Graphical Methods for Investigating the Finite-sample Properties of Confidence Regions: an application to long memory.

Christian de Peretti *
Department of Economics
University of Evry-Val-d'Essonne (France)

Carole Siani
Laboratoire d'Analyse des Systèmes de Santé (LASS)
University of Claude Bernard Lyon 1 (France)

February 27, 2006

Abstract

In the literature, there are not satisfactory methods for measuring and presenting the performance of confidence regions. In this paper, techniques for measuring effectiveness of confidence regions and for the graphical display of simulation evidence concerning the coverage and effectiveness of confidence regions are developed and illustrated. Three types of figures are discussed: called coverage plots, coverage discrepancy plots, and coverage effectiveness curves, that permits to show the "true" effectiveness, rather than a spurious nominal effectiveness. We demonstrate that when simulations are run to compute the coverage for only one confidence level, which is done for classical presentations in tables, all the information useful for building the coverage plot is present. Thus, there is absolutely no loss of computing time by using this method. These figures are used to illustrate the finite sample properties of long range dependence confidence regions. Particularly, we present and comment classical confidence intervals and confidence intervals based on inverting bootstrap tests for the long range dependence parameter in the ARFIMA models. Monte Carlo results on these confidence intervals for various situations are also presented. We show that classical confidence intervals have very poor performances, even the percentile-t interval, whereas confidence intervals based on inverting bootstrap tests have quite satisfactory performance. These intervals are then applied on the S&P500 index to illustrate a realistic case.

Keywords: Graphical method, confidence region, long memory, double bootstrap, inverting tests.

JEL Classification: C10, C13, C14, C15, C63.

*Correspondence to: Christian de Peretti. Address: Batiment Ile-de-France - Boulevard François Mitterrand - 91025 EVRY, FRANCE. Tel: +33 (0)1 69 47 71 95. Fax: +33 (0)1 69 47 70 50. Email: christian.deperetti@univ-evry.fr.

1 Introduction

To obtain evidence on the finite-sample properties of procedures giving confidence regions, econometricians generally resort to Monte Carlo methods. Unfortunately, in the literature there are not satisfactory methods for measuring and presenting the performance of confidence regions. In this paper, techniques for measuring the effectiveness of confidence regions and for the graphical display of simulation evidences as regards the coverage probability and effectiveness of confidence regions are developed and illustrated. These graphs convey much information, in a more easily assimilated form, than tables, *PP plots*, and *QQ plots* can do.

The conventional way to report the results of a Monte Carlo experiment is to tabulate the proportion of confidence regions that contain the true value of the "unknown" parameter for a confidence level of $90\%, 95\%$ and 99%. This approach has two disadvantages. First, the tables provide information about only a few points on the finite-sample distribution of the estimator, this can be an important limitation. Second, the tables require some effort to interpret, and they generally do not make it easy to see how changes in the sample size, the number of degrees of freedom and other factors can affect confidence region coverage probability. In addition to tabular presentation, *PP plots*, and *QQ plots* are also discussed here, and their poor capacity to provide readable and informative results is established. In this paper, we develop and advocate graphical methods that provide more information, and **yield graphs that are easy to interpret**. Dealing with the implementation, we demonstrate that when simulations are run to compute the coverage for only one confidence level, what is made for classical presentations in tables, all the information useful for building the coverage plot is present. Thus, there is absolutely **no loss of computing time by using this method**.

It is often desirable to compare the effectiveness of alternative confidence regions, but this can be difficult to do if all the regions do not have the correct coverage probability. If the values of effectiveness criteria are plotted against (nominal) confidence level, the result will not be very useful, since a method can have a good effectiveness curve due to a coverage distortion and not because of a real effectiveness. Unfortunately, this is what is often implicitly done when region effectiveness is reported in a table. For solving this problem, we propose a method that plots the effectiveness criterion against the coverage probability, *i.e.* the true confidence level; and then, the various methods can be compared. The choice of effectiveness criteria are also discussed.

In order to illustrate and motivate coverage plots and coverage effectiveness curves, these graphs are used to present the results of a study of the properties of confidence regions for long range dependence (denoted LRD). Various confidence regions based on Robinson's estimator (1995) are compared: the confidence region using the asymptotic distribution of the estimator, the percentile and the percentile-t confidence regions using the bootstrapped distribution, and the confidence region based on inverting bootstrapped Robinson's test (1995). Double bootstrap versions of the procedures also exist.

In Section 2, we present the graphs that we propose for experiments dealing with confidence region coverage probability. The use of coverage-effectiveness curves is also presented and discussed. Then, in Section 3, classical and inverting tests confidence intervals are applied on the estimation of the LRD parameter in the ARFIMA models. In Section 4, a number of Monte Carlo results are presented on the various LRD confidence intervals to illustrate the use of coverage plots, and the use of coverage-effectiveness curves. In Section 5, the long memory confidence intervals are applied to the S&P500

index. Section 6 concludes.

2 Graphical methods

2.1 The position of the problem

Let θ be the parameter vector of interest:

$$\theta \in \Theta \subset \mathbb{R}^k.$$

For instance, let us consider the following classical regression model:

$$\begin{aligned} y_t &= z_t\theta + \varepsilon_t \qquad t = 1, \ldots, T, \\ \varepsilon_t &\sim i.i.d.N(0, \sigma_\varepsilon^2), \end{aligned} \tag{1}$$

where y is the vector of the dependent variable, z is a vector of an explanatory variable (it is assumed to be stationary), θ is a scalar of unknown parameter, and ε the unobserved vector of error terms (assumed to be independent of z).

Let τ be the statistic used for constructing the confidence region:

$$\tau \equiv \tau_X(\theta),$$

where X is the observed finite sample. τ can also be a vector.

In our example, $X = (yz)$, and τ can be an estimator of θ and an estimator of the standard error of the estimator of θ: $\tau = (\hat{\theta}, \hat{\sigma}^2_{\hat{\theta}})$. If the OLS estimator is used, $\tau = \left(\frac{z'y}{z'z}, \frac{\hat{\sigma}^2_\varepsilon}{z'z}\right)$, consequently, τ depends on X, and thus on θ since $\frac{z'y}{z'z} = \theta + \frac{z'\varepsilon}{z'z}$. Here, τ depends indirectly on θ, but it can also depend directly on θ, for example in the case of a confidence region based on inverting tests.

The cumulative distribution function (CDF) of τ is denoted F_θ. The distribution of τ has to depend on θ for being able to estimate it.

In our example, $\tau \sim \left(N\left(\theta, \frac{\sigma_\varepsilon^2}{z'z}\right), \frac{\sigma_\varepsilon^2}{z'z}\chi^2_{T-1}\right)$, and thus, depends clearly on θ.

Let R be a confidence region for θ with confidence level [1] $1 - \alpha$:

$$R \equiv R\left(\tau, \{F_\theta^{-1}\}_{\theta\in\Theta}, 1 - \alpha\right).$$

However, there is an infinity of possibilities giving such a confidence region. Each of the possibilities corresponds to a different method. When the notation R is used in the following, we mean one of these methods.

Here are some examples of confidence regions for θ:

$$\begin{aligned} R_1 &= [\hat{\theta} + t_{\alpha/2}\hat{\sigma}_\theta, \hat{\theta} + t_{1-\alpha/2}\hat{\sigma}_\theta], \\ R_2 &= (-\infty, \hat{\theta} + t_{1-\alpha}\hat{\sigma}_\theta], \\ R_3 &= (-\infty, \hat{\theta} + t_{0.5-\alpha/2}\hat{\sigma}_\theta] \cup [\hat{\theta} + t_{0.5+\alpha/2}\hat{\sigma}_\theta, +\infty). \end{aligned}$$

[1] The confidence level for R is the probability of observing the true value of the parameter vector in the random region R, according some distribution F. See Davidson and MacKinnon (1993), chapter 5, for more explanations.

About $\{F_\theta^{-1}\}_{\theta\in\Theta}$, we do not need the whole CDF of τ here, but only of its studentised form, i.e. the $t(T-1)$ distribution. Consequently, let us consider $F_\theta = F = t(T-1)$. Since the studentised statistic is perfectly pivotal [2] in our example, F does not depend on θ. With these notations, $R_1 = [\hat{\theta} + F^{-1}(\alpha/2)\hat{\sigma}_\theta, \hat{\theta} + F^{-1}(1-\alpha/2)\hat{\sigma}_\theta]$. In general, the statistic of interest is not pivotal; this case will be treated in the following.

If θ_0 is the true value of the parameter vector θ that generates the random sample X, then

$$\forall\theta_0 \in \Theta, \forall\alpha \in [0,1], P\left(\theta_0 \in R(\tau(\theta_0), \{F_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha)\right) = 1-\alpha.$$

The graph of $\{(1-\alpha, P(\theta_0 \in R)); \alpha \in [0,1]\}$ is equal to the 45 degrees line. However, the family $\{F_\theta\}_{\theta\in\Theta}$ is not known in general.

Let our example be modified now by taking $z_t = y_{t-1}$. Since z is assumed stationary, θ has to be lower than one in absolute value. In this situation, τ does not follow a t distribution, but a more complicated distribution that depends on θ.

Therefore, the family $\{F_\theta\}_{\theta\in\Theta}$ has to be estimated and we denote its estimation by $\{\hat{F}_\theta\}_{\theta\in\Theta}$. $\hat{F}_\theta$ can be the asymptotic limit of $\{F_\theta\}_{\theta\in\Theta}$ as $T \longrightarrow \infty$, or it can be a distribution derived by bootstrapping, or it can also be some other approximations of F_θ (coming from a first order Taylor expansion, for instance).

In our example, the asymptotic distribution of τ is the $t(T-1)$ distribution. This distribution is chosen for the estimation of F_θ for all $\theta \in \Theta$.

Let us denote

$$\hat{R} \equiv R\left(\tau(\theta_0), \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha\right).$$

The difference between R and $\hat{R}$ can also come from an approximation in the analytical calculus of R. Since $\hat{F}_\theta$ is not exact, we have in general

$$P(\theta_0 \in \hat{R}) \neq 1-\alpha,$$

but we wish $\hat{R}$ such that

$$P(\theta_0 \in \hat{R}) \approx 1-\alpha,$$

i.e. that the graph of

$$\left\{(1-\alpha, P(\theta_0 \in \hat{R})); \alpha \in [0,1]\right\}$$

is near the 45 degrees line.

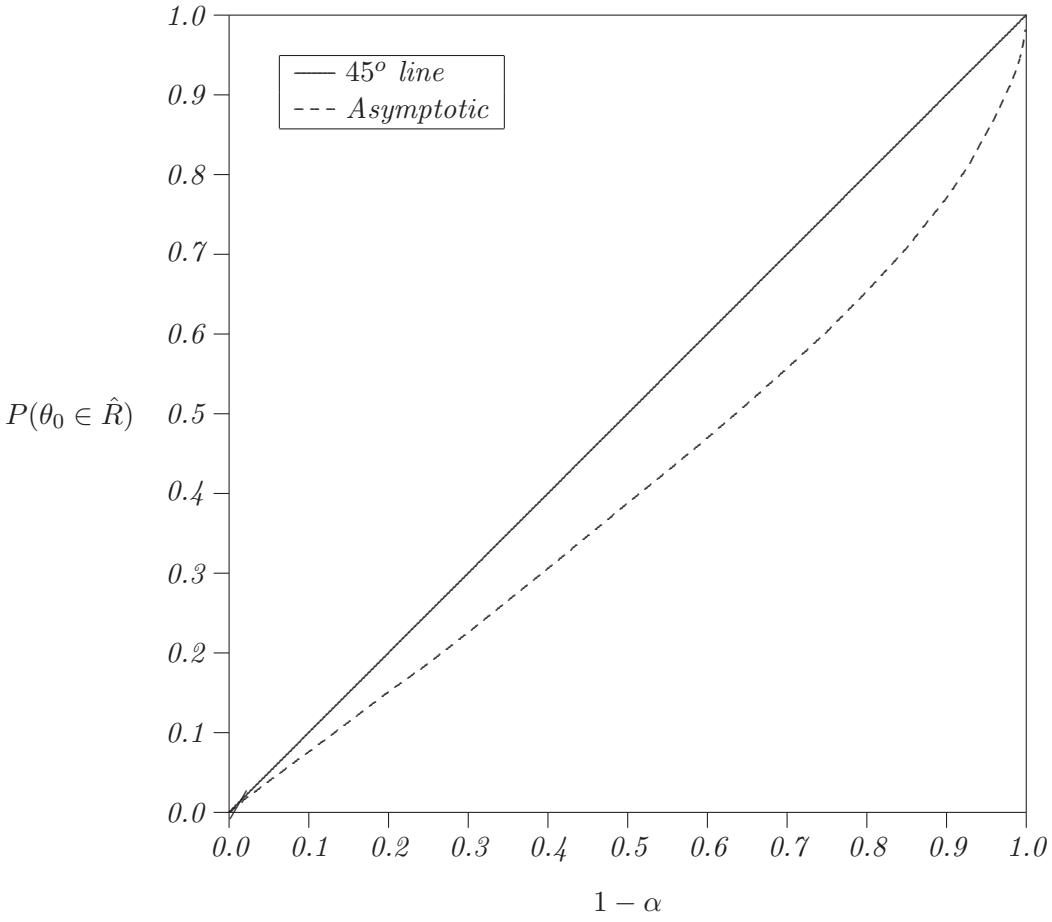
Figure 1 presents the graph of $P(\theta_0 \in \hat{R})$ against $1-\alpha$ for the asymptotic confidence interval for the long memory parameter in the ARFIMA(0,d,0) model when the long memory parameter $d = -0.4$ and the sample size $T = 256$ (see section 4).

$P(\theta_0 \in \hat{R})$ is the *coverage probability*, or just the *coverage*, of the random region $\hat{R}$. It is the true probability that the region will include, or cover, the true value of the parameter vector.

[2] A statistic is pivotal if it does not depend on the parameters of the model under the null.

Figure 1: *Case of ARFIMA(0,d,0) process*

$d = -0.4 \quad T = 256$



2.2 The Monte Carlo procedure

Consider a Monte Carlo experiment in which S realisations of the interest statistic $\tau(\theta)$ are generated using a data generating process (DGP) that is a special case of the model. We may denote these simulated values by $\tau_s(\theta), s \in \{1, \ldots, S\}$.

$P\left(\theta_0 \in R(\tau_s(\theta_0), \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha)\right)$ can be computed using the Monte Carlo experiment:

$$\hat{P}\left(\theta_0 \in R(\tau_s(\theta_0), \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha)\right) = \frac{1}{S}\sum_{s=1}^{S} \mathbb{I}\left(\theta_0 \in R(\tau_s(\theta_0), \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha)\right)$$

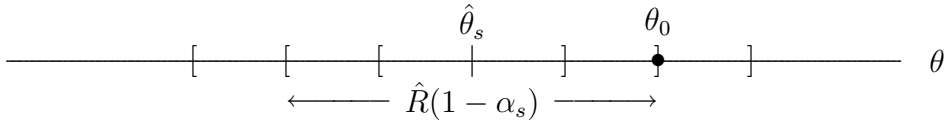
for S very large. $\mathbb{I}(\ .\ )$ denoted an indicator function that takes the value 1 if its argument is true an 0 otherwise.

Set $1-\alpha_s$ the value of $1-\alpha$ such that

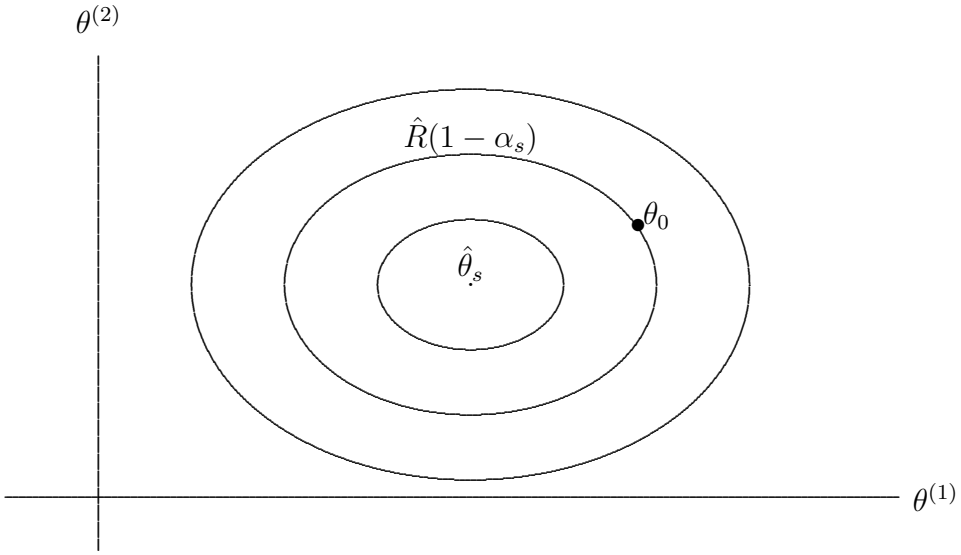
$$\theta_0 \in \partial R\left(\tau_s(\theta_0), \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha\right)$$

where ∂ represents the border of a set of values.

