Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Bootstrap Confidence Intervals and Hypotheses Tests in Structural Models

G. C. Franco, S. J. Koopman and R. C. Souza

Relatório Técnico RTP-03/2000

Relatório Técnico
Série Pesquisa

BOOTSTRAP CONFIDENCE INTERVALS AND HYPOTHESES TESTS IN STRUCTURAL MODELS

Glaura C. Franco - Dep. Estatística - UFMG

Siem J. Koopman - Department of Econometrics, Free University Amsterdam, Netherlands.

Reinaldo C. Souza - Dep. Engenharia Elétrica - PUC/RJ

SUMMARY

Bootstrap in time series models is not straightforward to implement, as in this case the observations are not independent. One of the alternatives is to bootstrap the residuals in order to obtain the bootstrap series and thus use these series for inference purposes. This work deals with the problem of assessing the accuracy of hyperparameters in structural models. We study the simplest case, the local level model, where the hyperparameters are given by the variances of the disturbance terms. As their distribution are not known, we employ the bootstrap to approximate the true distribution, using parametric and nonparametric approaches. Bootstrap standard deviations and confidence intervals are computed, using Monte Carlo simulations. A study into the efficiency of bootstrap tests in cases where the parameters lie on the boundary of the parameter space is also undertaken.

Keywords: State space, Unobserved components, Nonparametric and parametric Bootstrap.

1 Introduction

This paper discusses the use of bootstrap to approximate the exact distribution of the hyperparameters in the local level model. This one is the simplest structural model (see Harvey, 1989) for a time series, consisting of a trend component plus a random disturbance term. The hyperparameters, in this case, are the unknown variances of the disturbance terms, and we have access neither to their exact nor to their asymptotic distribution. Therefore, the bootstrap (Efron, 1979) appears as an alternative to calculate the empirical distribution of the hyperparameters, as well as their standard errors. Recently, there have been many works related to the use of bootstrap in time series models, an area where the bootstrap technique has to accomodate some slight changes, as the data are correlated over time. Papers such as the ones from Chatterjee (1986), Holbert and Son (1986), Li and Maddala (1996) and Souza and Neto (1996) brought some important contributions to the subject.

One of the objectives of this paper concerns the construction of bootstrap confidence intervals for the hyperparameters. The use of bootstrap to build nonparametric confidence intervals had a considerable growth in the last decades, mainly due to the outstanding developments in the computational area. Many authors have worked on the subject, employing bootstrap procedures to solve questions related to the accuracy of estimates in many distinct statistical problems. Important references in this field are Hall (1986, 1988), Beran (1987) and Schenker (1985), among others. In this work we calculate confidence intervals using four different bootstrap methods, proposed by Efron and Tibshirani (1986). The intervals are compared through their coverage probabilities.

Another aim of this paper relates to a general treatment for constructing bootstrap tests when parameters of nonstationary time series models lie on the boundary of the parameter space. We will give a test for trend stationary or difference stationary which have specific implications on how to transform a time series to stationarity. In many areas of econometrics and statistics the likelihood ratio (LR) test is used for testing linear restrictions. In standard situations, the LR test has attractive asymptotic properties. However, the asymptotic distribution of the LR test is not the usual χ^2 distribution when parameters under the null hypothesis lie on the boundary; see Gourieroux *et al* (1982), Rogers (1986) and Harvey (1989) for a general discussion. We will show that bootstrap methods are effective in computing the distribution of such likelihood-based test statistics under these circumstances. One of the disadvantages of the LR test is that we have to estimate the model under the null and alternative hypotheses. This fact can make the calculation of bootstrap LR tests very slow, depending on the number of bootstrap samples needed to attain the desired level. Thus, in an attempt to find a test which spends less time in the bootstrapping process and taking advantage of the results obtained by Koopman and Shephard (1992) concerning the use of the score vector for Gaussian state space models, we also bootstrap this test statistic in order to test the

hypothesis that the parameters of a structural model is equal to zero. This new test will be called Modified Score (MS) test and it is very easy to apply, only requiring the parameter estimation under the null hypothesis. Therefore, using the parametric and nonparametric approaches, we employ the bootstrap to calculate the size and power of the LR and MS tests.

The paper is outlined as follows. Section 2 defines the local level model, as stated in Harvey (1989) and some details about hyperparameter estimation are broached. In Section 3 we introduce the bootstrap technique and explain how to construct a bootstrap series in state-space form. Section 4 describes the four confidence intervals utilized in this work. Bootstrap tests are introduced in Section 5, where we show how to bootstrap the test statistics in order to have an approximation of their distributions and thus calculate the size and power of the tests. The simulation results for univariate time series following a local level model are presented in Section 6.

2 Local Level Model and Parameter Estimation

The simplest structural model is the local level model (LLM). In this model the level component, μ_t , follows the behaviour of a random walk process, where the current level is equal to the previous one plus a white noise disturbance.

Let y_t be a univariate time series of size n. The LLM is defined as in Harvey(1989) by

$$\begin{array}{lll} y_t = \mu_t + \epsilon_t, & \epsilon_t \sim 0, \sigma_\epsilon^2 & \text{(measurement equation)} \\ \mu_t = \mu_{t-1} + \eta_t, & \eta_t \sim 0, \sigma_\eta^2 & \text{(transition equation)} \end{array}$$

where ϵ_t and η_t are uncorrelated and the unknown quantities σ_ϵ^2 and σ_η^2 are called hyperparameters.

Written this way, the LLM is in its state-space form, allowing us to apply the Kalman filter in order to obtain the maximum likelihood estimator of the unknown parameters.

With the purpose of making the maximization easier to calculate, we can reparameterize this model in such a way that

$$Var(\epsilon_t) = \sigma_\epsilon^2 \qquad \text{and} \qquad Var(\eta_t) = q\sigma_\epsilon^2.$$

The quantity q is known as the signal-noise ratio, $q = \sigma_\eta^2/\sigma_\epsilon^2$. This slight change will enable us to concentrate in one of the hyperparameters out of the likelihood function, so that we have one less parameter to estimate.

