The Bootstrap

by

Russell Davidson

Department of Economics and CIREQ
McGill University
Montréal, Québec, Canada
H3A 2T7

AMSE and GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13236 Marseille cedex 02, France

email: **russell.davidson@mcgill.ca**

CREATES, February 2015

Definitions, Concepts, and Notation

The starting point for these definitions is the concept of **DGP**, by which is now meant a **unique recipe for simulation**. A DGP generates the virtual reality that our models use as a mirror for the reality of the economic phenomena we wish to study.

A **model** is a collection of DGPs. We need a model before embarking on any statistical enterprise. This is starkly illustrated by the theorems of Bahadur and Savage (1956). We must impose some constraints on the sorts of DGP we can envisage before any valid statistical conclusions can be drawn. Let $\mathbb{M}$ denote a model. Then $\mathbb{M}$ may represent a **hypothesis**. The hypothesis is that the true DGP, μ say, belongs to $\mathbb{M}$. Alternatively, we say that $\mathbb{M}$ is **correctly specified**.

Next, we almost always want to define a **parameter-defining mapping** θ. This maps the model $\mathbb{M}$ into a **parameter space** Θ, which is usually a subset of $\mathbb{R}^k$ for some finite positive integer k. For any DGP $\mu \in \mathbb{M}$, the k–vector $\theta(\mu)$, or θ_μ, is the **parameter vector** that corresponds to μ. Sometimes the mapping θ is one-one. This is the case with models estimated by maximum likelihood. More often, θ is many-one, so that a given parameter vector does not uniquely specify a DGP. Supposing the existence of θ implies that no **identification problems** remain to be solved.

In principle, a DGP specifies the probabilistic behaviour of all deterministic functions of the random data it generates – estimators, standard errors, test statistics, *etc*. If $\boldsymbol{y}$ denotes a data set, or **sample**, generated by a DGP μ, then a statistic $\tau(\boldsymbol{y})$ is a realisation of a random variable τ of which the distribution is determined by μ. A statistic τ is a **pivot**, or is **pivotal**, relative to a model $\mathbb{M}$ if its distribution under DGP $\mu \in \mathbb{M}$, $\mathcal{L}_\mu(\tau)$ say, is the same for all $\mu \in \mathbb{M}$.

We denote by $\mathbb{M}_0$ the set of DGPs that represents a null hypothesis we wish to test. The test statistic used is denoted by τ. Unless τ is a pivot with respect to $\mathbb{M}_0$, it has a different distribution under the different DGPs in $\mathbb{M}_0$, and it certainly has a different distribution under DGPs in the model, $\mathbb{M}$ say, that represents the alternative hypothesis. Here $\mathbb{M}_0 \subset \mathbb{M}$. It is conventional to suppose that τ is defined as a random variable on some suitable probability space, on which we define a different probability measure for each different DGP.

Rather than using this approach, we define a probability space $(\Omega, \mathcal{F}, P)$, with just one probability measure, P. Then we treat the test statistic τ as a stochastic process with as index set the set $\mathbb{M}$. We have

$$\tau \; : \; \mathbb{M} \times \Omega \to \mathbb{R}.$$

Since most of the discussion of these notes is couched in the language of simulation, the probability space can, for our present purposes, be taken to be that of a random number generator. A realisation of the test statistic is therefore written as $\tau(\mu, \omega)$, for some $\mu \in \mathbb{M}$ and $\omega \in \Omega$. Throughout the following discussion, we suppose that, under any DGP μ that we may consider, the distribution of the random variable $\tau(\mu, \cdot)$ is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$.

For notational convenience, we suppose that the range of τ is the $[0, 1]$ interval rather than the whole real line, and that the statistic takes the form of an approximate P value, which thus leads to rejection when the statistic is too small. Let $R : [0, 1] \times \mathbb{M}_0 \to [0, 1]$ be the CDF of τ under any DGP $\mu \in \mathbb{M}_0$:

$$R(x, \mu) = P\big\{\omega \in \Omega \,|\, \tau(\mu, \omega) \leq x\big\}. \tag{1}$$

Suppose that we have a statistic computed from a data set that may or may not have been generated by a DGP $\mu_0 \in \mathbb{M}_0$. Denote this statistic as t. Then the ideal P value that would give exact inference is $R(t, \mu_0)$. If t is indeed generated by μ_0, $R(t, \mu_0)$ is distributed as U(0,1), but not, in general, if t comes from some other DGP. This statistic is available by simulation only if τ is a pivot with respect to $\mathbb{M}_0$, since then we need not know the precise DGP μ_0. When it is available, it permits exact inference.

If τ is not pivotal with respect to $\mathbb{M}_0$, exact inference is no longer possible. If the DGP that generated t, μ_0 say, belongs to $\mathbb{M}_0$, then $R(t, \mu_0)$ is U(0,1). But this fact cannot be used for inference, since μ_0 is unknown.

The principle of the bootstrap is that, when we want to use some function or functional of an unknown DGP μ_0, we use an estimate in its place. Analogously to the stochastic process τ, we define the DGP-valued process

$$\beta \ : \ \mathbb{M} \times \Omega \to \mathbb{M}_0.$$

The estimate of μ_0, which we call the **bootstrap DGP**, is $\beta(\mu, \omega)$, where ω is the *same* realisation as in $t = \tau(\mu, \omega)$. We write $b = \beta(\mu, \omega)$. Then the bootstrap statistic that follows the U(0,1) distribution *approximately* is $R(t, b)$. Normally, the bootstrap principle must be implemented by a simulation experiment. Analogously to (1), we make the definition

$$\hat{R}(x, \mu) = \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}\big(\tau(\mu, \omega_j^*) < x\big), \tag{2}$$

where the ω_j^* are independent realisations of the random numbers needed to compute the statistic. Then, as $B \to \infty$, $\hat{R}(x, \mu)$ tends almost surely to $R(x, \mu)$. Accordingly, we estimate the bootstrap statistic by $\hat{R}(t, b)$.

The bootstrap is a very general statistical technique. The properties of the true unknown DGP that one wants to study are estimated as the corresponding properties of the bootstrap DGP. Thus the bootstrap can be the basis for estimating the bias, the variance, the quantiles, and so on, of an estimator, test statistic, or any other random quantity of interest. Although the bootstrap is most often implemented by simulation, conceptually simulation is not an essential element of the bootstrap.

Asymptotics

So far, the size of the samples generated by μ has not been mentioned explicitly. We denote the **sample size** by n. An **asymptotic theory** is an approximate theory based on the idea of letting n tend to infinity. This is a mathematical abstraction of course. We define an **asymptotic construction** as a construction, in the mathematical sense, of an infinite sequence $\{\mu_n\}$, $n = n_0, \ldots, \infty$, of DGPs such that μ_n generates samples of size n. Such sequences can then be collected into an **asymptotic model**, which we still denote as $\mathbb{M}$, and which can be thought of as a sequence of models $\mathbb{M}_n$.

Statistics, too, can be thought of as sequences $\{\tau_n\}$. The distribution of τ_n under μ_n is written as $\mathcal{L}_\mu^n(\tau)$. If this distribution tends in distribution to a limit $\mathcal{L}_\mu^\infty(\tau)$, then this limit is the **asymptotic** or **limiting distribution** of τ under μ. If $\mathcal{L}_\mu^\infty(\tau)$ is the same for all μ in an asymptotic model $\mathbb{M}$, then τ is an **asymptotic pivot** relative to $\mathbb{M}$.

The vast majority of statistics commonly used in econometrics are asymptotic pivots. t and F statistics, chi-squared statistics, Dickey-Fuller and associated statistics, *etc*, and even the dreaded Durbin-Watson statistic. All that matters is that the limiting distribution does not depend on unknown parameters.

Contrary to what many people have said and thought, the bootstrap is *not* an asymptotic procedure. It is possible to use and study the bootstrap for a fixed sample size without ever considering any other sample size. What is true, though, is that (current) bootstrap *theory* is almost all asymptotic.

Monte Carlo Tests

The simplest type of bootstrap test, and the only type that can be exact in finite samples, is called a **Monte Carlo test**. This type of test was first proposed by Dwass (1957). Monte Carlo tests are available whenever a test statistic is pivotal.

Suppose that we wish to test a null hypothesis represented by the model $\mathbb{M}_0$. Using real data, we compute a realisation t of a test statistic τ that is pivotal relative to $\mathbb{M}_0$. We then compute B independent bootstrap test statistics $\tau_j^* = \tau(\mu, \omega_j^*)$, $j = 1, \ldots, B$, using data simulated using *any* DGP $\mu \in \mathbb{M}_0$. Since τ is a pivot, it follows that the τ_j^* and t are independent drawings from one and the same distribution, *provided* that the true DGP, the one that generated t, also satisfies the null hypothesis.

The **empirical distribution function (EDF)** of the bootstrap statistics can be written as

$$\hat{F}(x) = \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}(\tau_j^* \leq x),$$

where $\mathrm{I}(\cdot)$ is the **indicator function**, with value 1 when its argument is true and 0 otherwise.

Imagine that we wish to perform a test at significance level α, where α might, for example, be .05 or .01, and reject the null hypothesis when the value of t is too small. Given the actual and simulated test statistics, we can compute a **bootstrap *P* value** as

$$\hat{p} = \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}(\tau_j^* < t).$$

Evidently, $\hat{p}$ is just the fraction of the bootstrap samples for which τ_j^* is smaller than t. If this fraction is smaller than α, we reject the null hypothesis. This makes sense, since t is extreme relative to the empirical distribution of the τ_j^* when $\hat{p}$ is small.

Now suppose that we sort the original test statistic t and the B bootstrap statistics τ_j^* in increasing order. Define the rank r of t in the sorted set in such a way that there are exactly r simulations for which $\tau_j^* < t$. Then r can have $B + 1$ possible values, $r = 0, 1, \ldots, B$, all of them equally likely under the null. The estimated P value $\hat{p}$ is then just r/B.