For example:



and also:



$1-\alpha_s$ can be called *critical coverage*. $1-\alpha_s$ can be multiple, but it is easy to assume that it is unique by assuming the natural hypothesis that R (and $\hat{R}$) is increasing with respect to $1-\alpha$ in the sense of inclusion. This hypothesis is obtained if R is optimised using the maximum likelihood principle, for instance, but not only.

It should be noted that

$$(1-\alpha_s \leq 1-\alpha) \Longleftrightarrow \left(\theta_0 \in \hat{R}(\tau_s(\theta_0), \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha)\right),$$

thus,

$$\hat{P}\left(\theta_0 \in R(\tau_s(\theta_0), \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha)\right) = \frac{1}{S}\sum_{s=1}^{S} \mathbb{I}(1-\alpha_s \in [0, 1-\alpha]).$$

In practice, we just count the proportion of $1-\alpha_s$ smaller or equal to $1-\alpha$. In fact, it is the empirical cumulative distribution function of the random variable:

$$1-\alpha_s \equiv 1-\alpha_s\left(\tau_s, \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, R, \theta_0\right).$$

All the graphs we discuss are based on the empirical cumulative distribution function of the confidence level $1-\alpha_s$ of $R(\tau_s, \{\hat{F}_\theta^{-1}\}_\theta \in \Theta, \ .\ )$ if the true value θ_0 of the parameter vector is just over ∂R (the border of R).

Let F_0 denotes the finite-sample distribution of $1-\alpha_s$ such that the coverage probability is equal to the confidence level, *i.e.*

$$F_0(1-\alpha) = P\left(\theta_0 \in R(\tau(\theta_0), \{F_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha)\right) = 1-\alpha.$$

Let F denote the generally unknown (true) finite-sample distribution of $1-\alpha_s$, *i.e.*

$$\begin{aligned} F(1-\alpha) &= P(1-\alpha_s \leq 1-\alpha), \\ &= P\left(\theta_0 \in R(\tau(\theta_0), \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha)\right), \end{aligned}$$

and $\hat{F}$, its Monte Carlo computation:

$$\begin{aligned}\hat{F}(1-\alpha) &= \hat{P}\left(\theta_0 \in R(\tau(\theta_0), \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha\right), \\ &= \frac{1}{S}\sum_{S=1}^{S} \mathbb{I}(1-\alpha_s \leq 1-\alpha).\end{aligned}$$

At any point x_i in the $(0,1)$ interval, it is defined by

$$\hat{F}(x_i) = \frac{1}{S}\sum_{S=1}^{S} \mathbb{I}(1-\alpha_s \leq x_i).$$

See appendix A for the choice of the x_i's.

2.3 Coverage Plots

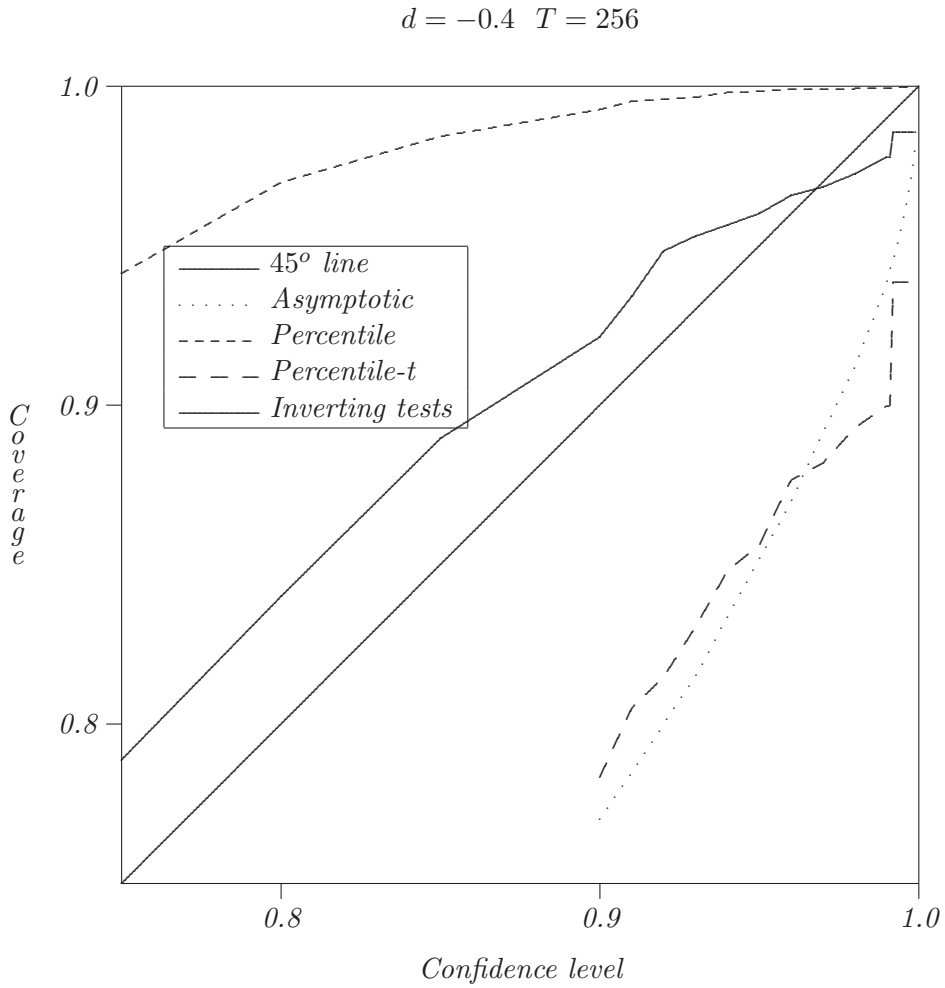
The simplest graph that we will discuss is a plot of $\hat{F}(x_i)$ against x_i. We shall refer to such a graph as a *coverage plot* since it presents the (true) coverage probability against the (nominal) confidence level. If F is correct, *i.e.* $F = F_0$, each of the $1-\alpha_s$ is distributed as an independent uniform random variable $U(0,1)$. Therefore, when $\hat{F}(x_i)$ is plotted against x_i, the resulting graph should be close to the 45 degrees line. Figure 1 presents in fact a *coverage plot*.

On the one hand, coverage plots make it very easy to distinguish confidence regions that perform badly. Moreover, because they show how a confidence region performs for all confidence levels, coverage plots are particularly useful for regions that both over-cover and under-cover the true value of the parameter. The corresponding potential disadvantage of coverage plots is that they can take up a lot of space on the page: For plotting the size against the level for a region, a two dimensional graph is required, whereas the table need only one line or column (one dimension). Nevertheless, plots for several regions can be put into the same graph, whereas a table need also two dimensions. Since, in most cases, we are primarily interested in reasonably high confidence levels, it may make sense to truncate the plot to some values of x more than zero, for instance, $x = 0.75$. It should be clear why coverage plots (and also coverage discrepancy plots defined in subsection 2.4) are often much more informative than tables of coverages at conventional confidence levels such as 0.95. First, why to choose the 0.95 level ? There is nothing special about this level: some investigators may prefer to use the 0.99 level or even the 0.999 level, while predescriptions are often performed at level of 0.75 or even higher. If the curves are plotted for various sample sizes or various parameter values, these plots provide a great deal of information about how the sample size or a parameter value affects the performances of confidences intervals.

> *Figure 2 presents the coverage plot for confidence intervals for the long memory parameter in the ARFIMA(0,d,0) model when the long memory parameter* $d = -0.4$ *and the sample size* $T = 256$ *(see section 4).*

On the other hand, coverage plots do not make it easy to see patterns in the behaviour of regions that perform satisfactory: coverage plots for all confidence regions that behave approximately well will look roughly like the 45 degrees lines. These plots are therefore not very useful for distinguishing among such confidence regions. In this case, we propose

Figure 2: *Coverages plots in the case of ARFIMA(0,d,0) process*



the coverage discrepancy plots in subsection 2.4. Obviously, since they use one dimension for confidence level, coverage plots (and coverage discrepancy plots) cannot use that dimension to represent something else, such as the value of a parameter or the sample size. However, the second dimension of the plots can be used either for plotting the curves for other confidence regions, or for plotting the curves for the same confidence region but for other parameter values or sample sizes.

It should be noted that for obtaining the whole plot (for all the confidence levels), only one simulation experiment [3] is necessary. For obtaining the coverage for only one confidence level, what is made for classical presentations in tables, a full simulation experiment has to be run, as for the coverage plot ! And for providing coverages for three confidence levels: 90%, 95%, and 99% in a table, what is done in most of the paper is to run three experiments ! Thus, **there is absolutely no loss of computing time by using coverage plots**. Moreover, some methods are very computational time consuming, for instance, the double bootstrap-t based on the Robinson's estimator (1995) confidence

[3] We call **one** experiment the set of S simulated series following **one** same and unique DGP and leading to only **one** computed result.

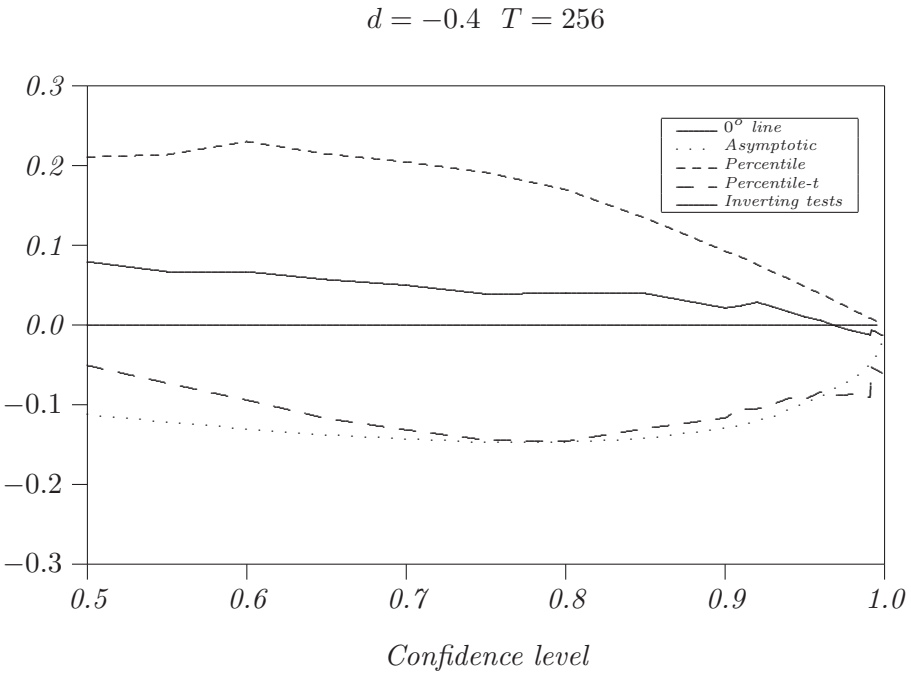
interval presented in subsection 3.4: first, for estimating the long memory parameter by Robinson's method (1995), an iterative optimising algorithm is necessary. Second, for estimating its standard deviation, a bootstrap loop is used in which a Robinson's estimation is run at each bootstrap iteration (this standard deviation is used for studentising the estimator). And third, for computing the distribution function of the studentised estimator, another bootstrap loop is run, in which a bootstrap loop for computing the standard deviation is run at each step. In this circumstance, a simulation method for providing performances that is not too much time consuming is very useful.

2.4 Coverage Discrepancy Plots

For dealing with confidence regions that are well-behaved, it is much more revealing to graph $\hat{F}(x_i) - x_i$ against x_i. We shall refer to such a graph as a *coverage discrepancy plot*. However, the information provided by the graph is partly spurious, reflecting experimental randomness. It is therefore natural to smooth the plots. In Davidson and MacKinnon (1998), they discuss semi-parametric methods for smoothing similar plots. Moreover, because there is no natural scale for the vertical axis, coverage discrepancy plots can be harder to interpret than coverage.

Figure 3 presents the same results than figure 2 using coverage discrepancy plot.

Figure 3: *Coverages discrepancy plots in the case of ARFIMA(0,d,0) process*



2.5 Coverage-Effectiveness Curves

Coverage plots and Coverage discrepancy plots are very useful for dealing with coverage probability, but they are not useful at all for dealing with confidence region effectiveness.

We will discuss graphical methods for comparing the effectiveness of competing regions using *coverage-effectiveness curves*. For an experiment (in which a given DGP is used), these curves can be constructed using the empirical cumulative distribution function of the critical coverage and the chosen effectiveness criterion. Effectiveness criteria will be discussed in the next subsection.

It is often desirable to compare the effectiveness of alternative confidence regions, but this can be difficult to do if all the regions do not have the correct coverage probability. If the values of an effectiveness criterion are plotted against (nominal) confidence level, the result will not be very useful, since claiming that a method is more satisfactory than another one on the basis of an effectiveness criterion has no sense if the methods suffer from different coverage distortions: for example, a criterion providing good results can be spurious due to a default of coverage [4]. Unfortunately, this is what is often implicitly done when region effectiveness is reported in a table.

In order to plot effectiveness against the (true) coverage probability, an experiment has to be performed, preferably using the same sequence of random numbers for each region, to avoid experimental errors. Let the points on the approximate empirical cumulative distribution function be denoted $\hat{F}(x)$, and let the estimated effectiveness for a confidence level of $x = 1 - \alpha$ be denoted $\hat{E}(x)$. They have to be evaluated at a pre-chosen set of points $\{x_i\}_{i=1,\ldots,N}$. As before, $F(x)$ is the probability of getting a critical coverage less than x. Similarly, $E(x)$ is the effectiveness for a confidence level of x. Tracing the locus of points $(F(x), E(x))$ as x varies from 0 to 1 thus generates a coverage-effectiveness curve on a correct coverage-adjusted basis. Plotting the points $(\hat{F}(x_i), \hat{E}(x_i))$, does exactly the same thing, except for experimental error due to the randomness of the Monte Carlo simulations (however, the experimental error converges to zero when the number of Monte Carlo replications goes to infinite). More precisely, it presents the effectiveness criterion against the coverage probability, *i.e.* the true confidence level; and then, the various methods can be compared.

Similarly to the coverage, calculating $\mathrm{E}(x_i)$ for a set $\{x_i\}_i$ is done from the same simulated series, and consequently, it is not necessary to run an additional experiment for each confidence level x_i, and thus, there is no loss of computing time. Moreover, in practice, the calculation for a set $\{x_i\}_i$ is often straightforward by matrix computation: for instance, for the asymptotic, percentile and percentile-t confidence intervals presented in Section 3, the calculus is written with only two lines in the Gauss program, and it takes an almost nil computing time compared to the one for the bootstrap loop.

There is one practical problem with drawing coverage-effectiveness curves by plotting $\hat{E}(x_i)$ against $\hat{F}(x_i)$. For regions that under- or over-cover severely, there may be a region of the coverage-effectiveness curve that is left out by a choice of values of x_i. For solving this problem, a very large number of Monte Carlo replications should have to be chosen, but it is not necessarily possible in practice because of the computing time. Nevertheless, if a region under- or over-covers severely, it cannot be chosen for practical uses, and thus, it is not useful to compute its "true" effectiveness by coverage-effectiveness curves.

[4] For example, if the effectiveness criterion for a confidence interval is the length of the interval, the length has to be as small as possible. However, if there is an error on the coverage of the confidence interval such that it is lower than the confidence level, the length of the confidence interval associated with this level corresponds to a lower coverage and therefore will be smaller than if there is no error since the length is decreasing with the confidence level. The length being smaller, the method seems having good performances, but it is spurious.

2.6 How to choose the effectiveness criterion?

The effectiveness criterion depends on the mathematical problematics under consideration, but also on the economic problematics.