Let $a_{t|t-1}$ denote the optimal estimator of μ_t at time $t-1$ and $P_{t|t-1}$ its covariance matrix. If we assume normality, the equations necessary to compute the Kalman filter (KF) in the LLM are given by

$$a_{t+1|t} = a_{t|t-1} + K_t v_t \tag{2.1a}$$
$$P_{t+1|t} = (1 - K_t)P_{t|t-1} + q \tag{2.1b}$$

where $v_t = y_t - a_{t|t-1}$, the one-step ahead forecast errors, are called innovations, with error variance $F_t = P_t + \sigma_\epsilon^2$ and $K_t = P_{t|t-1}/\left(P_{t|t-1}+1\right)$ is the Kalman gain.

Equations 2.1a and 2.1b are known as the innovations form representation of the observations.

The method employed to estimate the parameters in state-space models is the maximum likelihood. Let $p(y_t|Y_{t-1})$ be the conditional probability density function of a set of observations $y_1, ..., y_T$, where $Y_{t-1} = \{y_1, y_2, ..., y_{t-1}\}$. The likelihood function in this case is given by

$$L(y;q) = \prod_{t=1}^{T} p(y_t|Y_{t-1}).$$

If the observations are normally distributed, we can use the Kalman filter to calculate the mean and covariance of the conditional distribution, which will also be normal.

Thus, the likelihood function can be written as

$$\log L(q) = -\frac{T}{2} log 2\pi - \frac{1}{2}\sum_{t=1}^{T} \log F_t - \frac{1}{2}\sum_{t=1}^{T} \frac{v_t^2}{F_t}.$$

Maximizing this equation with respect to q yields the ML estimate, $\hat{q}$. As the likelihood is a nonlinear function of the hyperparameter q, we have to use numerical procedures to find the maximum. The apprach employed was the EM algorithm, that requires the calculation of the score vector

$$S(q^*) = \frac{\partial log L(q)}{\partial q}|_{q=q^*}$$

at each step.

Koopman and Shephard (1992) show that the computational effort in calculating the EM algorithm can be considerably reduced, if the aim is to estimate unknown values only in the covariance matrix, by calculating the exact score via the smoothed estimators of the disturbance terms. This procedure will always converge to a local maximum, as the log-likelihood function increases at each iteration, Harvey (1989).

As the hyperparameter q is always non-negative, it should be reparameterised to ensure that negative values will not come out. The transformation used was

$$q = exp\left(2\theta\left(q\right)\right)$$

as suggested in Koopman et al. (1995).

3 Bootstrap in State-Space Models

The bootstrap is a computer-intensive method introduced by Efron in 1979. It is used mainly to assess the acuracy of statistical estimates for which the exact distribution is not known, or is difficult to derive. This technique is essencially a non-parametric one, since it does not require any distributional assumptions about the data.

Roughly speaking, the unknown distribution is approximated by the empirical distribution of the original sample, and the data are resampled, with replacement, to obtain the bootstrap sample. The method can be described as follows (see Franco, 1998).

Let $X_1, ..., X_n$ be n independent random variables with common distribution function $F(\underset{\sim}{x})$. Assume that $\hat{F}(\underset{\sim}{x})$ is the correspondent empirical distribution, which puts probability $1/n$ on each of the observed values x_i, $i = 1, ..., n$, i.e., $\hat{F}(x) = \frac{\#(x_i \leq x)}{n}$. To construct the bootstrap sample we can draw, with replacement, a sample of n observations from this set, and denote it by $\underset{\sim}{x}^* = (x_1^*, ..., x_n^*)$. This will be the first bootstrap sample.

Suppose that our statistic of interest is given by $T(x_1, ..., x_n)$. We can repeat this procedure a large number of times, say B, and compute the value of T for each one of the bootstrap replications, that we denote by $T_b^*(\underset{\sim}{x}^*)$, $b = 1, ..., B$. Thus, we can use the empirical distribution of $T_b^*(\underset{\sim}{x}^*)$ as an approximation to the true distribution of the statistic $T(\underset{\sim}{x})$.

In time series, as the set of observations are not independent, this procedure must be slightly modified. So, instead of sampling from the original series, the idea is to fit a model to the data and then sampling from the residuals, which should be independent and identically distributed, if the model is well adjusted. If the distribution is not known, the bootstrap series can be calculated through the estimated residuals, thus employing the non-parametric approach. On the other hand, if we know the sample distribution, the parametric bootstrap can be used, by sampling the disturbances from this distribution and constructing the series recursively.

In the case of state-space models, after running the Kalman filter and estimating the model, we are able to get the estimated innovations, $\hat{v}_t$, which will be used to construct the bootstrap series, following the procedure of Stoffer and Wall (SW) (1991). The procedure is stated below.

Algorithm SW:

First estimate q and then run the Kalman filter, with $q = \hat{q}$. After that the following steps are performed:

i) Calculate the standardized innovations:

$$e_t = v_t / \sqrt{F_t}$$

ii) Obtain e_t^*; $1 \leq t \leq T$ by sampling e_t, with replacement, T times.

iii) Calculate the bootstrap data set y_t^* ; $1 \leq t \leq T$, using the innovations form representation, as follows:

Define the 2×1 vector $\xi_t = [a_{t+1|t}; y_t]'$

$$\xi_t = \begin{bmatrix} a_{t+1|t} \\ y_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a_{t|t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} K_t\sqrt{F_t} \\ \sqrt{F_t} \end{bmatrix} e_t$$

To construct the bootstrap data set, solve the equation above using e_t^* instead of e_t. Thus,

$$\begin{aligned} a^*_{t+1|t} &= a^*_{t|t-1} + K_t\sqrt{F_t}e^*_t \\ y^*_t &= a^*_{t|t-1} + \sqrt{F_t}e^*_t \end{aligned}$$

The initial conditions of the Kalman Filter are fixed at their given values and q is held fixed at $\hat{q}$.

iv) Repeat steps (ii) and (iii) B times. For each bootstrap series obtained, y_b^*, estimate the value of the signal-noise ratio, q, and call it $\hat{q}^{*b}$, $1 \leq b \leq B$.

4 Bootstrap Confidence Intervals

In this study we employ four methods involving the bootstrap to construct confidence intervals for the signal-noise ratio, q. These methods are proposed by Efron and Tibshirani (1986). The first procedure, the bootstrap-t has, according to another study of Efron and Tibshirani (1993), good theoretical coverage probabilities, but is not reliable in practice, while the percentile intervals, although being more reliable, does not have satisfactory coverage probabilities. Thus, they proposed two improved versions of the percentile method, the Bias-corrected (BC) and the Bias-Corrected and Accelerated (BCa) method. These later two procedures are shown to perform better than the former ones, in the sense that they can approach the exact confidence intervals, when such intervals can be calculated, and have accurate coverage probabilities in most situations, apart from some small sample cases.

