	$\sup F_T(5)$	2.8	3.2	5.0	15.0	43.2	6.4	6.4	9.4	18.4	47.4
	$UD\max$	3.6	4.6	7.0	19.0	48.6	6.0	6.8	11.6	22.2	49.4
	$WD\max$	2.6	3.0	7.0	19.4	54.4	6.2	6.8	12.6	23.6	57.0
0.20	$\sup F_T(1)$	4.2	5.0	6.0	10.2	19.6	4.6	6.2	7.0	10.2	19.6
	$\sup F_T(2)$	2.6	3.2	5.0	12.2	34.2	4.2	4.6	6.2	14.0	35.2
	$\sup F_T(3)$	2.2	2.6	4.4	11.4	33.6	3.8	4.4	6.4	14.2	35.2
	$UD\max$	4.2	4.4	6.0	13.4	34.2	4.4	5.6	8.0	14.8	35.8
	$WD\max$	3.6	3.2	5.6	13.4	39.4	4.6	5.2	8.6	16.8	41.4
0.25	$\sup F_T(1)$	4.4	4.6	4.8	8.8	17.2	4.6	4.8	5.8	8.8	17.2
	$\sup F_T(2)$	2.8	3.2	4.6	9.4	24.6	4.4	4.4	4.8	10.4	25.6
	$UD\max$	4.2	4.6	5.2	10.2	25.4	4.4	4.8	6.6	10.4	26.6
	$WD\max$	3.8	4.0	5.4	11.6	27.8	4.0	4.6	6.6	12.4	29.0

Table 5. Empirical size of the asymptotic tests for T = 250

$T = 250$		No heterogeneity in the errors					Heterogeneity in the errors				
		ρ					ρ				
ε	Tests	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
0.05	$\sup F_T(1)$	4.0	5.2	9.0	19.4	38.2	7.8	8.6	11.8	23.4	39.6
	$\sup F_T(2)$	0.8	2.8	12.6	34.8	71.4	12.0	15.4	22.6	40.4	72.2
	$\sup F_T(3)$	1.4	2.6	10.0	33.6	73.6	10.0	15.6	23.2	41.2	77.6
	$\sup F_T(4)$	2.0	2.8	8.8	36.2	79.8	11.0	14.0	24.2	46.2	82.8
	$\sup F_T(5)$	1.6	2.6	8.2	36.2	82.0	10.6	12.8	25.6	48.4	85.0
	$UD\max$	3.0	4.6	12.6	39.4	84.0	12.8	17.8	28.6	52.4	87.8
	$WD\max$	3.2	4.6	13.2	44.2	87.8	14.8	21.4	32.4	57.6	90.8
0.10	$\sup F_T(1)$	4.0	4.0	6.4	13.6	29.6	5.0	6.4	8.4	15.6	30.4
	$\sup F_T(2)$	2.2	3.0	6.8	23.8	58.2	6.0	8.0	12.2	28.4	60.4
	$\sup F_T(3)$	2.0	2.8	7.0	22.0	61.4	7.2	8.0	13.4	25.8	63.0
	$\sup F_T(4)$	1.8	2.4	6.4	19.8	65.2	6.2	7.6	12.2	27.8	67.2
	$\sup F_T(5)$	1.4	2.6	5.8	18.6	65.6	6.0	7.2	11.8	26.2	68.2
	$UD\max$	2.8	3.8	8.4	24.2	68.0	6.4	8.6	14.8	29.8	69.0
	$WD\max$	1.8	3.6	8.6	25.8	71.6	7.6	9.2	16.4	31.8	74.4
0.15	$\sup F_T(1)$	4.2	4.0	5.6	11.8	25.2	4.4	5.2	6.8	13.2	25.0
	$\sup F_T(2)$	2.8	2.2	5.2	16.0	43.0	4.0	4.2	6.4	20.0	45.8
	$\sup F_T(3)$	1.8	2.2	4.4	13.8	44.4	4.2	3.6	6.2	17.0	46.8
	$\sup F_T(4)$	1.6	2.2	3.8	13.8	45.6	3.2	3.8	5.6	16.6	48.2
	$\sup F_T(5)$	1.6	2.0	3.4	12.8	44.4	4.2	4.4	5.8	15.8	46.6
	$UD\max$	3.2	3.4	6.2	18.4	49.4	5.0	5.8	9.0	20.6	50.4
	$WD\max$	2.0	2.6	6.2	18.4	51.8	4.6	4.4	8.6	22.4	54.4
0.20	$\sup F_T(1)$	4.0	4.4	4.8	9.0	19.8	4.6	4.2	6.0	9.4	19.2
	$\sup F_T(2)$	2.6	2.8	5.0	12.8	34.2	3.0	4.0	5.2	13.4	35.8
	$\sup F_T(3)$	2.2	2.6	3.2	10.0	33.2	3.0	3.6	4.8	11.6	34.6
	$UD\max$	3.4	4.6	5.0	12.2	36.2	3.8	4.4	6.2	14.0	36.4
	$WD\max$	2.4	3.4	5.4	12.0	38.4	3.2	4.6	6.2	13.8	39.8
0.25	$\sup F_T(1)$	5.0	4.6	5.4	7.6	17.0	5.2	5.4	6.6	8.0	17.4
	$\sup F_T(2)$	2.8	3.4	4.8	8.0	25.0	4.2	3.8	5.4	8.0	26.0
	$UD\max$	4.6	4.6	5.4	9.0	26.2	5.4	4.4	6.4	9.6	26.0
	$WD\max$	3.8	3.8	5.2	10.4	27.2	4.6	4.2	5.4	10.6	27.4