Classically, a confidence region for a parameter θ is based on a studentised estimator (say $\hat{\theta}$) statistic that follows an unimodal distribution as a Gaussian or a Student distribution with mean θ. Let R_1 and R_2 be two $(1-\alpha)$-confidence regions:

$$\begin{aligned} R_1 &= \left[\hat{\theta} - c_1, \hat{\theta} + c_1\right], \\ R_2 &= \left(-\infty, \hat{\theta} - c_2\right] \cup \left[\hat{\theta} + c_2, +\infty\right). \end{aligned}$$

Both the intervals are correct, by definition, if their coverages are equal to $1-\alpha$, but R_2 does not contain likely values for θ. More rigorously, the probability that an elementary interval dx in R_1 contains the true value of θ is larger than if dx is in R_2. In fact, the statistic distribution can be used as a likelihood function, and optimising the interval according to this likelihood function is equivalent to minimising the length of the interval (the proof is trivial by contradiction). The length is measured using the same (mathematical) measure than the one from which the density function of the statistic is derived, *i.e.* Borel or Lebesgue measure in general for the continuous case. Dirac measure can be used for discrete parameters. More generally, even when the confidence region is not directly based on a statistic distribution, as for inverting tests based confidence regions, if two confidence regions have the same confidence level but two different lengths, the one that has the small length must has the largest probability of presence in dx for θ in an elementary interval dx.

The likelihood of the values for θ is not necessarily the only purpose when a confidence region is built. For instance, the confidence intervals based on a Wald statistic are not invariant under a nonlinear reparametrisation (see Davidson and MacKinnon (2001)). Thus, the confidence interval depends on how the model is written. Consequently, in certain situations, it can be preferable to use invariant confidence regions as the ones based on inverting LM or LR tests (see Davidson and MacKinnon (2001)).

Finally, the choice of the effectiveness criterion can also depend on the economic purpose. Let us consider the following example coming from Siani and Moatti (2003) in the context of health economic evaluations.

> *In cost-effectiveness analyses, which compare one or more treatment(s) with a standard treatment on the two-fold basis of cost and medical effects, health economists often use the incremental cost-effectiveness ratio (ICER) as a summary measure. The ICER statistic, in which a new therapy T_1 is compared with a standard therapy T_0, is defined by:*
>
> $$R = \frac{\mu_{C1} - \mu_{C0}}{\mu_{E1} - \mu_{E0}} = \frac{\mu_{\Delta C}}{\mu_{\Delta E}},$$
>
> *where μ is the true mean value of (subscripts) costs (C) and effects (E) for treatments number 1 and number 0* [5]*. The ICER can be estimated (among other possibilities) as follows: on the basis of data collected from two groups of patients, each undergoing one of the forms of therapy (group number 1, consisting of individuals*

[5] This ICER can be interpreted as the additional resources necessary to obtain a gain of one additional unit in health effects due to the use of treatment T_1 rather than treatment T_0.

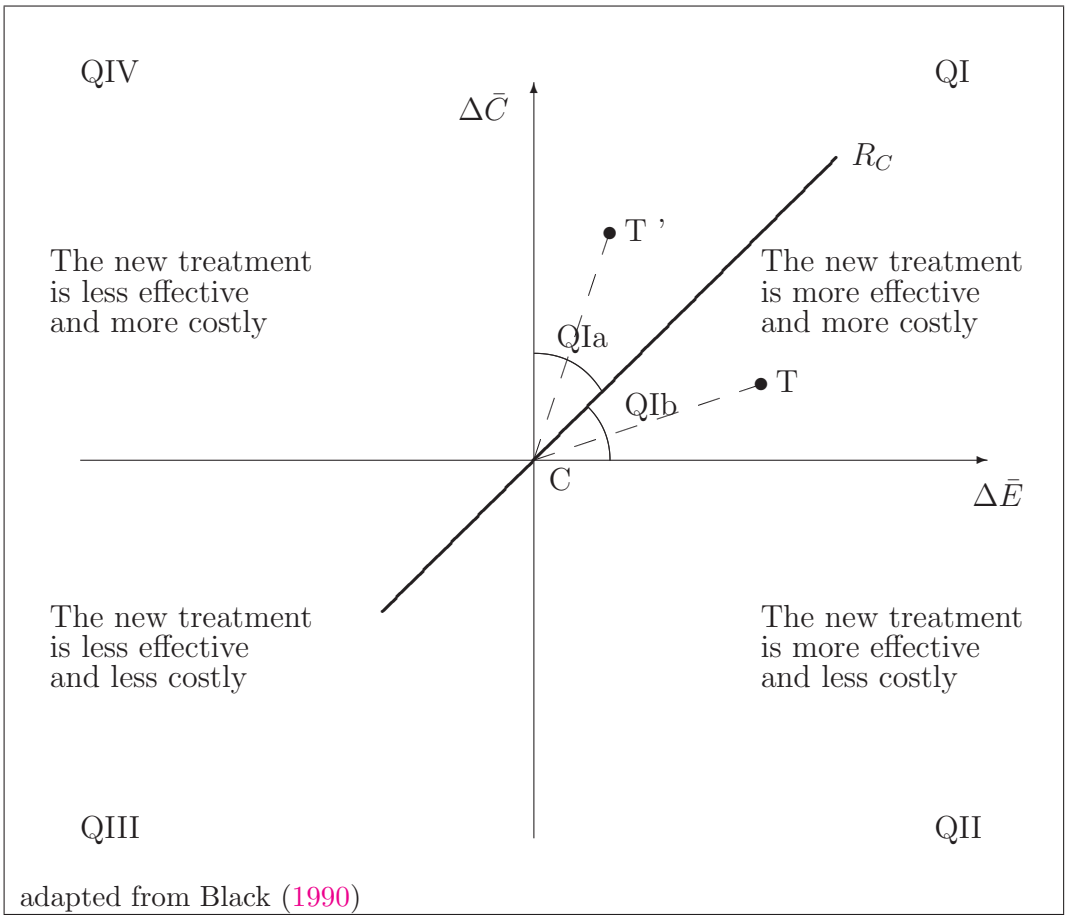
that underwent treatment T_1 and group number 0, consisting of individuals that underwent treatment T_0):

$$\hat{R} = \frac{\overline{C_1} - \overline{C_0}}{\overline{E_1} - \overline{E_0}} = \frac{\Delta\bar{C}}{\Delta\bar{E}} \ ,$$

where $\overline{C_1}$, $\overline{C_0}$ are the sample mean of the costs and $\overline{E_1}$, $\overline{E_0}$ in the two treatments arms are the sample mean of effects. The observed difference between the mean costs is denoted $\Delta\bar{C}$, respectively the observed difference between the mean effects is denoted $\Delta\bar{E}$.

The cost-effectiveness (CE) plane (see Black (1990)) presented in Figure 4, is often used to describe the decision-making rules which follow from the results of the CEA. The vertical, respectively horizontal, axis corresponds to $\Delta\bar{C}$, respectively $\Delta\bar{E}$.

Figure 4: The cost-effectiveness plane



The new treatment T_1 is represented by the point T in Figure 4 and the standard treatment T_0 is represented by the point C. It should be noted that the slope of the semi-straight line (CT) corresponds to the value of $\hat{R}$ [6]. In order to decide whether to adopt treatment T_1 rather than treatment T_0, it is necessary to introduce a ceiling ratio, denoted R_C, corresponding to some maximum value of the ICER that people are prepared to pay to achieve this additional effectiveness [7]. Thus, in quadrants

[6] In QI, the steeper this slope is, the greater the cost of an additional unit of effect and the less worthwhile the new therapy will be in comparison with the standard therapy.

[7] This ceiling ratio is assumed to be determined in an exogenous way.

QI and QII, if the ICER is lower than the ceiling ratio, then treatment T_1 should be adopted, and conversely, if the ICER is greater than the ceiling ratio, then treatment T_0 should be kept. In quadrants QIII and QIV, the opposite reasoning holds. In fact, what is really determinant at the decision-making level, is whether the point $(\Delta\bar{E}, \Delta\bar{C})$ is over or under the straight line associated with R_C.

A confidence interval should naturally have to be built from the distribution of $\hat{R}$. It should be noted that a confidence interval for R will be represented by a (angular) sector around $\hat{R}$ on the CE plane. Nevertheless, this kind of confidence intervals is very biased or has no mathematical sense when $\mu_{\Delta E}$ is statistically close to zero. However, in this latter case, there is absolutely no problem for taking decision-making on the basis of the $(\Delta\bar{E}, \Delta\bar{C})$ pair because the true question at the decision-making level is whether the confidence sector is under or above the straight line corresponding to the ceiling ratio on the CE plane (see Siani and Moatti (2003) and Siani and de Peretti (2004)). Fieller's method (1954) allows to solve perfectly this problem because it is directly based on the distribution of $(\Delta\bar{E}, \Delta\bar{C})$: Fieller's method can structurally provide confidence region of the form $]-\infty, R^U] \cup [R^L, +\infty[$ that simply corresponds to a sector (also around $\hat{R}$) containing the vertical axis and it permits to take decision. In this context, the length of the confidence region is not relevant for measuring the performances of confidence regions since regions having the form of the complement of an interval will have infinite length whereas they are quite efficient for decision-making. Only, the angle of the confidence sector is really informative since representing the uncertainty of the estimation (smaller the angle is, smaller the uncertainty will be, better the confidence region will be) [8].

In addition, the angle criterion is not disconnected from the likelihood principle: smaller the angle is, smaller the sector will be (in the sense of set inclusion, since the surface is infinite), and more the region will be likely with respect to the bivariate distribution of $(\Delta\bar{E}, \Delta\bar{C})$.

For concluding, the choice of the effectiveness criterion can be very different depending on the situation (mathematical or economical problematics).

2.7 How to chose the "best" method ?

The coverage plots and the effectiveness curves are very useful for choosing among methods that have reasonable coverage distortions: they permit to make *arbitrage* between the coverage distortion and the true effectiveness for each methods and then to chose the most appropriate. There is no criterion that permits to select the "best" method from the combination of both the whole coverage and the whole effectiveness curves. However, the following rule can be used:

1. First, select the methods that have not their coverage plots too far from the 45^o degree line. The coverage is the most important feature, since if the coverage error is too large, the method is false.

2. Among the methods selected in the first step, select the methods that have the best effectiveness curves. A large gain in the effectiveness can compensate the a small loss in the coverage. This largeness depends on the situation and on the decision-maker.

[8] The one-dimensional ICER confidence region is only a nonlinear transformation of the underlying two dimensional confidence sector (that is really of interest), and it is presented only for practical reason of presentation for decision makers.

2.8 Other common types of plots

In this subsection, two types of plots are presented and commented: the *PP plots* and the *QQ plots*. However, we conclude that both these plots give too poor information for analysing satisfactory the performances of confidence regions.

It would be more conventional to graph the CDF of the statistic τ instead of the CDF of its critical coverage $1 - \alpha_s$. The former can be graphed against the CDF of a hypothesised distribution, for example, the one used for the building of the confidence region. This graph would be what is often called a *PP plot*; see Wild and Gnanadesikan (1968). Note that the coverage plots are PP plots of critical coverage.

Another common type of plot is a *QQ plot*, in which the quantiles of τ are plotted against the quantiles of its hypothesised distribution. If the distribution of τ is closed to the hypothesised one, the plot will be close to the 45^o line. This approach, which has been used by Chesher and Spady (1991), among others, can yield useful information, because it shows the distortion between the hypothesised distribution and the true distribution. However, QQ plots have also disadvantages. One serious problem is that QQ plots have no natural scale for the axes. Thus, if the hypothesised distribution changes, so will that scale. This makes it difficult to plot on the same axes statistics that have different distributions. Moreover, although the *QQ plots* can certainly make clear that a confidence region does not work perfectly, these plots provide much less useful information than coverage plots. They also take up much more space than do coverage plots restricted to $[0.75, 1] \times [0.75, 1]$. It is extremely difficult, on the basis of a *QQ plot*, to see how the performance of a confidence region changes with a parameter or the sample size, something that is immediately obvious from the coverage plots and even from the coverage discrepancy plots. More details about criticisms of these plots can be found in Davidson and MacKinnon (1998) Section 2 in the context of test statistics.

In our view, however, there are two major problems with the use of *PP plots* and *QQ plots*. The first problem is that it is assumed that the confidence region is built from a hypothesised distribution of a statistic. This is not the case, for example, for confidence regions based on inverting tests (see subsection 3.5). Consequently, how to build the *PP plots* or the *QQ plots* for these regions? This criticism holds also in the context of tests since the acceptance region for a hypothesis test can be based on inverting confidence regions.

The second problem, that holds for all confidence regions and hypothesis tests, is that a procedure for providing a confidence region (for a parameter vector) or an acceptance region (for a test hypothesis) is not totally defined by the underlying statistic and its hypothesised distribution. A last step is missed: the building of the final region from both these elements. This last step is neither unique nor necessarily simple. From an one dimensional statistic distribution, for example, very different regions can be built: unilateral, bilateral. In the case of a bilateral interval, it is desirable to choose its limits by minimising the length of the interval. However, it is often too complicated and an equiprobable interval (*i.e.* having the same probability at the right and at the left) is generally preferred. In the case of a two dimensional statistic distribution, the region can be an ellipse (if there is no constraint), or a sector (as for the ICER estimation, see subsection 2.6), or a band (as for the net benefit estimation, see (1998; 1998)) depending on the economic problems. Thus, many different confidence regions and hypothesis tests can be defined from the statistic and its distribution. Consequently, presenting only the *PP plots* or the *QQ plots* can be insufficient for showing the performances of the methods:

they provide the error between the true distribution of the statistic and its hypothesised distribution, but it can be very difficult to see how this error affects the performances of the methods (i.e. the coverage and other criteria).

3 Long range dependence confidence regions

In order to illustrate and motivate coverage plots and coverage effectiveness curves, these graphs are used to present the results of a study of the properties of long range dependence confidence regions. Various confidence regions based on Robinson's estimator (1995) are compared: there are the confidence region using the asymptotic distribution of the estimator (subsection 3.2), the percentile (subsection 3.3) and the percentile-t (subsection 3.4) confidence regions using the bootstrapped distribution, and the confidence region based on inverting bootstrapped Robinson's test (1995) (subsection 3.5). Double bootstrap versions of the procedures also exist. See table 1 for a listing of these methods.

Table 1: Listing of the confidence region methods

	Classical	Inverting tests
Asymptotic	Wald confidence interval	inverting asymptotic tests
Bootstrap	Percentile	
Bootstrap-t	Percentile-t with asymptotic variance	Inverting Bootstrap tests
Double Bootstrap-t	Percentile-t with bootstrap variance	Inverting double bootstrap tests

Robinson's estimator (1995) is chosen rather than others, because, for example, Hurst and Lo methods have less satisfactory properties, and Higuchi's estimator (1988) seems to be consistent only for $d \in [-0.5, 0.5]$ whereas Robinson's estimator (1995) is consistent for $d \in [-0.5, 1]$.

Since Robinson's statistic (1995) is not exactly pivotal, bootstrapping will not perform perfectly. However, Robinson's statistic (1995) is asymptotically pivotal, which means that its distribution does not depend asymptotically on any nuisance parameters. In that case, bootstrapping should yield confidence regions that are accurate to higher order, in the sample size, than the confidence provided by asymptotic theory; see Beran (1988), Horowitz (1994), and Davidson and MacKinnon (1996b; 1996a).

In our Monte Carlo simulations, we deal with the univariate linear ARFIMA(p,d,q) model with Gaussian errors. Consequently, the error terms of the bootstrap samples in the methods presented in this section are obtained from a Gaussian distribution rather than by resampling from the residuals, since the error terms are normally distributed. This case is deliberately chosen for the following reason. The finite sample distribution of the statistic can suffer from two types of distortion from the asymptotic distribution:

1. The first type of distortion comes from the error terms that can be not Gaussian. This distortion due to non-Gaussian error terms is often quickly reduced thanks to

the limit central theorem (in many bootstrap studies, parametric bootstrap works similarly than nonparametric bootstraps).

2. The second type of distortion coming from the fact that the denominator of the studentised statistic is not independent from the numerator. Conversely, the second type of distortion can be more persistent with respect to the sample size, especially in the case of long range dependence time series (see de Peretti and Marimoutou (2002) where parametric bootstrap and nonparametric bootstraps work similarly but both have size distortions).

Thus, in our Monte Carlo experiments, we focus on the Gaussian case to show that asymptotic methods have serious problems that are not due to a misspecification of the error terms distribution but to the second type of distortion. In this situation, the use of bootstrap combined with inverting tests is greatly advised.