For each one of the methods described below, we first generate B bootstrap series $y^{*1}, y^{*2}, ..., y^{*B}$ and calculate the signal-noise ratio, $\hat{q}^*$, for each series. A short description of the methods used follows.

i) Bootstrap-t

After generating the series, we compute the statistic

$$Z^*(b) = \frac{\hat{q}^{*b} - \hat{q}}{\hat{se}^*(b)},$$

where $\hat{se}^*(b)$ is the estimated standard error of q^* for the bootstrap series y^{*b}.

This method constructs a table with the percentiles of the empirical distribution $Z^*(b)$, obtained after generating the B bootstrap samples.

The bootstrap-t confidence interval is given by

$$(\hat{q} - \hat{t}^{(1-\alpha/2)}.\hat{se} \ , \ \ \hat{q} - \hat{t}^{(\alpha/2)}.\hat{se}).$$

where $\hat{t}^{(\alpha/2)}$ and $\hat{t}^{(1-\alpha/2)}$ are , respectively,the $\alpha/2$ and $(1-\alpha/2)$ percentile of the empirical distribution of Z^* and $\hat{se}$ is the standard error of $\hat{q}$, which can be obtained through the bootstrap samples.

ii) Percentile

If we take the $\alpha/2$ and $1-\alpha/2$ percentiles of the bootstrap distribution of q^*, we can define the percentile interval by

$$[\hat{q}^{*(\alpha/2)}, \hat{q}^{*(1-\alpha/2)}]$$

Thus, in practice, after estimating the values of q for each of the B bootstrap series, we order them and take the $100.\alpha th$ ordered value as the lower interval point and the $100.(1-\alpha)th$ ordered value as the upper interval point.

iii) Bias-corrected (BC)

This method also uses the percentiles of the bootstrap distribution, but not exactly the αth and the $(1-\alpha)$th. Instead, it corrects these values for possible bias in the estimation of q, through a quantity $\hat{z}_0$ which measures the median bias of $\hat{q}$.

The Bias-corrected interval is defined by

$$[q^{*(\alpha_1)} \ , \ q^{*(1-\alpha_2)}]$$

where $\alpha_1 = \Phi\left(2\ \hat{z}_0 + z^{(\alpha/2)}\right)$ and $\alpha_2 = \Phi\left(2\ \hat{z}_0 + z^{(1-\alpha/2)}\right)$.

The function Φ is the cumulative distribution function of a $N(0,1)$ and $z^{(\alpha/2)}$ its $100.\alpha th$ percentile point. The value of $\hat{z}_0$ is calculated using the proportion of $\hat{q}$ in the bootstrap samples that are smaller than the signal-noise ratio estimated in the original series,

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\hat{q}^*(b) < \hat{q}}{B}\right)$$

iv)Bias-corrected and accelerated (BCa)

The BCa interval, besides calculating the quantity $\hat{z}_0$, also calculates the *accelaration*, given by

$$a = \frac{\sum_{i=1}^{T}(\hat{q}_{(.)} - \hat{q}_{(i)})^3}{6\left[\sum_{i=1}^{T}(\hat{q}_{(.)} - \hat{q}_{(i)})^2\right]^{3/2}}$$

Thus, the same way we computed the Bias-corrected interval, we calculate the BCa interval,

$$[q^{*(\alpha_1)} \ , \ q^{*(1-\alpha_2)}]$$

but at this time with $\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha/2)}}{1 - a(\hat{z}_0 + z^{(\alpha/2)})}\right)$ and

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha/2)}}{1 - a(\hat{z}_0 + z^{(1-\alpha/2)})}\right).$$

5 Test statistics and bootstrap tests

5.1 Likelihood ratio test

The LR test for the null hypothesis

$$H_0 : \psi = \psi_0$$

is given by

$$LR = -2[\log L(y; \hat{\psi}_0) - \log L(y; \hat{\psi})], \tag{5.1}$$

where $\hat{\psi}$ is the maximum likelihood estimate of ψ under the unrestricted model and $\hat{\psi}_0$ is the estimate for the model under the null hypothesis. The LR statistic is asymptotically distributed as χ^2_m under the null where m is the number of restrictions imposed by H_0. However, when elements of ψ_0 lie on the boundary of the parameter space of ψ, the asymptotic distribution is the sum of χ^2 distributions under the null and alternative hypotheses. It is shown by Chernoff (1954) that if a single element of ψ is constrained to lie on the boundary, the LR statistic is asymptotically distributed as $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ instead of χ^2_1, where χ^2_0 is a degenerate distribution with all its mass at the origin, and χ^2_1 is the one degree of freedom Chisquare distribution. In the context of the simple LLM the following complication arises. The parameter vector is

$$\psi = \begin{bmatrix} \sigma^2_\varepsilon \\ \sigma^2_\eta \end{bmatrix},$$

which is clearly bounded to non-negative values. The hypothesis

$$H_0 : \sigma^2_\eta = 0,$$

will imply a LR statistic with the distribution $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$. However, the maximum likelihood estimator $\hat{\psi}$ will not have a joint asymptotic normal distribution under the hypothesis $\sigma^2_\eta = 0$ since some regularity condition is not valid, see Harvey (1989, Section 5.1.2). For such a case the LR statistic will not have an asymptotic χ^2 distribution.

At this point we should mention the problem raised by Feng and McCulloch (1996). The distribution implied by the bootstrap method is not valid when the maximum likelihood estimator under the alternative hypothesis is not a consistent estimator if the null hypothesis is true. However, in the Appendix we show that under the null hypothesis $\hat{\psi}_0$ converges to ψ_0 in probability so that the test sizes will match the fixed nominal levels. We may conclude by emphasizing that it is appropriate to bootstrap the LR statistic in cases where the asymptotic distribution of the test statistic can not be obtained.

