EPEE
UNIVERSITE D'EVRY
VAL D'ESSONNE

**DOCUMENT DE RECHERCHE**

**EPEE**

**CENTRE D'ETUDE DES POLITIQUES ECONOMIQUES DE L'UNIVERSITE D'EVRY**

# Graphical Methods for Investigating the Finite-sample Properties of Confidence Regions: A Gap in the Literature? A New Proposal

*Christian de PERETTI & Carole SIANI*

**08 - 01**

# Graphical Methods for Investigating the Finite-sample Properties of Confidence Regions: A Gap in the Literature? A New Proposal

**Christian de Peretti** *
*Centre d'Etudes des Politiques Economiques* (EPEE)
Department of Economics
University of Evry Val d'Essonne (France)

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

February 23, 2008

**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 as regards 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 prove that when simulations are run to compute the coverage for only one confidence level, which is usually done in the literrature for classical presentations in tables, all the information useful for computing the coverage for all the levels is present. Thus, there is absolutely no loss of computing time by using this method, whereas it provides more information than the corresponding tabular presentations. These figures are used to illustrate the finite sample properties of autoregressive parameter confidence regions in the context of AR(1) processes. Particularly, asymptotic, percentile, and percentile-t confidence intervals, as well as confidence intervals based on inverting bootstrap tests are presented and commented. Monte Carlo results assessing the performance of 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. An application is made on stock market indices.

*Correspondence to: Christian de Peretti. Address: Bâtiment Île-de-France - 4, 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.

**Keywords:** Monte Carlo experiments, Graphical methods, confidence regions, inverting tests, bootstrap.

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

# 1 Introduction

To obtain evidence on the finite-sample properties of procedures giving confidence regions, Monte Carlo methods are generally use by econometricians. Unfortunately, in the literature there are no 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 providing more information, and yielding graphs that are easy to interpret. Dealing with the implementation of the methods, we prove that when simulations are run to compute the coverage for only one confidence level, which is usually done in the literature for classical presentations in tables, all the information useful for computing the coverage for all the levels is present. Thus, there is absolutely **no loss of computing time by using this method**, whereas they give more information than the corresponding tabular presentations.

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 the effectiveness of a method 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 probability of containing the true value of the parameter; and then, the various methods can be compared. The choice of effectiveness criteria is also discussed.

In order to illustrate and motivate coverage plots, coverage discrepancy plots, and coverage effectiveness curves, these graphs are used to present the results of a study dealing with the properties of confidence regions for the autoregressive parameter in AR model. Various confidence regions 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 test. Double bootstrap versions of the procedures are also briefly discussed.

section 2 presents the graphs we propose for experiments dealing with confidence region coverage probability. The use of coverage-effectiveness curves is also proposed and discussed. For illustrating the use of the graphical methods, they are applied to various confidence intervals for the autoregressive parameter in the AR models in section 3. Then, the number of Monte Carlo results on the various confidence intervals are presented in section 4. In section 5, the autoregressive parameter confidence intervals are applied to stock market indices. section 6 concludes.

# 2 Graphical methods

## 2.1 Background

Let $\theta$ 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 (assumed to be stationary), $\theta$ is a scalar of an unknown parameter, and $\varepsilon$ the unobserved vector of error terms (assumed to be independent of $z$).*

Let $\tau$ be the statistic used for constructing the confidence region:

$$\tau \equiv \tau(\theta; X),$$

where $X$ is the observed finite sample. $\tau$ can also be a vector.

*In our example, $X = (y\ z)$, and $\tau$ can be an estimator of $\theta$ and an estimator of the standard error of the estimator of $\theta$: $\tau = (\hat{\theta}, \hat{\sigma}^2_{\hat{\theta}})$. If the OLS estimator is used, $\tau = \left(\frac{z'y}{z'z}, \frac{\hat{\sigma}^2_\varepsilon}{z'z}\right)$, consequently, $\tau$ depends on $X$, and thus on $\theta$ since $\frac{z'y}{z'z} = \theta + \frac{z'\varepsilon}{z'z}$.*

The cumulative distribution function (CDF) of $\tau$ is denoted $F_\theta$. The distribution of $\tau$ has to depend on $\theta$ for being able to make inference on its value.

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

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

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

However, there is an infinite number of functions $R$ giving such a confidence region. Each function $R$ corresponds to a different method. When the notation $R$ is used in the following, we mean one of these methods, without any indexation.

[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.

*Here are some examples of confidence regions for $\theta$:*

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

*About the family $\{F_\theta^{-1}\}_{\theta\in\Theta}$, we do not need the whole CDF of $\tau$ here, but only its studentised form, i. e. the Student distribution with $T-1$ degrees of freedom, denoted $t_{T-1}$. Consequently, let us consider $F_\theta \equiv F$ the CDF of the $t_{T-1}$ distribution. Since the studentised statistic is perfectly pivotal [2] in our example, $F$ does not depend on $\theta$. With these notations, $R_1$ can be rewritten:*

$$R_1 = \left[\hat{\theta} + F^{-1}\left(\frac{\alpha}{2}\right)\hat{\sigma}_\theta, \hat{\theta} + F^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_\theta\right].$$

*In general, the statistic of interest is not exactly pivotal, but only asymptotically pivotal; this case will be treated in the following.*

## 2.2 Graphical representation

If $\theta_0$ is the true value of the parameter vector $\theta$ that generates the random sample $X$, then:

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

Consequently, the plot of $\{(1-\alpha, \mathbb{P}\{\theta_0 \in R\}); \alpha \in [0,1]\}$ is equal to the 45 degrees line. However, the family $\{F_\theta\}_{\theta\in\Theta}$ is generally unknown, since it generally depends on the unknown parameter $\theta$.

*Let our example be modified now by taking $z_t = y_{t-1}$. Since $z$ is assumed to be stationary, the absolute value of $\theta$ has to be lower than one. In this situation, $\tau$ does not follow a $t_{T-1}$ distribution, but a more complicated distribution depending on $\theta$.*

Therefore, the family $\{F_\theta\}_{\theta\in\Theta}$ has to be estimated and its estimator is denoted 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 $F_\theta$ (coming from a first order Taylor expansion, for instance).

*In our example, the asymptotic limit of $\{F_\theta\}_{\theta\in\Theta}$ is the $t_{T-1}$ distribution, that is independent to $\theta$ (the statistic $\tau$ is asymptotically pivotal). This distribution is chosen for estimating $F_\theta$ for all $\theta \in \Theta$.*

Let us denote:

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

being the confidence region calculated with the estimated statistic distribution rather than the true one. In practice, the difference between $R$ and $\hat{R}$ can also come from an approximation in the analytical calculus of $R$.

$\mathbb{P}\{\theta_0 \in \hat{R}\}$ 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. Since $\hat{F}_\theta$ is not exact, we have in general:

$$\mathbb{P}\{\theta_0 \in \hat{R}\} \neq 1-\alpha.$$

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

However, if the estimator of $\{F_\theta\}_{\theta\in\Theta}$ is consistent, it is expected that:

$$\mathbb{P}\{\theta_0 \in \hat{R}\} \approx 1-\alpha.$$

Consequently, the plot of:

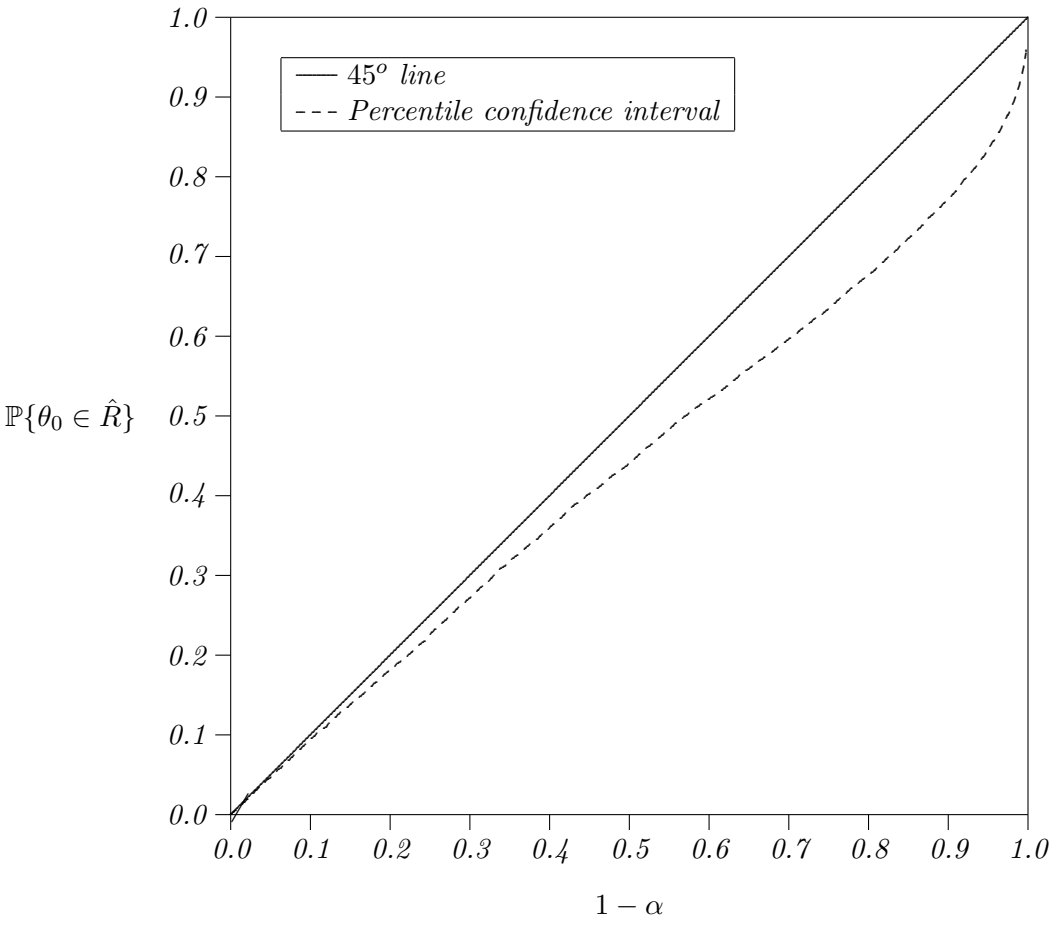
$$\left\{\left(1-\alpha, \mathbb{P}\{\theta_0 \in \hat{R}\}\right); \alpha \in [0,1]\right\}$$

is near the 45 degrees line.

*Figure 1 presents the graph of* $\mathbb{P}\{\theta_0 \in \hat{R}\}$ *against* $1-\alpha$ *for the confidence interval obtained with the percentile bootstrap method for the autoregressive parameter in the AR(1) model when the autoregressive parameter* $\theta = 0.9$ *and the sample size* $T = 8$.

Figure 1: *Case of AR(1) process*

$\theta = 0.9 \quad T = 8$




## 2.3 Use of Monte Carlo simulations

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

(a DGP is a special case of model parameter values). We may denote these simulated values by $\tau_s(\theta), s \in \{1, \ldots, S\}$.

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

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

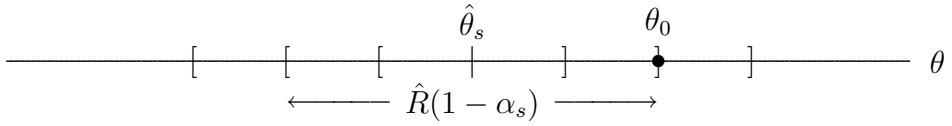
for S very large. $\mathbb{I}(\ .\ )$ denoted an indicator function that takes the value 1 if its argument is true an 0 otherwise.