In practice, macroeconomic series are often non-Gaussian and financial series are almost always strongly non Gaussian. The bootstrap procedure can be adapted to this kind of data by the use of nonparametric methods. In our case, when the replications of the series are generated, we just have to draw the bootstrap error terms in the empirical distribution of the residuals of the estimation of the series rather than in a Gaussian distribution (see section 5 for an application of this nonparametric bootstrap, and Davidson (1998) and appendix B.1 for examples and details of other nonparametric bootstrap distributions).

3.1 The model and the estimator

Robinson's estimator (1995) pertains to the normal linear ARFIMA(p,d,q) model [9]:

$$\phi(L)(1-L)^d x_t = \theta(L)\varepsilon_t \qquad t \in \{1,\dots,T\}, \tag{2}$$

$$\{\varepsilon_t\} \sim i.i.d.N(0,\sigma^2), \tag{3}$$

where

- ϕ and θ are polynomials that have not necessarily all roots outside the unit circle,
- $\sigma^2 < \infty$,
- L is the lag operator,
- d is the differencing parameter and takes a real value.

In some circumstances, a long-memory process may be approximated by a fractionally integrated model; hence estimating for long-memory can be done by an estimation of d.

For applying Robinson's procedure (1995), a truncation point m has to be chosen and we choose the rule proposed by de Peretti and Marimoutou (2002):

$$m = \sqrt{T \times 2}.$$

[9] The ARFIMA model is presented in detail by Granger and Joyeux (1980) and Hosking (1981).

3.2 The classical asymptotic confidence region

The classical asymptotic confidence region for the long range dependence parameter d is obtained from Robinson's estimator (1995) as follows:

$$[d_1, d_2] = \left[\hat{d} - \hat{\sigma}(\ \hat{d}\ )\ q_{\alpha/2}\ ,\ \hat{d} + \hat{\sigma}(\ \hat{d}\ )\ q_{1-\alpha/2}\right],$$

where:

- $\hat{d}$ is the estimate of d,
- $\hat{\sigma}(\hat{d})$ is the asymptotic estimate of the standard deviation of $\hat{d}$, that is $0.5/\sqrt{m}$, (see Robinson (1995)),
- q_α is the α-quantile of the asymptotic distribution of the studentised $\hat{d}$ statistic, that is the normal distribution, (see Robinson (1995)).

Since $\hat{d}$ is asymptotically Gaussian only for $d_0 \in \left[\frac{1}{2}, \frac{3}{4}\right)$ (see Velasco (1999)), the procedure is valid only for this range of values.

For using our graphical methods, we need to calculate the confidence level $1-\alpha_s$ such that d_0, the true value for d, is included in the border of the confidence region ∂R, that is $\{d_1, d_2\}$ here. Consequently, we obtain:

$$1 - \alpha_s = F\left(\left(\frac{d_0 - \hat{d}}{\hat{\sigma}(\ \hat{d}\ )}\right)^2\right)$$

where F is the CDF of a χ^2 variable with one degree of freedom.

Dealing with the effectiveness, denoted $E(1-\alpha)$ in subsection 2.5, we measure it by the confidence interval length expectation. It should be noted that the length of confidence intervals depends on the confidence level $1-\alpha$. Consequently, the length expectation, corrected or not by the technique proposed in subsection 2.5, has to be calculated for all the confidence levels. The length of the asymptotic confidence interval is:

$$e(1-\alpha) = 2\ q_{1-\alpha/2}\hat{\sigma}(\ \hat{d}\ ).$$

For each replicated series, the length of the confidence interval is calculated for each level $1-\alpha$. For the whole Monte Carlo experiment, *i.e.* for all the replicated series, the expectations for each level are estimated by the sample mean of the lengths $e(1-\alpha)$.

In the computational program, a set of values for the confidence level is defined:

$$\{0.001, 0.002, \ldots, 0.009, 0.01, 0.02, \ldots, 0.09, 0.1, 0.2, \ldots$$
$$\ldots, 0.9, 0.91, 0.92, \ldots, 0.99, 0.991, 0.992, \ldots, 0.999\}.$$

For each of these values, the vector $e(1-\alpha)$ is computed (with matrix computation, it is very straightforward).

3.3 The percentile confidence interval

For a general presentation of *percentile* and percentile-t methods, see Davidson and MacKinnon (1993), Efron and Tibshirani (1993), Hall (1992), Hjorth (1994), and Shao and Tu (1995).

Let $\{\hat{d}^b\}_{b=1}^B$ denotes the set of statistics simulated by bootstrap, where B is the number of bootstrap replications. The *percentile* confidence interval is given by

$$(b_1, b_2) = \left(\hat{d}^{[B\alpha/2]}, \hat{d}^{[B(1-\alpha)/2]}\right],$$

where $[\ .\ ]$ is the rank statistic of the integer part.

The level such that $d_0 \in \partial R$ is equal to:

$$1 - \alpha_s = 2\min\{pv, 1 - pv\},$$

where

$$pv = \frac{1}{B}\sum_{b=1}^{B} I(\hat{d}^b \leq d_0).$$

Since $\hat{d}$ is consistent only for $d_0 \in \left[-\frac{1}{2}, 1\right)$ (see Velasco (1999)), the procedure is valid only for this range of values.

More details about the bootstrap procedure have to be provided. We go through the following steps:

- First, the data must be estimated using a model as close as possible to the real data. This model will be used as a *data generating process* (DGP) for generating simulated samples, and then, the simulated replications $\hat{d}^b$ of the statistic $\hat{d}$. We propose here to estimate an ARFIMA(1,d,1) model [10]. For the parameter d, the value $\hat{d}$ is chosen (estimated in a semi-parametric way here by Robinson's estimator). The remaining parameters are estimated by maximum likelihood conditionally to $\hat{d}$. We prefer this procedure rather than estimating all the parameters by an one step estimation by maximum likelihood, because we need the estimation of d in the ARFIMA(1,d,1) model in the same way as the estimation of d for the confidence interval. Otherwise, the replications $\hat{d}^b$ can follow a probability law too different from the probability law that generates $\hat{d}$. Since the aim of the bootstrap procedure is to estimate the distribution of $\hat{d}$, the method can be biased.

- The second step of the bootstrap procedure is to generate B replications $\{x^b\}_{b=1}^B$ of the time series x. B has to be sufficiently large to avoid random effect of the bootstrap experiment: at least several hundreds, several thousands is desirable. It is also desirable that $(1-\alpha)(B+1)$ be an integer (see Davidson and MacKinnon (1993)). Each x^b is generated from a Gaussian ARFIMA(1,d,1) with the previously estimated parameters (see section 5 for the non-Gaussian bootstrap). For each x^b, d is estimated by Robinson's procedure, providing $\hat{d}^b$, a bootstrap replication of $\hat{d}$. And then, the critical coverage is calculated using the set of $\{\hat{d}^b\}_b$.

For easily calculating the effectiveness, here the length of the interval, we propose the following method. The length of the interval is given by

$$e(1-\alpha) = \hat{d}^{[B(1-\alpha)/2]} - \hat{d}^{[B\alpha/2]}.$$

Rather than calculating $e(1-\alpha)$ for a predetermined set of $1-\alpha$, as in subsection 3.2, we prefer calculating each possible different length since the number of possibilities is finite,

[10] A more sophisticated estimate can be done by estimating an ARFIMA(p,d,q) model and choosing p and q by a selecting criterion as AIC or BIC, but the large flexibility of the ARFIMA(1,d,1) model is sufficient for our illustration here.

because the bootstrap distribution is discrete. The corresponding levels are calculated only later. Using matrix computation, it is very straightforward: the vector of lengths is:

$$e = \hat{d}^{\star}(\text{trunc}(B/2+1) : B) - \hat{d}^{\star}(\text{ceil}(B/2) : 1),$$

where $\hat{d}^{\star}$ is the ordered vector of bootstrap replications. The function *trunc* gives the greatest integer smaller than the argument and the function *ceil* give the smallest integer greater than the argument. The associated confidence levels are the sequence:

$$\left\{\frac{2b}{B+1}\right\}_b \text{ where } b \in \left\{0, \ldots, \text{trunc}\left(\frac{B+1}{2}\right) - 1\right\}.$$

3.4 Percentile-t confidence interval

The *percentile-t* procedure is similar to the percentile procedure, but rather than using directly the estimator of d, the studentised form of this statistic is used.

$$\tau = \frac{\hat{d} - d}{\hat{\sigma}(\hat{d})},$$

$\hat{\sigma}(\hat{d})$ will be discussed latter. This leads a to statistic that is asymptotically pivotal, and this permits to obtain a higher rate of convergence for the method. τ is bootstrapped to obtain B replications, denoted τ^b, as follows:

$$\tau^b = \frac{\hat{d}^b - \hat{d}}{\hat{\sigma}(\hat{d}^b)}.$$

The *percentile-t* confidence interval is

$$\left[\hat{d} - \tau^{[1-\alpha/2]}\hat{\sigma}(\hat{d}), \hat{d} - \tau^{[\alpha/2]}\hat{\sigma}(\hat{d})\right)$$

The coverage corresponding to $d_0 \in \partial R$ is equal to:

$$1 - \alpha_s = 1 - 2\min\{pv, 1 - pv\},$$

where

$$pv = \frac{1}{B}\sum_{b=1}^{B} \mathrm{I}(\hat{\tau} \leq \tau_0),$$

and

$$\tau_0 = \frac{\hat{d} - \mathbf{d_0}}{\hat{\sigma}(\hat{d})}.$$

The length for the $(1-\alpha)$-confidence interval is:

$$\left(\tau^{[1-\alpha/2]} - \tau^{[\alpha/2]}\right)\hat{\sigma}(\hat{d}).$$

As for percentile procedure, the *percentile-t* procedure is valid only for $d_0 \in \left[-\frac{1}{2}, 1\right)$.

The DGP for replicating τ is determined in the same way as for the percentile as well as the lengths of the intervals.

$\hat{\sigma}(\hat{d})$, the estimated variance of $\hat{d}$, is calculated in two different ways: the first one is the asymptotic estimate (see Subsection 3.2), the second one is the bootstrap estimate. For

obtaining the bootstrap estimator of $\sigma(\hat{d})$, the time series are estimated and replicated B_2 times in the same way as previously. B_2 can be different from B. Since it is less important, it is generally taken smaller than B. For each replicated series, $\hat{d}$ is computed leading to a set of $\left\{\hat{d}^b\right\}_b$ from which the standard error is computed. It should be noted that when the *percentile-t* method is used, and thus replicated series are generated for computing the test P value, the bootstrap estimator of $\sigma(\hat{d})$ must be applied on each bootstrap replication of the series for obtaining the studentised statistics. This leads to replications of replicated series. This method is often called *double bootstrap*.

A last remark about the *percentile-t* method: even if it has a better asymptotic convergence rate than the percentile method, in finite sample, the studentisation can produce a statistic that is farther from pivotal than the original statistic. This instability can be catastrophic, see among others Li and Maddala (1996), Berkowitz and Kilian (2000), Davidson (2000), and Siani and Moatti (2003). Thus, we have to check by Monte Carlo experiments that this method remains stable in finite sample.

3.5 Confidence region based on inverting tests

For a general presentation of confidence intervals based on inverting tests, see Davidson and MacKinnon (1993) Chapter 5, and for confidence intervals based on inverting bootstrap tests, see Davidson and MacKinnon (2001).

Let $T_{1-\alpha}(d')$ denote the result of a test for $H_0 : d = d'$ against $H_1 : d \neq d'$.

$$T_{1-\alpha}(d') = \begin{cases} 1 & \text{if } d = d' \text{ is retained} \\ 0 & \text{otherwise} \end{cases}.$$

The confidence region built by inverting tests for a confidence level of $1-\alpha$ is defined as following: d' is in the region if and only if the test retains $d = d'$ for a significance level of $1-\alpha$, *i.e.*

$$d' \in \hat{R}(1-\alpha) \Longleftrightarrow T_{1-\alpha}(d') = 1.$$

Let $p(d')$ denote the P value of the test for $d = d'$. We have

$$T_{1-\alpha}(d') = 1 \Longleftrightarrow p(d') \geq \alpha.$$

In our case, the region is an interval. Consequently, it can be defined by both its limits, say $d_{\inf}$ and $d_{\sup}$ (corresponding respectively to the lower and the upper limit). These limits correspond to both the values of d' such that $p(d') = \alpha$ (see Davidson and MacKinnon (1993) Chapter 5). For computing the P value, see subsections 3.5.1 and 3.5.2.

For using our graphical methods, the critical confidence level $1-\alpha_s$ has to be calculated. This level $1-\alpha_s$ is such that $d_0 = d_{\inf}^{(s)}$ or $d_0 = d_{\sup}^{(s)}$, where $^{(s)}$ indicates that the values come from a simulated series. But it is known that $p(d_{\inf}^{(s)}) = 1-\alpha = p(d_{\sup}^{(s)})$, thus

$$1-\alpha_s = p(d_0),$$

where $p(d_0)$ is the P value of the underlying test at $d = d_0$. Fortunately, for confidence regions based on inverting tests, the critical confidence level is very easy to calculate.

With regard to the effectiveness of this method, a problem arises since the test is not defined for all values for d. For computing the confidence interval limits, the values d' for d such that $p(d') = \alpha$ have to be found. Classically, $p(d)$ is a function that increases from 0 to 1 and then decreases from 1 to 0 when d goes from $-\infty$ to $+\infty$. In our situation, the

method is only defined for $d_0 \in \left[-\frac{1}{2}, 1\right)$, thus, the function reaches its minimum values for $d_0 = -\frac{1}{2}$ and $d_0 = 1$ and the defined confidence levels go from 0 to a certain maximum value, say $1 - \alpha_{\max}$, that is lower than 1.

Since $p(d)$ is random, depending on the realisation of the observed time series, $1-\alpha_{\max}$ will be random, going from 0 to 1. The consequence is that for all predefined confidence levels $1 - \alpha$, there is a positive probability that the confidence interval is not defined, and thus the expectation of the interval length is also undefined. It should be noted that this does not mean that the method does not work. In fact, it says rigorously that there is no sufficient information in the time series for allowing an estimation whereas the asymptotic confidence interval gives meaningless limits and the percentile(-t) methods give always numerical limits based on computational proportions that can have no sense. For illustrating all the same the effectiveness of confidence interval based on inverting tests, we detail results for two particular series in section 5.

3.5.1 Inverting asymptotic tests

More precisely, if the asymptotic Wald test based on an estimator of d (Robinson's one, or another estimator) is used, the P value can be built as following [11]

$$p(d') = 1 - F_{d'}\left(\tau(d')\right),$$

where

$$\tau(d') = \left(\frac{\hat{d} - d'}{\hat{\sigma}(\hat{d})}\right)^2,$$

$F_{d'}$ is the asymptotic cdf of $\tau(d')$, and $\sigma(\hat{d})$ is the standard deviation of $\hat{d}$. Any explanations must be given for why F is indexed by d' since we can think that $\tau(d')$ follows simply a χ^2_1. If $d' \in (-\frac{1}{2}, \frac{3}{4})$, the Robinson's estimator is Gaussian under the null that $d = d'$, and $F_{d'}$ is simply the cdf of a χ^2_1 random variable. In this case, the confidence interval built by inverting asymptotic tests is equal to the classical asymptotic confidence interval. However, when $1 - \alpha$ goes to 1, then $d_{\inf}$ goes to $-\infty$ and $d_{\sup}$ goes to $+\infty$, and we necessarily fall in the non Gaussian case, and F_d is not the cdf of a χ^2.

It is not very useful to develop this method only for obtaining an extension of the classical asymptotic confidence interval (see Subsection 3.2) for the case where d_0 is not included in $\left[-\frac{1}{2}, \frac{3}{4}\right)$. Moreover, confidence regions based on inverting bootstrap tests (see Subsection 3.5.2) will provide better approximation than with asymptotic tests. Consequently, this method is not presented.

3.5.2 Inverting bootstrap tests

We consider here bootstrap Wald tests based on an estimator of d (Robinson's one, or another estimator). More precisely, for the Monte Carlo experiments, the parametric (single and double) bilateral bootstrap test based on Robinson's estimator is used: see its description in Appendix B.1 [12]. Obviously, as for percentile methods, nonparametric

[11] The P value will be more complicated for bootstrap because it takes into account the asymmetry of the statistic distribution: see subsection B.1.2.