5.2 Modified Score test

An alternative test statistic can be based on the score vector, the first derivative of the log-likelihood function with respect to the parameter vector ψ, that is

$$s(\psi) = \frac{\partial \log L(y;\psi)}{\partial \psi}. \tag{5.2}$$

The score vector can be evaluated at different points but it is clear that when it is evaluated at the maximum likelihood estimate $\hat{\psi}$, the score vector is zero. Thus, we define the Modified Score (MS) test

$$s(\hat{\psi}) = \left.\frac{\partial \log L(y;\psi)}{\partial \psi}\right|_{\psi=\hat{\psi}} = 0.$$

In the same way, when the null hypothesis is true we should have

$$s(\psi_0) = 0,$$

but this value tends to be large in absolute value when the null hypothesis is not true. The asymptotic distribution of $s(\hat{\psi}_0)$ is not available when the null hypothesis is true for the same reasons as those for the LR statistic.

The MS statistic has a considerable computational advantage. The LR statistic requires to estimate ψ twice (unrestricted and for the null hypothesis) while for the score statistic we only need to compute it for ψ_0 or $\hat{\psi}_0$. The evaluation of the score test can be done numerically. Koopman and Shephard (1992) have given a device to compute the score exactly using a simple smoothing algorithm which can be applied after the Kalman filter.

5.3 Bootstrap tests

In the following we assume that ψ is already estimated, under the alternative, using the available data set $y_1, \ldots, y_n$. To construct a bootstrap test we estimate ψ_0, that is ψ under the null hypothesis. When the distributional properties of the model are known, a parametric bootstrap sample $y_1^*, \ldots, y_n^*$ can be generated by sampling from the appropriate distributions. For the LLM with normal densitites we create y_t^* by

$$\begin{aligned} y_t^* &= \mu_t^* + \varepsilon_t^*, \qquad \varepsilon_t^* \sim \mathrm{N}(0, \sigma_\varepsilon^2), \\ \mu_{t+1}^* &= \mu_t^* + \eta_t^*, \qquad \eta_t^* \sim \mathrm{N}(0, \sigma_\eta^2), \end{aligned} \tag{5.3}$$

for $t = 1, \ldots, n$ and with $\mu_1^* \sim \mathrm{N}(a, P)$. The variances σ_ε^2 and σ_η^2 are based on $\hat{\psi}_0$ (or $\hat{\psi}$ if this is appropriate). When no assumptions are made about the distribution, we generate a nonparametric bootstrap with the details given in Section 3. Using the bootstrap series y_t^*, the test statistic is computed for $\hat{\psi}_0^*$ which refers to the estimate for ψ_0 using the bootstrap series y_t^*. For the LR statistic we also need to estimate ψ unrestricted using the bootstrap series y_t^*, that is $\hat{\psi}^*$. It should be noted that estimation of ψ may rely on regression methods or on maximum likelihood methods which may require numerical optimisation routines, see the discussion in Section 2. This procedure will be repeated a

number of times to obtain a bootstrap sample of B test statistics where B is typically 99, 499 or 999.

Instead of calculating critical values from the empirical distribution of the bootstrap sample of test statistics, we follow the suggestion of Davidson and MacKinnon (1997) and calculate the significance levels directly. We find the significance level more informative than knowing whether the statistics exceeds some critical value. This method also implies considerable savings in computer time.

The significance level, or the so-called P value, for the test statistic of the observations $y_1, \ldots, y_n$ is obtained from the bootstrap sample of test statistics by

$$P = \frac{1}{B+1}\left[1 + \sum_{i=1}^{B} I\left(\hat{\tau} \leq \hat{\tau}_i^*\right)\right], \tag{5.4}$$

where $\hat{\tau}$ is the test statistic computed using $y_1, \ldots, y_n$, $\hat{\tau}_i^*$ is the test statistic computed using the i-th bootstrap sample and $I(\cdot)$ is the usual indicator function.

The performance of the bootstrap test statistic can be assessed using a Monte Carlo simulation where we look at the size and the power of the test. The procedure is simple for obtaining the size and power but it can be computationally demanding. The procedure works as usual when distributional assumptions are made:

1) Generate y_t under the null hypothesis (using $\hat{\psi}_0$) for $t = 1, \ldots, n$.

2) Simulate B bootstrap series using $\hat{\psi}_0$ which is obtained using the data generated in step1.

3) Calculate the corresponding P value.

Repeat steps 1) to 3) R times and compute the size of the test as

$$\frac{1}{R}\sum_{r=1}^{R} I(p^{(r)} < \alpha),$$

where $p^{(r)}$ is the P value of the r-th simulation and α is a fixed nominal level, e.g. $\alpha = 0.05$.

The power of the test is obtained in the same way but now we generate the series in step 1. under the alternative (unconstrained) hypothesis (using $\hat{\psi}$). Note that in step 2) the bootstrap series will still be based on $\hat{\psi}_0$.

6 Simulation Studies

6.1 Empirical distribution of $\hat{q}$

The performance of bootstrap estimates for the signal-noise ratio was investigated by carrying out a series of simulations in the local level model. The results turned out to be

a very satisfactory ones, compared to the standard asymptotics, in the sense that they have more accurate standard errors (Table 1). Besides, the empirical distribution of $\hat{q}$ is skewed, as we can observe through the graphics built in Figure 1, and thus the bootstrap can provide better confidence intervals than the ones of asymptotic theory.

In order to compare the behaviour of the bootstrap estimates, we generated series of three diferent sizes: small ($T = 30$), medium ($T = 50$) and large ($T = 100$), and for each size, we assigned two values to the signal-noise ratio, $q = 0.10$ and 1.00. The empirical bootstrap distribution was constructed using $B = 1000$ replications.

Following Stoffer and Wall (1991) we also estimate, for each case, the value of q using a parametric Monte Carlo procedure. According to the authors, this is an attempt to approximate the true distribution of $\hat{q}$. The idea is to generate a series with q fixed at its true value and repeat this procedure a large number of times, estimating the parameter through maximum likelihood estimation. They claim that 1000 simulations is sufficient for the empirical distribution resembles the exact one.

The asymptotic standard errors were computed using the following formula, given by Harvey (1989),

$$a.s.d.(q) = \sqrt{\frac{2}{\sum_{j=0}^{T-1} \frac{1}{(g_{\epsilon j})^2} - \frac{1}{T}\left(\sum_{j=0}^{T-1} \frac{1}{g_{\epsilon j}}\right)^2}}$$

where $g_{\epsilon j} = q + 2\left(1 - cos\frac{2\pi j}{T}\right)$. Thus the asymptotic standard deviation of q depends only on the true value of q and the series lenght.