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

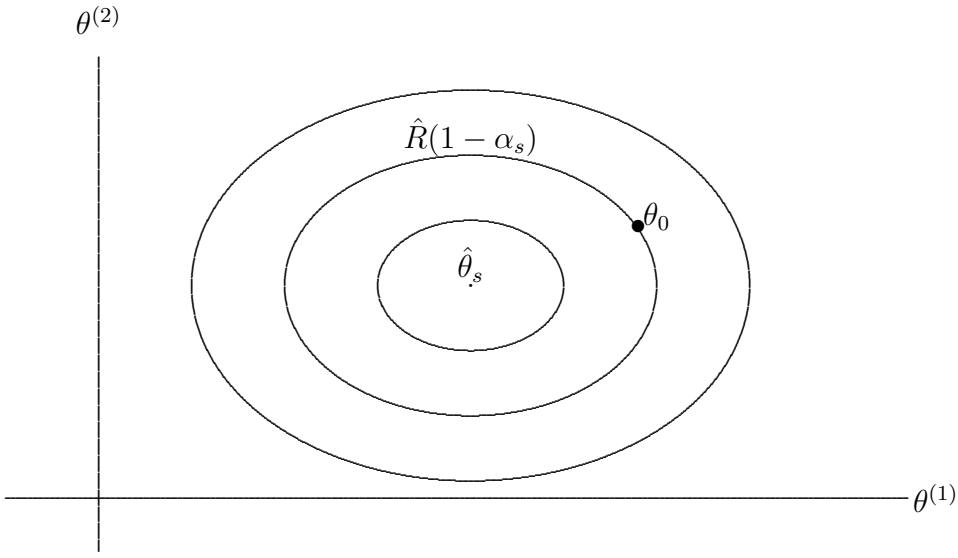
$$\theta_0 \in \partial R\left(\tau_s(\theta_0), \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha\right)$$

where $\partial$ represents the border of the set of values.

*For example, if $\theta$ is scalar, and thus the confidence region is an interval around $\hat{\theta}_s$, the parameter estimate for the simulated sample number $s$, the situation can be illustrated by the following figure:*




*$\theta$ can also be a two dimensional parameter, and thus the confidence region can be an ellipse, and in this case, the situation is illustrated by the following figure:*




$1-\alpha_s$ can be called the *critical coverage*. If the confidence region is defined for all the values for $1-\alpha$, at least one $1-\alpha_s$ necessarily exists. However, $1-\alpha_s$ is not necessarily unique, depending on the form of the confidence region. But it is easy to get the uniqueness of $1-\alpha_s$ by assuming the natural hypothesis that $R$ (and $\hat{R}$) is increasing with respect to $1-\alpha$ in the sense of set inclusion. This hypothesis is obtained if $R$ is optimised using the maximum likelihood principle, for instance, but not only.

It should be noted that

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

where $\bar{R}$ denotes $R$ and its border (if $1-\alpha_s = 1-\alpha$, then $\theta_0 \in \partial R$). Thus:

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

In practice, we just count the proportion of $1-\alpha_s$ smaller or equal to $1-\alpha$. In fact, $\hat{\mathbb{P}}\{\theta_0 \in \hat{R}\}$ is the empirical CDF of $1-\alpha_s$ viewed as a random variable. $1-\alpha_s$ can be viewed as a random variable since it depends on $\tau_s$ that is random, and on $\{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}$ that can also be random because of the estimate:

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

$1-\alpha_s$ depends also on $R$ in the sense of the way in which the confidence region is built from $\tau_s$ and $\{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}$. If $R$ is not analytically known, and has to be numerically computed, it can also be a source of error.

All the graphs we will discuss are based on the empirical CDF of the confidence level $1-\alpha_s$ of $R(\tau_s, \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta},\ .\ )$ if the true value $\theta_0$ of the parameter vector is just in $\partial R$ (the border of $R$).

Let $F$ denote the generally unknown (true) finite-sample CDF of $1-\alpha_s$:

$$\begin{aligned} F(1-\alpha) &= \mathbb{P}\left\{1-\alpha_s \leq 1-\alpha\right\} && \forall\alpha\in\mathbb{R}, \\ &= \mathbb{P}\left\{\theta_0 \in \bar{R}\left(\tau(\theta_0), \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha\right)\right\} && \forall\alpha\in[0,1]. \end{aligned}$$

If the computation of $R$ is exact, and $\{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}$ is known, the coverage probability is equal to the confidence level:

$$\begin{aligned} F(1-\alpha) &= \mathbb{P}\left\{\theta_0 \in \bar{R}\left(\tau(\theta_0), \{F_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha\right)\right\}, \\ &\equiv F_0(1-\alpha), \\ &= 1-\alpha. \end{aligned}$$

$F_0$ is in fact the CDF of a uniform random variable $U(0,1)$. Consequently, $(1-\alpha_s)_{s=1}^S \sim i.i.d.U(0,1)$. $F$ can be computed by Monte Carlo experiments as follows:

$$\begin{aligned} \hat{F}(1-\alpha) &= \hat{\mathbb{P}}\left\{\theta_0 \in \bar{R}\left(\tau(\theta_0), \{\hat{F}_\theta^{-1}\}_{\theta\in\Theta}, 1-\alpha\right)\right\} && \forall\alpha\in[0,1], \\ &= \frac{1}{S}\sum_{s=1}^{S}\mathbb{I}(1-\alpha_s \leq 1-\alpha) && \forall\alpha\in\mathbb{R}. \end{aligned}$$

Only the points in the $(0,1)$ interval are of interest since $1-\alpha \in [0,1]$. At any point $x_i$ in the $(0,1)$ interval, $\hat{F}$ is defined by:

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

When $S$ is large, storage space can be conserved by evaluating the EDF 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. A parsimonious way to choose the $x_i$ is:

$$\begin{aligned} x_i &= 0.001, 0.002, \ldots, 0.009, 0.01, 0.02, \ldots, 0.09, 0.1, 0.15, \ldots \\ &\quad \ldots, 0.85, 0.9, 0.91 \ldots, 0.98, 0.99, 0.991, \ldots, 0.998, 0.999 \qquad (N = 53) \end{aligned} \tag{2}$$

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

## 2.4 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 exact, $F = F_0$, the CDF of a $U(0,1)$ variable. 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*.

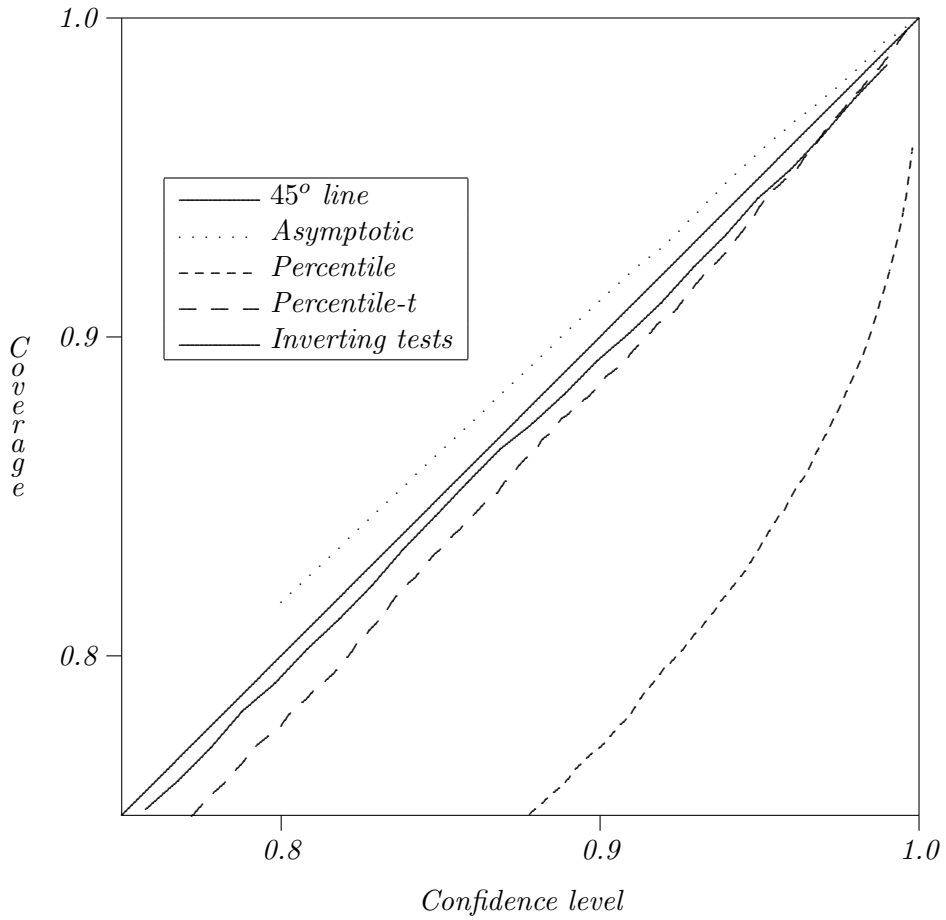
On the one hand, coverage plots make it very easy to distinguish confidence regions that perform badly from the others. Moreover, since they show how a confidence region performs for all confidence levels, these coverage plots are particularly useful for regions that both over-cover and under-cover the true value of the parameter. The corresponding potential disadvantage of such plots is that they can take up a lot of space on the page: for plotting the coverage against the confidence 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 2.5) 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 regions.

> *Figure 2 presents the coverage plots for confidence intervals calculated with various methods (asymptotic, percentile bootstrap, percentile-t bootstrap, based on inverting bootstrap tests) for the autoregressive parameter in the AR(1) model when the autoregressive parameter $\theta = 0.9$ and the sample size $T = 8$ (the mehtods are presented in section 4).*
>
> *It is easy to observe that the percentile bootstrap confidence interval presents a serious coverage distortion, since the corresponding curve is very far from the 45 degree line. However, it is difficult to distinguish between the three other methods, since they seem to perform correctly.*

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, and seem confused. These

Figure 2: *Coverage plots in the case of AR(1) process*

$\theta = 0.9 \quad T = 8$




plots are therefore not very useful for distinguishing among such confidence regions. In this case, we propose to use the coverage discrepancy plots detailed in subsection 2.5. Obviously, since coverage plots (and coverage discrepancy plots) use one dimension for confidence level, they 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 for calculating confidence regions are very computational time consuming, for instance, the double bootstrap-t based on some nonlinear estimators confidence interval: first, for estimating the parameter by a nonlinear method,

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

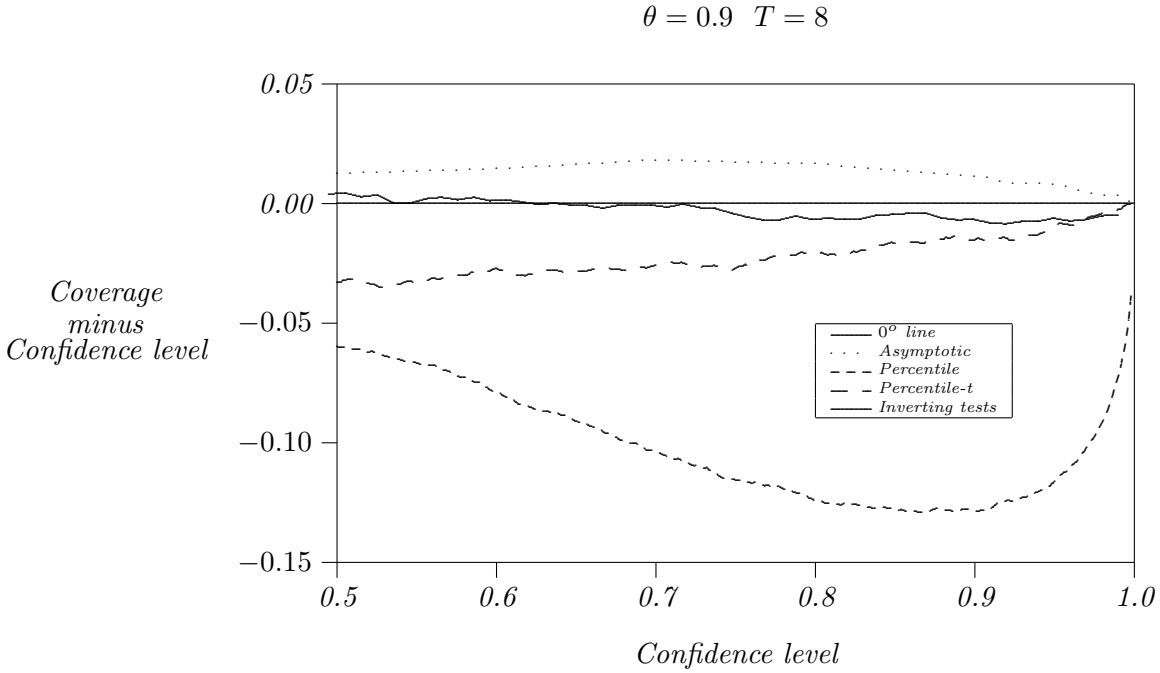
an iterative optimising algorithm is necessary. Second, for estimating its standard deviation, a bootstrap loop is used in which the estimator is run at each bootstrap iteration (this standard deviation is used for studentising the estimator). And third, for computing the distribution function of the 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.