The Monte Carlo test rejects if $r/B < \alpha$, that is, if $r < \alpha B$. Under the null, the probability that this inequality is satisfied is the proportion of the $B+1$ possible values of r that satisfy it. If we denote by $\lfloor \alpha B \rfloor$ the largest integer that is no greater than αB, then, assuming that αB is not an integer, there are exactly $\lfloor \alpha B \rfloor + 1$ such values of r, namely, $0, 1, \ldots, \lfloor \alpha B \rfloor$. Thus the probability of rejection is $(\lfloor \alpha B \rfloor + 1)/(B+1)$. We want this probability to be exactly equal to α. For that to be true, we require that

$$\alpha(B+1) = \lfloor \alpha B \rfloor + 1.$$

Since the right-hand side above is the sum of two integers, this equality can hold only if $\alpha(B+1)$ is also an integer. In fact, it is easy to see that the equation holds whenever $\alpha(B+1)$ is an integer. Suppose that $\alpha(B+1) = k$, k an integer. Then $\lfloor \alpha B \rfloor = k - 1$, and so

$$\Pr(r < \alpha B) = \frac{k-1+1}{B+1} = \frac{k}{B+1} = \frac{\alpha(B+1)}{B+1} = \alpha.$$

In that case, therefore, the rejection probability under the null, that is, the probability of Type I error, is precisely α, the desired significance level.

Of course, using simulation injects randomness into this test procedure, and the cost of this randomness is a loss of power. A test based on $B = 99$ simulations will be less powerful than a test based on $B = 199$, which in turn will be less powerful than one based on $B = 299$, and so on; see Jöckel (1986) and Davidson and MacKinnon (2000). Notice that all of these values of B have the property that $\alpha(B+1)$ is an integer whenever α is an integer percentage like .01, .05, or .10.

Examples

Exact pivots, as opposed to asymptotic pivots, can be hard to find. They exist with the **classical normal linear model**, but most, like t and F tests, have distributions that are known analytically, and so neither bootstrapping nor simulation is necessary. But there exist a few cases in which there is a pivotal statistic of which the distribution under the null is unknown or else intractable.

Consider the classical normal linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2),$$

The $1 \times k$ vector of regressors $\boldsymbol{X}_t$ is the t^{th} row of the $n \times k$ matrix $\boldsymbol{X}$. $\boldsymbol{X}$ is treated as a *fixed* property of the null model. Thus every DGP belonging to this model is completely characterised by the values of the parameter vector $\boldsymbol{\beta}$ and the variance σ^2. Any test statistic the distribution of which does not depend on these values is a pivot for the null model. In particular, a statistic that depends on $\boldsymbol{y}$ only through the OLS residuals and is invariant to the scale of $\boldsymbol{y}$ is pivotal.

The first example is the Durbin-Watson test for serial correlation. ('Nuf said!) A better example is the estimated autoregressive parameter $\hat{\rho}$ that is obtained by regressing the t^{th} residual $\hat{u}_t$ on its predecessor $\hat{u}_{t-1}$. The estimate $\hat{\rho}$ can be used as a test for serial correlation of the disturbances. Evidently,

$$\hat{\rho} = \frac{\sum_{t=2}^{n} \hat{u}_{t-1}\hat{u}_t}{\sum_{t=2}^{n} \hat{u}_{t-1}^2}.$$

Since $\hat{u}_t$ is proportional to σ, there are implicitly two factors of σ in the numerator and two in the denominator. Thus $\hat{\rho}$ is independent of the scale factor σ.

Implementation

Since the bootstrap DGP can be any DGP in the null model, we choose the simplest such DGP, with $\boldsymbol{\beta} = \mathbf{0}$ and $\sigma^2 = 1$. It can be written as

$$y_t^* = u_t^*, \quad u_t^* \sim \text{NID}(0, 1).$$

For each of B bootstrap samples, we then proceed as follows:

1. Generate the vector $\boldsymbol{y}^*$ as an n–vector of IID standard normal variables.
2. Regress $\boldsymbol{y}^*$ on $\boldsymbol{X}$ and save the vector of residuals $\hat{\boldsymbol{u}}^*$.
3. Compute ρ^* by regressing $\hat{u}_t^*$ on $\hat{u}_{t-1}^*$ for observations 2 through n.

Denote by ρ_j^*, $j = 1, \ldots, B$, the bootstrap statistics obtained by performing the above three steps B times. We now have to choose the alternative to our null hypothesis of no serial correlation. If the alternative is positive serial correlation, then we perform a **one-tailed test** by computing the bootstrap P value as

$$\hat{p} = \frac{1}{B} \sum_{j=1}^{B} \text{I}(\rho_j^* > \hat{\rho}).$$

This P value is small when $\hat{\rho}$ is positive and sufficiently large, thereby indicating positive serial correlation.

However, we may wish to test against both positive and negative serial correlation. In that case, there are two possible ways to compute a P value corresponding to a **two-tailed test**. The first is to assume that the distribution of $\hat{\rho}$ is symmetric, in which case we can use the bootstrap P value

$$\hat{p} = \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}\big(|\rho_j^*| > |\hat{\rho}|\big).$$

This is implicitly a symmetric two-tailed test, since we reject when the fraction of the ρ_j^* that exceed $\hat{\rho}$ in absolute value is small. Alternatively, if we do not assume symmetry, we can use

$$\hat{p} = 2 \min\left(\frac{1}{B} \sum_{j=1}^{B} \mathrm{I}(\rho_j^* \leq \hat{\rho}),\ \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}(\rho_j^* > \hat{\rho}) \right).$$

In this case, for level α, we reject whenever $\hat{\rho}$ is either below the $\alpha/2$ quantile or above the $1 - \alpha/2$ quantile of the empirical distribution of the ρ_j^*. Although tests based on these different P values are both exact, they may yield conflicting results, and their power against various alternatives will differ.

The power of a bootstrap test depends on B. If to any test statistic we add random noise independent of the statistic, we inevitably reduce the power of tests based on that statistic. Note that the bootstrap P value $\hat{p}$ is an estimate of the **ideal bootstrap P value**

$$p = \operatorname*{plim}_{B \to \infty} \hat{p},$$

When B is finite, $\hat{p}$ differs from p because of random variation in the bootstrap samples. This random variation is generated in the computer, and is therefore completely independent of the random variable τ. The bootstrap testing procedure incorporates this random variation, and in so doing it reduces the power of the test.

Power loss is illustrated in Figure 1. It shows power functions for four tests at the .05 level of a null hypothesis with only 10 observations. All four tests are exact, as can be seen from the fact that, in all cases, power equals .05 when the null is true. When it is not, there is a clear ordering of the four curves, depending on the number of bootstrap samples used. The loss of power is quite modest when $B = 99$, but it is substantial when $B = 19$.

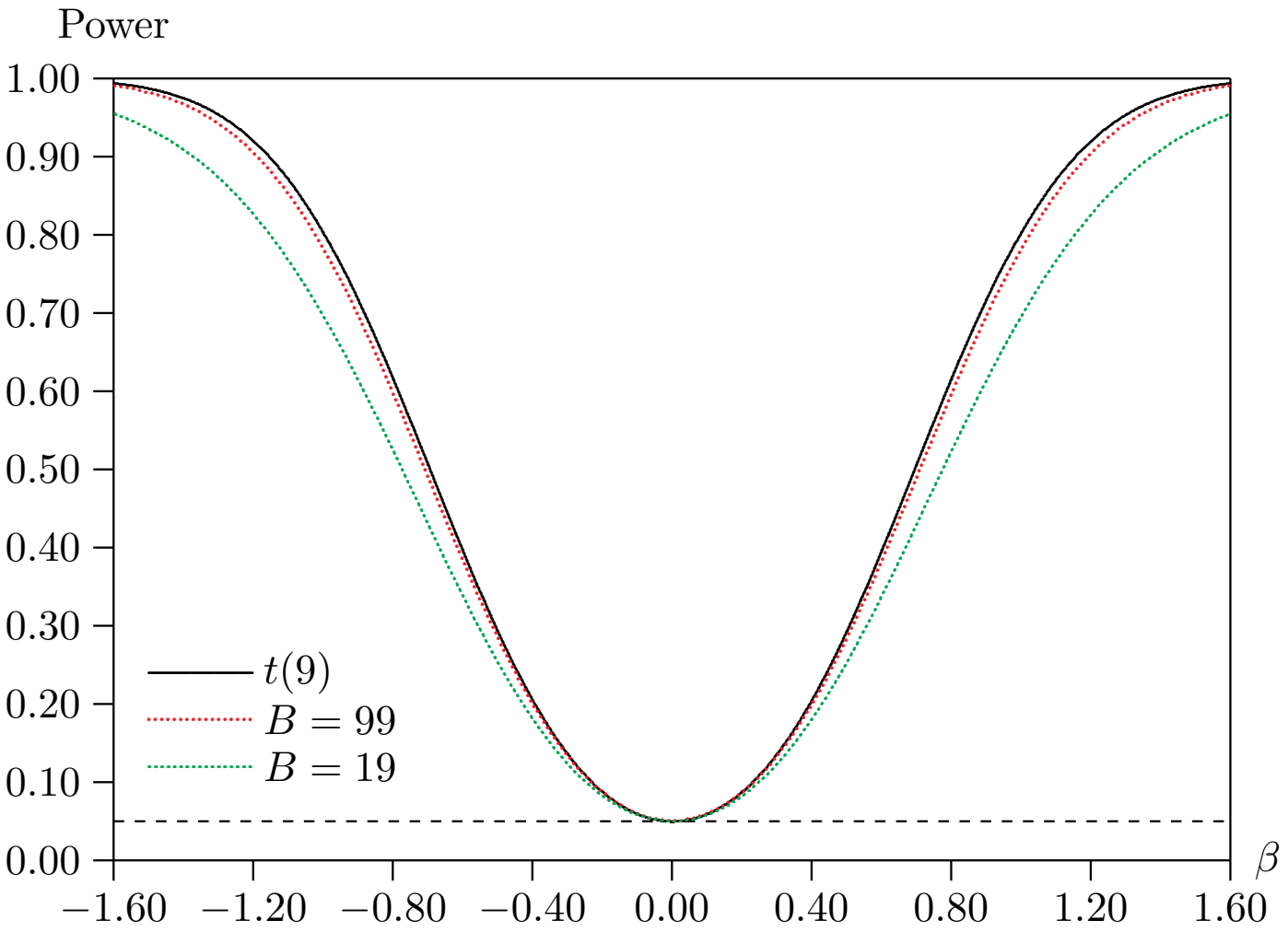


Figure 1: Power loss with finite B.