The results are presented in Table 1, for all sample sizes.

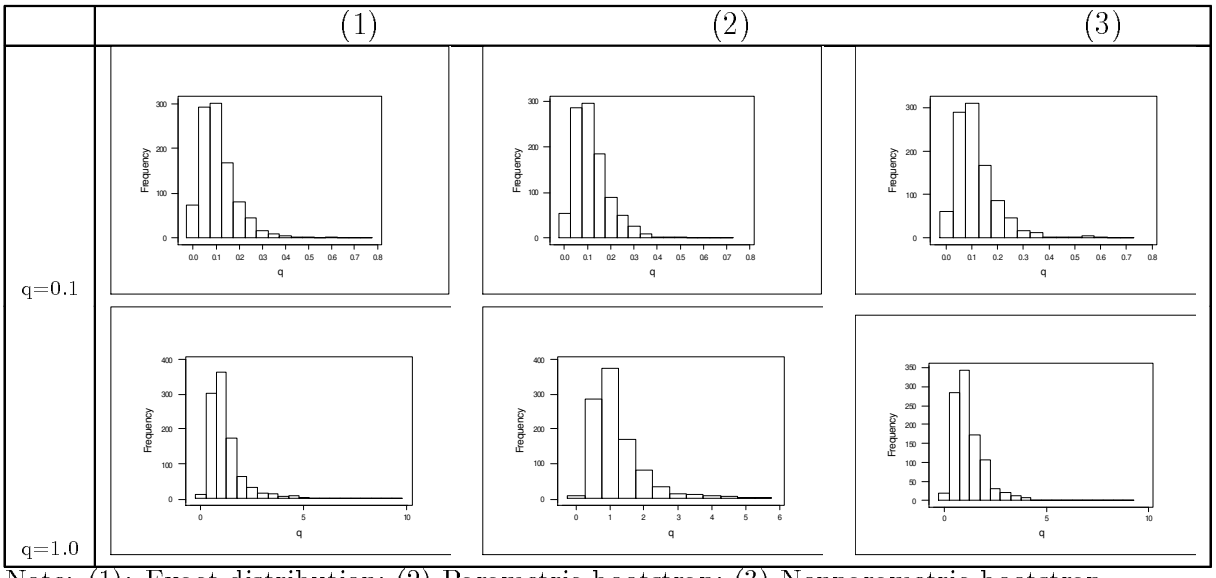
Table 1. Mean and standard deviation of q in 1000 replications for the exact, parametric and nonparametric bootstrap and the original distribution

T		PMC		BOOTSTRAP				ORIGINAL	
			nominal	Parametric		Nonparametric			
	q	Mean	S.E.	Mean	S.D.	Mean	S.D.	Estimate	a.s.e.
30	0.10	0.1339	0.1693	0.1158	0.1495	0.1333	0.1629	0.0942	0.1450
	1.00	1.3596	1.6142	1.2509	1.5166	1.2681	1.3838	0.9501	1.0230
50	0.10	0.1250	0.1273	0.1224	0.1255	0.1191	0.1198	0.1021	0.0982
	1.00	1.3139	1.1795	1.4980	1.4581	1.5584	1.4207	1.0801	0.7796
100	0.10	0.1117	0.0824	0.1165	0.0786	0.1154	0.0837	0.1057	0.0641
	1.00	1.2055	0.9097	1.1866	0.7621	1.1907	0.7915	1.0329	0.5457

The columns under the title "PMC " were calculated using the parametric Monte Carlo to get the exact distribution of q. The columns under "BOOTSTRAP " were obtained through parametric and nonparametric bootstrap. The last two columns are the estimate of q, for the original series used to generate the bootstrap samples, and its asymptotic standard error, respectively. As we can check from this table, bootstrap estimates of the standard errors have a satisfactory perform when compared to the exact distribution, even for values of T as small as 30. In all situations, the bootstrap estimates remain closer to the true value than the asymptotic ones, which tend to be always below the

nominal standard error. Another fact that must be emphasized is the good performance of the nonparametric bootstrap which, in some cases, gets closer to the nominal standard error than the parametric bootstrap. The strong asymmetry to the right observed from the empirical distribution of q (see Figure 1) can explain the reason why the asymptotic estimates tend to be systematically smaller than the nominal standard errors.

Figure 1. Empirical distribution of q for 1000 repetitions of the experiment (T=100)



Note: (1): Exact distribution; (2) Parametric bootstrap; (3) Nonparametric bootstrap

6.2 Confidence Intervals

We built 90% confidence intervals for q using the four methods proposed in Section 4. The sample sizes are the same stated earlier and the values assigned to q are 0.10 and 1.00. Two measures were used to describe the intervals, the length (l) and shape (f) of the intervals, given by

$$l = up - lp \qquad \text{and} \qquad f = \frac{up - \hat{q}}{\hat{q} - lp}$$

where lp is the lower bound and up the upper bound of the interval.

The performance of each method is compared through the simulated coverage probability (CP) of the interval, calculated repeating the experiment 100 times and computing the number of times that the interval contains the true parameter value. The results for $q = 0.10$ and $q = 1.00$ are presented in Tables 2 and 3, respectively. The lower and upper points for each interval are calculated as the mean of the 100 repetitions of each experiment.

The four methods show coverage probabilities close to the 90% nominal level, for all sample sizes, and for parametric and nonparametric bootstrap. A fail in the bootstrap-t procedure is to calculate negative values for the lower point of the intervals, such as in $q = 0.10$ an $q = 1.00$ ($T = 30$). Besides, the empirical distribution of q is highly skewed,

a fact that can be verified by the histograms and the values of s (shape of the intervals), and thus we are more inclined to use the BC and BCa methods to build the confidence intervals.