## 2.5 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 that graph is partly spurious, reflecting experimental randomness. It is therefore natural to smooth the plots. In Davidson and MacKinnon [1998], the authors 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 as for Figure 2 using coverage discrepancy plot.*

Figure 3: *Coverage discrepancy plots in the case of AR(1) process*




## 2.6 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 property. We will discuss graphical methods for comparing the effectiveness of competing regions using *coverage-effectiveness curves*. For a Monte Carlo experiment (in which a

given DGP is used), these curves can be constructed using the empirical CDF of the critical coverage $1-\alpha_s$ and the empirical CDF of a 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 the (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.

> *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.*

Unfortunately, this is what is often implicitly done in the context of statistical tests for example, when the power of the tests is reported in tables.

In order to plot effectiveness criterion 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 CDF be denoted $\hat{F}(x)$, and let the estimated effectiveness criterion 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 criterion 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 $\left(\hat{F}(x_i), \hat{E}(x_i)\right)$, 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.

However, 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 (that is shown in the coverage plots), this region cannot be chosen for practical uses, and thus, it is not useful to compute

its "true" effectiveness criterion by coverage-effectiveness curves for all the coverages.

## 2.7 Choosing the effectiveness criterion

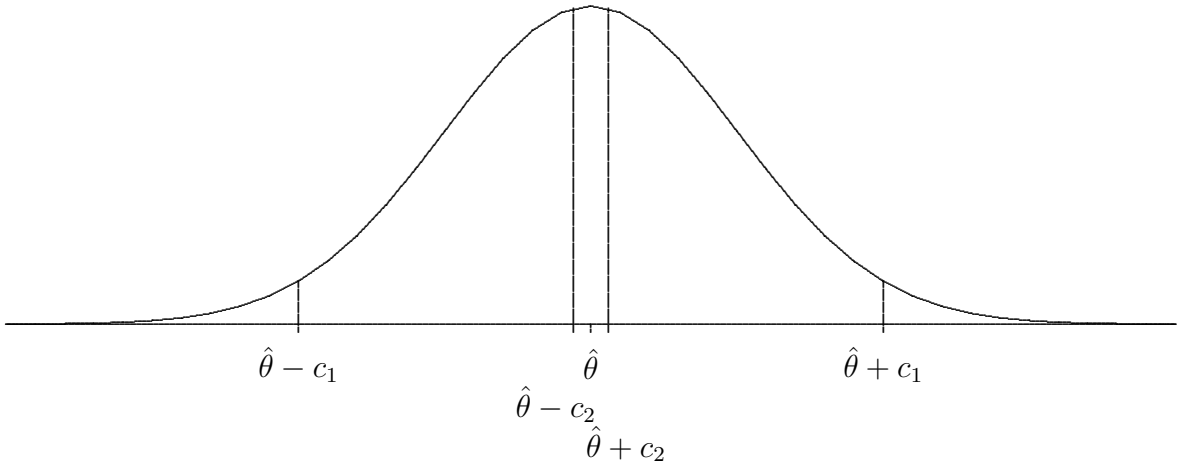
In this subsection, we will show that comparing confidence region methods only on the basis of their coverage property is not enough: two confidence regions can have the same coverage property whereas one can be preferable to the other. However, this preferable relation among the methods is not straightforward, and has to be discussed: it depends on the mathematical problematics under consideration, but also on the economic problematics. Once the mathematical and economic objectives are clearly determined, a corresponding effectiveness criterion reflecting the preference can be defined and used for measuring the effectiveness of the methods.

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

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

These regions are illustrated in Figure 4.

Figure 4: Two confidence regions with the same confidence level




Both these intervals are correct if their coverages are equal to $1-\alpha$. However, $R_2$ does not contain the most likely values for $\theta$. More rigorously, the probability that an elementary interval $dx$ in $R_1$ contains the true value of $\theta$ is larger than if $dx$ lies 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 (Borel or Lebesgues 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 confidence regions based on inverting tests, if two confidence regions have the same confidence level but two different lengths,

the one that has the small length must have the largest probability of presence for $\theta$ in an elementary interval $dx$.

The likelihood of the values for $\theta$ 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 reparameterisation (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, rather than more straightforward but not invariant confidence regions, as the ones based on Student statistics or Wald tests.

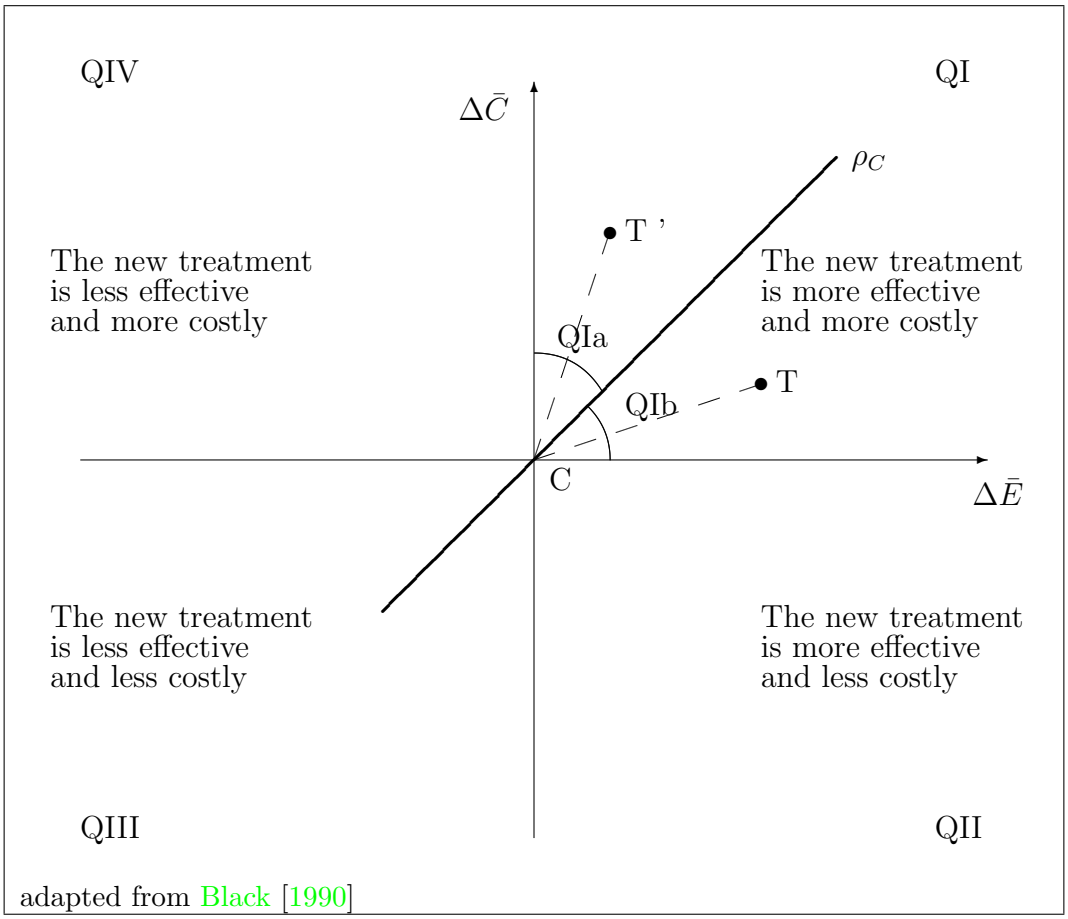
Finally, the choice of the effectiveness criterion can also depend on the economic purpose. Let us consider the following example coming from Siani and Moatti [2003] in the context of health economic evaluations.

> *Let us consider the following example. In cost-effectiveness analyses (CEA), 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:*
>
> $$\rho = \frac{\mu_{C1} - \mu_{C0}}{\mu_{E1} - \mu_{E0}} = \frac{\mu_{\Delta C}}{\mu_{\Delta E}} \ ,$$
>
> *where $\mu$ is the theoretical mean value of costs ($C$) and effects ($E$) for treatments number 1 and number 0. 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 $T_0$. 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 that underwent treatment $T_1$ and group number 0, consisting of individuals that underwent treatment $T_0$):*
>
> $$\hat{\rho} = \frac{\overline{C_1} - \overline{C_0}}{\overline{E_1} - \overline{E_0}} = \frac{\Delta\bar{C}}{\Delta\bar{E}} \ ,$$
>
> *where $\overline{C_1}$, $\overline{C_0}$ are the sample mean of the costs and $\overline{E_1}$, $\overline{E_0}$ are the sample mean of effects in the two treatments arms. 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 5, 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 $\Delta\bar{E}$. The new treatment $T_1$ is represented by the point $T$ in Figure 5 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 $\hat{\rho}$ [4]. In order to decide whether to adopt treatment $T_1$ rather than treatment $T_0$, it is necessary to introduce a ceiling ratio, denoted $\rho_C$, corresponding to some maximum value of the ICER that people are prepared to pay to achieve this additional effectiveness [5]. Thus, in quadrants QI and QII, if the ICER is lower than the ceiling ratio, then treatment $T_1$ should be*

[4] 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.

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

Figure 5: The cost-effectiveness plane




*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 $\rho_C$.*

*A confidence interval for $\rho$ would naturally be built from the distribution of $\hat{\rho}$. It should be noted that a confidence interval for $\rho$ will be represented by a (angular) sector around $\hat{\rho}$ 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 [1954]'s method 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 $]-\infty, \rho^U] \cup [\rho^L, +\infty[$ that simply corresponds to a sector (also around $\hat{\rho}$) 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). 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.*

*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).

## 2.8 Chosing the confidence region method with respect to the graphical representations

The coverage plots and the effectiveness curves are very useful for choosing among methods that have reasonable coverage distortions: they permit to make an *arbitrage* between the coverage distortion and the true effectiveness of the methods, and then to chose the most appropriate one. 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 rules 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 wrong.

2. Among the methods selected in the first step, select the methods that have the best effectiveness curves.

3. Among the methods that do not present a too large coverage distortion, and that have a good effectiveness, select the most appropriate on the double basis of coverage and effectiveness: a large gain in the effectiveness can compensate the a small loss in the coverage. The lenght of the gain in effectiveness required for compensating a coverage distortion depends on the situation and on the decision-maker.

## 2.9 Other types of plots of the literature

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 satisfactorily the performances of confidence regions.

It would be more conventional to graph the CDF of the statistic $\tau$ instead of the CDF of its critical coverage $1 - \alpha_s$. The CDF of the statistic $\tau$ 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]. It is worth to note that the coverage plot is the *PP plot* of the critical coverage random variable $1 - \alpha_s$.

Another common type of plot is a *QQ plot*, in which the quantiles of $\tau$ are plotted against the quantiles of its hypothesised distribution. If the distribution of $\tau$ 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. Figure 6 presents an example of a *QQ plot* for a $\chi^2(1)$ statistic based on Fieller's Theorem (Fieller [1954]) used for building a confidence region for the ICER (see subsection 2.7).