Many common test statistics for serial correlation, heteroskedasticity, skewness, and excess kurtosis in the classical normal linear regression model are pivotal, since they depend on the regressand only through the least squares residuals $\hat{\boldsymbol{u}}$ in a way that is invariant to the scale factor σ. The Durbin-Watson d statistic is a particularly well-known example. We can perform a Monte Carlo test based on d just as easily as a Monte Carlo test based on $\hat{\rho}$, and the two tests should give very similar results. Since we condition on $\boldsymbol{X}$, the infamous upper and lower bounds from the classic tables of the d statistic are quite unnecessary.

With modern computers and appropriate software, it is extremely easy to perform a variety of exact tests in the context of the classical normal linear regression model. These procedures also work when the disturbances follow a non-normal distribution that is known up to a scale factor; we just have to use the appropriate distribution in step 1 above. For further references and a detailed treatment of Monte Carlo tests for heteroskedasticity, see Dufour, Khalaf, Bernard, and Genest (2004).

The Parametric Bootstrap

If the model $\mathbb{M}_0$ that represents the null hypothesis can be estimated by maximum likelihood (ML), there is a one-one relation between the parameter space of the model and the DGPs that belong to it. For any fixed admissible set of parameters, the likelihood function evaluated at those parameters is a probability density. Thus there is one and only one DGP associated with the set of parameters. By implication, the only DGPs in $\mathbb{M}$ are those completely characterised by a set of parameters.

If the model $\mathbb{M}_0$ actually is estimated by ML, then the ML parameter estimates provide an asymptotically efficient estimate not only of the true parameters themselves, but also of the true DGP. It makes sense therefore if the bootstrap DGP is chosen as the DGP in $\mathbb{M}_0$ characterised by the ML parameter estimates. In this case we speak of a *parametric bootstrap*.

In microeconometrics, models like probit and logit are commonly estimated by ML. These are of course just the simplest of microeconometric models, but they are representative of all the others for which it is reasonable to suppose that the data can be described by a purely parametric model. We use the example of a binary choice model to illustrate the parametric bootstrap.

A binary choice model

Suppose that a binary dependent variable y_t, $t = 1, \ldots, n$, takes on only the values 0 and 1, with the probability that $y_t = 1$ being given by $F(\boldsymbol{X}_t\boldsymbol{\beta})$, where $\boldsymbol{X}_t$ is a $1 \times k$ vector of exogenous explanatory variables, $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameters, and F is a function that maps real numbers into the $[0, 1]$ interval. For probit, F is the CDF of the standard normal distribution; for logit, it is the CDF of the logistic distribution.

The contribution to the loglikelihood for the whole sample made by observation t is

$$\mathrm{I}(y_t = 1) \log F(\boldsymbol{X}_t\boldsymbol{\beta}) + \mathrm{I}(y_t = 0) \log\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big),$$

Suppose now that the parameter vector $\boldsymbol{\beta}$ can be partitioned into two subvectors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, and that, under the null hypothesis, $\boldsymbol{\beta}_2 = \mathbf{0}$. The *restricted* ML estimator, that is, the estimator of the subvector $\boldsymbol{\beta}_1$ only, with $\boldsymbol{\beta}_2$ set to zero, is then an asymptotically efficient estimator of the only parameters that exist under the null hypothesis.

Although asymptotic theory is used to convince us of the desirability of the ML estimator, the bootstrap itself is a purely finite-sample procedure. If we denote the restricted ML estimate as $\tilde{\boldsymbol{\beta}} \equiv [\tilde{\boldsymbol{\beta}}_1 \vdots \mathbf{0}]$, the bootstrap DGP can be represented as follows.

$$y_t^* = \begin{cases} 1 & \text{with probability } F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}}), \text{ and} \\ 0 & \text{with probability } 1 - F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}}). \end{cases}, \qquad t = 1, \dots, n.$$

Here the usual notational convention is followed, according to which variables generated by the bootstrap DGP are starred. Note that the explanatory variables $\boldsymbol{X}_t$ are *not* starred. Since they are assumed to be exogenous, it is not the business of the bootstrap DGP to regenerate them; rather they are thought of as fixed characteristics of the bootstrap DGP, and so are used unchanged in each bootstrap sample. Since the bootstrap samples are exactly the same size, n, as the original sample, there is no need to generate explanatory variables for any more observations than those actually observed.

It is easy to implement the above bootstrap DGP. A **random number** m_t is drawn, using a random number generator, as a drawing from the uniform $\mathrm{U}(0, 1)$ distribution. Then we generate y_t^* as $\mathrm{I}\big(m_t \leq F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})\big)$. Most matrix or econometric software can implement this as a vector relation, so that, after computing the $n-$vector with typical element $F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})$, the vector $\boldsymbol{y}^*$ with typical element y_t^* can be generated by a single command.

Recursive simulation

In dynamic models, the implementation of the bootstrap DGP may require **recursive simulation**. Let us now take as an example the very simple autoregressive time-series model

$$y_t = \alpha + \rho y_{t-1} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \quad t = 2, \dots, n.$$

Here the dependent variable y_t is continuous, unlike the binary dependent variable above. The model parameters are α, ρ, and σ^2. However, even if the values of these parameters are specified, we still do not have a complete characterisation of a DGP. Because the defining relation is a recurrence, it needs a starting value, or initialisation, before it yields a unique solution. Thus, although it is not a parameter in the usual sense, the first observation, y_1, must also be specified in order to complete the specification of the DGP.

ML estimation of the model is the same as estimation by ordinary least squares (OLS) omitting the first observation. If the recurrence represents the null hypothesis, then we would indeed estimate α, ρ, and σ by OLS. If the null hypothesis specifies the value of any one of those parameters, requiring for instance that $\rho = \rho_0$, then we would use OLS to estimate the model in which this restriction is imposed:

$$y_t - \rho_0 y_{t-1} = \alpha + u_t,$$

with the same specification of the disturbances u_t.

The bootstrap DGP is then the DGP contained in the null hypothesis that is characterised by the restricted parameter estimates, and by some suitable choice of the starting value, y_1^*. One way to choose y_1^* is just to set it y_1, the value in the original sample. In most cases, this is the best choice. It restricts the model by fixing the initial value. A bootstrap sample can now be generated recursively, starting with y_2^*. For all $t = 2, \ldots, n$, we have

$$y_t^* = \tilde{\alpha} + \tilde{\rho} y_{t-1}^* + \tilde{\sigma} v_t^*, \quad v_t^* \sim \text{NID}(0, 1).$$

Often, one wants to restrict the possible values of ρ to values strictly between -1 and 1. This restriction makes the series y_t **asymptotically stationary**, by which we mean that, if we generate a very long sample from the recurrence, then towards the end of the sample, the distribution of y_t becomes independent of t, as also the joint distribution of any pair of observations, y_t and y_{t+s}, say. Sometimes it make sense to require that the series y_t should be stationary, and not just asymptotically stationary, so that the distribution of every observation y_t, including the first, is always the same. It is then possible to include the information about the first observation into the ML procedure, and so get a more efficient estimate that incorporates the extra information. For the bootstrap DGP, y_1^* should now be a random drawing from the stationary distribution.

The Golden Rules of Bootstrapping

If a test statistic τ is asymptotically pivotal for a given model $\mathbb{M}$, then its distribution should not vary too much as a function of the specific DGP, μ say, within that model. It is usually possible to show that the distance between the distribution of τ under the DGP μ for sample size n and that for infinite n tends to zero like some negative power of n, commonly $n^{-1/2}$. The concept of "distance" between distributions can be realised in various ways, some ways being more relevant for bootstrap testing than others.

Heuristically speaking, if the distance between the finite-sample distribution for any DGP $\mu \in \mathbb{M}$ and the limiting distribution is of order $n^{-\delta}$ for some $\delta > 0$, then, since the limiting distribution is the same for all $\mu \in \mathbb{M}$, the distance between the finite-sample distributions for two DGPs μ_1 and μ_2 in $\mathbb{M}$ is also of order $n^{-\delta}$. If now the distance between μ_1 and μ_2 is also small, in some sense, say of order $n^{-\varepsilon}$, it should be the case that the distance between the distributions of τ under μ_1 and μ_2 should be of order $n^{-(\delta+\varepsilon)}$.

Arguments of this sort are used to show that the bootstrap can, in favourable circumstances, benefit from **asymptotic refinements**. The form of the argument was given in a well-known paper of Beran (1988). No doubt wisely, Beran limits himself in this paper to the outline of the argument, with no discussion of formal regularity conditions. It remains true today that no really satisfying general theory of bootstrap testing has been found to embody rigorously the simple idea set forth by Beran. Rather, we have numerous piecemeal results that prove the existence of refinements in specific cases, along with other results that show that the bootstrap does not work in other specific cases. Perhaps the most important instance of negative results of this sort, often called **bootstrap failure**, applies to bootstrapping when the true DGP generates data with a heavy-tailed distribution; see Athreya (1987) for the case of infinite variance.

A technique that has been used a good deal in work on asymptotic refinements for the bootstrap is **Edgeworth expansion** of distributions, usually distributions that become standard normal in the limit of infinite sample size. The standard reference to this line of work is Hall (1992), although there is no shortage of more recent work based on Edgeworth expansions. Whereas the technique can lead to useful theoretical insights, it is unfortunately not very useful as a quantitative explanation of the properties of bootstrap tests. In concrete cases, the true finite-sample distribution of a bootstrap P value, as estimated by simulation, can easily be further removed from an Edgeworth approximation to its distribution than from the asymptotic limiting distribution.