Table 2: 90% Confidence Intervals for q using the 4 methods (case $q = 0.10$)

	$T = 30$		$T = 50$		$T = 100$	
Method	Par	NPar	Par	NPar	Par	NPar
Boot-t	$CP = 0.85$ $[-.027; .336]$ $(t = .363)$ $(f = 1.99)$	$CP = 0.89$ $[-.017; .351]$ $(t = .368)$ $(f = 2.31)$	$CP = 0.84$ $[-.031; .223]$ $(t = .254)$ $(f = .91)$	$CP = 0.82$ $[-.026; .230]$ $(t = .256)$ $(f = 1.00)$	$CP = 0.88$ $[-.004; .186]$ $(t = .190)$ $(f = .73)$	$CP = 0.86$ $[-.001; .185]$ $(t = .186)$ $(f = .74)$
Percent	$CP = 0.84$ $[0; .418]$ $(t = .418)$ $(f = 3.44)$	$CP = 0.89$ $[0; .438]$ $(t = .438)$ $(f = 3.65)$	$CP = 0.83$ $[.005; .332]$ $(t = .327)$ $(f = 2.37)$	$CP = 0.82$ $[.007; .337]$ $(t = .330)$ $(f = 2.47)$	$CP = 0.88$ $[.023; .253]$ $(t = .230)$ $(f = 1.78)$	$CP = 0.87$ $[.024; .253]$ $(t = .229)$ $(f = 1.80)$
BC	$CP = 0.88$ $[.030; .625]$ $(t = .595)$ $(f = 8.27)$	$CP = 0.92$ $[.025; .639]$ $(t = .614)$ $(f = 7.87)$	$CP = 0.82$ $[.052; .489]$ $(t = .437)$ $(f = 7.72)$	$CP = 0.82$ $[.054; .495]$ $(t = .441)$ $(f = 8.17)$	$CP = 0.88$ $[.045; .294]$ $(t = .249)$ $(f = 3.10)$	$CP = 0.86$ $[.047; .284]$ $(t = .237)$ $(f = 8.17)$
BCa	$CP = 0.88$ $[.029; .471]$ $(t = .442)$ $(f = 5.78)$	$PC = 0.92$ $[.024; .610]$ $(t = .586)$ $(f = 7.35)$	$PC = 0.82$ $[.051; .464]$ $(t = .413)$ $(f = 7.08)$	$PC = 0.82$ $[.053; .467]$ $(t = .414)$ $(f = 7.43)$	$PC = 0.88$ $[.044; .286]$ $(t = .242)$ $(f = 2.92)$	$PC = 0.86$ $[.046; .268]$ $(t = .222)$ $(f = 2.72)$

Note: CP: interval Coverage Probability

l: interval lenght

S: interval shape

Table 3: 90% C.I. for q using the 4 methods (case $q = 1.00$)

	$T = 30$		$T = 50$		$T = 100$	
Method	Par	NPar	Par	NPar	Par	NPar
Boot-t	$CP = 0.90$ $[-.241; 2.543]$ $(t = 2.784)$ $(f = 1.34)$	$CP = 0.88$ $[-.013; 2.796]$ $(t = 2.809)$ $(f = 1.92)$	$CP = 0.93$ $[.316; 2.532]$ $(t = 2.216)$ $(f = 1.9)$	$CP = 0.92$ $[.368; 2.586]$ $(t = 2.218)$ $(f = 2.11)$	$CP = 0.87$ $[.374; 1.943]$ $(t = 1.569)$ $(f = 1.38)$	$CP = 0.89$ $[.400; 1.952]$ $(t = 1.552)$ $(f = 1.45)$
Percent	$CP = 0.88$ $[.061; 4.538]$ $(t = 4.477)$ $(f = 4.03)$	$CP = 0.88$ $[0.116; 4.693]$ $(t = 4.577)$ $(f = 4.49)$	$CP = 0.94$ $[.258; 4.768]$ $(t = 4.510)$ $(f = 4.48)$	$CP = 0.94$ $[.279; 4.896]$ $(t = 4.617)$ $(f = 4.76)$	$CP = 0.89$ $[.426; 4.466]$ $(t = 4.040)$ $(f = 5.66)$	$CP = 0.90$ $[.435; 3.226]$ $(t = 2.791)$ $(f = 3.67)$
BC	$CP = 0.89$ $[.393; 5.316]$ $(t = 4.923)$ $(f = 7.84)$	$CP = 0.90$ $[.309; 5.110]$ $(t = 4.801)$ $(f = 6.49)$	$CP = 0.93$ $[.338; 4.206]$ $(t = 3.868)$ $(f = 4.21)$	$CP = 0.93$ $[.376; 4.260]$ $(t = 3.884)$ $(f = 4.52)$	$CP = 0.90$ $[.483; 2.793]$ $(t = 2.31)$ $(f = 3.20)$	$CP = 0.91$ $[.480; 2.793]$ $(t = 2.313)$ $(f = 3.18)$
BCa	$PC = 0.89$ $[.386; 5.087]$ $(t = 4.701)$ $(f = 7.33)$	$PC = 0.90$ $[.301; 4.858]$ $(t = 4.557)$ $(f = 6.02)$	$PC = 0.93$ $[.386; 4.079]$ $(t = 3.751)$ $(f = 3.99)$	$PC = 0.93$ $[.366; 4.077]$ $(t = 3.711)$ $(f = 4.20)$	$PC = 0.90$ $[.474; 2.690]$ $(t = 2.216)$ $(f = 2.96)$	$PC = 0.91$ $[.476; 2.666]$ $(t = 2.19)$ $(f = 2.93)$

Note: CP: interval Coverage Probability

l: interval lenght

S: interval shape

6.3 Size and power of the bootstrap tests

The tests were compared through their performances related to the size and power obtained in Monte Carlo simulations.

The algorithm to compute the significance level is as follows:

i) Generate the series by random sampling from the model under the null hypothesis and estimate σ_ϵ^2.
ii) **Parametric bootstrap:** Using the value of σ_ϵ^2 estimated in step (i), generate the bootstrap series by random sampling from the model under the null hypothesis and compute the LR and MS statistics. **Nonparametric bootstrap:** First, bootstrap the centered innovations obtained after estimating the model under the null hypothesis, from the series generated in step (i). Then calculate the bootstrap series employing the procedure of Stoffer and Wall (1991) and compute the LR and MS statistics.

According to Hall (1986), if the nominal coverage probability is a multiple of $(B+1)^{-1}$, we can reduce the error of coverage probability from $O(n^{-1/2})$ to $O(n^{-1})$. Therefore, the number of bootstrap replications was set equal to 399, in order to get a 0.05-level critical value from the 95−percentile of the empirical distribution of the test statistics.