Figure 6: Example of *QQ plot*

*QQ plots* have disadvantages. One serious problem is that *QQ plots* have no natural scale for the axes: the scale depends on the distributions. Thus, if the hypothesised distribution changes for example, the associated scale will change too. This makes it difficult to plot on the same axes statistics that have different distributions. They also take up much more space than do coverage plots restricted to $[0.75, 1] \times [0.75, 1]$. Moreover, 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 Section 2 of Davidson and MacKinnon [1998] 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 3.4). 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 but also for all 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, under constraint, .... In the case of a bilateral interval for example, 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 2.7), or a band (as for the net benefit estimation, see Stinnet and Mullahy [1998], Tambour et al. [1998]) depending on the economic problems. Thus, many different confidence regions and hypothesis tests can be defined from the statistic and its distribution. Consequently, although the *QQ plots* can certainly make clear that a confidence region does not work perfectly, presenting only the *PP plots* or the *QQ plots* can be insufficient for showing the performance 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). To conclude, *PP plots* and *QQ plots* provide much less useful information than coverage plots.

# 3 Application to various confidence regions for the autoregressive parameter

In order to illustrate and motivate coverage plots and coverage effectiveness curves, these graphs are used to present the results of a study on the autoregressive parameter confidence regions properties in the context of univariate linear AR(1) processes with Gaussian errors:

$$y_t = \rho y_{t-1} + \varepsilon_t \qquad t \in \{1, \ldots, T\}, \tag{3}$$

$$\{\varepsilon_t\} \sim i.i.d. N(0, \sigma^2), \tag{4}$$

where $|\rho| < 1$ such that the series is stationary and invertible, and $\sigma^2 < \infty$.

Various confidence regions based on the *Ordinary Least Squares* (OLS) estimator are compared: the confidence region using the asymptotic distribution of the estimator (subsection 3.1), the confidence regions using the percentile (subsection 3.2) and the percentile-t (subsection 3.3) methods based on the bootstrapped distributions, and the confidence region based on inverting bootstrapped $t$ tests (subsection 3.4). See table 1 for a listing of these methods.

Since the $t$ statistic is not exactly pivotal in the context of AR processes, bootstrapping will not perform perfectly. However, the $t$ statistic 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,a]. 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

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 |

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

2. The second type of distortion comes from the fact that the denominator of the studentised statistic is not independent from the numerator. Conversely to the first type of distortion, the second type of distortion can be more persistent with respect to the sample size. In this case, parametric bootstrap and nonparametric bootstraps perform similarly but both have size distortions in finite sample.

Thus, in our Monte Carlo experiments, we prefer to focus on the case of Gaussian errors 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. 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. Nonparametric versions of the test are not used. That will permit to show the gain of using inverting tests confidence regions compared to not inverting tests, without noisy error due to nonparametric estimation. 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 replicated series are generated, we just have to draw the bootstrap error terms in the empirical distribution of the residuals of the estimated series rather than in a Gaussian distribution (see Davidson [1998] for examples and details of other nonparametric bootstrap methods).

The various confidence regions are presented in the following subsections, as well as how to use the graphical representation for measuring their performance.

## 3.1 Asymptotic confidence region

The classical asymptotic confidence region for the autoregressive parameter $\rho$ is obtained from the OLS estimator as follows:

$$[\rho_-^{as}, \rho_+^{as}] = \left[\hat{\rho} + \hat{\sigma}(\ \hat{\rho}\ )\ \hat{F}^{-1}\left(\frac{\alpha}{2}\right)\ ,\ \hat{\rho} + \hat{\sigma}(\ \hat{\rho}\ )\ \hat{F}^{-1}\left(1 - \frac{\alpha}{2}\right)\right],$$

where:

- $\hat{\rho}$ is the OLS estimator of $\rho$,
- $\hat{\sigma}(\hat{\rho})$ is the standard error of $\hat{\rho}$,

- $\hat{F}$ is the CDF of the asymptotic distribution of the $t$ statistic, that is the Student distribution with $T-1$ degrees of freedom, denoted $t(T-1)$.

For using our graphical methods, we need to calculate the confidence level $1-\alpha_s$ such that $\rho_0$, the true value for $\rho$, is included in the border of the confidence region $\partial R$, that is $\{\rho_1, \rho_2\}$ here. Consequently, we obtain:

$$1-\alpha_s = \hat{F}\left(\left(\frac{\rho_0 - \hat{\rho}_s}{\hat{\sigma}(\ \hat{\rho}_s\ )}\right)^2\right)$$

where $\hat{\rho}_s$ is the estimated value for $\rho$ for the simulated series number $s$.

Dealing with the effectiveness, denoted $E(1-\alpha)$ in subsection 2.6, 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 2.6, has to be calculated for all the confidence levels. The length of the asymptotic confidence interval for the simulated series number $s$ is:

$$e(1-\alpha; s) = 2\ \hat{F}^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}(\ \hat{\rho}_s\ ).$$

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 expectation for each level is estimated by the sample mean of the lengths $e(1-\alpha, s)$:

$$\hat{E}(1-\alpha) = \frac{1}{S}\sum_{s=1}^{S} e(1-\alpha, s).$$

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 every values for $1-\alpha$, the vector $e(\ .\ , s)$ is computed (the computation is very straightforward using matrix programming).

A secondary effectiveness criterion can also be used: the confidence interval length standard deviation. This criterion is estimated by:

$$\hat{E}_2(1-\alpha) = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\left[e(1-\alpha, s) - \hat{E}(1-\alpha)\right]^2}.$$

## 3.2 Percentile bootstrap confidence interval

The bootstrap procedure for calculating a percentile confidence interval can be summarised as follows:

1. The real data sample has to be estimated using a model as close as possible to the reality (for instance an Gaussian AR(1) model here).

2. This model, with the previously estimated parameters, is used as a *Data Generating Process* (DGP) for generating $B$ simulated samples $\{y^b\}_{b=1}^{B}$, corresponding to $B$ bootstrap replications of the real data sample $y$. The number of replications $B$ has

to be chosen sufficiently large to avoid random effect of the bootstrap experiment: at least several hundreds, but several thousands is desirable (see Davidson and MacKinnon [2000]). It is also desirable that $(1-\alpha)(B+1)$ be an integer (see Davidson and MacKinnon [1993]).

3. The estimator $\hat{\rho}$ is computed for each bootstrap sample $y^b$, and is denoted $\hat{\rho}_b^\star$. Consequently, the procedure provides $B$ bootstrap replications of $\hat{\rho}$. The aim of the bootstrap procedure is to get the probability distribution of $\hat{\rho}$ from their replications $(\hat{\rho}_b^\star)_{b=1}^B$.

4. Consequently, the percentile confidence interval is calculated using the set of $\{\hat{\rho}_b^\star\}_{b=1}^B$:

$$(\rho_-^p, \rho_+^p] = \left(\hat{\rho}^\star_{(\lfloor B\frac{\alpha}{2}\rfloor)}, \hat{\rho}^\star_{(\lfloor B(1-\frac{\alpha}{2})\rfloor)}\right],$$

where $(\ .\ )$ in $\hat{\rho}^\star_{(\ .\ )}$ is the ordered statistic, and $\lfloor\ .\ \rfloor$ is the integer part.

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].

The confidence level such that $\rho_0 \in \partial R$ is equal to:

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

where

$$pv = \frac{1}{B}\sum_{b=1}^{B} I(\hat{\rho}_b^\star \leq \rho_0).$$

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

$$e(1-\alpha) = \hat{\rho}^\star_{(\lfloor B(1-\frac{\alpha}{2})\rfloor)} - \hat{\rho}^\star_{(\lfloor B\frac{\alpha}{2}\rfloor)} \qquad \forall\alpha \in [0,1].$$

It should be noted that conversely to the asymptotic confidence interval in subsection 3.1, the number of different possible values for the percentile interval length is finite, since the bootstrap probability distribution is discrete. Consequently, rather than calculating the interval length $e(1-\alpha)$ for a predetermined set of values for $1-\alpha$ (as for the asymptotic interval), we prefer to calculate each possible different length and their corresponding confidence levels. Using matrix programming, computing the vector of lengths $e$ is very straightforward:

$$e = \begin{pmatrix} \hat{\rho}^\star_{(\lfloor\frac{B}{2}+1\rfloor)} \\ \hat{\rho}^\star_{(\lfloor\frac{B}{2}+1\rfloor+1)} \\ \vdots \\ \hat{\rho}^\star_{(B-1)} \\ \hat{\rho}^\star_{(B)} \end{pmatrix} - \begin{pmatrix} \hat{\rho}^\star_{(\lceil\frac{B}{2}\rceil)} \\ \hat{\rho}^\star_{(\lceil\frac{B}{2}\rceil-1)} \\ \vdots \\ \hat{\rho}^\star_{(2)} \\ \hat{\rho}^\star_{(1)} \end{pmatrix}$$

where $\lceil\ .\ \rceil$ is the smaller integer larger than the argument. The associated confidence levels are the sequence:

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

### 3.3 Percentile-t bootstrap confidence interval

The percentile-t procedure is similar to the percentile procedure, but rather than using directly the estimator of $\rho$, the studentised form of that estimator is used:

$$\tau = \frac{\hat{\rho} - \rho}{\hat{\sigma}(\hat{\rho})}.$$

$\hat{\sigma}(\hat{\rho})$ is the estimated variance of $\hat{\rho}$, and can be calculated in several ways. In this paper, we only consider the estimator coming from the OLS regression, that is valid asymptotically (see subsection 3.1). A bootstrap estimator can also be considered [6], however, this kind of bootstrap is very time consuming. That is the reason why double bootstrapping is not chosen in the context of this paper, dealing with Monte Carlo experiments. Studentising the estimator leads to a statistic asymptotically pivotal, permitting to obtain a higher rate of convergence for the bootstrap method. $\tau$ is bootstrapped for obtaining $B$ replications, denoted $\tau_b^\star$, as follows:

$$\tau_b^\star = \frac{\hat{\rho}_b^\star - \hat{\rho}}{\hat{\sigma}(\hat{\rho}_b^\star)} \qquad \text{for } b \in \{1, \ldots, B\}.$$

The DGP for replicating $\tau$ is determined in the same way as for the percentile method. The percentile-t confidence interval is:

$$\left[\rho_-^{pt}, \rho_+^{pt}\right) = \left[\hat{\rho} - \tau^\star_{\left(\lfloor B\left(1-\frac{\alpha}{2}\right)\rfloor\right)}\hat{\sigma}(\hat{\rho}), \hat{\rho} - \tau^\star_{\left(\lfloor B\frac{\alpha}{2}\rfloor\right)}\hat{\sigma}(\hat{\rho})\right)$$

The coverage corresponding to $\rho_0 \in \partial R$ is equal to:

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

where

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

and

$$\tau_0 = \frac{\hat{\rho} - \rho_{\mathbf{0}}}{\hat{\sigma}(\hat{\rho})}.$$

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

$$e(1-\alpha) = \left(\tau^\star_{\left(\lfloor B\left(1-\frac{\alpha}{2}\right)\rfloor\right)} - \tau^\star_{\left(\lfloor B\frac{\alpha}{2}\rfloor\right)}\right)\hat{\sigma}(\hat{\rho}) \qquad \forall \alpha \in [0,1].$$

The lengths of the intervals are determined in the same way as for the percentile method.

It should be noted that even if the method has a better asymptotic convergence rate than the percentile method, in finite sample, studentising the estimator can produce a statistic that is farther from a 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.

[6] For obtaining the bootstrap estimator of $\sigma(\hat{\rho})$, each bootstrap time series is estimated and replicated $B_2$ times in the same way as previously. $B_2$ can be different from $B$. $B_2$ is generally taken smaller than $B$ since it is less important. For each replicated series, $\hat{\rho}$ is computed leading to a set of $\{\hat{\rho}_b^\star\}_b$ 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 $\sigma(\hat{\rho})$ 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*.