Rules for bootstrapping

All these theoretical caveats notwithstanding, experience has shown abundantly that bootstrap tests, in many circumstances of importance for applied econometrics, are much more reliable than tests based on asymptotic theories of one sort or another. The bootstrap DGP will henceforth be denoted by b. Since in testing the bootstrap is used to estimate the distribution of a test statistic under the null hypothesis, the first golden rule of bootstrapping is:

Golden Rule 1:

> The bootstrap DGP b must belong to the model $\mathbb{M}_0$ that represents the null hypothesis.

It is not always possible, or, even if it is, it may be difficult to obey this rule in some cases, as we will see with confidence intervals. In that case, we may use the common technique of changing the null hypothesis so that the bootstrap DGP that is to be used does satisfy it.

If, in violation of this rule, the null hypothesis tested by the bootstrap statistics is not satisfied by the bootstrap DGP, a bootstrap test can be wholly lacking in power. Test power springs from the fact that a statistic has different distributions under the null and the alternative. Bootstrapping under the alternative confuses these different distributions, and so leads to completely unreliable inference, even in the asymptotic limit.

Whereas Golden Rule 1 must be satisfied in order to have an asymptotically justified test, Golden Rule 2 is concerned rather with making the probability of rejecting a true null with a bootstrap test as close as possible to the significance level. It is motivated by the argument of Beran discussed earlier.

Golden Rule 2:

> Unless the test statistic is pivotal for the null model $\mathbb{M}_0$, the bootstrap DGP should be as good an estimate of the true DGP as possible, under the assumption that the true DGP belongs to $\mathbb{M}_0$.

How this second rule can be followed depends very much on the particular test being performed, but quite generally it means that we want the bootstrap DGP to be based on estimates that are *efficient* under the null hypothesis.

Once the sort of bootstrap DGP has been chosen, the procedure for conducting a bootstrap test based on simulated bootstrap samples follows the following pattern.

(i) Compute the test statistic from the original sample; call its realised value t.

(ii) Determine the realisations of all other data-dependent things needed to set up the bootstrap DGP b.

(iii) Generate B bootstrap samples using b, and for each one compute a realisation of the bootstrap statistic, τ_j^*, $j = 1, \dots B$. It is prudent to choose B so that $\alpha(B+1)$ is an integer for all interesting significance levels α, typically 1%, 5%, and 10%.

(iv) Compute the simulated bootstrap P value as the proportion of bootstrap statistics τ_j^* that are more extreme than t. For a statistic that rejects for large values, for instance, we have

$$P_{\text{bs}} = \frac{1}{B} \sum_{j=1}^{B} \text{I}(\tau_j^* > t),$$

where $\text{I}(\cdot)$ is an indicator function, with value 1 if its Boolean argument is true, and 0 if it is false.

The bootstrap test rejects the null hypothesis at significance level α if $P_{\text{bs}} < \alpha$.

Resampling

Our subsequent analysis of the properties of the bootstrap relies on the assumption of the absolute continuity of the distribution of the test statistic for all $\mu \in \mathbb{M}$. Even when a parametric bootstrap is used, absolute continuity does not always pertain. For instance, the dependent variable of a binary choice model is a discrete random variable, and so too are any test statistics that are functions of it. However, since the discrete set of values a test statistic can take on rapidly becomes very rich as sample size increases, it is reasonable to suppose that the theory of the previous section remains a good approximation for realistic sample sizes.

Another important circumstance in which absolute continuity fails is when the bootstrap DGP makes use of **resampling**. Resampling was a key aspect of the original conception of the bootstrap, as set out in Efron's (1979) pioneering paper.

Basic Resampling

Resampling is valuable when it is undesirable to constrain a model so tightly that all of its possibilities are encompassed by the variation of a finite set of parameters. A classic instance is a regression model where one does not wish to impose the normality of the disturbances. To take a concrete example, let us look again at the autoregressive model, relaxing the condition on the disturbances so as to require only IID disturbances with expectation 0 and variance σ^2.

The old bootstrap DGP satisfies Golden Rule 1, because the normal distribution is plainly allowed when all we specify are the first two moments. But Golden Rule 2 incites us to seek as good an estimate as possible of the unknown distribution of the disturbances. If the disturbances were observed, then the best non-parametric estimate of their distribution would be their EDF. The unobserved disturbances can be estimated, or proxied, by the residuals from estimating the null model. If we denote the empirical distribution of these residuals by $\hat{F}$, the bootstrap DGP would be

$$y_t^* = \tilde{\alpha} + \tilde{\rho} y_{t-1}^* + u_t^*, \quad u_t^* \sim \text{IID}(\hat{F}), \quad t = 2, \ldots, n.$$

where the notation indicates that the bootstrap disturbances, the u_t^*, are IID drawings from the empirical distribution characterised by the EDF $\hat{F}$.

The term *resampling* comes from the fact that the easiest way to generate the u_t^* is to sample from the residuals at random with replacement. The residuals are thought of as sampling the true DGP, and so this operation is called "resampling". For each $t = 2, \ldots, n$, one can draw a random number m_t from the U(0, 1) distribution, and then obtain u_t^* by the operations:

$$s = \lfloor 2 + (n-1)m_t \rfloor, \quad u_t^* = \tilde{u}_s,$$

where the notation $\lfloor x \rfloor$ means the greatest integer not greater than x. For m_t close to 0, $s = 2$; for m_t close to 1, $s = n$, and we can see that s is uniformly distributed over the integers $2, \ldots, n$. Setting u_t^* equal to the (restricted) residual $\tilde{u}_s$ therefore implements the required resampling operation.

More sophisticated resampling

But is the empirical distribution of the residuals really the best possible estimate of the distribution of the disturbances? Not always. Consider an even simpler model, one with no constant term:

$$y_t = \rho y_{t-1} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2).$$

When this is estimated by OLS, or, if the null hypothesis fixes the value of ρ, in which case the "residuals" are just the observed values $y_t - \rho_0 y_{t-1}$, the residuals do not in general sum to zero, precisely because there is no constant term. But the model requires that the expectation of the disturbance distribution should be zero, whereas the expectation of the empirical distribution of the residuals is their mean. Thus using this empirical distribution violates Golden Rule 1.

This is easily fixed by replacing the residuals by the deviations from their mean, and then resampling these centred residuals. But now what about Golden Rule 2?

The variance of the centred residuals is the sum of their squares divided by n:

$$V = \frac{1}{n} \sum_{t=1}^{n} (\tilde{u}_t^2 - \bar{u})^2,$$

where $\bar{u}$ is the mean of the uncentred residuals. But the unbiased estimator of the variance of the disturbances is

$$s^2 = \frac{1}{n-1} \sum_{t=1}^{n} (\tilde{u}_t^2 - \bar{u})^2.$$

More generally, in any regression model that uses up k degrees of freedom in estimating regression parameters, the unbiased variance estimate is the sum of squared residuals divided by $n - k$. What this suggests is that what we want to resample is a set of *rescaled* residuals, which here would be the $\sqrt{n/(n-k)}\tilde{u}_t$. The variance of the empirical distribution of these rescaled residuals is then equal to the unbiased variance estimate.

Of course, some problems are scale-invariant. Indeed, test statistics that are ratios are scale invariant for both the autoregressive models we have considered under the stationarity assumption. For models like these, therefore, there is no point in rescaling, since bootstrap statistics computed with the same set of random numbers are unchanged by scaling. This property is akin to pivotalness, in that varying some, but not all, of the parameters of the null model leaves the distribution of the test statistic invariant. In such cases, it is unnecessary to go to the trouble of estimating parameters that have no effect on the distribution of the statistic τ.

Example: A poverty index

In some circumstances, we may wish to affect the values of more complicated functionals of a distribution. Suppose for instance that we wish to perform inference about a poverty index. An IID sample of individual incomes is available, drawn at random from the population under study, and the null hypothesis is that a particular poverty index has a particular given value. For concreteness, let us consider one of the FGT indices, defined as follows; see Foster, Greer, and Thorbecke (1984).

$$\Delta^\alpha(z) = \int_0^z (z-y)^{\alpha-1} \, \mathrm{d}F(y).$$

Here z is interpreted as a poverty line, and F is the CDF of income. We assume that the poverty line z and the parameter α are fixed at some prespecified values. The obvious estimator of $\Delta^\alpha(z)$ is just

$$\hat{\Delta}^\alpha(z) = \int_0^z (z-y)^{\alpha-1} \, \mathrm{d}\hat{F}(y),$$

where $\hat{F}$ is the EDF of income in the sample. For sample size n, we have explicitly that

$$\hat{\Delta}^\alpha(z) = \frac{1}{n} \sum_{i=1}^{n} (z-y_i)_+^{\alpha-1},$$

where y_i is income for observation i, and $(x)_+$ denotes $\max(0, x)$.

Since $\hat{\Delta}^{\alpha}(z)$ is just the mean of a set of IID variables, its variance can be estimated by

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}(z - y_i)_+^{2\alpha-2} - \left(\frac{1}{n}\sum_{i=1}^{n}(z - y_i)_+^{\alpha-1}\right)^2.$$

A suitable test statistic for the hypothesis that $\Delta^{\alpha}(z) = \Delta_0$ is then

$$t = \frac{\hat{\Delta}^{\alpha}(z) - \Delta_0}{\hat{V}^{1/2}}.$$

With probability 1, the estimate $\hat{\Delta}^{\alpha}(z)$ is not equal to Δ_0. If the statistic t is bootstrapped using ordinary resampling of the data in the original sample, this fact means that we violate Golden Rule 1. The simplest way around this difficulty, as mentioned after the statement of Golden Rule 1, is to change the null hypothesis tested by the bootstrap statistics, testing rather what is true under the resampling DGP, namely $\Delta^{\alpha}(z) = \hat{\Delta}^{\alpha}(z)$. Thus each bootstrap statistic takes the form

$$t^* = \frac{(\Delta^{\alpha}(z))^* - \hat{\Delta}^{\alpha}(z)}{(V^*)^{1/2}}.$$

Here $(\Delta^{\alpha}(z))^*$ is the estimate computed using the bootstrap sample, and V^* is the variance estimator computed using the bootstrap sample. Golden Rule 1 is saved by the trick of changing the null hypothesis for the bootstrap samples, but Golden Rule 2 would be better satisfied if we could somehow impose the real null hypothesis on the bootstrap DGP.