In order to establish a basis of comparison for the performance of the bootstrap LR and MS tests, we are going to compute a locally most powerful invariant test (NM test) introduced by Nyblom and Mäkeläinen (1983), to test the hypothesis $H_0 : \sigma_\eta^2 = 0$. For the LLM, the test statistic is calculated as

$$NM = \frac{\sum_{t=1}^{T} \left[\sum_{s=t}^{T} (y_s - \bar{y})\right]^2}{(T-1)\sum_{t=1}^{T}(y_s - \bar{y})^2} \tag{5.1}$$

The asymptotic distribution of this test statistic is given in the above-mentioned paper and, for a significance level of 0.05, we have the value 0.461.

To calculate the size for LR, MS and NM tests, we repeat the following procedure $R = 1000$ times.

1. Generate the series by random sampling from the model under the null hypothesis.
   a) Compute the LR statistic. Call it $LR_{obs}^{(r)}$.
   b) Compute the MS statistic. Call it $S_{obs}^{(r)}$.
   c) Compute the NM statistic. Call it $NM_{obs}^{(r)}$.

2. (only for the LR and MS tests) Simulate bootstrap series $y_1^{*b}, y_2^{*b}, ..., y_n^{*b}$, b = 1, ..., 399 from H_0, using the parameters estimated in step 1). Calculate the test statistics LR_b^* and S_b^* for each series and obtain the simulation approximation to the P-value,

$$p_{LR}^{(r)} = \frac{\#\left(LR_b^* \geq LR_{obs}^{(r)} + 1\right)}{B+1}$$

$$p_S^{(r)} = \frac{\#\left(S_b^* \geq S_{obs}^{(r)} + 1\right)}{B+1}$$

3. Compute the sizes of the tests as

$$SZ(LR) = \frac{1}{R}\sum_{r=1}^{R} I(p_{LR}^{(r)} < \alpha)$$

$$SZ(S) = \frac{1}{R}\sum_{r=1}^{R} I(p_S^{(r)} < \alpha)$$

$$SZ(NM) = \frac{1}{R}\sum_{r=1}^{R} I(NM_{obs}^{(r)} > nm_\alpha)$$

nm_α is the critical point of level α for the NM statistic, obtained by Nyblom and Mäkeläinen (1983).

To calculate the power, we use the same algorithm described when calculating the size of the tests, with the modification that in step (i) the series is generated from the model under the alternative hypothesis.

Our study analyses two different configurations for the distribution of the disturbance term ϵ_t. The choices are (a) Normal with mean 0 and variance $\sigma_\epsilon^2 = 0.10$. (b) t-Student with 5 degrees of freedom. We consider three values for the series sizes: small ($T = 30$), medium ($T = 50$) and large ($T = 100$). Tests of nominal level 0.05 and 0.10 were conducted for each of the cases described above.

The results are presented in Table 4, where the indexes p and np in the LR and MS tests denote parametric and nonparametric bootstrap, respectively.

Table 4. Empirical level of the Bootstrap and NM tests

Level		Normal					t(5)				
α	T	LR_p	LR_{np}	Sco_p	Sco_{np}	NM	LR_p	LR_{np}	Sco_p	Sco_{np}	NM
0.05	30	0.055	0.048	0.064	0.066	0.067	0.045	0.034	0.044	0.042	0.042
	50	0.049	0.042	0.035	0.054	0.050	0.047	0.042	0.055	0.053	0.052
	100	0.056	0.051	0.044	0.052	0.056	0.050	0.040	0.051	0.048	0.046
0.10	30	0.114	0.105	0.124	0.117	0.126	0.089	0.075	0.090	0.093	0.086
	50	0.094	0.097	0.101	0.107	0.104	0.099	0.091	0.115	0.105	0.104
	100	0.109	0.104	0.097	0.103	0.106	0.090	0.093	0.099	0.092	0.103

As we can observe, even when the sample size is as small as 30, the probability of correctly accepting H_0 are near the nominal size α, for all procedures. Besides, the behaviour of the bootstrap tests are equivalent to the NM test, with the parametric bootstrap LR being slightly better, mainly when the sample size is small. An important remark concerns the MS test, which is computationally faster than LR tests, and presents empirical levels quite as good as the latter. We have also to emphasize the results obtained with nonparametric bootstrap tests which, without making any assumptions about the distribution of the disturbances, achieve empirical levels very close to the nominal ones.

The difference between choosing a normal or a t-distribution for the disturbance ϵ_t does not seem to affect the results in LLM.

We assign three values to σ_η^2 in the calculation of the power: 0.001, 0.01 and 0.1. The results are summarized in Table 5. In the study of power we can clearly notice the superiority of bootstrap tests in relation to NM tests, the only exception being the nonparametric MS. For series of size $T = 100$, even for values of σ_η^2 as small as 0.001, we can reject the null hypothesis in about 60% of times. For series of small size, we need σ_η^2 be at least 0.01 to get the same power. Again, the parametric bootstrap LR performs better than the others, but the nonparametric LR and parametric MS are very close to it and, in some cases, the parametric MS is still better, as for ϵ_t Normal, $T = 30$ and $\sigma_\eta^2 = 0.1$, or ϵ_t Normal, $T = 50$ and $\sigma_\eta^2 = 0.001$. For the power we can observe some differences in the results between the Normal and the t-distribution, with the former one showing a better performance.

Table 5. Empirical power of the 0.05 level Bootstrap and NM tests

T		Normal					t(5)				
	σ_η^2	LR_p	LR_{np}	Sco_p	Sco_{np}	NM	LR_p	LR_{np}	Sco_p	Sco_{np}	NM
30	.001	0.151	0.130	0.147	0.143	0.141	0.122	0.108	0.121	0.122	0.123
	.01	0.588	0.555	0.588	0.489	0.522	0.462	0.437	0.460	0.426	0.428
	.1	0.919	0.908	0.945	0.755	0.796	0.899	0.886	0.899	0.731	0.775
50	.001	0.286	0.278	0.296	0.286	0.290	0.205	0.193	0.191	0.193	0.189
	.01	0.803	0.794	0.773	0.708	0.726	0.721	0.708	0.704	0.653	0.665
	.1	0.995	0.993	0.995	0.921	0.936	0.976	0.970	0.967	0.876	0.897
100	.001	0.621	0.619	0.591	0.570	0.568	0.516	0.512	0.489	0.479	0.483
	.01	0.982	0.981	0.961	0.925	0.926	0.944	0.944	0.926	0.889	0.892
	.1	1.000	1.000	1.000	0.983	0.987	1.000	1.000	1.000	0.982	0.985

Appendix

Feng and McCulloch (1996) show, in Theorem 1, that the restricted maximum likelihood estimator $\hat{\psi}$ converges to $\psi_0^* \left(\hat{\psi}\right) \in \Omega_0$ with probability one, where Ω_0 is the space parameter where H_0 holds, if, and only if, some regularity conditions are satisfied.