[12] Our procedure is inspired from the test of de Peretti (2003), but this last one is reduced to the case where the null hypothesis is "no long memory" ($d_0 = 0$), whereas our procedure extends to the case $d_0 \in \left[-\frac{1}{2}, 1\right)$. So, for more discussions about double and single bilateral bootstrap tests for long memory, see de Peretti (2003).

versions of the test can be used. In our experiments, nonparametric versions of the test are not used. We restrict our attention to Gaussian procedures for permitting to see the gain of the use of inverting tests confidence regions compared to not inverting tests, without noisy error due to nonparametric estimation. Nevertheless, the basic nonparametric test (only resampling) is applied in Section 5 and the details of this method and of others are in appendix B.1. Both the single and the double bootstraps are discussed, since, even if the double bootstrap is asymptotically better than the single bootstrap, it has the disadvantage of having a computing time B_2 times larger than the single version, where B_2 is the number of bootstrap replications used for the computation of the standard deviation of the test statistic. Finally, the unilateral version of the test is not presented, since it has absolutely no advantage compared to the bilateral version.

3.5.3 Note on the LRD estimation of differentiated series

When the values for d such that $p(d) = 1 - \alpha$ are sought for building the confidence interval based on inverting tests, it can be frustrating to stop at -0.5 and 1. A solution that appears naturally is to differentiate the series such that they are stationary for example. It can be done by successive ADF tests, differentiating the series at each step, until the series looks stationary. If the series is differentiated n times, the LRD estimate is then defined as equal to the LRD estimate of the differentiated series plus n: $\hat{d}(\Delta^n x) + n$. The problem is that the LRD estimate on the original series (when it can be calculated) is in general unequal to $\hat{d}(\Delta^n x) + n$ because the bias coming from the estimation on the original series is in general different from the bias coming from the estimation on the differentiated series (even if both the estimators are consistent). This difference causes jumps into the P value function, leading to problematic resolution of the equation $p(d) = 1 - \alpha$. Consequently, we advise to not use differentiation for inverting tests without furthur improvement of this methodology in the future.

4 Monte Carlo experiments

All the experiments deal with Gaussian ARFIMA(p,d,q) processes. Since the constant term is only a location parameter and does not influence the regions performances, it is set to zero. Similarly, σ, the standard deviation of the model, is only a scale parameter and does not affect at all the performances, is set to one. The test statistic depends then on the parameters (ϕ, d, θ) and T in Equation 2. In a first step, we focus on the coefficient d and the sample size T, setting ϕ and θ to zero. $T = 2^n$ is used, where n is an integer. Each experiment is run with $S = 10,000$ replications of the time series, using the same random numbers for avoiding additional experimental errors (variance). For each replication in each experiment, the same random numbers for bootstrapping are used for avoiding random effect in the result of the bootstrap method between the replications.

4.1 Case of ARFIMA(0,d,0) processes

We pick combinations of d and T in table 2 to investigate. B is the number of bootstrap replications [13]. It should be noted that the values for B chosen for these experiments are sufficient for illustrating the methods since in our case the distortions found in the results

[13] B is chosen to meet the capacity of Gauss software to store a $T \times B$ matrix which is used in matrix computation for these simulation experiments.

Table 2: choice of d and T

Case	d	T	B	B_2
1.1	-0.4	256	488	0
1.2	0	512	244	0
1.3	0.4	256	488	0
1.4	0.8	128	976	0

are not due to a gap of bootstrap replications (even if increasing B can improve a little the results) but to large distortions of the distribution of the test statistic depending on the parameter values. Since the bootstrap methods have to estimate these parameters, an error in the estimations lead to an error in the coverage. The plot of the P value function with respect to various parameters makes it clear (see de Peretti (2003)). However, for a real data study, there is not the same constraint than for Monte Carlo experiments, and we advise to use much more bootstrap replications than used in our experiments, since the user may want more precision for the coverage. B_2 is the number of bootstrap replications used in the estimation of the variance in the *double bootstrap*. It is taken equal to 0 for the moment (instead, the asymptotic estimation of the variance is used).

4.1.1 Coverage plots

The standard deviation of the coverage plots is equal to

$$\sqrt{\frac{c(1-\alpha)(1-c(1-\alpha))}{S}} \tag{4}$$

where $c(1-\alpha)$ is the coverage. Thus, when $1-\alpha$ goes from 0 to 1, and for $S = 10,000$, the standard deviation goes from 0 to 0.005.

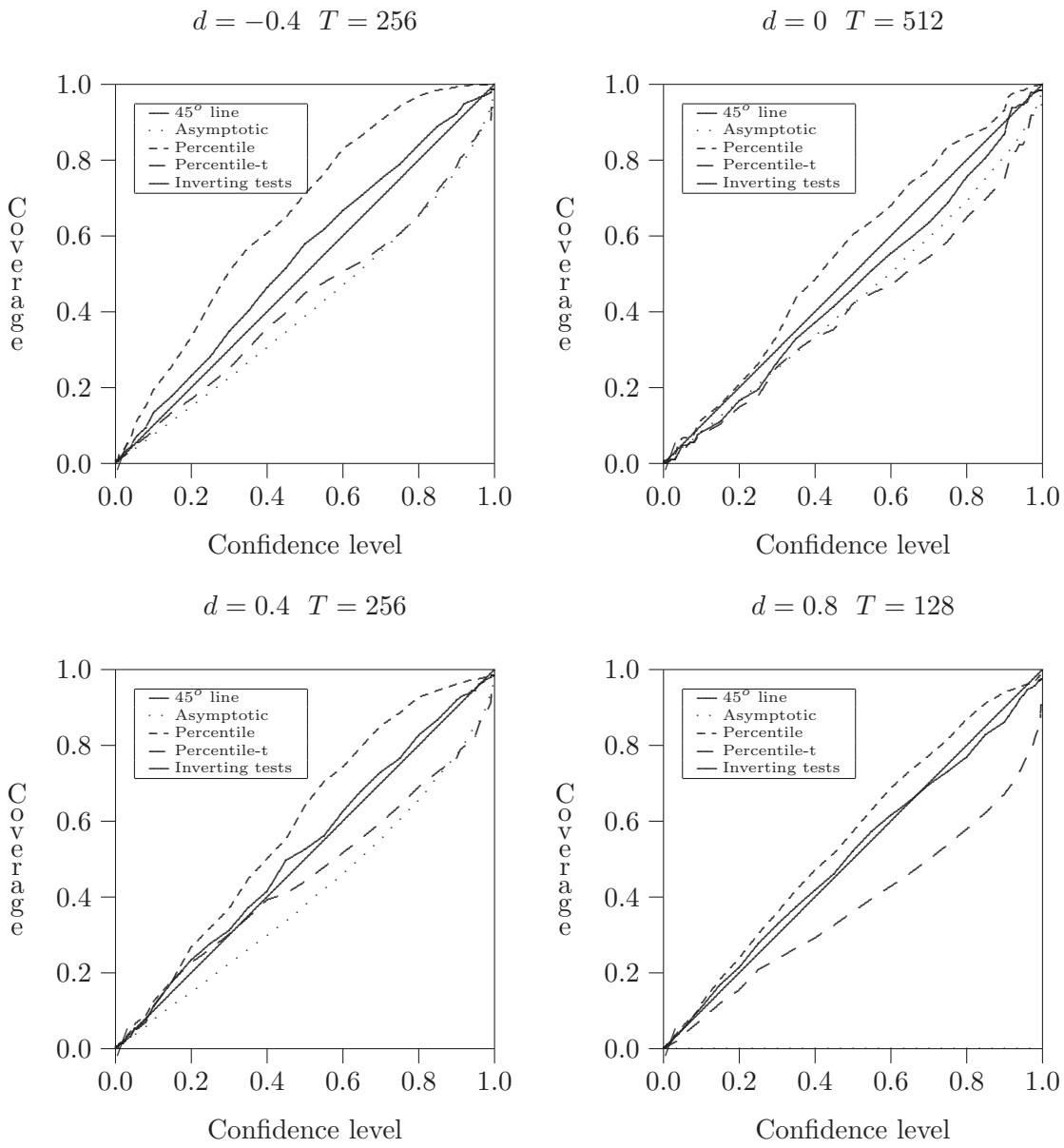
Figure 5 presents the coverage plots of the asymptotic confidence interval, the percentile interval, the (single) percentile-t interval, and the interval based on inverting (single) bootstrap tests for the four cases of the parameters described in table 2. Table 3 presents the same results than the third graph in figure 5 (*i.e.* the case 1.3 of the parameter in table 2) but using a tabular presentation. Is this table pleasant to read ? Figure 5 shows clearly that the confidence interval based on inverting bootstrap tests is

Table 3: Coverages in the case of an ARFIMA(0,d,0) process
$d = 0.4 \quad T = 256$

	Confidence levels			
Method	75%	90%	95%	99%
Asymptotic	0.60	0.77	0.85	0.94
Percentile	0.89	0.96	0.97	0.99
Percentile-t	0.64	0.77	0.82	0.91
Inverting tests	0.77	0.92	0.95	0.98

by far the best method on the basis of coverage accuracy criterion in all the situations.

Figure 5: Coverages plots in the case of an ARFIMA(0,d,0) process



Percentile and percentile-t methods have large coverage distortions, that are as large as for the asymptotic confidence interval, except for the case $d = 0$ (*i.e.* i.i.d.) where the asymptotic method dominates them a little (there is no curve for the asymptotic method in the case of $d = 0.8$ since it is not applicable in this situation). It is not surprising that the percentile method does not perform for estimating the long memory parameter, since the percentile does not correct the bias in the estimates whereas it is known that the long memory estimators are generally very biased. What is more surprising is the unsatisfactory result for the percentile-t method that normally corrects the bias. The unsatisfactory result is probably due to the studentisation of the statistic that does not bring it closer to pivotal (closer to pivotal the statistic is, better the bootstrap methods will perform) since for the single bootstrap, only the asymptotic standard error of the statistic if used for studentising it, and for Robinson's method (1995), this asymptotic standard error is only a constant depending on the sample size (see Robinson (1995)). Double bootstrap should lead to more satisfactory results for the percentile-t methods, but also for confidence interval based on inverting double bootstrap tests, that will still dominate all the methods.

It should be noted that the specification of the data model is not necessarily well chosen by the bootstrap DGP. For example, in this paper, the bootstrap DGP is an ARFIMA(1,d,1) process whereas the true DGPs of the first set of simulations are ARFIMA(0,d,0) processes, and the true DGPs of the second set of simulations are ARFIMA(1,d,0) processes. Nevertheless, our Monte Carlo experiments are more realistic if the true DGP be used, since in practice the real DGP (the one of nature) is not known, and the orders of the ARFIMA(p,d,q) model have to be chosen *a priori* (the ARFIMA(1,d,1) model is often chosen) or estimated by any methods (AIC or BIC criteria or by inference tests).

Since the differences between the confidence regions are very clear with coverage plots, the use of coverage discrepancy plots is not necessary in this situation.

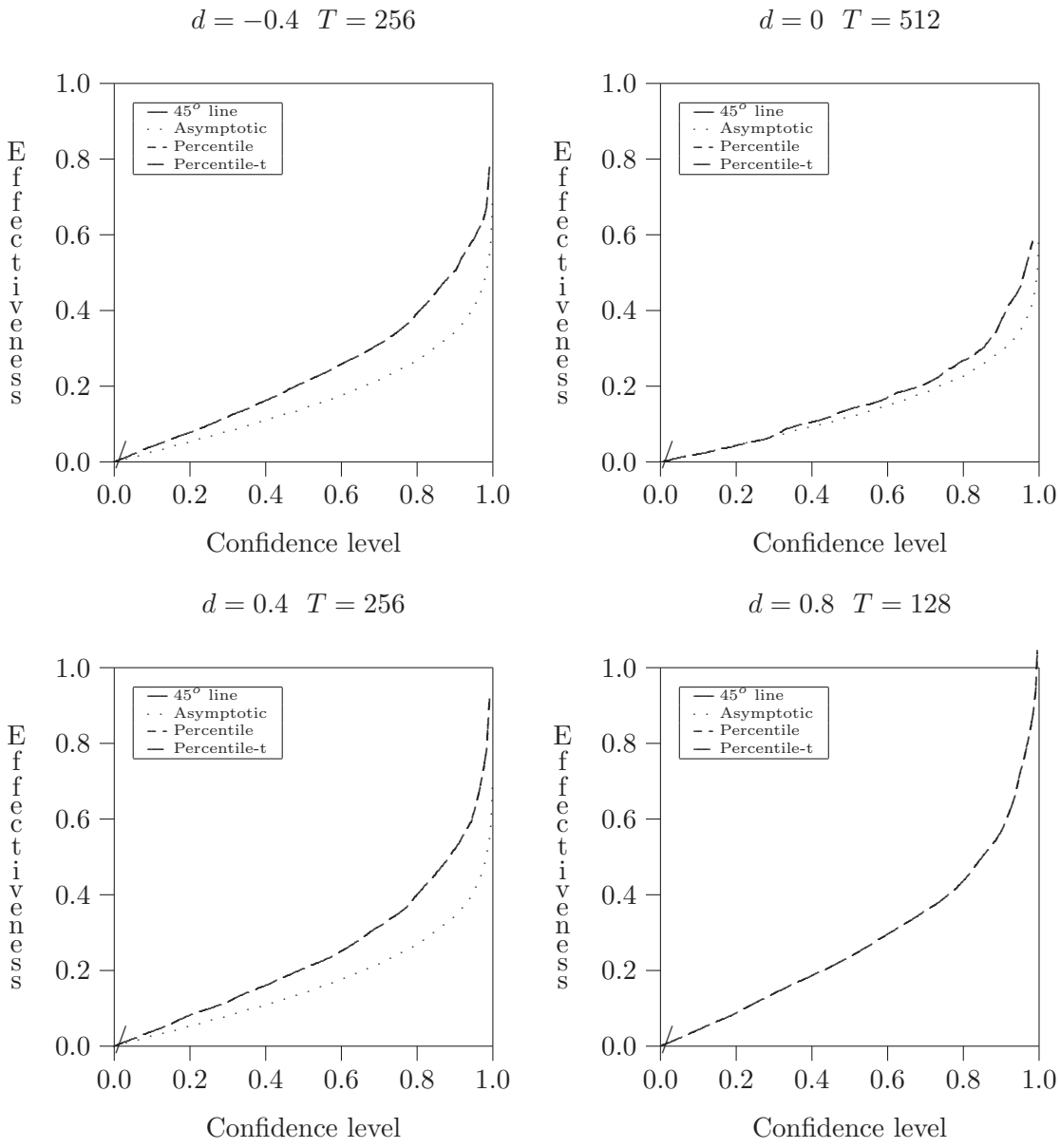
4.1.2 Coverage-effectiveness curves

Figure 6 presents the confidence level-effectiveness curves of the asymptotic confidence interval, the percentile interval, the (single) percentile-t interval for the four cases of the parameters of the table 2. The results for the interval based on inverting (single) bootstrap tests are not presented here because of the difficulty of estimating the length of the interval (see subsection 3.5). Several results will be provided later about this point. On the basis of confidence level-effectiveness curves, the percentile and the percentile-t seem to have the same effectiveness and to be dominated by the asymptotic method that has a lower average length.

However, before concluding hastily, let us have a look to figure 7 that presents the coverage-effectiveness curves for the intervals obtained with the same methods. Following the results on figure 7, the percentile method has the most satisfactory "true" effectiveness. For these three methods, this information is not very interesting since these methods have too large coverage distortions to be used in practice. What the graphics in figure 7 can tell us is only that if the coverage distortions could be corrected, the percentile method will have the most satisfactory effectiveness compared to both other methods. Nevertheless, this kind of graphic is very useful for choosing among methods that have reasonable coverage distortions: it permits to make *arbitrage* between the coverage distortion and the true effectiveness for each methods and then to chose the most appropriate.