Weighted resampling

A way to impose the null hypothesis with a resampling bootstrap is to resample with unequal weights. Ordinary resampling assigns a weight of n^{-1} to each observation, but if different weights are assigned to different observations, it is possible to impose various sorts of restrictions. This approach is suggested by Brown and Newey (2002).

A non-parametric technique that shares many properties with parametric maximum likelihood is **empirical likelihood**; see Owen (2001). In the case of an IID sample, the empirical likelihood is a function of a set of non-negative probabilities p_i, $i = 1, \ldots, n$, such that $\sum_{i=1}^{n} p_i = 1$. The empirical loglikelihood, easier to manipulate than the empirical likelihood itself, is given as

$$\ell(\boldsymbol{p}) = \sum_{i=1}^{n} \log p_i.$$

Here $\boldsymbol{p}$ denotes the n–vector of the probabilities p_i. The idea now is to maximise the empirical likelihood subject to the constraint that the FGT index for the reweighted sample is equal to Δ_0. Specifically, $\ell(\boldsymbol{p})$ is maximised subject to the constraint

$$\sum_{i=1}^{n} p_i (z - y_i)_+^{\alpha-1} = \Delta_0.$$

With very small sample sizes, it is possible that this constrained maximisation problem has no solution with non-negative probabilities. In such a case, the **empirical likelihood ratio** statistic would be set equal to ∞, and the null hypothesis rejected out of hand, with no need for bootstrapping. In the more common case in which the problem can be solved, the bootstrap DGP resamples the original sample with observation i resampled with probability p_i rather than n^{-1}. The use of empirical likelihood for the determination of the p_i means that these probabilities have various optimality properties relative to any other set of probabilities satisfying the desired constraint. Golden Rule 2 is satisfied.

The best algorithm for weighted resampling appears to be little known in the econometrics community. It is described in Knuth (1998). Briefly, for a set of probabilities p_i, $i = 1, \ldots, n$, two tables of n elements each are set up, containing the values q_i, with $0 < q_i \leq 1$, and y_i, where y_i is an integer in the set $1, \ldots, n$. In order to obtain the index j of the observation to be resampled, a random number m_i from $\mathrm{U}(0, 1)$ is used as follows.

$$k_i = \lceil nm_i \rceil, \quad r_i = k_i - nm_i, \quad j = \begin{cases} k_i & \text{if } r_i \leq q_i, \\ y_i & \text{otherwise.} \end{cases}$$

For details, readers are referred to Knuth's treatise.

Confidence Intervals

A confidence interval for some scalar parameter θ consists of all values θ_0 for which the hypothesis $\theta = \theta_0$ cannot be rejected at some specified level α. Thus we can construct a confidence interval by "inverting" a test statistic. If the finite-sample distribution of the test statistic is known, we obtain an **exact confidence interval**. If, as is more commonly the case, only the asymptotic distribution of the test statistic is known, we obtain an **asymptotic confidence interval**, which may or may not be reasonably accurate in finite samples. Whenever a test statistic based on asymptotic theory has poor finite-sample properties, a confidence interval based on that statistic has poor coverage.

To begin with, suppose that we wish to base a confidence interval for the parameter θ on a family of test statistics that have a distribution or asymptotic distribution like the χ^2 or the F distribution under their respective nulls. Statistics of this type are always positive, and tests based on them reject their null hypotheses when the statistics are sufficiently large. Such tests are often equivalent to two-tailed tests based on statistics distributed as standard normal or Student's t. Let us denote the test statistic for the hypothesis that $\theta = \theta_0$ by the random variable $\tau(\theta_0, \boldsymbol{y})$.

For each θ_0, the test consists of comparing the realised $\tau(\theta_0, \boldsymbol{y})$ with the level-α critical value of the distribution of the statistic under the null. If we write the critical value as c_α, then, for any θ_0, we have by the definition of c_α that

$$\Pr{}_{\theta_0}\big(\tau(\theta_0, \boldsymbol{y}) \leq c_\alpha\big) = 1 - \alpha.$$

For θ_0 to belong to the confidence interval obtained by inverting the family of test statistics $\tau(\theta_0, \boldsymbol{y})$, it is necessary and sufficient that

$$\tau(\theta_0, \boldsymbol{y}) \leq c_\alpha.$$

Thus the limits of the confidence interval can be found by solving the equation

$$\tau(\theta, \boldsymbol{y}) = c_\alpha$$

for θ. This equation normally has two solutions. One of these solutions is the upper limit, θ_u, and the other is the lower limit, θ_l, of the confidence interval that we are trying to construct.

A random function $\tau(\theta, \boldsymbol{y})$ is said to be pivotal for $\mathbb{M}$ if, when it is evaluated at the true value θ_μ corresponding to some DGP $\mu \in \mathbb{M}$, the result is a random variable whose distribution does not depend on μ. Pivotal functions of more than one model parameter are defined in exactly the same way. The function is merely asymptotically pivotal if only the asymptotic distribution is invariant to the choice of DGP.

Suppose that $\tau(\theta, \boldsymbol{y})$ is an exactly pivotal function. Then the confidence interval contains the true parameter value θ_μ with probability exactly equal to $1 - \alpha$, whatever the true parameter value may be.

Even if it is not an exact pivot, the function $\tau(\theta, \boldsymbol{y})$ must be asymptotically pivotal, since otherwise the critical value c_α would depend asymptotically on the unknown DGP in $\mathbb{M}$, and we could not construct a confidence interval with the correct coverage, even asymptotically. Of course, if c_α is only approximate, then the coverage of the interval differs from $1 - \alpha$ to a greater or lesser extent, in a manner that, in general, depends on the unknown true DGP.

Asymptotic confidence intervals

To obtain more concrete results, let us suppose that

$$\tau(\theta_0, \boldsymbol{y}) = \big((\hat{\theta} - \theta_0)/s_\theta\big)^2,$$

where $\hat{\theta}$ is an estimate of θ, and s_θ is the corresponding standard error, that is, an estimate of the standard deviation of $\hat{\theta}$. Thus $\tau(\theta_0, \boldsymbol{y})$ is the square of the t statistic for the null hypothesis that $\theta = \theta_0$. The asymptotic critical value c_α is the $1 - \alpha$ quantile of the $\chi^2(1)$ distribution.
The equation for the limits of the confidence interval are

$$\big((\hat{\theta} - \theta)/s_\theta\big)^2 = c_\alpha.$$

Taking the square root of both sides and multiplying by s_θ then gives

$$|\hat{\theta} - \theta| = s_\theta\, c_\alpha^{1/2}.$$

As expected, there are two solutions, namely

$$\theta_l = \hat{\theta} - s_\theta\, c_\alpha^{1/2} \quad \text{and} \quad \theta_u = \hat{\theta} + s_\theta\, c_\alpha^{1/2},$$

and so the asymptotic $1 - \alpha$ confidence interval for θ is

$$\big[\hat{\theta} - s_\theta\, c_\alpha^{1/2},\ \ \hat{\theta} + s_\theta\, c_\alpha^{1/2}\big].$$

We would have obtained the same confidence interval if we had started with the asymptotic t statistic $\tau(\theta_0, \boldsymbol{y}) = (\hat{\theta} - \theta_0)/s_\theta$ and used the $\mathrm{N}(0,1)$ distribution to perform a two-tailed test. For such a test, there are two critical values, one the negative of the other, because the $\mathrm{N}(0,1)$ distribution is symmetric about the origin. These critical values are defined in terms of the quantiles of that distribution. The relevant ones are $z_{\alpha/2}$ and $z_{1-\alpha/2}$, the $\alpha/2$ and the $1-(\alpha/2)$ quantiles, since we wish to have the same probability mass in each tail of the distribution. Note that $z_{\alpha/2}$ is negative, since $\alpha/2 < 1/2$, and the median of the $\mathrm{N}(0,1)$ distribution is 0. By symmetry, it is the negative of $z_{1-(\alpha/2)}$. The equation with two solutions is replaced by two equations, each with just one solution, as follows:

$$\tau(\theta, \boldsymbol{y}) = \pm c.$$

The positive number c can be defined either as $z_{1-(\alpha/2)}$ or as $-z_{\alpha/2}$. The resulting confidence interval $[\theta_l,\ \theta_u]$ can thus be written in two different ways:

$$\big[\hat{\theta} + s_\theta\, z_{\alpha/2},\ \ \hat{\theta} - s_\theta\, z_{\alpha/2}\big] \quad \text{and} \quad \big[\hat{\theta} - s_\theta\, z_{1-(\alpha/2)},\ \ \hat{\theta} + s_\theta\, z_{1-(\alpha/2)}\big].$$

Asymmetric confidence intervals

The confidence intervals so far constructed are **symmetric** about the point estimate $\hat{\theta}$. The symmetry is a consequence of the symmetry of the standard normal distribution and of the form of the test statistic.

It is possible to construct confidence intervals based on two-tailed tests even when the distribution of the test statistic is not symmetric. For a chosen level α, we wish to reject whenever the statistic is too far into either the right-hand or the left-hand tail of the distribution. Unfortunately, there are many ways to interpret "too far" in this context. The simplest is probably to define the rejection region in such a way that there is a probability mass of $\alpha/2$ in each tail. This is called an **equal-tailed confidence interval**. Two critical values are needed for each level, a lower one, c_α^-, which is the $\alpha/2$ quantile of the distribution, and an upper one, c_α^+, which is the $1-(\alpha/2)$ quantile. A realised statistic t leads to rejection at level α if either $t < c_\alpha^-$ or $t > c_\alpha^+$. This leads to an **asymmetric confidence interval**.