The first two conditions, named the dimension of the parameter space and the unicity of the density function $f(y, \psi)$ under H_0 are straightforward in the case of testing the hypothesis $H_0 : \sigma_\eta^2 = 0$. It remains to prove that

1. The expectation of the score function $\frac{\partial}{\partial \psi_i} \log f(y, \psi)$ under H_0 is equal to zero;

2. The information matrix $I_{ij}(\psi_0) = -E_{\psi_0}\left[\frac{\partial^2}{\partial \psi_i \partial \psi_j}\{\log f(y, \psi)\}\right]$, $i, j = 1, ..., p$ is positive definite for any $\psi_0 \in \Omega_0$ and

3. The third derivatives of the likelihood exist and are continuous in the neighbourhood of the true parameter value.

To verify the assumptions above we have, with $Y_{t-1} = \{y_1, ..., y_{t-1}\}$:

1. $E_{\psi_0}\left[\frac{\partial}{\partial\psi_i}\left\{\log f\left(y,\psi\right)\right\}\right] = E_{\psi_0}\left[-\frac{1}{2}tr\left[\left(F_t^{-1}\frac{\partial F_t}{\partial\psi_i}\right)\left(\mathrm{I}-F_t^{-1}v_tv_t'\right)\right]\right] - E_{\psi_0}\left[\left(\frac{\partial v_t}{\partial\psi_i}\right)'F_t^{-1}v_t\right]$

   As the only random variables are the elements of the innovation vector v_t and its first derivative (Harvey, 1989),

   $E\left(\frac{\partial v_t'}{\partial\psi_i}v_t\,|Y_{t-1}\right) = \frac{\partial v_t'}{\partial\psi_i}E\left(v_t\,|Y_{t-1}\right) = 0 \quad$ and $\quad E\left(v_tv_t'\,|Y_{t-1}\right) = F_t$

   imply that $E_{\psi_0}\left[\frac{\partial}{\partial\psi_i}\left\{\log f\left(y,\psi\right)\right\}\right] = 0.$

2. To calculate I_{ij} we use the fact that

$$E\left(\frac{\partial^2 v_t'}{\partial\psi_i\partial\psi_j}v_t\,|Y_{t-1}\right) = 0$$

   so we obtain

   $\mathrm{I}_{ij}\left(\psi_0\right) = \frac{1}{2}\sum_{t=1}^{n}\left[tr\left[F_t^{-1}\frac{\partial F_t}{\partial\psi_i}F_t^{-1}\frac{\partial F_t}{\partial\psi j}\right]\right] + E\left[\sum_{t=1}^{n}\left(\frac{\partial v_t}{\partial\psi_i}\right)'F_t^{-1}\frac{\partial v_t}{\partial\psi j}\right], \qquad i,j=1,...,p.$

   As F_t is positive definite, so is I_{ij}.

3. To check this one we can observe that the third derivatives of the likelihood will be a summation of terms involving the derivatives of F_t and v_t with relation to each one of the parameters $\psi_i, \quad i=1,2,3$. The whole expression can be found in Franco (1998). Thus, the continuity follows from the continuity of F_t and v_t.

REFERENCES

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, **74**, 3, 457-468.

Chatterjee, S. (1986). Bootstrapping ARMA models: some simulations. *IEEE Transactions on Systems, Managements and Cybernetics*, **16**, 294-297.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, **25**, 573-8.

Davidson, R and MacKinnon, J.G. (1997). Bootstrap testing in nonlinear models. *GREQAM Document de Travail No. 97A05.*

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1-26.

Efron, B., Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy *Statistical Science*, **1**, 1, 54-77.

Efron, B., Tibshirani, R. (1993). *An introduction to the bootstrap.* New York: Chapman and Hall.

Feng, Z.D. and McCulloch, C.E. (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society: series B*, **3**, 609-617.

Franco, G.C. (1998). Bootstrap em modelos estruturais: construção de intervalos de confiança e testes de hipóteses. Ph.D. thesis, Department of Electrical Engineering,

Pontifícia Universidade Católica, Rio de Janeiro.

Gourieroux, C., Holly, A. and Montfort, A. (1982). Lielihood ratio test, Wald test and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica*, **50**, 63-80.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *The Annals of Statistics*, **14**, 125-129.

_____ (1988). Theoretical comparison of bootstrap confidence intervals (with discussions). *The Annals of Statistics*, **16**, 927-953.

Harvey, A.C. (1989). *Forecasting, structural time series models and the Kalman filter.* Cambridge: Cambridge University Press.

Holbert, D. and Son, M.S. (1986). Bootstrapping a time series model: some empirical results. *Comm. Statist.*, **1**, 3669-3691.

Koopman, S.J., Harvey, A.C., Doornik, J.A. and Shephard, N. (1995). *STAMP 5.0: structural time series analyser, modeller and predictor.* London: Chapman and Hall.

Koopman, S.J. and Shephard, N. (1992). Exact score for time series models in state space form. *Biometrika*, **79**, 823-826.

Li, H. and Maddala, G.S. (1996). Bootstrapping time series models. *Econometric Reviews*, **15**, 115-158.

Nyblom, J. and Mäkeläinen, T. (1983). Comparisons of tests for the presence of random walk coefficients in a simple linear model. *Journal of the American Statistical* Associa*tion*, **78**, 384.

Rogers, A.J. (1986). Modified Lagrange multiplier tests for problems with one-sided alternatives. *Journal of Econometrics*, **31**, 341-61.

Schenker, N. (1985). Qualms about bootstrap confidence intervals. *JASA*, **80**, 360-361.

Souza, R.C. and Neto, A.C. (1996). A bootstrap simulation study in ARMA(p,q) structures. *Journal of Forecasting*, **15**, 4, 343-353.

Stoffer, D.S. and Wall, K.D. (1991). Bootstrapping state-space models: gaussian maximum likelihood estimation and the Kalman filter. *JASA*, **86**, 1024-1033.