Table 6. Empirical size of the parametric bootstrap tests for T = 50

$T = 50$		No heterogeneity in the errors					Heterogeneity in the errors				
		ρ					ρ				
ε	Tests	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
0.05	$\sup F_T(1)$	3.6	3.2	4.0	4.2	5.0	2.8	4.0	3.8	4.6	4.6
	$\sup F_T(2)$	4.6	3.6	3.4	3.4	2.0	2.4	3.4	4.8	3.0	3.2
	$\sup F_T(3)$	4.2	3.4	4.4	5.4	3.2	3.4	4.2	4.4	4.4	4.4
	$\sup F_T(4)$	4.2	4.0	4.8	3.0	3.8	3.0	4.4	3.6	2.6	5.2
	$\sup F_T(5)$	4.8	4.2	5.0	4.2	3.8	4.2	5.0	4.4	5.4	3.8
	$UD\max$	4.4	4.6	4.4	4.4	4.2	3.6	5.0	4.2	4.0	3.8
	$WD\max$	4.6	4.8	4.4	4.4	4.4	3.6	5.0	4.6	4.2	3.8
0.10	$\sup F_T(1)$	3.8	3.6	5.2	5.2	5.8	6.0	4.6	5.8	5.2	6.4
	$\sup F_T(2)$	6.2	5.4	5.8	5.4	6.2	6.4	7.4	7.0	6.4	5.0
	$\sup F_T(3)$	6.2	7.0	6.0	5.0	4.4	6.6	8.0	8.6	5.2	7.0
	$\sup F_T(4)$	6.2	6.6	6.0	4.8	4.6	7.2	8.2	7.2	4.8	6.8
	$\sup F_T(5)$	7.0	6.4	6.4	4.4	4.8	6.8	7.8	7.8	5.6	6.6
	$UD\max$	4.8	5.0	5.0	4.8	6.4	7.0	7.2	6.2	5.8	6.8
	$WD\max$	4.6	5.4	5.0	4.8	5.4	6.6	7.0	7.6	6.2	6.0
0.15	$\sup F_T(1)$	4.6	4.2	6.2	4.8	6.6	4.2	5.0	7.2	4.2	6.0
	$\sup F_T(2)$	4.8	4.2	5.6	5.4	6.6	6.2	6.0	7.2	4.8	5.8
	$\sup F_T(3)$	5.8	5.0	5.4	5.8	6.8	6.2	5.6	7.4	4.6	5.4
	$\sup F_T(4)$	5.4	5.0	5.0	5.2	5.2	6.0	5.2	7.4	4.8	5.2
	$\sup F_T(5)$	5.4	4.8	5.0	5.8	4.8	5.8	5.4	6.8	4.4	5.0
	$UD\max$	4.2	4.0	5.6	5.6	6.4	6.0	4.8	6.6	4.0	6.2
	$WD\max$	4.0	3.8	6.0	5.4	5.8	6.4	5.2	6.8	4.0	6.4
0.20	$\sup F_T(1)$	4.6	4.8	4.8	5.4	6.0	4.6	4.6	5.4	4.8	5.8
	$\sup F_T(2)$	3.8	3.8	5.4	4.8	8.0	5.0	4.0	5.6	4.2	7.2
	$\sup F_T(3)$	4.2	4.0	5.4	5.8	7.2	4.6	4.6	5.4	5.8	6.6
	$UD\max$	4.0	5.0	5.0	5.4	6.4	4.2	4.6	6.2	5.2	7.2
	$WD\max$	4.2	4.2	5.8	5.8	6.0	4.6	4.4	6.0	4.4	7.0
0.25	$\sup F_T(1)$	5.4	4.2	4.2	4.4	4.8	5.4	4.6	4.4	4.6	5.4
	$\sup F_T(2)$	4.0	4.4	5.4	5.4	6.6	5.0	5.0	5.6	5.0	6.4
	$UD\max$	5.0	4.6	5.4	5.4	6.0	4.0	4.0	4.8	4.6	6.2
	$WD\max$	5.0	4.4	5.0	5.8	6.2	4.0	5.2	6.0	5.0	6.2