## 3.4 Confidence region based on inverting tests

Let $T_\alpha(\rho')$ denote the result of a test for $H_0 : \rho = \rho'$ against $H_1 : \rho \neq \rho'$ at significance level $\alpha$:

$$T_\alpha(\rho') = \begin{cases} 0 & \text{if } \rho = \rho' \text{ is retained at level } \alpha, \\ 1 & \text{otherwise.} \end{cases}$$

The confidence region built by inverting tests for a $1-\alpha$ confidence level is defined as follows: $\rho'$ belongs to the region if and only if the test retains $\rho = \rho'$ at the $\alpha$ significance level, *i.e.*:

$$\rho' \in \hat{R}(1-\alpha) \Longleftrightarrow T_\alpha(\rho') = 0.$$

Let $p(\rho')$ denote the P value of the test for the hypothesis $\rho = \rho'$. We have:

$$T_\alpha(\rho') = 0 \Longleftrightarrow p(\rho') \geq \alpha.$$

Here, the region will be an interval. Consequently, it can be defined by both its limits, say $\rho_{\text{inf}}$ and $\rho_{\text{sup}}$ (corresponding respectively to the lower and the upper limit). These limits correspond to both the values of $\rho'$ such that $p(\rho') = \alpha$ (see Davidson and MacKinnon [1993] Chapter 5). For computing the P value, see subsections 3.4.1 and 3.4.2. 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].

For using our graphical methods, the critical confidence level $1-\alpha_s$ has to be calculated. In the context of inverting tests, $1-\alpha_s$ is such that $\rho_0 = \rho_{\text{inf}}^{(s)}$ or $\rho_0 = \rho_{\text{sup}}^{(s)}$, where $^{(s)}$ indicates that the values come from the simulated series number $s$. But it is known that:

$$p\left(\rho_{\text{inf}}^{(s)}\right) = p\left(\rho_{\text{sup}}^{(s)}\right) = \alpha,$$

thus:

$$\alpha_s = p(\rho_0).$$

Fortunately, for confidence regions based on inverting tests, the critical confidence level is very easy to calculate, since only one P value has to be calculated; no sequential search for $\rho_{\text{inf}}^{(s)}$ or $\rho_{\text{sup}}^{(s)}$ has to be run.

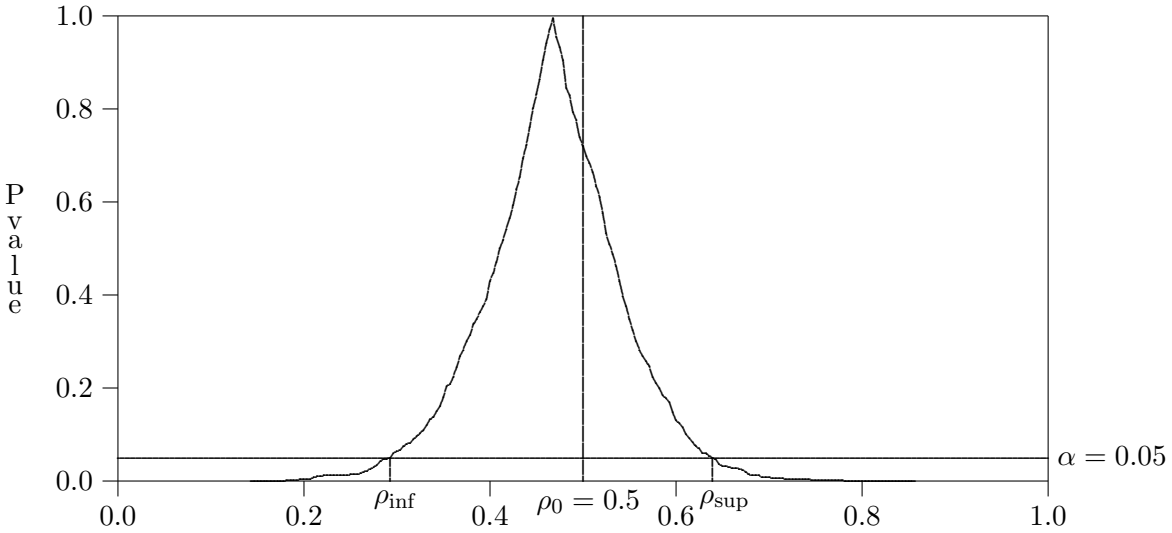
As regards the effectiveness of this method, a computing time problem arises since the limits of the region, $\rho_{\text{inf}}^{(s)}$ and $\rho_{\text{sup}}^{(s)}$, has to be calculated. For computing the confidence interval limits, the values $\rho'$ for $\rho$ such that $p(\rho') = \alpha$ have to be found. Classically, $p(d)$ is a function that increases from 0 to 1 and then decreases from 1 to 0 when $\rho$ goes from $-\infty$ to $+\infty$. For illustrating the P value function, Figure 7 present the P value of the bootstrap test presented in subsubsection 3.4.2 using $B = 999$ bootstrap replications. The null hypothesis of the test is $\rho = \rho'$. The time series are generated using $\rho_0 = 0.5$, $T = 100$, and $\sigma^2 = 1$. The value 1 for the P value function is not necessarily reached if the number of bootstrap replications $B$ is not large enough.

### 3.4.1 Inverting asymptotic tests

More precisely, if the asymptotic $t$ test based on an estimator of $\rho$ is used, the P value can be built as follows:

$$p(\rho') = 1 - \hat{F}_{\rho'}\left[\tau(\rho')\right] \equiv 1 - \hat{F}\left[\tau(\rho')\right],$$

Figure 7: P value function of bootstrap test for the value of $\rho$ in case of AR(1) process with $\rho_0 = 0.5$ and $T = 100$




where

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

$\hat{F}_{\rho'} = \hat{F}$ is the asymptotic cdf of $\tau(\rho')$, and $\hat{\sigma}(\hat{\rho})$ is the standard error of $\hat{\rho}$. However, a confidence interval based on inverting asymptotic tests leads to the same interval than the classical asymptotic confidence interval. Consequently, the performance of this method is not studied here.

### 3.4.2 Inverting bootstrap tests

While the $t$ test based on the asymptotic distribution is asymptotically valid, it is not exact in finite samples, and so, it is natural to "bootstrap" it. The parametric bilateral bootstrapped $t$ tests based on the OLS estimator of $\rho$ is considered here. This procedure can be summarised as follows:

1. Compute the test statistic (the $t$ statistic here) on the real data sample, which will be denoted $\hat{\tau}$.

2. Estimate the AR(1) model by OLS under the null $H_0 : \rho = \rho'$, for obtaining the model parameter estimates (only $\hat{\sigma}^2$ here) and the residuals $\hat{\varepsilon}$ [7]. As for the percentile-t confidence interval, the standard error of $\hat{\rho}$, denoted $\hat{\sigma}(\hat{\rho})$, used for studentising $\hat{\rho}$, can be calculated in several different ways: one of them is the estimator coming from the OLS estimate, the second one is the bootstrap estimator. Again, the OLS estimator is chosen.

[7] In the context of a more general model ARMA(p,q), $p$ and $q$ can be determined 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 the parameters.

3. Draw $B$ sets of bootstrap error terms, $\varepsilon^b$. There are numerous ways to draw the error terms. In the context of our Monte Carlo experiments, the parametric way (parametric bootstrap) is chosen: the $\varepsilon_t^b$ are independent draws from the $N(0, \hat{\sigma}^2)$ distribution. Nonparametric bootstrap methods can also be used (see Davidson [1998]) [8].

4. Use the $B$ sets of bootstrap error terms $\varepsilon^b$ generated in step 3 for generating $B$ bootstrap samples $y^b$. The elements of $y^b$ should be generated from the recurrence equation:

$$y_t^b = \rho' y_{t-1}^b + \varepsilon_t^b \qquad t \in \{1, \ldots, T\}, \tag{6}$$
$$y_0^b = y_0 \qquad \text{the initial value .} \tag{7}$$

5. For each bootstrap sample $y^b$, compute the $t$ statistic, denoted $\tau_b^\star$.

6. Then, a bootstrap P value can be computed by the following formula (see Davidson and MacKinnon [1993]):

$$\hat{p}(\hat{\tau}^2) = \frac{1}{B} \sum_{b=1}^{B} I((\tau_b^\star)^2 > \hat{\tau}^2). \tag{8}$$

This is the P value for an unilateral test corresponding in fact to a bilateral test assuming (implicitly) the symmetry of the statistic distribution. However, the size distortion of the statistic distribution is not necessarily symmetric. Thus, the P value for a bilateral asymmetric test, accounting for the distribution asymmetry, is preferred:

$$\hat{p}(\hat{\tau}) = 2 \min\{\hat{p}_{uni}(\hat{\tau}), 1 - \hat{p}_{uni}(\hat{\tau})\}, \tag{9}$$

where

$$\hat{p}_{uni}(\hat{\tau}) = \frac{1}{B} \sum_{b=1}^{B} I(\tau_b^\star > \hat{\tau}). \tag{10}$$

[8] Three nonparametric bootstrap methods for generating the $\varepsilon_t^b$ are described below:

- The simplest nonparametric bootstrap, called $b_1$: the $\varepsilon_t^b$ are obtained by re-sampling with replacement from the vector of $\{\hat{\varepsilon}_t\}_{t=\hat{p}+1}^T$.
- A slightly more complicated from of nonparametric bootstrap called $b_2$: the $\varepsilon^b$ are generated by re-sampling with replacement from the vector

$$\left\{ \sqrt{\frac{T}{T - 2\hat{p} - 1}} \left( \hat{\varepsilon}_t - \frac{1}{T - \hat{p}} \sum_{i=\hat{p}+1}^{T} \hat{\varepsilon}_i \right) \right\}_{t=\hat{p}+1}^{T}. \tag{5}$$

- The most complicated nonparametric bootstrap, called $b_3$: the $\varepsilon^b$ are generated by re-sampling from the vector with typical element $\tilde{\varepsilon}_t$ constructed as follows:
  (a) 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 $(1-L)^{-d_0}\hat{\phi}(L)^{-1}\hat{\theta}(L)$;
  (b) divide each element of $\hat{\varepsilon}$ by $\sqrt{1 - d_t}$;
  (c) re-centre the resulting vector;
  (d) re-scale it so that it has variance $\hat{\sigma}_\varepsilon^2$.

  This type of procedure is advocated in Weber [1984].

Further considerations about this P value can be found in Chapter 5 of Davidson and MacKinnon [1993] dealing. The bootstrap P value is more complicated than for asymptotic test because it takes into account the asymmetry of the statistic distribution. The unilateral version of the test is not presented here, since it has absolutely no advantage compared to the bilateral version.

For the conception of bootstrap tests see Efron [1979], for its development, see Davidson and MacKinnon [1993], and for further analysis, see Davidson and MacKinnon [1996b,a].

# 4 Monte Carlo experiments

All the experiments deal with Gaussian AR(1) processes. Since the constant term is only a location parameter and does not influence the confidence region performance, this parameter is set to zero. Similarly, the standard deviation of the model, $\sigma$, is only a scale parameter and does not affect at all the performance; consequently, this parameter is set to one. The test statistic depends then on the autoregressive parameter $\rho$ and on the sample size $T$ in Equation 3. $T = 2^n$ is used, where $n$ is an integer. We pick combinations of $\rho$ and $T$ to investigate: $\rho \in \{-1, -0.5, 0, 0.5, 1, 1.5, 2\}$ and $T \in \{4, 8, 16, 32\}$. Each Monte Carlo experiment is run with $S = 10,000$ replications of the time series.

The bootstrap DGP is also a Gaussian AR(1) process, estimated by OLS under the null hypothesis [9]. All the bootstrap methods are run using $B = 999$. Bootstrap methods are run using the same random numbers for avoiding random effect in the result of the bootstrap method between the Monte Carlo replications. That point is also important in the construction of the P value function for searching the confidence interval limits based on inverting bootstrap tests for making the function smoother. It should be noted that in our case the distortions found in the results are due to a distribution distortion of the test statistic depending on the autoregressive parameter value. Since the bootstrap methods have to estimate these parameters, an error in the estimations leads to an error in the coverage.

## 4.1 Coverage plots

The standard deviation of the coverage plots is equal to

$$\sqrt{\frac{c(1-\alpha)\left(1-c(1-\alpha)\right)}{S}}, \tag{11}$$

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

Figure 8 presents the coverage plots of the asymptotic confidence interval, the percentile interval, the (bilateral) percentile-t interval, and the interval based on inverting (bilateral) bootstrap tests.