If we denote by F the CDF used to calculate critical values or P values, the P value associated with a statistic t should be $2F(t)$ if t is in the lower tail, and $2\big(1-F(t)\big)$ if it is in the upper tail. In complete generality, the P value is

$$p = 2\min\big(F(t), 1-F(t)\big).$$

Consider a one-dimensional test with a rejection region containing a probability mass of α_1 in the left tail of the distribution and α_2 in the right tail, for an overall level of α, where $\alpha = \alpha_1 + \alpha_2$. Let q_{α_1} and $q_{1-\alpha_2}$ be the α_1 and $(1-\alpha_2)$–quantiles of the distribution of $(\hat{\theta} - \theta)/s_\theta$, where θ is the true parameter. Then

$$\Pr(q_{\alpha_1} \leq (\hat{\theta} - \theta)/s_\theta \leq q_{1-\alpha_2}) = 1 - \alpha.$$

The inequalities above are equivalent to

$$\hat{\theta} - s_\theta q_{1-\alpha_2} \leq \theta \leq \hat{\theta} - s_\theta q_{\alpha_1},$$

and from this it is clear that the confidence interval $[\hat{\theta} - s_\theta q_{1-\alpha_2}, \hat{\theta} - s_\theta q_{\alpha_1}]$ contains the true θ with probability α. Note the somewhat counter-intuitive fact that the *upper* quantile of the distribution determines the *lower* limit of the confidence interval, and *vice versa*.

Bootstrap confidence intervals

If $\tau(\theta, \boldsymbol{y})$ is an approximately pivotal function for a model $\mathbb{M}$, its distribution under the DGPs in $\mathbb{M}$ can be approximated by the bootstrap. For each one of a set of bootstrap samples, we compute the parameter estimate, θ^* say, for each of them. Since the true value of θ for the bootstrap DGP is $\hat{\theta}$, we can use the distribution of $\theta^* - \hat{\theta}$ as an estimate of the distribution of $\hat{\theta} - \theta$. In particular, the $\alpha/2$ and $(1-\alpha/2)$–quantiles of the distribution of $\theta^* - \hat{\theta}$, $q^*_{\alpha/2}$ and $q^*_{1-\alpha/2}$ say, give the **percentile confidence interval**

$$C^*_\alpha = [\hat{\theta} - q^*_{1-\alpha/2},\ \hat{\theta} - q^*_{\alpha/2}].$$

For a one-sided confidence interval that is open to the right, we use $[\hat{\theta} - q^*_{1-\alpha},\ \infty[$, and for one that is open to the left $]-\infty,\ \hat{\theta} - q^*_\alpha]$.

The percentile interval is very far from being the best bootstrap confidence interval. The first reason is that, in almost all interesting cases, the random variable $\hat{\theta} - \theta$ is not even approximately pivotal. Indeed, conventional asymptotics give a limiting distribution of $\mathrm{N}(0, \sigma^2_\theta)$, for some asymptotic variance σ^2_θ. Unless σ^2_θ is constant for all DGPs in $\mathbb{M}$, it follows that $\hat{\theta} - \theta$ is not asymptotically pivotal.

For this reason, a more popular bootstrap confidence interval is the **percentile–*t*** interval. Now we suppose that we can estimate the variance of $\hat{\theta}$, and so base the confidence interval on the **studentised** quantity $(\hat{\theta}-\theta)/s_\theta$, which in many circumstances is asymptotically standard normal, and hence asymptotically pivotal. Let $q_{\alpha/2}$ and $q_{1-\alpha/2}$ be the relevant quantiles of the distribution of $(\hat{\theta}-\theta)/\hat{\sigma}_\theta$, when the true parameter is θ. Then

$$\Pr\left(q_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\hat{\sigma}_\theta} \leq q_{1-\alpha/2}\right) = 1-\alpha.$$

If the quantiles are estimated by the quantiles of the distribution of $(\theta^*-\hat{\theta})/\sigma^*_\theta$, where σ^*_θ is the square root of the variance estimate computed using the bootstrap sample, we obtain the percentile–t confidence interval

$$C^*_\alpha = [\hat{\theta} - \hat{\sigma}_\theta q^*_{1-\alpha/2},\ \hat{\theta} - \hat{\sigma}_\theta q^*_{\alpha/2}].$$

In many cases, the performance of the percentile–t interval is much better than that of the percentile interval. For a more complete discussion of bootstrap confidence intervals of this sort, see Hall (1992).

Equal-tailed confidence intervals are not the only ones than can be constructed using the percentile or percentile–t methods. Recall that critical values for tests at level α can be based on the α_1 and $(1-\alpha_2)-$quantiles for the lower and upper critical values provided that $\alpha_1 + \alpha_2 = \alpha$. A bootstrap distribution is rarely symmetric about its central point (unless it is deliberately so constructed). The α_1 and α_2 that minimise the distance between the α_1-quantile and the $(1-\alpha_2)-$quantile under the constraint $\alpha_1 + \alpha_2 = \alpha$ are then not $\alpha/2$ and $1-\alpha/2$ in general. Using the α_1 and α_2 obtained in this way leads to the *shortest* confidence interval at confidence level $1-\alpha$.

The confidence interval takes a simple form only because the test statistic is a simple function of θ. This simplicity may come at a cost, however. The statistic $(\hat{\theta}-\theta)/\hat{\sigma}_\theta$ is a **Wald statistic**, and it is known that Wald statistics may have undesirable properties. The worst of these is that such statistics are not invariant to nonlinear reparametrisations. Tests or confidence intervals based on different parametrisations may lead to conflicting inference. See Gregory and Veall (1985) and Lafontaine and White (1986) for analysis of this phenomenon.

Confidence Intervals that Respect Golden Rule 2

Earlier, we argued against the use of Wald statistics, either for hypothesis testing or for constructing confidence intervals. But even if we use a Lagrange multiplier statistic, based on estimation of the null hypothesis, it can be argued that Golden Rule 2 is still not satisfied. One problem is that, in order to construct a confidence set, it is in principle necessary to consider an infinity of null hypotheses. In practice, provided one is sure that a confidence set is a single, connected, interval, then it is enough to locate the two values of θ that satisfy $\tau(\theta) = q_{1-\alpha}$.

Where Golden Rule 2 is not respected is in the assumption that the distribution of $\tau(\theta)$, under a DGP for which θ is the true parameter, is the same for all θ. If this happens to be the case, the statistic is called pivotal, and there is no further problem. But if the statistic is only approximately pivotal, its distribution when the true θ is an endpoint of a confidence interval is not the same as when the true parameter is the point estimate $\hat{\theta}$. The true parameter for the bootstrap DGP, however, is $\hat{\theta}$.

For Golden Rule 2 to be fully respected, the equation that should be solved for endpoints of the confidence interval is

$$\tau(\theta) = q_{1-\alpha}(\theta),$$

where $q_{1-\alpha}(\theta)$ is the $(1-\alpha)$–quantile of the distribution of $\tau(\theta)$ when θ is the true parameter. If θ is the only parameter, then it is possible, although usually not easy, to solve the equation by numerical methods based on simulation. In general, though, things are even more complicated. If, besides θ, there are other parameters, that we can call nuisance parameters in this context, then according to Golden Rule 2, we should use the best estimate possible of these parameters under the null for the bootstrap DGP. So, for each value of θ considered in a search for the solution of the equation that defines the endpoints, we should re-estimate these nuisance parameters under the constraint that θ is the true parameter, and then base the bootstrap DGP on θ and these restricted estimates. This principle underlies the so-called **grid bootstrap** proposed by Hansen (1999). It is, not surprisingly, very computationally intensive, but Hansen shows that it yields satisfactory results for an autoregressive model where other bootstrap confidence intervals give unreliable inference. Davidson and MacKinnon have recently studied bootstrap confidence intervals in the context of the weak-instrument model we looked at earlier, and find that the computational burden is not excessive, while performance is considerably better than with confidence intervals that use only one bootstrap DGP.

The Bootstrap Discrepancy

Unlike a Monte Carlo test based on an exactly pivotal statistic, a bootstrap test does not in general yield exact inference. This means that there is a difference between the actual probability of rejection and the nominal significance level of the test. We can define the **bootstrap discrepancy** as this difference, as a function of the true DGP and the nominal level. In order to study the bootstrap discrepancy, we suppose, without loss of generality, that the test statistic t is already in approximate P value form. Rejection at level α is thus the event $t < \alpha$.

We introduce two functions of the nominal level α of the test and the DGP μ. The first of these is the **rejection probability function**, or RPF. The value of this function is the true rejection probability under μ of a test at level α, and for some fixed finite sample size n. It is defined as

$$R(\alpha, \mu) \equiv \Pr_\mu(t < \alpha) = P\big(\tau(\mu, \omega) < \alpha\big).$$

Throughout, as mentioned earlier, we assume that, for all $\mu \in \mathbb{M}$, the distribution of $\tau(\mu, \cdot)$ has support $[0, 1]$ and is absolutely continuous with respect to the uniform distribution on that interval.