Table 7. Empirical size of the nonparametric bootstrap tests for T = 50

$T = 50$		No heterogeneity in the errors					Heterogeneity in the errors				
		ρ					ρ				
ε	Tests	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
0.05	$\sup F_T(1)$	5.2	4.6	4.8	3.8	4.4	4.4	4.0	4.8	4.2	4.0
	$\sup F_T(2)$	6.0	6.6	5.0	5.2	5.2	6.4	6.4	5.6	5.2	4.2
	$\sup F_T(3)$	5.4	5.2	4.6	5.0	5.2	5.6	6.0	5.6	4.4	4.4
	$\sup F_T(4)$	4.8	5.4	6.2	5.0	4.8	6.0	5.8	7.2	6.0	4.2
	$\sup F_T(5)$	4.6	4.8	5.0	4.4	5.8	5.6	5.2	6.2	4.8	5.6
	$UD\max$	5.6	5.2	5.0	5.2	5.6	5.2	4.6	7.4	5.0	5.2
	$WD\max$	5.8	5.4	4.8	5.4	5.4	5.2	4.8	7.0	5.0	5.0
0.10	$\sup F_T(1)$	3.6	4.0	4.4	4.0	5.6	3.4	4.0	4.0	3.6	5.6
	$\sup F_T(2)$	4.4	5.4	5.6	4.8	5.6	6.0	4.8	5.8	7.4	5.4
	$\sup F_T(3)$	4.4	6.4	5.2	6.0	5.0	5.0	5.8	6.0	6.4	5.8
	$\sup F_T(4)$	4.6	5.0	5.6	4.4	4.6	4.2	5.4	4.6	5.6	5.0
	$\sup F_T(5)$	4.6	5.2	4.4	4.6	5.4	4.6	4.6	4.0	5.6	6.2
	$UD\max$	4.8	4.8	3.6	4.4	5.2	4.0	5.8	5.6	5.4	5.2
	$WD\max$	5.2	4.2	3.4	4.2	5.2	5.0	5.8	4.4	5.2	6.2
0.15	$\sup F_T(1)$	4.2	3.8	3.8	5.0	7.2	3.8	4.4	3.4	4.8	7.0
	$\sup F_T(2)$	4.4	5.2	6.0	4.6	5.0	4.4	5.6	5.8	5.6	5.6
	$\sup F_T(3)$	4.8	6.0	6.8	5.8	5.0	5.8	6.0	6.2	6.0	5.0
	$\sup F_T(4)$	4.2	4.6	5.4	4.8	4.6	5.0	5.2	5.6	6.0	6.0
	$\sup F_T(5)$	4.0	4.6	4.2	4.0	5.0	4.2	5.0	5.4	5.4	6.4
	$UD\max$	4.6	4.6	4.0	4.4	6.0	4.8	4.6	5.2	5.8	5.8
	$WD\max$	4.0	4.4	4.6	4.4	5.8	4.8	4.6	5.4	6.2	6.2
0.20	$\sup F_T(1)$	4.2	3.6	4.4	4.6	8.2	3.4	3.4	4.2	4.8	8.2
	$\sup F_T(2)$	4.6	4.4	5.2	5.2	6.2	4.2	4.0	4.6	4.8	6.2
	$\sup F_T(3)$	3.6	5.4	5.4	5.0	6.6	4.0	4.8	5.2	4.4	6.2
	$UD\max$	4.4	4.4	5.0	4.6	7.0	4.2	3.8	4.2	4.6	6.8
	$WD\max$	4.6	4.6	4.6	5.2	6.2	4.0	4.4	4.6	5.2	6.2
0.25	$\sup F_T(1)$	3.8	3.8	4.2	5.4	9.0	3.2	3.6	4.4	4.8	8.8
	$\sup F_T(2)$	5.2	3.8	5.0	5.2	6.0	4.8	4.2	4.8	5.0	6.0
	$UD\max$	4.0	4.2	4.8	5.6	7.6	3.4	4.2	5.0	5.4	7.4
	$WD\max$	4.2	4.2	5.8	4.8	7.0	4.0	3.6	5.0	5.6	7.2