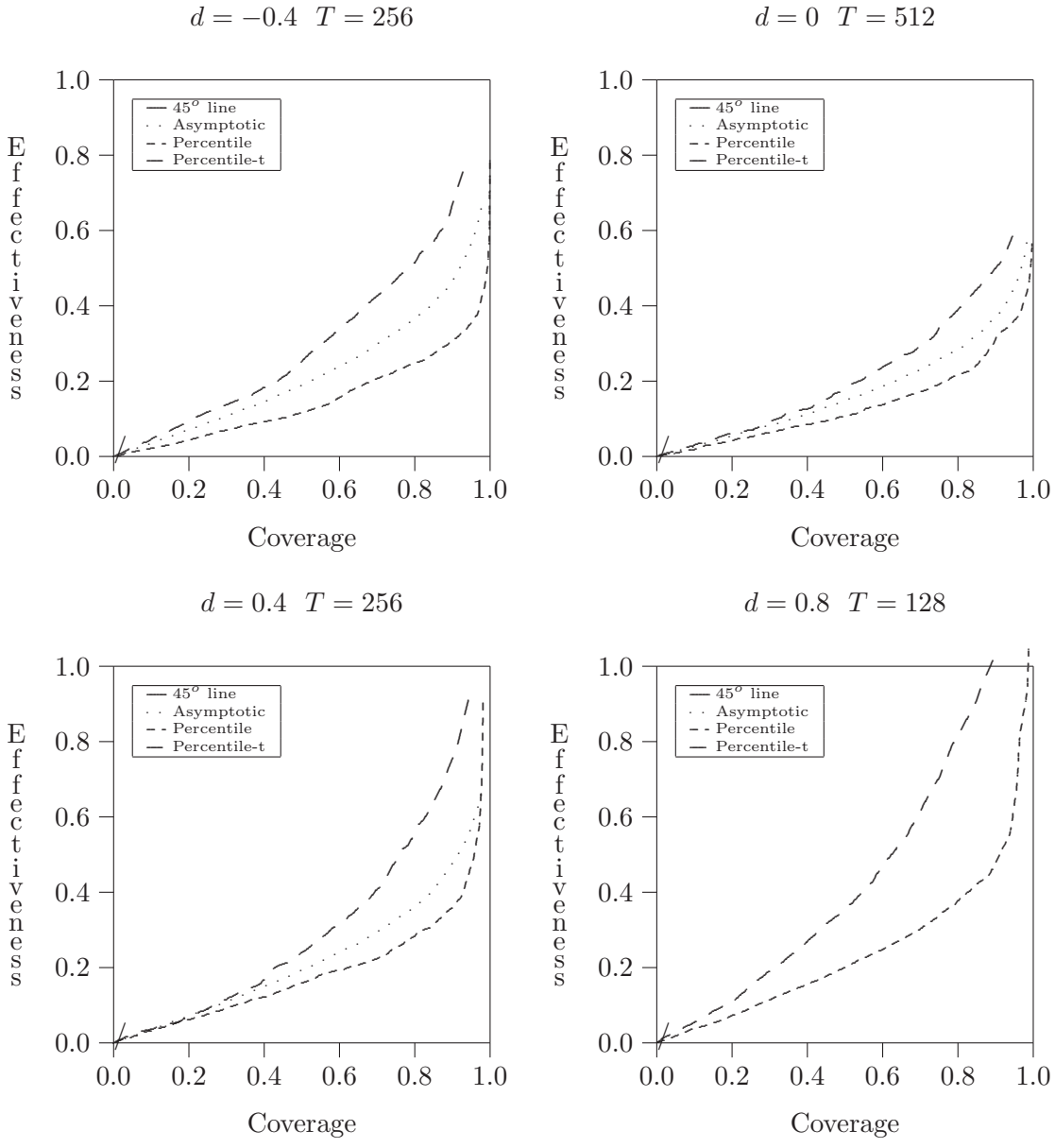
In addition, the standard errors of the lengths of the confidence intervals are plotted

Figure 6: Confidence level-effectiveness curves in the case of an ARFIMA(0,d,0) process



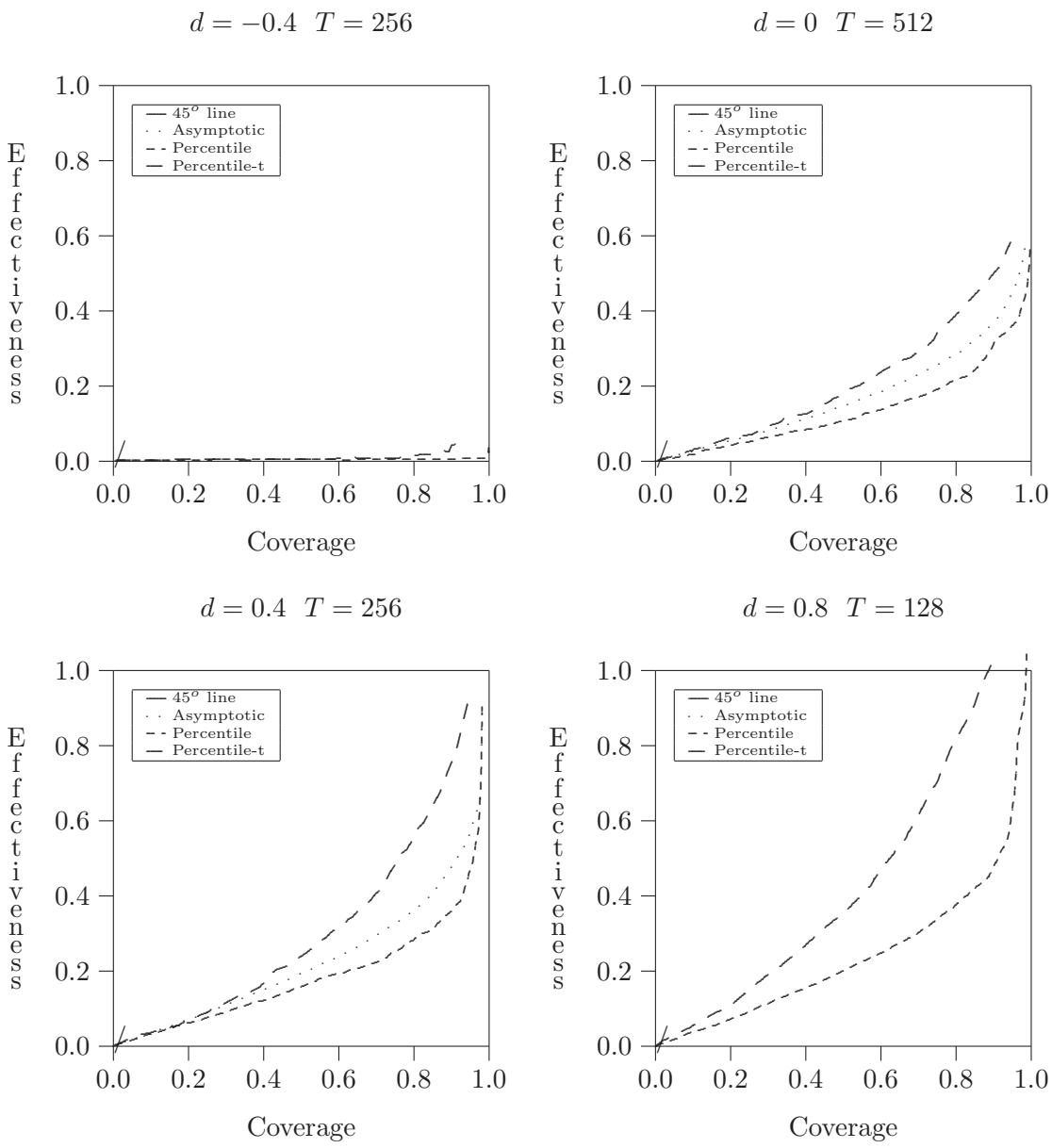
The effectiveness criterion is the average length here.

Figure 7: Coverage-effectiveness curves in the case of an ARFIMA(0,d,0) process



The effectiveness criterion is the average length here.

Figure 8: Coverage-'Standard error of the length' curves in the case of an ARFIMA(0,d,0) process



The effectiveness criterion is the standard error of the length here.

Figure 9: Coverages plots in the case of an ARFIMA(1,d,0) process

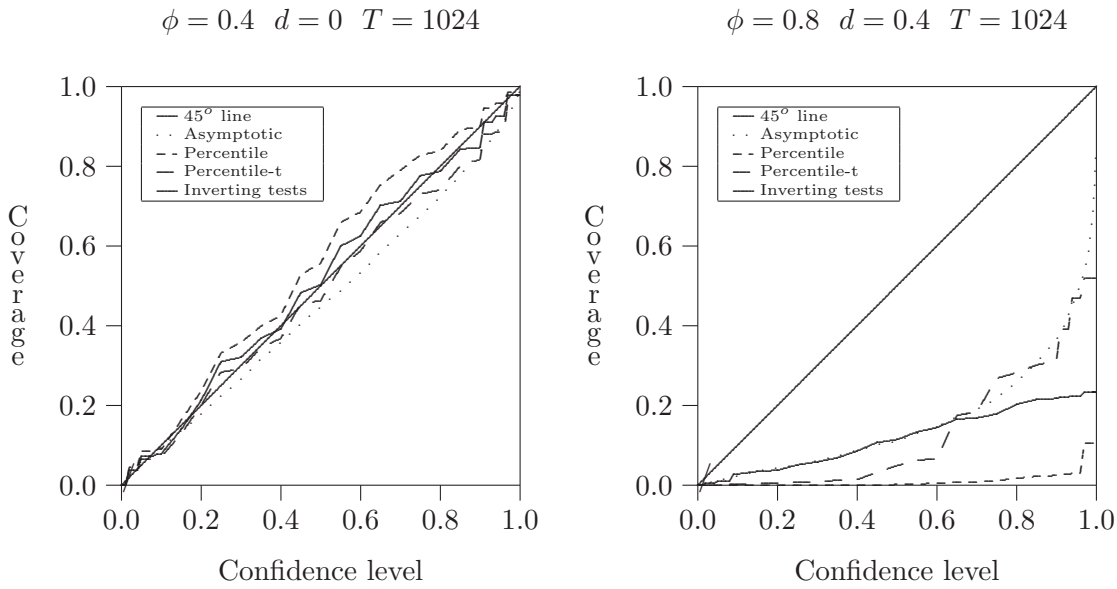
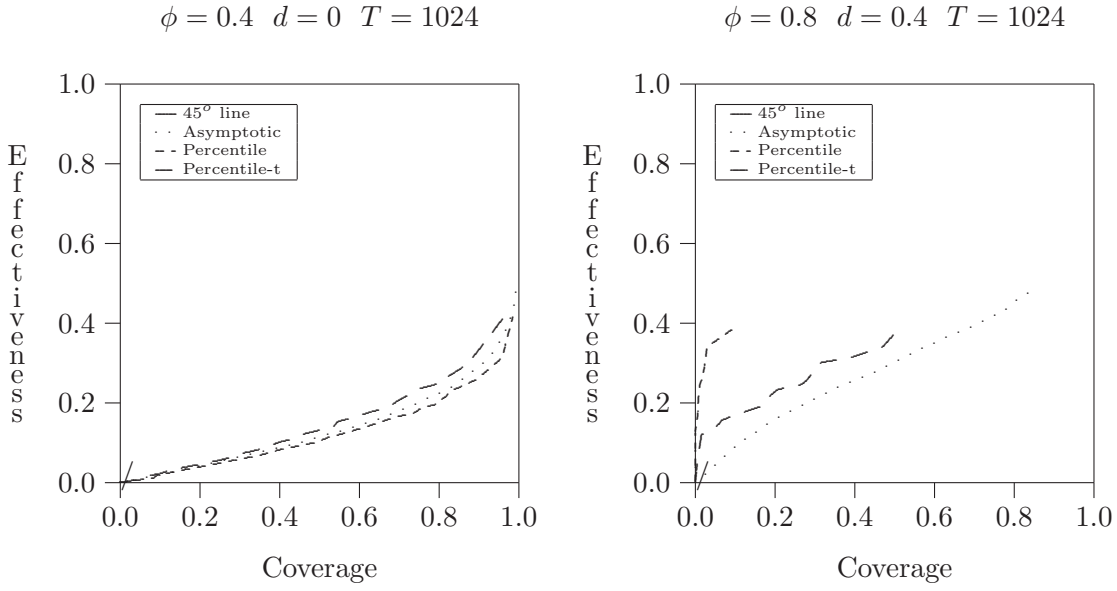


Figure 10: Coverage-effectiveness curves in the case of an ARFIMA(0,d,0) process



The effectiveness criterion is the average length here.

against their coverages. This leads to what can be called *Coverage-'Standard error of the length' curves* that can be viewed as coverage-effectiveness curves for a secondary effectiveness criterion (see figure 8). The same conclusion than previously holds. However, the graphics on figure 8 allow to see that the larger d is, the larger the standard error of the length. This is unsatisfactory since for large values of d, even if the average length of an interval is satisfactory, the probability to have a large length is great, and thus there is a big uncertainty on the value for the true value for d. This disadvantage cannot be viewed only on the basis of average length, but also on the basis of length standard error.

For the length of confidence interval based on inverting tests, see the analysis in subsection 5.3.

4.2 Case of ARFIMA(1,d,0) processes

We choose combinations of d and T in table 4 for the investigation. Figure 9 presents the

Table 4: choice of ϕ, d, and T

Case	ϕ	d	T	B
2.1	0.4	0	1024	122
2.2	0.8	0.4	1024	122

coverage plots of the asymptotic confidence interval, the percentile interval, the (single) percentile-t interval, and the interval based on inverting (single) bootstrap tests for the two cases of the parameters described in table 4). Figure 10 presents the coverage-effectiveness curves for the three first methods. For $\phi = 0.4 \quad d = 0 \quad T = 1024$, the same conclusion than for the previous subsection holds. Conversely, the case $\phi = 0.8 \quad d = 0.4 \quad T = 1024$ shows catastrophic results which remind us that when the short memory parameters are too strong (close to unit root), the short memory property is confused with long memory property.

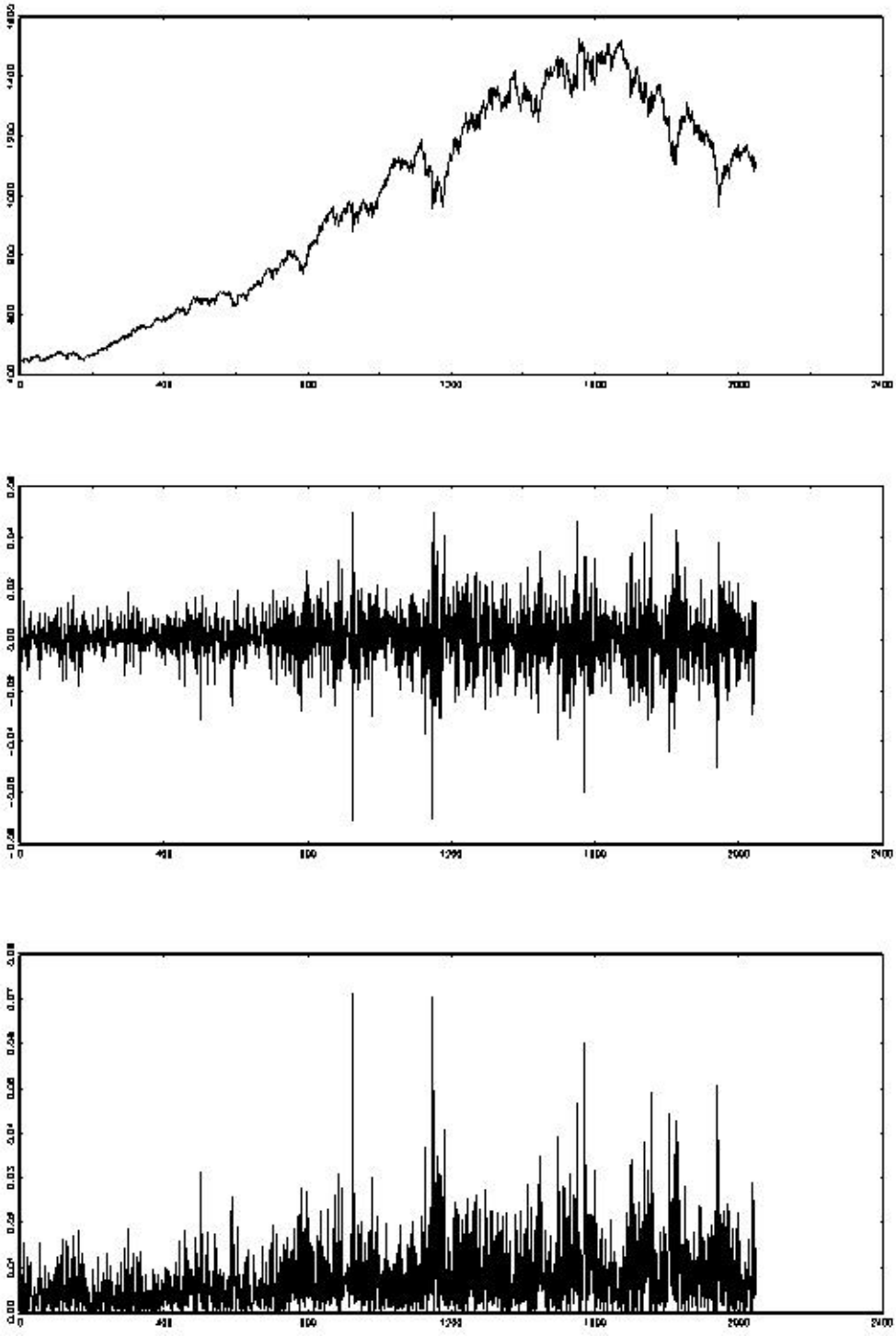
5 Application to the S&P500 index

We propose to investigate the long memory properties of the daily S&P500 index from 01/01/1994 to 13/02/2002. The number of observation is 2048. Data comes from Terrance Mills' book. Figure 11 plots the S&P500 index in level. Table 5 presents some descriptive statistics.