Table 2 presents the same results than the third graph in Figure 8 (*i.e.* the case $\rho_0 = 0.5$) but using a tabular presentation. Is this table pleasant to read?

[9] In practice, it should be noted that the specification of the data model is not necessarily well chosen by the bootstrap DGP, since the real DGP (the one of Nature) is not known. The orders of the model have to be chosen *a priori* (the ARMA(1,1) model is often chosen), or estimated by any methods (AIC or BIC criteria or by inference tests).

Figure 8: Coverages plots for AR(1) processes, $T = 16$

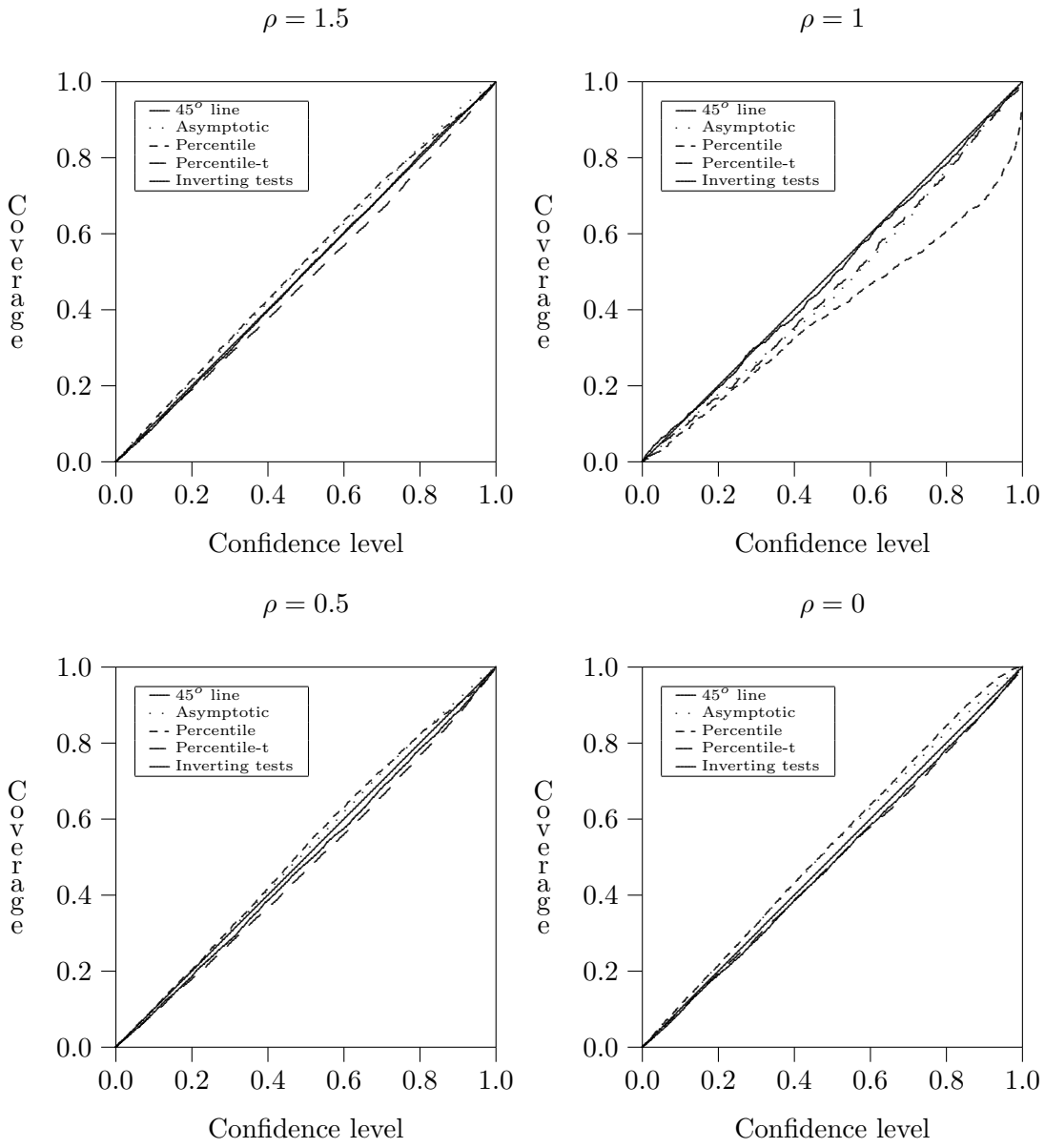



Table 2: Coverages in the case of an Gaussian AR(1) process with $T = 16$, and $\rho_0 = 0.5$

| | Confidence levels | | | |
|---|---|---|---|---|
| Methods | 75% | 90% | 95% | 99% |
| Asymptotic | 0.7729 | 0.9142 | 0.9598 | 0.9926 |
| Percentile | 0.7769 | 0.9016 | 0.9403 | 0.9846 |
| Percentile-t | 0.7140 | 0.8734 | 0.9334 | 0.9857 |
| Inverting tests | 0.7319 | 0.8808 | 0.9385 | 0.9880 |

Since the methods perform correctly, distinguishing the performance of the methods is difficult. Consequently, the coverage discrepancy plots are resorted to examining the performance of the methods.

## 4.2 Coverage discrepancy plots

Figure 9 presents the coverage discrepancy plots of the methods.

Figure 9 shows clearly that the confidence interval based on inverting bootstrap tests is by far the best method on the basis of coverage accuracy criterion in all the situations. Percentile and percentile-t methods have large coverage distortions, that are as large as for the asymptotic confidence interval. It is not surprising that the percentile method does not perform correctly for estimating the autoregressive parameter, since the percentile method does not account for correct the estimate bias. What is more surprising is the unsatisfactory result for the percentile-t method that does normally correct the estimate 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. This asymptotic standard error is only a constant depending on the sample size. 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 other methods.

## 4.3 Confidence level-effectiveness curves

Figure 10 presents the confidence level-effectiveness curves of the confidence intervals. Here, the effectiveness criterion is the average length that should be as small as possible.

On the basis of confidence level-effectiveness curves, in the case of $\rho = 1.5$, the percentile and the percentile-t methods seem to present the smallest average length. In the case of $\rho = 1$, it is the asymptotic and the percentile-t methods that seem to perform better. In the cases of $\rho = 0.5$ and $\rho = 0$, the percentile method presents the best result. However, these results can be spurious, since a negative distortion in the coverage of a method induces a smaller average length of the corresponding confidence interval. For examining the "true" effectiveness of the methods, the coverage-effectiveness curves are used.

## 4.4 Coverage-effectiveness curves

Figure 11 presents the coverage-effectiveness curves for the confidence intervals obtained with the same methods. Following the results provided by Figure 11, the percentile method has the most satisfactory "true" effectiveness in cases of $\rho = 1.5$ and $\rho = 0$, followed by the inverting tests method. In the cases of $\rho = 1$ and $\rho = 0.5$, the inverting tests method presents the best effectiveness.

However, the difference between the average length of the various methods is not very large. And since the percentile method cannot be retained because of its large coverage distortion, the inverting tests method should be retained, on the double basis of the coverage and the average length criteria.