Table 8.a. True power of the asymptotic tests for small break size

$T = 100$	No heterogeneity in the errors					Heterogeneity in the errors				
	ε					ε				
Tests	0.05	0.10	0.15	0.20	0.25	0.05	0.10	0.15	0.20	0.25
$\sup F_T(1)$	79.6	81.2	83.4	84.6	85.4	80.0	81.2	83.4	84.0	84.8
$\sup F_T(2)$	79.6	79.8	81.4	83.0	84.6	83.4	83.2	83.0	84.4	84.4
$\sup F_T(3)$	79.6	81.8	82.4	81.6	–	85.2	85.4	84.2	84.2	–
$\sup F_T(4)$	81.2	82.6	82.0	–	–	85.6	84.8	84.0	–	–
$\sup F_T(5)$	82.2	81.4	80.8	–	–	89.2	86.8	84.0	–	–
$UD\max$	88.4	87.6	87.2	86.8	86.6	93.4	90.2	88.6	87.2	87.2
$WD\max$	90.2	89.6	87.8	86.8	86.8	93.8	92.4	89.8	88.2	87.6
$\sup F_T(2\|1)$	17.4	10.2	8.0	5.2	2.4	24.6	13.8	8.6	5.2	2.4
$\sup F_T(3\|2)$	9.4	3.0	1.8	0.8	0.0	14.6	5.2	2.6	1.0	0.0
$\sup F_T(4\|3)$	6.2	2.2	0.0	0.0	–	11.2	2.8	0.2	0.0	–
$\sup F_T(5\|4)$	3.6	0.8	0.0	–	–	7.8	1.2	0.0	–	–
$\sup F_T(6\|5)$	4.6	0.2	0.0	–	–	7.8	0.2	0.0	–	–

Table 8.b. True power of the asymptotic tests for large break size

$T = 100$	No heterogeneity in the errors					Heterogeneity in the errors				
	ε					ε				
Tests	0.05	0.10	0.15	0.20	0.25	0.05	0.10	0.15	0.20	0.25
$\sup F_T(1)$	99.0	99.2	99.2	99.4	99.6	99.2	99.2	99.4	99.4	99.6
$\sup F_T(2)$	97.6	97.8	98.2	98.4	98.6	97.4	98.2	98.4	98.2	98.8
$\sup F_T(3)$	95.2	97.0	98.0	97.8	–	96.6	97.0	98.0	98.0	–
$\sup F_T(4)$	95.4	97.2	97.8	–	–	96.4	97.8	97.8	–	–
$\sup F_T(5)$	96.2	97.0	97.8	–	–	97.8	97.0	98.0	–	–
$UD\max$	99.2	99.2	99.2	99.2	99.6	99.8	99.2	99.4	99.4	99.6
$WD\max$	99.0	99.0	99.2	99.2	99.2	99.6	99.0	99.2	99.4	99.4
$\sup F_T(2\|1)$	16.4	12.0	9.8	6.8	3.5	23.6	13.6	11.2	7.0	3.8
$\sup F_T(3\|2)$	9.6	4.4	1.2	0.2	0.0	13.2	4.8	1.4	0.2	0.0
$\sup F_T(4\|3)$	5.2	0.8	0.0	0.0	–	8.8	1.2	0.0	0.0	–
$\sup F_T(5\|4)$	3.4	0.6	0.0	–	–	7.0	0.6	0.0	–	–
$\sup F_T(6\|5)$	2.6	0.2	0.0	–	–	4.8	0.4	0.0	–	–