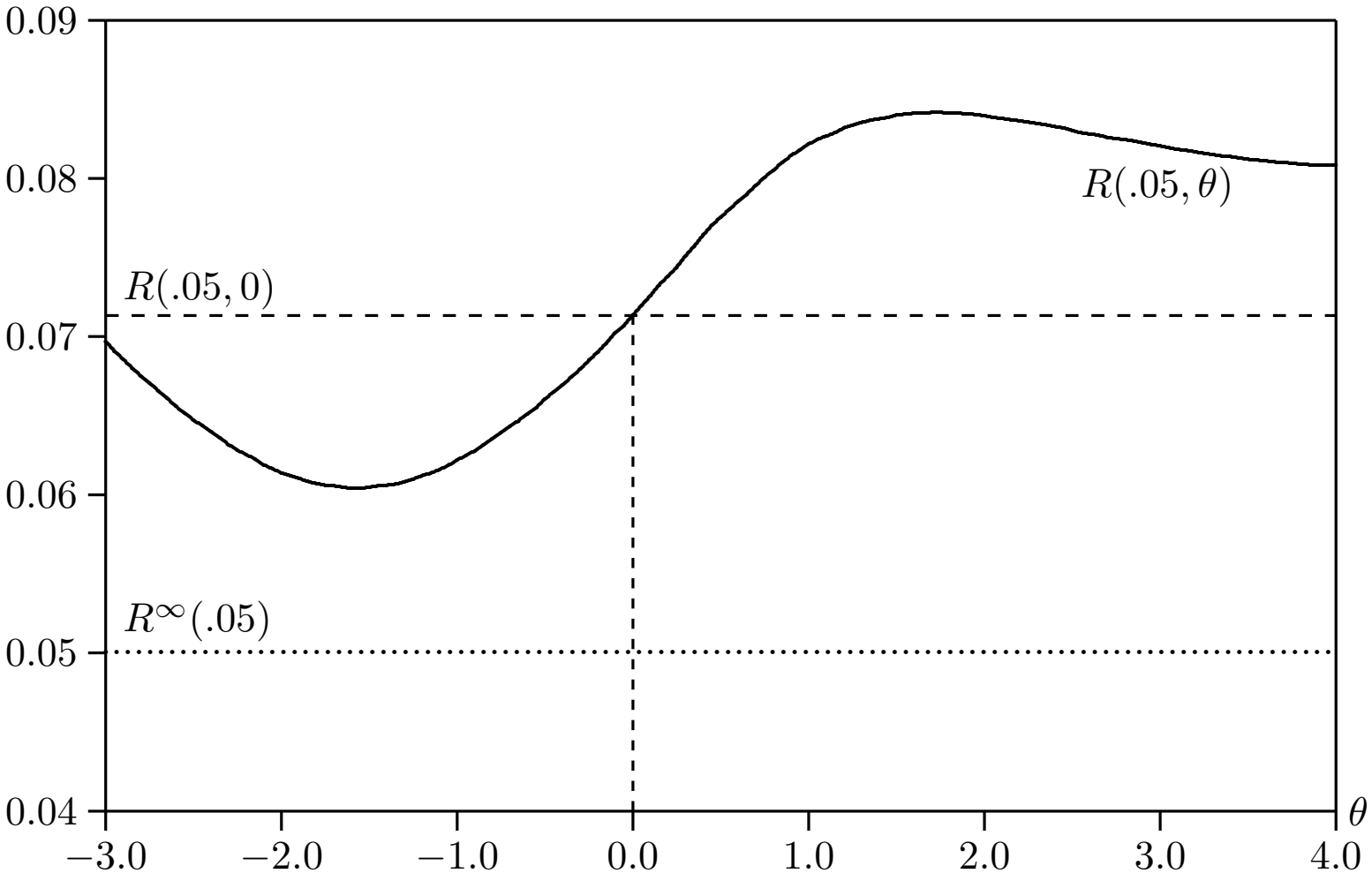


Figure 2: A Rejection Probability Function

For given μ, $R(\alpha, \mu)$ is just the CDF of $\tau(\mu, \omega)$ evaluated at α. The inverse of the RPF is the **critical value function**, or CVF, which is defined implicitly by the equation

$$\Pr_\mu\big(t < Q(\alpha, \mu)\big) = \alpha.$$

It is clear from this that $Q(\alpha, \mu)$ is the α–quantile of the distribution of t under μ. In addition, given that the distributions are continuous, we have that

$$R\big(Q(\alpha, \mu), \mu\big) = Q\big(R(\alpha, \mu), \mu\big) = \alpha$$

for all α and μ.

In what follows, we will abstract from simulation randomness, and assume that the distribution of t under the bootstrap DGP is known exactly. The bootstrap critical value at level α is $Q(\alpha, b)$; recall that b denotes the bootstrap DGP. This is a random variable which would be non-random and equal to α if τ were exactly pivotal. If τ is approximately (for example, asymptotically) pivotal, realisations of $Q(\alpha, b)$ should be close to α. This is true whether or not the true DGP belongs to the null hypothesis, since the bootstrap DGP β does so, according to the first Golden Rule. The bootstrap discrepancy under a DGP $\mu \in \mathbb{M}$ arises from the possibility that, in a finite sample, $Q(\alpha, b) \neq Q(\alpha, \mu)$.

Rejection by the bootstrap test is the event $t < Q(\alpha, b)$. Applying the increasing transformation $R(\cdot, b)$ to both sides, we see that the bootstrap test rejects whenever

$$R(t, b) < R\big(Q(\alpha, b), b\big) = \alpha.$$

Thus the bootstrap P value is just $R(t, b)$. This can be interpreted as a bootstrap test statistic. The probability under μ that the bootstrap test rejects at nominal level α is

$$\Pr{}_\mu\big(t < Q(\alpha, b)\big) = \Pr{}_\mu\big(R(t, b) < \alpha\big).$$

We define two random variables that are deterministic functions of the two random elements, t and b, needed for computing the bootstrap P value $R(t, b)$. The first of these random variables is distributed as $\mathrm{U}(0, 1)$ under μ; it is

$$p \equiv R(t, \mu).$$

The uniform distribution of p follows from the fact that $R(\cdot, \mu)$ is the CDF of t under μ and the assumption that the distribution of t is absolutely continuous on the unit interval for all $\mu \in \mathbb{M}$. The second random variable is

$$r \equiv R(Q(\alpha, b), \mu).$$

We may rewrite the event which leads to rejection by the bootstrap test at level α as $R(t,\mu) < R(Q(\alpha,b),\mu)$, by acting on both sides of the inequality $t < Q(\alpha,b)$ by the increasing function $R(\cdot,\mu)$. This event becomes simply $p < r$. Let the CDF of r under μ conditional on the random variable p be denoted as $F(r\,|\,p)$. Then the probability under μ of rejection by the bootstrap test at level α is

$$\begin{aligned}\mathrm{E}(\mathrm{I}(p<r)) &= \mathrm{E}\big(\mathrm{E}(\mathrm{I}(p<r)\,|\,p)\big) = \mathrm{E}\big(\mathrm{E}(\mathrm{I}(r>p)\,|\,p)\big)\\ &= \mathrm{E}\big(1-F(p\,|\,p)\big) = 1-\int_0^1 F(p\,|\,p)\,\mathrm{d}p,\end{aligned}$$

since the marginal distribution of p is $\mathrm{U}(0,1)$.

A useful expression for the bootstrap discrepancy is obtained by defining the random variable $q \equiv r - \alpha$. The CDF of q conditional on p is then $F(\alpha + q \,|\, p) \equiv G(q \,|\, p)$. The RP minus α is

$$1 - \alpha - \int_0^1 G(p - \alpha \,|\, p)\, \mathrm{d}p.$$

Changing the integration variable from p to $x = p - \alpha$ gives for the bootstrap discrepancy

$$\begin{aligned} & 1 - \alpha - \int_{-\alpha}^{1-\alpha} G(x \,|\, \alpha + x)\, \mathrm{d}x \\ &= 1 - \alpha - \Big[x\, G(x \,|\, \alpha + x)\Big]_{-\alpha}^{1-\alpha} + \int_{-\alpha}^{1-\alpha} x\, \mathrm{d}G(x \,|\, \alpha + x) \\ &= \int_{-\alpha}^{1-\alpha} x\, \mathrm{d}G(x \,|\, \alpha + x), \end{aligned}$$

because $G(-\alpha \,|\, 0) = F(0 \,|\, 0) = 0$ and $G(1 - \alpha \,|\, 1) = F(1 \,|\, 1) = 1$.

To a very high degree of approximation, the expression above can often be replaced by

$$\int_{-\infty}^{\infty} x\, \mathrm{d}G(x \,|\, \alpha), \tag{3}$$

that is, the expectation of q conditional on p being at the margin of rejection at level α. In cases in which p and q are independent or nearly so, it may even be a good approximation just to use the unconditional expectation of q.

The random variable r is the probability that a statistic generated by the DGP μ is less than the $\alpha-$quantile of the bootstrap distribution, conditional on that distribution. The expectation of r minus α can thus be interpreted as the bias in rejection probability when the latter is estimated by the bootstrap. The actual bootstrap discrepancy, which is a non-random quantity, is the expectation of $q = r - \alpha$ conditional on being at the margin of rejection. The approximation (3) sets the margin at the $\alpha-$quantile of τ under the true DGP μ, while the exact expression takes account of the fact that the margin is in fact determined, not by μ, but by the bootstrap DGP b.

If the statistic τ is asymptotically pivotal, the random variable q tends to zero under the null as the sample size n tends to infinity. This follows because, for an asymptotically pivotal statistic, the limiting value of $R(\alpha, \mu)$ for given α is the same for all $\mu \in \mathbb{M}$, and similarly for $Q(\alpha, \mu)$. Let the limiting functions of α alone be denoted by $R^\infty(\alpha)$ and $Q^\infty(\alpha)$. Under the assumption of an absolutely continuous distribution, the functions R^∞ and Q^∞ are inverse functions, and so, as $n \to \infty$, $r = R(Q(\alpha, b), \mu)$ tends to $R^\infty(Q^\infty(\alpha)) = \alpha$, and so $q = r - \alpha$ tends to zero in distribution, and so also in probability.