Figure 9: Coverage discrepancy plots for AR(1) processes, $T = 16$

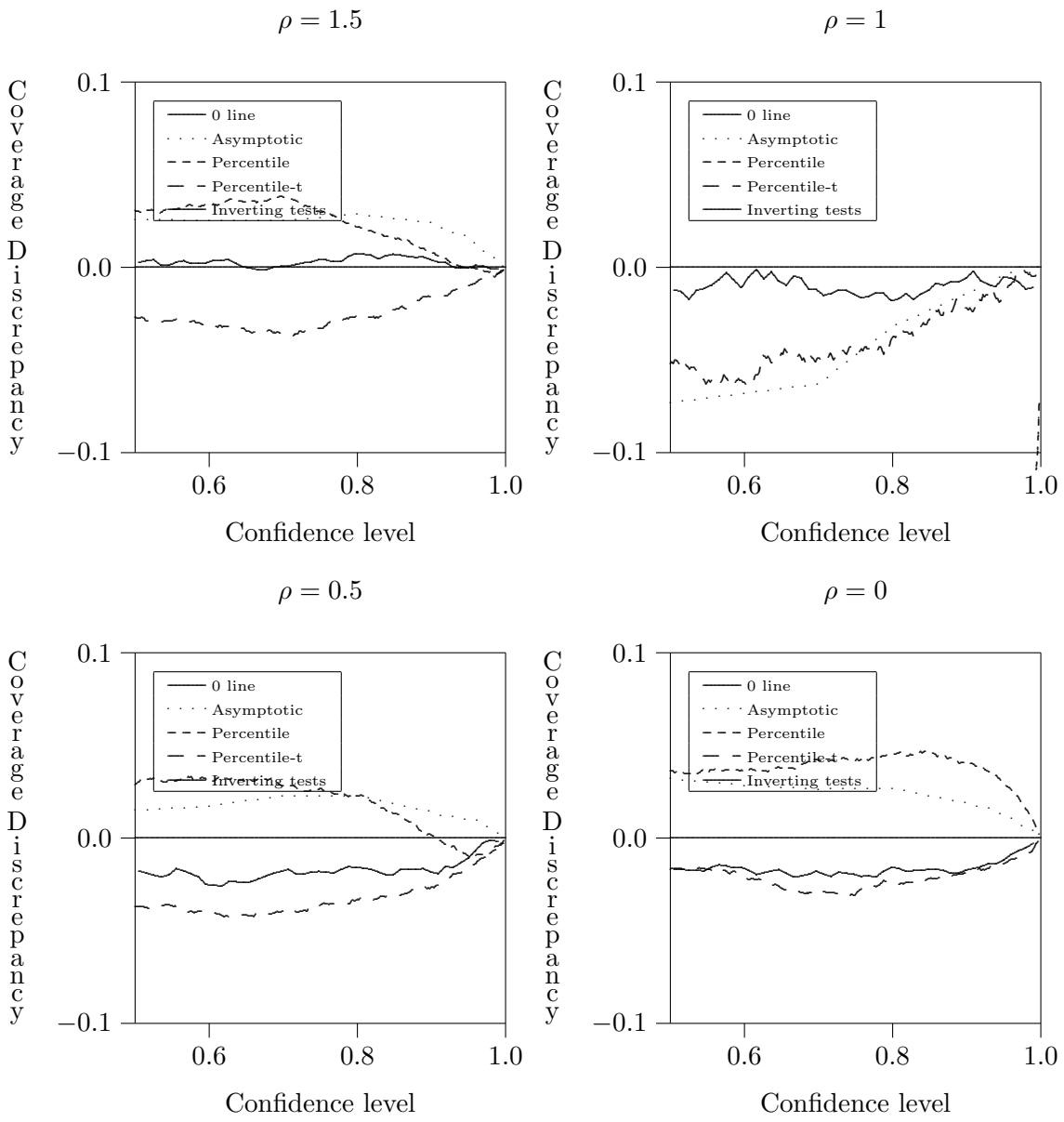


Figure 10: Confidence level-effectiveness curves for AR(1) processes, $T = 16$

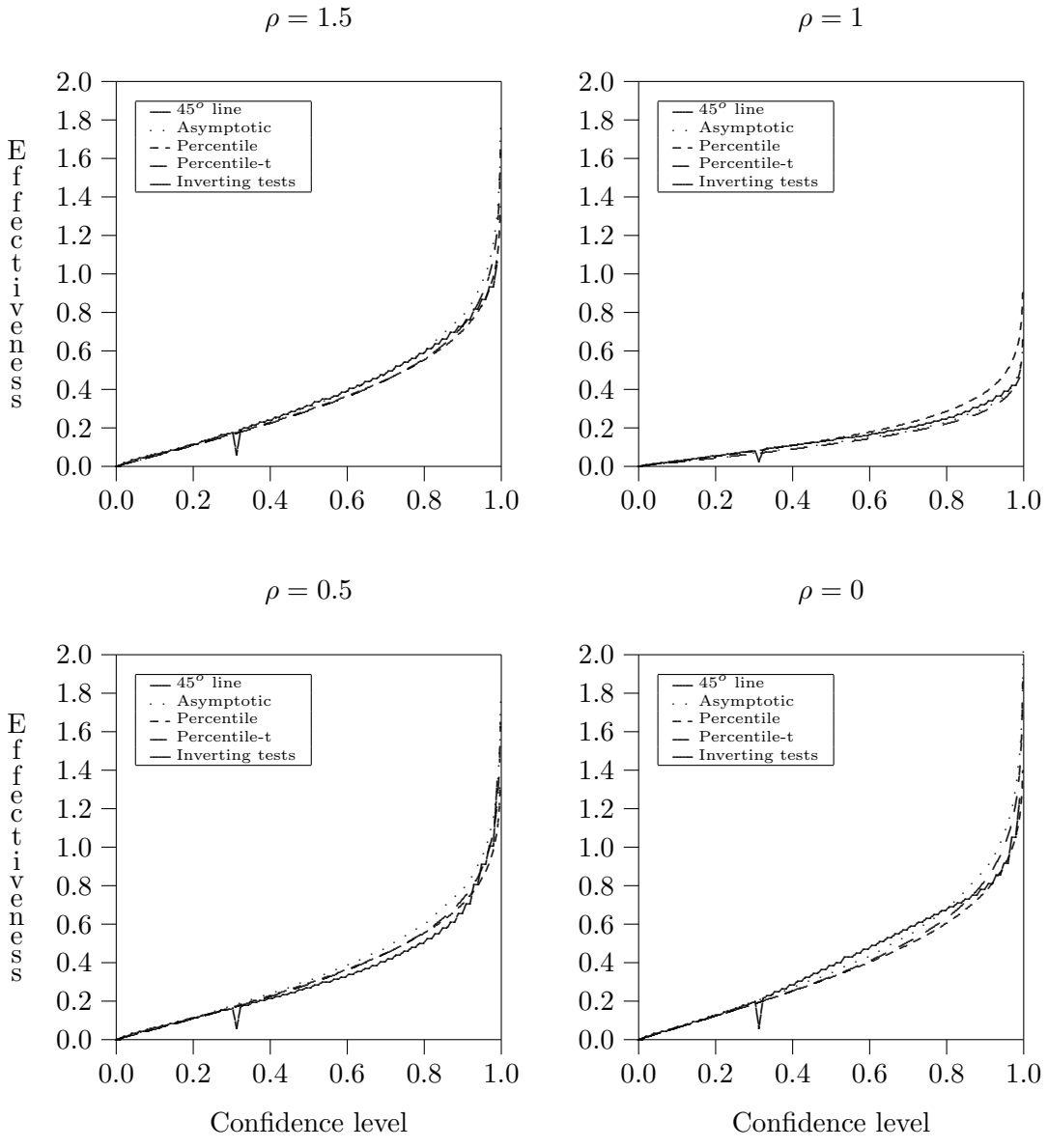


Figure 11: Coverage-effectiveness curves for AR(1) processes, $T = 16$

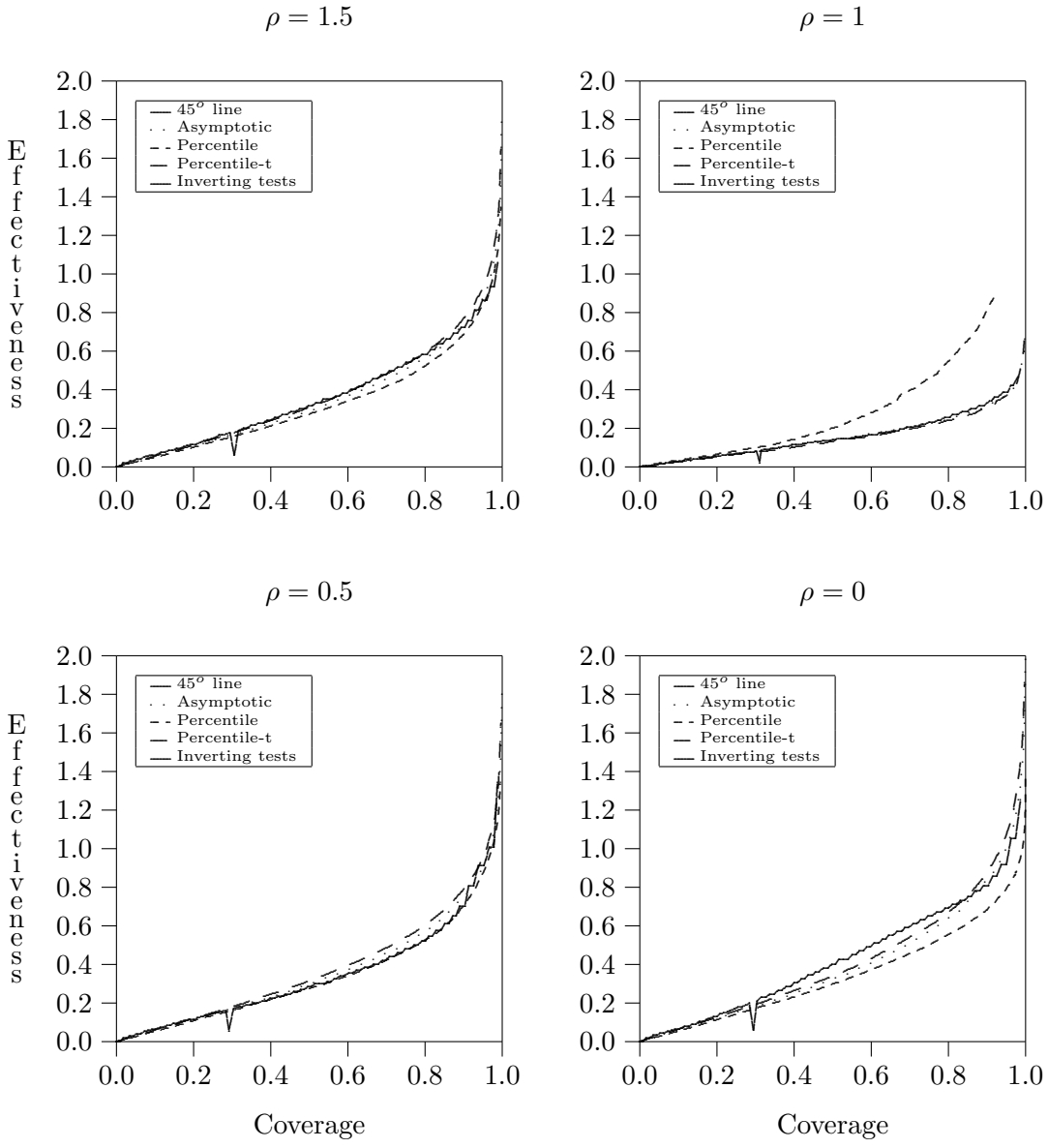
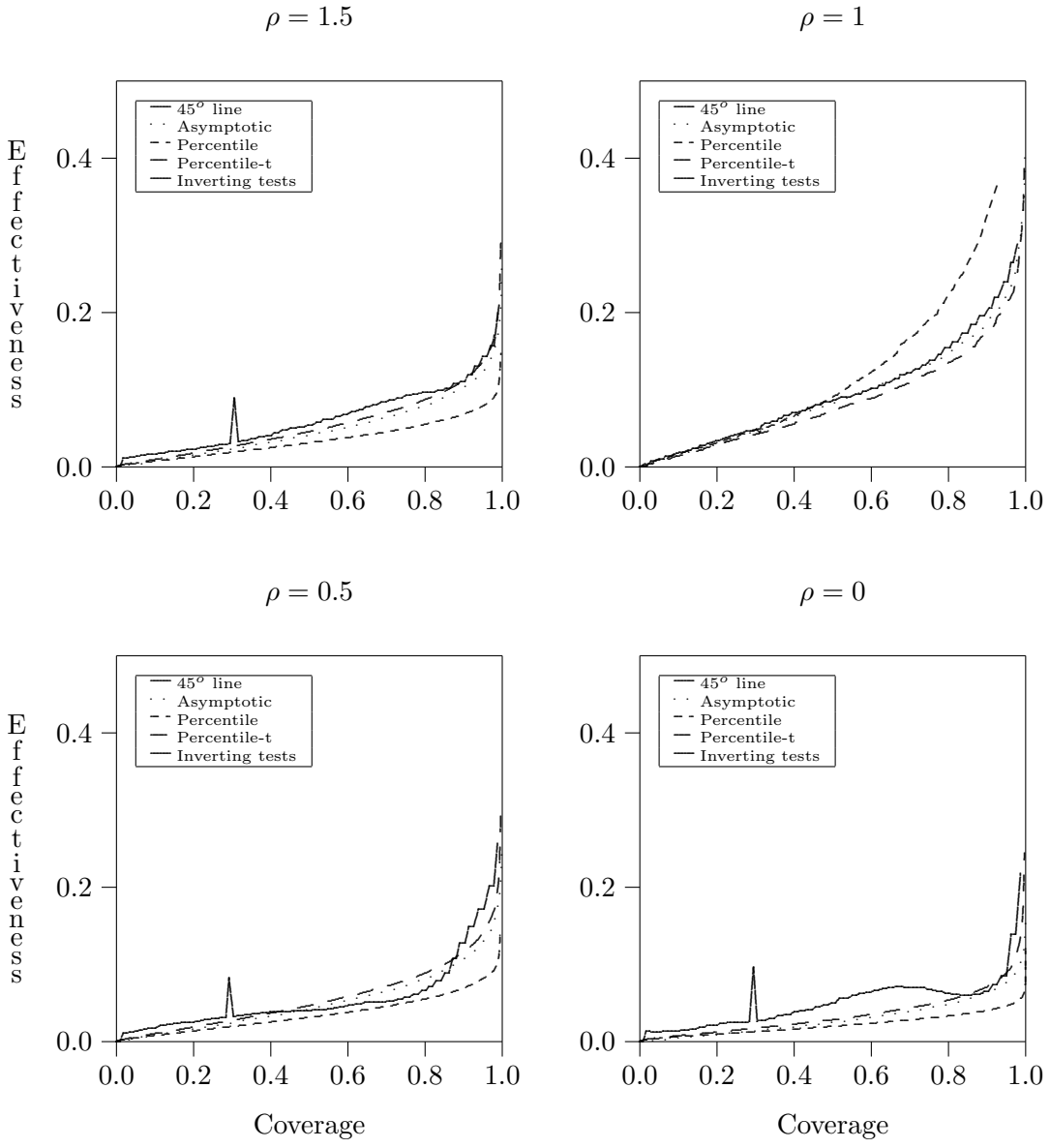


Figure 12: Coverage-'Standard error of the length' curves for AR(1) processes, $T = 16$