Table 9.a. True power of the parametric bootstrap tests for small break size

$T = 100$	No heterogeneity in the errors					Heterogeneity in the errors				
	ε					ε				
Tests	0.05	0.10	0.15	0.20	0.25	0.05	0.10	0.15	0.20	0.25
$\sup F_T(1)$	32.2	43.6	53.6	60.4	67.4	15.6	33.2	45.6	57.2	65.2
$\sup F_T(2)$	23.0	31.0	40.6	48.4	60.0	10.6	19.8	31.4	42.8	55.0
$\sup F_T(3)$	19.6	30.8	41.4	49.8	–	9.6	18.0	33.2	45.0	–
$\sup F_T(4)$	18.4	31.8	42.2	–	–	10.0	19.6	36.2	–	–
$\sup F_T(5)$	18.2	30.8	43.6	–	–	9.4	21.2	35.6	–	–
$UD\max$	22.6	36.4	47.0	56.2	64.2	8.2	21.0	34.6	51.2	61.2
$WD\max$	22.4	33.8	44.4	54.6	62.2	8.4	20.2	33.0	49.2	59.4
$\sup F_T(2\|1)$	4.6	5.4	4.4	5.2	4.0	4.6	4.8	5.2	5.2	4.2

Table 9.b. True power of the parametric bootstrap tests for large break size

$T = 100$	No heterogeneity in the errors					Heterogeneity in the errors				
	ε					ε				
Tests	0.05	0.10	0.15	0.20	0.25	0.05	0.10	0.15	0.20	0.25
$\sup F_T(1)$	95.8	97.4	98.4	99.2	99.4	87.8	96.2	97.2	98.8	99.4
$\sup F_T(2)$	87.6	92.4	95.8	98.0	98.6	53.0	79.4	91.4	96.4	98.2
$\sup F_T(3)$	83.2	90.4	96.0	98.0	–	45.6	75.0	90.8	95.8	–
$\sup F_T(4)$	79.8	90.2	95.6	–	–	42.2	73.6	89.4	–	–
$\sup F_T(5)$	77.2	89.2	95.2	–	–	36.6	73.8	90.0	–	–
$UD\max$	93.4	96.6	97.8	98.4	99.4	65.0	91.0	96.6	98.2	99.4
$WD\max$	93.0	96.2	97.6	98.6	99.0	57.0	89.0	95.4	97.6	98.6
$\sup F_T(2\|1)$	5.8	4.4	5.4	6.4	4.4	4.4	4.2	5.4	5.6	4.4

Table 10.a. True power of the nonparametric bootstrap tests for small break size

$T = 100$	No heterogeneity in the errors					Heterogeneity in the errors				
	ε					ε				
Tests	0.05	0.10	0.15	0.20	0.25	0.05	0.10	0.15	0.20	0.25
$\sup F_T(1)$	32.4	45.8	55.6	64.0	70.0	13.4	33.8	48.6	60.4	67.6
$\sup F_T(2)$	25.2	32.8	43.4	53.2	62.0	11.6	20.4	32.4	45.6	58.4
$\sup F_T(3)$	21.6	30.6	44.4	54.2	–	10.8	18.8	35.2	46.0	–
$\sup F_T(4)$	20.8	29.8	42.4	–	–	10.0	19.6	35.0	–	–
$\sup F_T(5)$	20.0	30.0	43.2	–	–	10.0	21.2	35.0	–	–
$UD\max$	24.0	40.2	51.0	61.2	70.0	8.2	22.6	40.2	54.4	65.2
$WD\max$	21.8	37.6	47.8	58.8	67.2	7.8	21.0	37.8	51.4	63.6
$\sup F_T(2\|1)$	4.8	3.8	4.0	3.0	3.6	5.0	4.2	3.6	2.8	3.4

Table 10.b. True power of the nonparametric bootstrap tests for large break size

$T = 100$	No heterogeneity in the errors					Heterogeneity in the errors				
	ε					ε				
Tests	0.05	0.10	0.15	0.20	0.25	0.05	0.10	0.15	0.20	0.25
$\sup F_T(1)$	96.8	97.8	98.4	98.8	99.4	90.0	96.0	98.0	98.2	99.2
$\sup F_T(2)$	88.8	95.0	97.4	97.8	97.8	55.2	84.0	93.6	97.0	97.4
$\sup F_T(3)$	87.0	93.0	96.4	98.0	–	47.0	81.8	91.2	96.2	–
$\sup F_T(4)$	83.0	90.8	96.4	–	–	42.8	75.6	91.4	–	–
$\sup F_T(5)$	79.8	90.8	96.8	–	–	35.8	75.2	92.0	–	–
$UD\max$	95.0	97.6	98.0	98.6	99.2	65.4	93.4	97.4	98.2	98.8
$WD\max$	94.2	96.6	98.0	98.4	98.8	56.0	90.2	96.2	98.2	98.2
$\sup F_T(2\|1)$	4.6	3.8	3.8	4.0	3.8	4.6	3.8	3.6	3.6	3.6