Table 5: Descriptive statistics for the S&P500 index.

Time series	Mean	Standard error	Skewness	Kurtosis
Returns	0.0004	0.0107	-0.3055	7.2131
Absolute returns	0.0075	0.0075	2.2990	12.3470
Squared returns	0.0001	0.0003	8.6066	116.5850

Figure 11: S&P500 index, returns, and absolute returns



5.1 The S&P500 returns

Since the S&P500 series (denoted x_t) is clearly nonstationary, the returns of the series

$$r_t = \ln\left(\frac{x_t}{x_{t-1}}\right) \tag{5}$$

will be used in the following (see figure 11).

For studying the stationarity of the S&P500 returns, parametric and nonparametric bootstrapped versions of the augmented of Dickey-Fuller (ADF) tests are used, with unilateral and bilateral bootstrap P values (see Appendix C). The number of bootstrap replications being 999. For selecting the number of augmentations, the residuals from the ADF regressions are tested for serial correlation using Ljung-Box and Box-Pierce tests (from 1 to 8 lags) until they look like white noise. For the S&P500 returns, the number of augmentations is 0. All tests have been conducted by using our own programs via Gauss software. All the P values are close to zero (they are not presented) concluding that the series is clearly stationary.

For studying the normality hypothesis, three tests for normality are used. The first procedure tests for the nullity of the skewness. The second one tests whether the kurtosis is equal to 3. Finally, the last one tests for both simultaneously: it is the Jarque-Bera test. Since the critical values of the Jarque-Bera test are not exact in finite sample due to the correlation between the estimated skewness and the estimated remind, and since the finite sample distribution for all the tests are not equal to their asymptotic distributions, we propose to use bootstrap versions of these tests (see appendix D). Of course, it is a parametric bootstrap since the null hypothesis is the normality of the series, but in addition, we allow the Gaussian process under the null to be autocorrelated because if it is not taken into account, the autocorrelation amplifies the distortion of the test statistic distribution compared to the asymptotic distribution. Unilateral and bilateral bootstrap P values are also used (except for the asymptotic tests, since the asymptotic distributions are symmetric: both the P values are the same, and for Bera-Jarque's test because it is unilateral by construction). The number of bootstrap replications is 999. There is a lag of the dependent variable in the regressor set. The number of lags is chosen by Ljung-Box and Box-Pierce tests using 1 to 8 lags until the residuals seem independent. Again, all the P values are close to zero, showing that the series is strongly non-Gaussian. Consequently, nonparametric methods have to be used in the following.

The presence of long memory in the series is estimated using a confidence interval based on Robinson's test. The classical asymptotic confidence interval is used. In addition, nonparametric percentile and percentile-t are also used. The nonparametric bootstraps are based on the empirical distribution of the residuals (they resample the residuals). There are also single and double versions of bootstrap. Table 6 presents the confidence intervals for the long memory parameter in the S&P500 returns. The methods are the asymptotic one, (single) percentile, (single) percentile-t, and inverting (single bootstrap) tests. The number of bootstrap replication is $B = 999$. Since zero is in all the interval, the results suggests that there is no long memory in the returns of the index.

5.2 The S&P500 volatility

We propose to explore the long memory property of the S&P500 volatility. The following measure of the volatility is used:

$$v_{1,t} = |r_t|. \tag{6}$$

Table 6: Long memory confidence interval on the S&P500 returns

Confidence interval	Lower limit	Upper limit	Interval length
Asymptotic	-0.0893	0.1556	0.2449
Percentile	-0.0926	0.0877	0.1803
Percentile-t	-0.0875	0.0928	0.1803
Inverting tests	-0.0968	0.104	0.2008

Two other measures are often used in practice:

$$\begin{aligned} v_{2,t} &= r_t^2, \\ v_{3,t} &= \ln(|r_t|). \end{aligned}$$

$v_{3,t}$ cannot be used here because of zero values for some observations of r_t. The absolute returns are presented in figure 11 .

For these real data, our confidence intervals have to be modified. For the return series, the bootstrap *Data Generating Process* (DGP) is an ARFIMA(p,d,q) process. But for the volatility series, a heteroskedastic model have to be used. Consequently, the bootstrap DGP has to be changed from an ARFIMA(p,d,q) process to a FIGARCH(p,d,q) process. More precisely, for the volatility measured by the absolute value of the returns, the bootstrap DGP is the absolute value of a FIGARCH(p,d,q) process. The same reasoning holds for the volatility measured by the logarithm absolute returns or the squared returns. The results are presented in table 7. The long memory property is clearly found in the

Table 7: Long memory confidence interval on the absolute returns

Confidence interval	Lower limit	Upper limit	Interval length
Asymptotic	0.3320	0.5770	0.2450
Percentile	0.2261	0.5404	0.3143
Percentile-t	0.3686	0.6830	0.3144
Inverting tests	0.3528	0.7791	0.4263

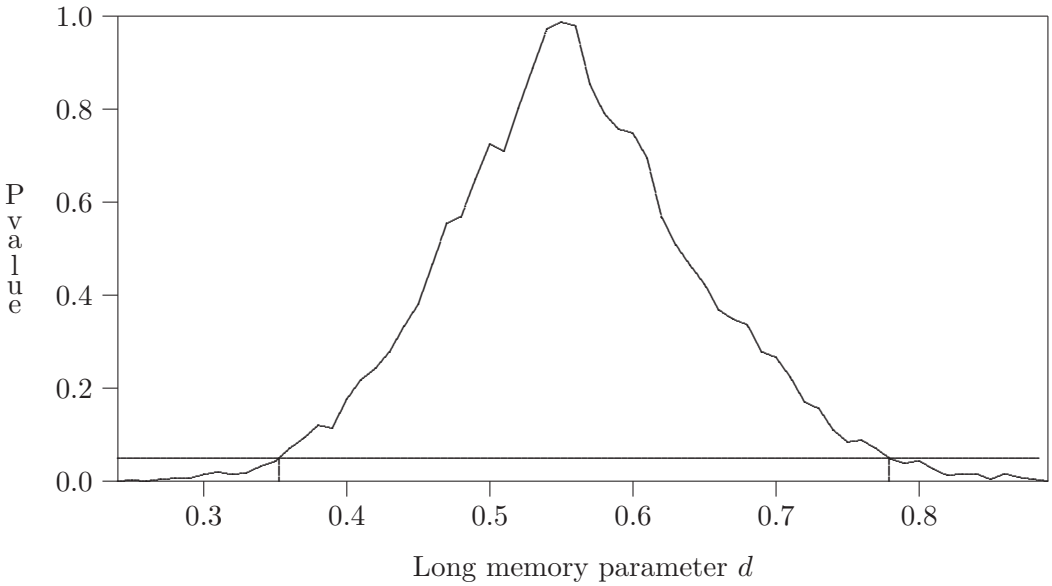
The number of bootstrap replication is $B = 999$.

volatility. The possible values for the long memory parameter are between the stationary and the nonstationary cases.

5.3 Length of confidence interval based on inverting tests: case of S&P500 volatility

The limits of the confidence interval based on inverting (single bootstrap) tests presented in table 7 are determined by the values for d such that the test P value is equal to 5%. The P value function with respect to d is presented in figure 12. Table 7 shows

Figure 12: P value function corresponding to confidence interval based on inverting tests
Case of volatility series



that the confidence interval based on inverting tests is the largest. However, the Monte Carlo experiments suggest that inverting tests leads to much more accurate interval on the basis of the coverage. They also suggest that the asymptotic and the percentile-t intervals undercover the LRD parameter, that can explain by their smaller length.

6 Conclusion

Monte Carlo experiments are a valuable tool for obtaining information about the properties of confidence region procedures in finite samples. However, the rich detail in the results they provide can be difficult to apprehend if they are presented in the usual tabular form. In this paper, we have discussed several graphical techniques that can make the principal results of an experiment immediately obvious, namely, coverage plots, coverage discrepancy plots (which may optionally be smoothed), and coverage effectiveness curves. All of these techniques are based on the construction of an estimated cumulative distribution function of the (true) coverage associated with some confidence regions and on effectiveness criteria that were discussed, without loss of computing time compared to classical tabular presentation.

These techniques were illustrated by presenting the results of a number of experiments concerning long memory confidence regions. We think that results, which are entirely presented in graphical form, are of interest and provide more information in a more easily assimilable fashion than a tabular presentation or *QQ plots* could possibly have done. The results show that percentile and percentile-t methods does not perform well at all, conversely to confidence interval based on inverting bootstrap tests that works quasi-perfectly compared to the confidence intervals obtained with the previous methods: the coverage is much more close to the confidence level.

References

Beran, R. 1988. Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, **83(403)**, 687–697. 3

Berkowitz, & Kilian. 2000. Recent developments on bootstrapping time series. *Econometrics Reviews*, **19**, 49–54. 3.4

Black, W. C. 1990. The cost-effectiveness plane: a graphic representation of cost-effectiveness. *Medical decision Making*, **10(3)**, 212–215. 5, 4

Chesher, A., & Spady, R. 1991. Asymptotic expansions of the information matrix test statistic. *Econometrica*, **59(3)**, 787–815. 2.8

Davidson, R. 1998. Notes on the bootstrap. *GREQAM*. 3, B.1.1

Davidson, R. 2000. Comments on 'Recent developments in bootstrapping time series' by Berkowitz and Kilian. *Econometrics Reviews*, **19**, 49–54. 3.4

Davidson, R., & MacKinnon, J. 1998. Graphical methods for investigating the size and the power of hypothesis tests. *The Manchester School*, **66**, 1–22. 2.4, 2.8

Davidson, R., & MacKinnon, J. G. 1993. *Estimation and inference in economics*. Oxford University Press. New York. 1, 3.3, 10, 3.5, B.1, B.1.2, B.1.2, 7c, 14

Davidson, R., & MacKinnon, J. G. 1996a. The Size Distortion of Bootstrap Tests. *GREQAM Working Paper No 96A15 and Queen's Institute for Economic Research Discussion Paper No 937*. 3, B.1

Davidson, R., & MacKinnon, J. G. 1996b. the power of bootstrap tests. *Queen's University Institute for Economic Research, Discussion Paper 937*. 3, B.1

Davidson, R., & MacKinnon, J. G. 2001. Improving the reliability of bootstrap confidence intervals. *Université de Montréal*, **Conference Resampling Methods in Econometrics**. 2.6, 3.5

de Peretti, C. 2003. Bilateral Bootstrap Tests for Long Memory: An Application to the Silver Market. *Computational Economics*, Forthcoming. 12, 13, B.1

de Peretti, C., & Marimoutou, V. 2002. Are the long memory tests really effective? *GREQAM Working Paper*, **No 02A14**. 2, 9

Efron, B. 1979. Bootstrap methods: another look at the Jacknife. *Annals of Statistics*, **7**, 1–26. B.1

Efron, B., & Tibshirani, R.J. 1993. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57, Chapman and Hall. London. 3.3

Fieller, E. C. 1954. Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B*, **16**, 175–183. 7

Granger, C.W.J., & Joyeux, R. 1980. An introduction to long-memory time series models and fractional integration. *Journal of Time Series Analysis*, **1(1)**, 15–29. 9

Hall, P. 1992. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag. New York. 3.3

Hauser, M. A., Pötscher, B. M., & E., Reschenhofer. 1999. Measuring persistence in aggregate output: ARMA models, fractionally integrated ARMA models and non-parametric procedures. *Empirical Economics*, **24**, 243–269. 2

Higuchi, T. 1988. Approach to an irregular time series on the basis of the fractal theory. *Physica*, **D 31**, 277–283. 3

Hjorth, J.S.U. 1994. *Computer Intensive Statistical Methods*. Chapman and Hall. London. 3.3

Horowitz, J. L. 1994. Bootstrap-based critical values for the information matrix test. *Journal of Econometrics*, **61(2)**, 395–411. 3

Hosking, J.R.M. 1981. Fractional differencing. *Biometrika*, **68**, 165–176. 9

Li, H., & Maddala, G. S. 1996. Bootstrapping time series models. *Econometric Reviews*, **15**, 297–318. 3.4

Robinson, P.M. 1995. Gaussian semiparametric estimation of long range dependence. *The Annals of Statistics*, **23**, 1048–1072. 1, 3, 3, 3, 3.1, 9, 3.2, 4.1.1

Shao, J., & Tu, D. 1995. *The Jackknife and Bootstrap*. Springer-Verlag. New York. 3.3

Siani, C., & de Peretti, C. 2004. Is Fieller's method applicable in all the situations ? *Health Economics*, **forthcoming**. 7

Siani, C., & Moatti, J. P. 2003. The handling of uncertainty in economic evaluations of health care strategies. *Revue d'Epidémiologie et de Santé Publique (Frensh)*, **51**, 255–276. 2.6, 7, 3.4

Stinnet, A. A., & Mullahy, J. 1998. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making*, **18**, S68–S80. 2.8

Tambour, M., Zethraeus, N., & Johannesson, M. 1998. A note on confidence intervals in cost-effectiveness analysis. *International Journal of Technology Assessment in Health Care*, **14(3)**, 467–471. 2.8

Velasco, Carlos. 1999. Non-stationary log-periodogram regression. *Journal of Econometrics*, **91**, 325–371. 3.2, 3.3

Weber, N.C. 1984. On resampling techniques for regression models. *Statistics and Probability Letters*, **2**, 275–278. 4

Wild, M. B., & Gnanadesikan, R. 1968. Probability plotting methods for the analysis of data. *Biometrika*, **33(1)**, 1–17. 2.8

Appendix

A Details about the graphical representation

When S is large, storage space can be conserved by evaluating the edf (1) only at N points $x_i, i = 1, \ldots, N$, which should be chosen in advance so as to provide a reasonable snapshot of the $(0, 1)$ interval, or of that part of it which is of interest. As parsimonious way to choose the x_i is

$$\begin{aligned} x_i &= 0.001, 0.002, \ldots, 0.009, 0.01, 0.02, \ldots, 0.09, 0.1, 0.15, \ldots \\ &\quad \ldots, 0.85, 0.9, 0.91 \ldots, 0.98, 0.99, 0.991, \ldots, 0.998, 0.999 \qquad (N = 53) \end{aligned} \tag{7}$$

There are extra points near 0 and 1 in order to ensure that we do not miss any unusual behaviour in the tails.

B Details on the LRD confidence intervals

B.1 Bilateral (single) bootstrap tests for long memory

For more details in the case of the null of no long memory, see de Peretti (2003). While the asymptotic tests are asymptotically valid, the tests based on the asymptotic distributions are not exact in finite samples, and so, it is natural to "bootstrap" them. For the conception of the bootstrap see Efron (1979), for its development, see Davidson and MacKinnon (1993), and for further analysis, see Davidson and Mackinnon (1996b; 1996a).

B.1.1 The bootstrap procedure

The procedure is as follows:

1. Compute the test statistic (Hurst, Lo, Robinson, Higuchi, Jensen, or others), which will be denoted $\hat{\tau}$.

2. Estimate the ARFIMA(p,d,q) model by maximum likelihood under the null H_0: $d = d_0$, for obtaining the model parameters $(\hat{\phi}, \hat{\theta}, \hat{\sigma}_\varepsilon^2)$ and the residuals $\hat{\varepsilon}$. For our simulations, $p = 1$ and $q = 1$ are chosen, but for estimating real data, we strongly advice users to determine p and q by information criterions or other efficient methods because an error in the choice of p and/or q generally yields to large bias in the estimation and the inference of d (see Hauser *et al.* (1999).