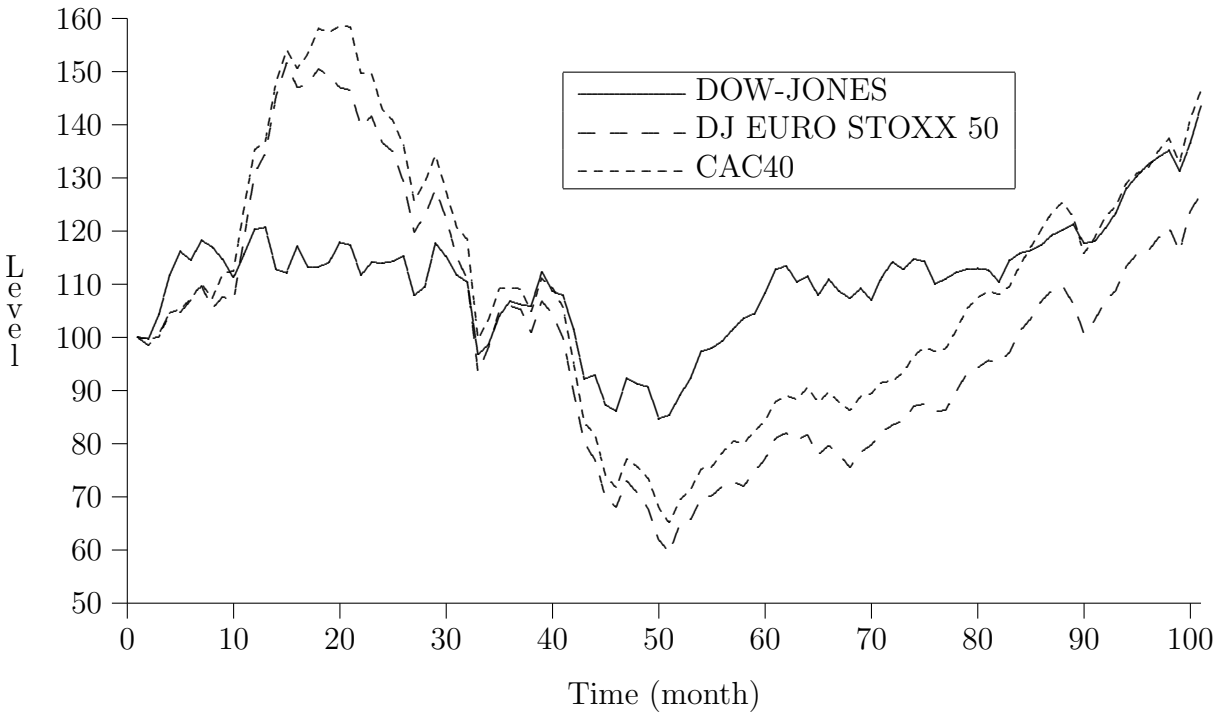
### 4.5 Secondary coverage-effectiveness curves

In addition, the standard errors of the lengths of the confidence intervals are plotted 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. Here, the effectiveness criterion is the standard error of the length. The results are presented in Figure 12. The percentile method displays the smallest standard error of the length except in the case of $\rho = 1$ where the method has a large average length. But again, the difference of standard error between the methods is not very large.

## 5 Application to Empirical Data

As an illustration of the results presented in section 4, the various confidence intervals are applied to some stock market indices. The series used in this section come from the French *National Institute of Statistics and Economic Studies* (INSEE) [10]. The indices are DOW-JONES, DJ EURO STOXX 50, and CAC40. The series are monthly averages from January 1999 to May 2007. A 100 basis at January 1999 is used. Low frequency observations are used to make possible the detection of some tendencies in the data. Indeed, because of the financial markets efficiency, there is no short run tendency in the series, and assessing the autocorrelation coefficient in high frequency financial data does not permit any forecasting of the indices. The series are presented in Figure 13.

Figure 13: Stock market indices




Since the series are integrated of order 1, for detecting any tendencies in these data,

[10] The data are available on the web site of the INSEE: http://www.insee.fr/.

their log returns are studied. The log returns are computed as follows:

$$r_t = \ln\left(\frac{p_t}{p_{t-1}}\right),$$

where $p_t$ is the level of the index at time $t$. A preliminary study of the data is made by regressing the returns on the constant term and four lags, using Gauss 7.0 software. The results are presented in Tables 3–5.

Table 3: DOW-JONES index returns autoregression

| Variable | Estimate | Standard Error | $t$-value | Prob $>\|t\|$ | Standardized Estimate | Cor with Dep Var |
|---|---|---|---|---|---|---|
| CONSTANT | 0.002696 | 0.003508 | 0.768518 | 0.444 | — | — |
| Lag 1 | 0.085151 | 0.103423 | 0.823320 | 0.412 | 0.084785 | 0.063579 |
| Lag 2 | -0.154826 | 0.102694 | -1.507650 | 0.135 | -0.156269 | -0.125833 |
| Lag 3 | 0.067268 | 0.102605 | 0.655605 | 0.514 | 0.068128 | 0.026161 |
| Lag 4 | -0.169509 | 0.102066 | -1.660786 | 0.100 | -0.171669 | -0.146856 |

Table 4: DJ EURO STOXX 50 index returns autoregression

| Variable | Estimate | Standard Error | $t$-value | Prob $>\|t\|$ | Standardized Estimate | Cor with Dep Var |
|---|---|---|---|---|---|---|
| CONSTANT | 0.001470 | 0.004521 | 0.325085 | 0.746 | — | — |
| Lag 1 | 0.290403 | 0.103951 | 2.793645 | 0.006 | 0.290012 | 0.300998 |
| Lag 2 | 0.037733 | 0.108892 | 0.346514 | 0.730 | 0.037466 | 0.135791 |
| Lag 3 | 0.095115 | 0.110113 | 0.863797 | 0.390 | 0.094143 | 0.106588 |
| Lag 4 | -0.128197 | 0.105552 | -1.214537 | 0.228 | -0.126913 | -0.061578 |

Table 5: CAC40 index returns autoregression

| Variable | Estimate | Standard Error | $t$-value | Prob $>\|t\|$ | Standardized Estimate | Cor with Dep Var |
|---|---|---|---|---|---|---|
| CONSTANT | 0.002273 | 0.004339 | 0.523862 | 0.602 | — | — |
| Lag 1 | 0.260726 | 0.103956 | 2.508050 | 0.014 | 0.259918 | 0.276399 |
| Lag 2 | 0.053758 | 0.106350 | 0.505484 | 0.614 | 0.053373 | 0.153321 |
| Lag 3 | 0.184828 | 0.107351 | 1.721725 | 0.089 | 0.182814 | 0.200278 |
| Lag 4 | -0.139564 | 0.105287 | -1.325561 | 0.188 | -0.137983 | -0.025731 |

The results suggest that DOW-JONES index is a weak white noise process, and that DJ EURO STOXX 50 index and CAC40 index can be modelled by AR(1) processes without constant term. Consequently, the methods for computing the confidence interval for the autoregressive parameter can be applied. The autoregressive parameter on each return series is first computed, and is presented in Table 6.

The asymptotic interval, the percentile interval, the percentile-t interval, and the interval based on inverting tests are computed on each return series. For the interval

Table 6: Autoregressive parameter

| Series | Parameter estimate | Standard error |
|---|---|---|
| DOW-JONES index returns | 0.112 | 0.101 |
| DJ EURO STOXX 50 index returns | 0.302 | 0.096 |
| CAC40 index returns | 0.281 | 0.097 |

based on inverting tests, bootstrapped $t$-tests are used. For the percentile and percentile-t intervals, as well as for the bootstrapped $t$-tests, the number of bootstrap replications is chosen to be equal to 9999, using the same set of random numbers to avoid additional random errors. The results are presented in Tables 7–9.

Table 7: DOW-JONES index returns confidence interval

| Method | lower limit | upper limit |
|---|---|---|
| Asymptotic | -0.087 | 0.311 |
| Percentile | -0.084 | 0.300 |
| Percentile-t | -0.086 | 0.309 |
| Inverting tests | -0.087 | 0.315 |

Table 8: DJ EURO STOXX 50 index returns confidence interval

| Method | lower limit | upper limit |
|---|---|---|
| Asymptotic | 0.113 | 0.491 |
| Percentile | 0.099 | 0.475 |
| Percentile-t | 0.115 | 0.496 |
| Inverting tests | 0.116 | 0.498 |

Table 9: CAC40 index returns confidence interval

| Method | lower limit | upper limit |
|---|---|---|
| Asymptotic | 0.090 | 0.472 |
| Percentile | 0.080 | 0.455 |
| Percentile-t | 0.094 | 0.475 |
| Inverting tests | 0.090 | 0.470 |

The results dealing with the confidence intervals of the autocorrelation coefficient of the return indices confirm that DJ EURO STOXX 50 index and CAC40 index can be forecasted on the long run. However, the set of possible values for the autocorrelation coefficient is large, and consequently large mistakes can be made in the forecasting. In addition, it has to be checked that transaction costs do not cancel any possible speculative gains.

# 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 we named coverage plots, coverage discrepancy plots (which may possibly be smoothed), and coverage effectiveness curves. These techniques can make the principal results of a Monte Carlo experiment immediately obvious, since the results are entirely presented in graphical form, and they provide more information (without loss of computing time) in a more easily assimilable fashion than a classical tabular presentation or *QQ plots* could possibly do. All the graphics are based on the construction of the cumulative distribution function of the (true) coverage associated with some confidence regions and on effectiveness criteria. The cumulative distribution function of the coverage is computed by Monte Carlo experiments. The effectiveness criteria were discussed in detail: they depend on the mathematical purpose, but also on the economical objective. This kind of graphics 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 method, and then to chose the most appropriate.

These techniques were illustrated by presenting the results of a number of experiments concerning autoregressive parameter confidence regions. The results show that percentile and percentile-t methods does not perform correctly, conversely to confidence interval based on inverting bilateral bootstrap tests that displays much less coverage distortion than the previous methods: the coverage is closer to the (nominal) confidence level.

# References


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

Berkowitz and Kilian. Recent developments on bootstrapping time series. *Econometrics Reviews*, 19:49–54, 2000. 6

W. C. Black. The cost-effectiveness plane: a graphic representation of cost-effectiveness. *Medical decision Making*, 10(3):212–215, 1990. 2.7, 5

A. Chesher and R. Spady. Asymptotic expansions of the information matrix test statistic. *Econometrica*, 59(3):787–815, 1991. 2.9

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

R. Davidson. Comments on 'recent developments in bootstrapping time series' by berkowitz and kilian. *Econometrics Reviews*, 19:49–54, 2000. 6

R. Davidson and J. MacKinnon. Graphical methods for investigating the size and the power of hypothesis tests. *The Manchester School*, 66:1–22, 1998. 2.5, 2.9

R. Davidson and J. G. MacKinnon. *Estimation and inference in economics*. Oxford University Press, 1993. New York. 1, 2, 3.2, 3.4, 6, 6, 3.4.2

R. Davidson and J. G. MacKinnon. The size distortion of bootstrap tests. *GREQAM Working Paper No 96A15 and Queen's Institute for Economic Research Discussion Paper No 937*, 1996a. 3, 3.4.2

R. Davidson and J. G. MacKinnon. the power of bootstrap tests. *Queen's University Institute for Economic Research, Discussion Paper 937*, 1996b. 3, 3.4.2

R. Davidson and J. G. MacKinnon. Bootstrap tests: how many bootstraps? *Econometric Reviews*, 19, 2000. 2

R. Davidson and J. G. MacKinnon. Improving the reliability of bootstrap confidence intervals. *Université de Montréal*, Conference Resampling Methods in Econometrics, 2001. 2.7, 3.4

B. Efron. Bootstrap methods: another look at the jacknife. *Annals of Statistics*, 7:1–26, 1979. 3.4.2

B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57, Chapman and Hall, 1993. London. 3.2

E. C. Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B*, 16:175–183, 1954. 5, 2.9

P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, 1992. New York. 3.2

J.S.U. Hjorth. *Computer Intensive Statistical Methods*. Chapman and Hall, 1994. London. 3.2

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

H. Li and G. S. Maddala. Bootstrapping time series models. *Econometric Reviews*, 15: 297–318, 1996. 6

J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer-Verlag, 1995. New York. 3.2

C. Siani and C. de Peretti. Is fieller's method applicable in all the situations ? *Health Economics*, forthcoming, 2004. 5

C. Siani and J. P. Moatti. The handling of uncertainty in economic evaluations of health care strategies. *Revue d'Epidémiologie et de Santé Publique (Frensh)*, 51:255–276, 2003. 2.7, 5, 6

A. A. Stinnet and J. Mullahy. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making*, 18:S68–S80, 1998. 2.9

M. Tambour, N. Zethraeus, and M. Johannesson. A note on confidence intervals in cost-effectiveness analysis. *International Journal of Technology Assessment in Health Care*, 14(3):467–471, 1998. 2.9

N.C. Weber. On resampling techniques for regression models. *Statistics and Probability Letters*, 2:275–278, 1984. 8

M. B. Wild and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 33(1):1–17, 1968. 2.9