3. Draw B sets of bootstrap error terms, ε^b, and use them to generate B bootstrap samples x^b. There are numerous ways to drawn the error terms, four of which are described below. The elements of x^b should be generated from the equation

$$x_t = (1 - L)^{-d_0} \hat{\phi}(L)^{-1} \hat{\theta}(L) \varepsilon_t^b \qquad t \in \{1, \ldots, T\}. \tag{8}$$

   In practice, we do as the GAUSS procedure for generating ARFIMA series: we calculate the MA(q) filtre corresponding to $(1 - L)^{-d_0} \hat{\phi}(L)^{-1} \hat{\theta}(L)$ with a very large ordre q (T for example), and we apply it to ε^b for obtaining x_t. For that, of course, we have to generate ε^b with a larger sample size ($2 \times T$ for example).

4. For each bootstrap sample, compute the statistic (Hurst, Lo, Robinson, Higuchi, Jensen, or others), denoted τ^b, with x^b instead of x.

5. Then, compute the estimated bootstrap P value: see equation 10 or equations 11–12.

I examine four ways for generating the ε_t^b (see Davidson (1998)):

1. The parametric bootstrap, called b_0: the ε_t^b are independent draws from the $N(0, \hat{\sigma}_\varepsilon^2)$ distribution.

2. The simplest nonparametric bootstrap, called b_1: the ε_t^b are obtained by re-sampling with replacement from the vector of $\{\hat{\varepsilon}_t\}_{t=\hat{p}+1}^T$.

3. A slightly more complicated from of nonparametric bootstrap called b_2: the ε^b are generated by re-sampling with replacement from the vector

$$\left\{ \sqrt{\frac{T}{T-2\hat{p}-1}} \left(\hat{\varepsilon}_t - \frac{1}{T-\hat{p}} \sum_{i=\hat{p}+1}^{T} \hat{\varepsilon}_i \right) \right\}_{t=\hat{p}+1}^{T}. \tag{9}$$

4. The most complicated nonparametric bootstrap, called b_3: the ε^b are generated by re-sampling from the vector with typical element $\tilde{\varepsilon}_t$ constructed as follows:

   - let d_t be the t^{th} diagonal element of $P_{[(1-L)^{-d_0}\hat{\phi}(L)^{-1}\hat{\theta}(L)]}$, the matrix projecting onto the space spanned by $(1-L)^{-d_0}\hat{\phi}(L)^{-1}\hat{\theta}(L)$;
   - divide each element of $\hat{\varepsilon}$ by $\sqrt{1-d_t}$;
   - re-centre the resulting vector;
   - re-scale it so that it has variance $\hat{\sigma}_\varepsilon^2$.

   This type of procedure is advocated in Weber (1984).

B.1.2 The choice of the bootstrap P value

By making a large number of drawings of bootstrap statistics τ^b, a bootstrap P value can be computed by the following formula:

$$\hat{p}_{bil}(\hat{\tau}^2) = \frac{1}{B} \sum_{b=1}^{B} I((\tau^b)^2 > \hat{\tau}^2), \tag{10}$$

see Davidson and MacKinnon (1993). This formula corresponds to an unilateral test. This sort of formulae is often associated with symmetric bilateral tests. However, the size distortion is not necessarily symmetric. Thus, I prefer to use the following formula:

$$\hat{p}(\hat{\tau}) = 2\min\{\hat{p}_{uni}(\hat{\tau}), 1 - \hat{p}_{uni}(\hat{\tau})\}, \tag{11}$$

where

$$\hat{p}_{uni}(\hat{\tau}) = \frac{1}{B} \sum_{b=1}^{B} I(\tau^b > \hat{\tau}), \tag{12}$$

that corresponds to a bilateral (asymmetric) test. Further considerations about this P value can be found in Chapter 5 of Davidson and MacKinnon (1993) dealing with confidence intervals.

B.2 Bilateral double bootstrap tests for long memory

The standard error of $\hat{d}$, say $\hat{\sigma}(\hat{d})$, used for studentising $\hat{d}$ can be calculated in two different ways: the first one is the asymptotic estimate (see subsection 3.2), the second one is the bootstrap estimate. For obtaining the bootstrap estimator of $\sigma(\hat{d})$, the time series are estimated and replicated B_2 times in the same way as previously: for each replicated series, $\hat{d}$ is computed leading to a set of $\left\{\hat{d}^b\right\}_b$ from which the standard error is computed. It should be noted that when this test is used, and thus replicated series are generated, the bootstrap estimator of $\sigma(\hat{d})$ must be applied on each replicated series for obtaining the test statistics leading to replications of replicated series. This method is often called *double bootstrap*.

C Bootstrapped ADF tests

The following variables have to be defined:

- B, the number of bootstrap replications,
- p, the number of augmentations in the ADF regressions.

B has to be chosen as large as possible, depending on the characteristics of the computer. The choice of p is more difficult. We recall that the ADF regressions are:

$$\begin{aligned}
\Delta y_t &= && && \alpha y_{t-1} + \beta_1 \Delta y_{t-1} + \ldots + \beta_p \Delta y_{t-p} + e_t, \\
\Delta y_t &= && constant + && \alpha y_{t-1} + \beta_1 \Delta y_{t-1} + \ldots + \beta_p \Delta y_{t-p} + e_t, \\
\Delta y_t &= && constant + && trend + \alpha y_{t-1} + \beta_1 \Delta y_{t-1} + \ldots + \beta_p \Delta y_{t-p} + e_t,
\end{aligned}$$

where y_t is the time series, and e_t are the error terms, t goed from 1 to T. We propose the following procedure for chosing p: The procedure starts at $p = 0$. The residuals of each ADF regressions with p augmentations are tested for independence using both the Ljung-Box's and Box-Pierce's tests. The number of autocorrelation coefficients taken into account for the Ljung-Box's and Box-Pierce's tests go from 1 to 8. If the residuals are not independent, p is incremented by 1 until the residuals look independent.

The steps of the bootstrapped ADF test are the following:

1. The Student test statistics for α for each ADF regression are computed. Let the statistics be denoted t_α. At this step, the residuals can be kept to be tested for independence.

2. The boostrap procedure needs a DGP for generating simulated samples under the null. This DGP is determined by estimating the model under the null using the data and the OLS procedure.

3. The bootstrap loop starts now. The simulated error terms, denoted e_t^b, are generated for a sample. There are four ways for generating the simulated error terms:

   (a) Parametric bootstrap: The simulated error terms are drawn from the normal distribution
   $$e_t^b \sim N(0, s^2),$$
   where s is the standard error of the error terms estimated from the ADF regression using the data.

(b) Basic nonparametric boostrap: The simulated error terms are drawn by ...

(c) Nonparametric bootstrap with corrected degree of freedom: since $\mathrm{E}(\hat{e}_t^2) \neq \mathrm{E}(e_t^2)$ where $\hat{e}_t^2$) are the residuals of the ADF regression, but $\mathrm{E}(\hat{e}_t^2) =$

For our program, the parametric and the second nonparametric bootstrap are chosen.

4. The simulated time series under the null, denoted $(y_t^b)_t$, is generated recursively using both the following steps:

   (a) first, define Δy_t^b recursively:

$$\begin{aligned}
\Delta y_t^b &= & & \hat{\beta}_1 \Delta y_{t-1}^b + \ldots + \hat{\beta}_p \Delta y_{t-p}^b + e_t^b, \\
\Delta y_t^b &= \widehat{constant} + & & \hat{\beta}_1 \Delta y_{t-1}^b + \ldots + \hat{\beta}_p \Delta y_{t-p}^b + e_t^b, \\
\Delta y_t^b &= \widehat{constant} + & \widehat{trend} + & \hat{\beta}_1 \Delta y_{t-1}^b + \ldots + \hat{\beta}_p \Delta y_{t-p}^b + e_t^b,
\end{aligned}$$

   The p first values for y_t^b can be chosen equal to the the p first values of y_t, interpreted as initial conditions. (Another way is to choose them randomly.)

   (b) second, compute y_t^b:

$$y_t^b = y_1 + \sum_{i=2}^{t} y_i^b.$$

   y_1 is an initial condition.

5. The Student test statistics for α for each ADF regression are computed using the simulated series $(y_t^b)_t$. Let the statistics be denoted t_α^b.

6. The steps 3–5 are done again B times. A set of statistics t_α^b, $b = 1, \ldots, B$, is then obtained for each the three ADF regressions, and for each both the parametric and nonparametric bootstraps (thus there are six statistics).

7. The bootstrap P value is finally computed depending on the test hypothesis:

   (a) If the null hypothesis $H_0 : \alpha = 0$ is tested against the alternative hypothesis $H_1 : \alpha < 0$, the classical P value is

$$p_{\mathrm{uni}} = \frac{1}{B} \sum_{b=1}^{B} I(t_\alpha^b \leq t_\alpha),$$

   where I is the indicator function. This P value corresponds to an unilateral test.

   (b) If the null hypothesis $H_0 : \alpha = 0$ is tested against the alternative hypothesis $H_1 : \alpha \neq 0$, the classical bootstrap P value is

$$p_{\mathrm{bil_{sym}}} = \frac{1}{B} \sum_{b=1}^{B} I(|t_\alpha^b| \geq |t_\alpha|).$$

   This P value corresponds to a bilateral test.

(c) In the case where the null hypothesis $H_0 : \alpha = 0$ is tested against the alternative hypothesis $H_1 : \alpha \neq 0$, we also propose the following bootstrap P value:

$$p_{\text{bil}_{\text{asym}}} = 2\min\{p_{\text{uni}}, 1 - p_{\text{uni}}\}.$$

This P value also corresponds to a bilateral test, but it takes into account the asymmetry of the statistic distribution in addition. This P value can be found in Davidson and MacKinnon (1993), chapter 5, in the context of confidence regions.

In our program, the two last P values are used.

8. Finally, a significance level is chosen and compared to the P values. If a P value is lower to the significance level, H_1 is retained, otherwise H_0 is retained.

D Bootstrapped tests for normallity

It is first assumed that the processes that will be tested are stationary and independent.

D.1 The test statistics

First, the estimated centered moments are presented:

$$\begin{aligned}
\hat{\mu}_1(y) &= \frac{1}{T}\sum_{t=1}^{T} y_t, \\
\hat{\mu}_2^c(y) &= \frac{1}{T}\sum_{t=1}^{T} (y_t - \mu_1(y))^2, \\
\hat{\mu}_3^c(y) &= \frac{1}{T}\sum_{t=1}^{T} (y_t - \mu_1(y))^3, \\
\hat{\mu}_4^c(y) &= \frac{1}{T}\sum_{t=1}^{T} (y_t - \mu_1(y))^4,
\end{aligned}$$

where T is the sample size, and y is the time series. The estimated skewness and estimated the kurtosis are defined as following:

$$\begin{aligned}
\widehat{\text{sk}}(y) &= \frac{\hat{\mu}_3^c(y)}{(\hat{\mu}_2^c(y))^{3/2}}, \\
\widehat{\text{ku}}(y) &= \frac{\hat{\mu}_4^c(y)}{(\hat{\mu}_2^c(y))^{2}}.
\end{aligned}$$

From the skewness and the kurtosis, three statistics can be built:

- a statistic based on the skewness:

$$t_{\text{sk}}(y) = \frac{\widehat{\text{ku}}(y)}{\sqrt{(6/T)}},$$

- a statistic based on the kurtosis:

$$t_{\text{ku}}(y) = \frac{\widehat{\text{ku}}(y) - 3}{\sqrt{(}24/T)},$$

- and a statistic based on both the the skewness and the kurtosis: the Jarque-Bera statistic (see Jarque and Bera (**?**)):

$$t_{\text{jb}}(y) = \left(t_{\text{sk}}(y)\right)^2 + \left(t_{\text{ku}}(y)\right)^2.$$

D.2 The asymptotic tests

Under the null hypothesis H_0:"y is a Gaussian process", the asymptotic theory establishes that:

$$t_{\text{sk}}(y) \sim N(0,1), \tag{13}$$
$$t_{\text{ku}}(y) \sim N(0,1), \tag{14}$$
$$t_{\text{jb}}(y) \sim \chi^2(2). \tag{15}$$

However, in finite sample, the two first statistics do not follow the normal distribution, even if the time series is Gaussian. Consequently, the third statistic do not follow the $chi^2(2)$ distribution. This is a first reason for bootstrapping them. In addition, Jarque-Bera statistic follows asymptotically a $chi^2(2)$ because of asymptotic independence of $t_{\text{sk}}(y)$ and $t_{\text{ku}}(y)$. However, in finite sample, these two components are dependent, leading to a second error in Jarque-Bera test. This is a second reason for bootstrapping Jarque-Bera statistic, since the bootstrap techniques are naturally able to take into account this dependence.

D.3 The bootstrap tests

Let B denotes the number of bootstrap replications. B has to be chosen as large as possible, depending on the characteristics of the computer. y_t is the time series, t goed from 1 to T. The steps of the bootstrapped test for normality are the following:

1. The test statistics defined in equations 13–15 are computed. Let $\tau(y)$ denote the chosen statistic.

2. The boostrap procedure needs a DGP for generating simulated samples under the null that is simply $H_0 : y_t \sim \text{i.i.d.N}(\mu_1, (\mu_2^c)^2)$. This DGP is determined by estimating the model under the null using the data, *i.e.* by estimating μ_1 nad μ_2^c. This estimation can be made simply by $\hat{\mu}_1(y)$ and $\hat{\mu}_2^c(y)$. Of course, it is a parametric bootstrap since the null hypothesis is the normality of the process.

3. The bootstrap loop starts now. The simulated time series under the null, denoted $(y_t^b)_t$, is generated by independently drawing from the normal distribution:

$$y_t^b \sim N\left(\hat{\mu}_1(y), (\hat{\mu}_2^c(y))^2\right),$$

   It is necessarilly a parametric bootstrap.

4. The three test statistics are computed using the simulated series $(y_t^b)_t$.

5. The steps 3–4 are done again B times. A set of bootstrap statistics is then obtained for each original statistic.

6. The bootstrap P value is finally computed: the classical bootstrap P value is

$$p_{\text{sym}} = \frac{1}{B}\sum_{b=1}^{B} I(|\tau(y^b)^2| \geq |\tau(y)^2|).$$

We also propose the following bootstrap P value:

$$p_{\text{asym}} = 2\min\{p_{uni}, 1 - p_{uni}\},$$

where

$$p_{\text{uni}} = \frac{1}{B}\sum_{b=1}^{B} I(|\tau(y^b)| \leq |\tau(y)|).$$

This P value takes into account the asymmetry of the statistic distribution in addition [14] This P value can be found in Davidson and MacKinnon (1993), chapter 5, in the context of confidence regions. This P value can be used only for $t_{\text{sk}}(y)$ and $t_{\text{ku}}(y)$, but not for $t_{\text{jb}}(y)$ because p_{uni} cannot be compute for $t_{\text{jb}}(y)$ since there are two underlying statistics. An extension of the asymmetric P value could be done for two dimensions, but in practice, it is very difficult.

7. Finally, a significance level is chosen and compared to the P values. If a P value is lower to the significance level, H_1 is retained, otherwise H_0 is retained.

D.4 Extension to linear dependence

We propose here an extension of the bootstrap tests to linearly dependent processes. For the asymptotic tests, the time series has to be independent, since the skewness and the kurtosis depend on the dependence parameters. However, if the linear dependence is specified in the bootstrap procedure, the bootstrap tests can applyed easily to linearly dependent processes, if they are stationary. If the linear dependence is not taken into account, the autocorrelation amplifies the distortion of the test statistic distribution compared to the asymptotic distribution. Two steps change: the DGP estimation under the null, and the simulated series generation. We propose the following procedure:

2 The null is now $y_t \sim$ Gaussian $AR(p)$. This DGP is determined by estimating the model under the null using the data and the OLS procedure. p is determined by the AIC and SIC criteria. Let the estimated DGP be denoted $\widehat{AR}(\hat{p})$. $\hat{p}$ can be chosen by criteria as AIC and SIC, and also by Ljung-Box and Box-Pierce tests (using 1 to 8 lags for example) until the residuals seem independent.

3a The simulated error terms, denoted e_t^b, are generated for a sample. The simulated error terms are drawn from the normal distribution

$$e_t^b \sim N(0, s^2),$$

where s is the standard error of the error terms estimated from the AR(p) regression using the data.

[14] This bilateral bootstrap P value is not useful for the asymptotic tests, since the asymptotic distributions are symmetric: both the P values are the same, or unilateral.

3b The simulated time series under the null, denoted $(y_t^b)_t$, is generated recursively using both the Gaussian $\widehat{AR}(\hat{p})$ model.

D.5 Monte Carlo results

The Monte Carlo results, that are not presented here, suggest that the bootstrap tests are much more reliable than the asymptotic tests. The bootstrap tests are not perfect, and suffer from size distortion, but only for very small sample size.