Suppose now that the random variables q and p are independent. Then the conditional CDF $G(\cdot \mid \cdot)$ is just the unconditional CDF of q, and the bootstrap discrepancy is the unconditional expectation of q. The unconditional expectation of a random variable that tends to 0 can tend to 0 more quickly than the variable itself, and more quickly than the expectation conditional on another variable correlated with it. Independence of q and p does not often arise in practice, but approximate (asymptotic) independence occurs regularly when the parametric bootstrap is used along with ML estimation of the null hypothesis. It is a standard result of the asymptotic theory of maximum likelihood that the ML parameter estimates of a model are asymptotically independent of the classical test statistics used to test the null hypothesis that the model is well specified against some parametric alternative. In such cases, the bootstrap discrepancy tends to zero faster than if inefficient parameter estimates are used to define the bootstrap DGP. This argument, which lends support to Golden Rule 2, is developed in Davidson and MacKinnon (1999).

Approximations to the Bootstrap Discrepancy

It is at this point more convenient to suppose that the statistic is approximately standard normal, rather than uniform on $[0,1]$. The statistic is denoted $\tau_N(\mu,\omega)$ to indicate this. Under the null hypothesis that τ_N is designed to test, we suppose that its distribution admits a valid **Edgeworth expansion**. The expansion takes the form

$$R_N(x,\mu) = \Phi(x) - n^{-1/2}\phi(x)\sum_{i=1}^{\infty} e_i(\mu)\mathrm{He}_{i-1}(x).$$

Here ϕ is the density of the N(0,1) distribution, $\mathrm{He}_i(\cdot)$ is the Hermite polynomial of degree i, and the $e_i(\mu)$ are coefficients that are at most of order 1 as the sample size n tends to infinity. The Edgeworth expansion up to order n^{-1} then truncates everything of order lower than n^{-1}.

The first few Hermite polynomials are $\mathrm{He}_0(x) = 1$, $\mathrm{He}_1(x) = x$, $\mathrm{He}_2(x) = x^2 - 1$, $\mathrm{He}_3(x) = x^3 - 3x$, $\mathrm{He}_4(x) = x^4 - 6x^2 + 3$. The $e_i(\mu)$ can be related to the moments or cumulants of the statistic τ_N as generated by μ by means of the equation

$$n^{-1/2}e_i(\mu) = \frac{1}{i\,!}\mathrm{E}\big(\mathrm{He}_i(\tau_N(\mu,\omega))\big).$$

The bootstrap DGP, $b = \beta(\mu, \omega)$, is realised jointly with $t = \tau_N(\mu, \omega)$, as a function of the same data. We suppose that the CDF of the bootstrap statistic can also be expanded, with the $e_i(\mu)$ replaced by $e_i(b)$, and so the CDF of the bootstrap statistics is $R_N(x, b)$. We consider a one-tailed test based on τ_N that rejects to the left. Then the random variable $p = R_N(t, \mu)$ is approximated by the expression

$$\Phi(t) - n^{-1/2}\phi(t)\sum_{i=1}^{\infty} e_i(\mu)\mathrm{He}_{i-1}(t)$$

truncated so as to remove all terms of order lower than n^{-1}. Similarly, the variable q is approximated by $R'_N(Q_N(\alpha, \mu), \mu)\big(Q_N(\alpha, b) - Q_N(\alpha, \mu)\big)$, using a Taylor expansion where R'_N is the derivative of R_N with respect to its first argument.

It is convenient to replace μ and b as arguments of R_N and Q_N by the sequences $\boldsymbol{e}$ and $\boldsymbol{e}^*$ of which the elements are the $e_i(\mu)$ and $e_i(b)$ respectively. Denote by $\mathbf{D}_e R_N(x, \boldsymbol{e})$ the sequence of partial derivatives of R_N with respect to the components of $\boldsymbol{e}$, and similarly for $\mathbf{D}_e Q_N(\alpha, \boldsymbol{e})$. Then, on differentiating the identity $R_N(Q_N(\alpha, \boldsymbol{e}), \boldsymbol{e}) = \alpha$, we find that

$$R'_N(Q_N(\alpha, \boldsymbol{e}), \boldsymbol{e})\mathbf{D}_e Q_N(\alpha, \boldsymbol{e}) = -\mathbf{D}_e R_N(Q_N(\alpha, \boldsymbol{e}), \boldsymbol{e}).$$

To leading order, $Q_N(\alpha, \boldsymbol{e}^*) - Q_N(\alpha, \boldsymbol{e})$ is $\mathbf{D}_e Q_N(\alpha, \boldsymbol{e})(\boldsymbol{e}^* - \boldsymbol{e})$, where the notation implies a sum over the components of the sequences. Thus the variable q can be approximated by

$$-\mathbf{D}_e R_N(Q_N(\alpha, \boldsymbol{e}), \boldsymbol{e})(\boldsymbol{e}^* - \boldsymbol{e}).$$

The Taylor expansion above is limited to first order, because, in most ordinary cases, $Q_N(\alpha, b) - Q_N(\alpha, \mu)$ is of order n^{-1}. This is true if, as we expect, the $e_i(b)$ are root-n consistent estimators of the $e_i(\mu)$. We see that component i of $\mathbf{D}_e R(x, \boldsymbol{e})$ is $-n^{-1/2}\phi(x)\mathrm{He}_{i-1}(x)$. To leading order, $Q_N(\alpha, \boldsymbol{e})$ is just z_α, the α-quantile of the N(0,1) distribution. Let $l_i = n^{1/2}(e_i(b) - e_i(\mu))$. In regular cases, the l_i are of order 1 and are asymptotically normal. Further, let $\gamma_i(\alpha) = \mathrm{E}(l_i \,|\, p = \alpha)$. Then the approximation of the bootstrap discrepancy at level α is a truncation of

$$n^{-1}\phi(z_\alpha)\sum_{i=1}^{\infty}\mathrm{He}_{i-1}(z_\alpha)\gamma_i(\alpha).$$

The Edgeworth expansion is determined by the coefficients $e_i(\mu)$. These coefficients are enough to determine the first four moments of a statistic τ_N up to the order of some specified negative power of n. Various families of distributions exist for which at least the first four moments can be specified arbitrarily subject to the condition that there exists a distribution with those moments. An example is the **Pearson family** of distributions. A distribution which matches the moments given by the $e_i(\mu)$, truncated at some chosen order, can then be used to approximate the function $R_N(x, \mu)$ for both the DGP μ and its bootstrap counterpart b. An approximation to the bootstrap discrepancy can then formed in the same way as for the Edgeworth expansion, with a different expression for $\mathbf{D}_e R_N(z_\alpha, \boldsymbol{e})$.

Both of these approaches to approximating the bootstrap distribution and the bootstrap discrepancy lead in principle to methods of bootstrapping without simulation. The approximations of the bootstrap distributions have an analytic form, which depends on a small number of parameters, such as the third and fourth moments of the disturbances. These parameters could be estimated directly from the data. Then the analytic approximation to the bootstrap distribution could be used instead of the bootstrap distribution as estimated by simulation. Unfortunately, Edgeworth expansions are often not true CDFs. But it will be interesting to see how well Pearson distributions might work.

Estimating the Bootstrap Discrepancy

Brute force

The conventional way to estimate the bootstrap rejection probability (RP) for a given DGP μ and sample size n by simulation is to generate a large number, M say, of samples of size n using the DGP μ. For each replication, a realisation $t_m \equiv \tau(\mu, \omega_m^*)$ of the statistic is computed from the simulated sample, along with a realisation $b_m \equiv \beta(\mu, \omega_m^*)$ of the bootstrap DGP. Then B bootstrap samples are generated using b_m, and bootstrap statistics τ_{mj}^*, $j = 1, \ldots, B$ are computed. The realised estimated bootstrap P value for replication m is then

$$\hat{p}_m \equiv \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}(\tau_{mj}^* < t_m),$$

where we assume that the rejection region is to the left. The estimate of the RP at nominal level α is the proportion of the $\hat{p}_m$ that are less than α. The whole procedure requires the computation of $M(B+1)$ statistics and M bootstrap DGPs. The bootstrap statistics τ_{mj}^* are realisations of a random variable that we denote as τ^*. As a stochastic process, we write $\tau^* = \tau\big(\beta(\mu, \omega), \omega^*\big)$, which makes it plain that τ^* depends on two (sets of) random numbers; hence the notation τ_{mj}^*.

If one wishes to compare the RP of the bootstrap test with that of the underlying asymptotic test, a simulation estimate of the latter can be obtained directly as the proportion of the t_m less than the asymptotic α level critical value. Of course, estimation of the RP of the asymptotic test by itself requires the computation of only M statistics.

Let us assume that $B = \infty$, and consider the ideal bootstrap P value, that is, the probability mass in the distribution of the bootstrap statistics in the region more extreme than the realisation t of the statistic computed from the real data. For given t and b, we know that the ideal bootstrap P value is $R(t, b)$. Thus, as a stochastic process, the bootstrap P value can be expressed as

$$\begin{aligned} p(\mu, \omega) &= R\big(\tau(\mu, \omega), \beta(\mu, \omega)\big) \\ &= \int_\Omega \mathrm{I}\big(\tau(\beta(\mu, \omega), \omega^*) < \tau(\mu, \omega)\big)\, \mathrm{d}P(\omega^*). \end{aligned}$$

The inequality in the indicator function above can be rewritten as $\tau^* < t$ in more compact, if not completely unambiguous, notation.

We denote the CDF of $p(\mu, \omega)$ by $R_1(x, \mu)$, so that

$$R_1(x, \mu) = \mathrm{E}\Big(\mathrm{I}\big(R(\tau(\mu, \omega), \beta(\mu, \omega)) \leq x\big)\Big).$$

The fast approximation

It is shown in Davidson and MacKinnon (2007) that, under certain conditions, it is possible to obtain a much less expensive approximate estimate of the bootstrap RP, as follows. As before, for $m = 1, \dots, M$, the DGP μ is used to draw realisations t_m and b_m. In addition, b_m is used to draw a *single* bootstrap statistic τ^*_m. The τ^*_m are therefore IID realisations of the variable τ^*. We estimate the RP as the proportion of the t_m that are less than $\hat{Q}^*(\alpha)$, the α quantile of the τ^*_m. This yields the following estimate of the RP of the bootstrap test:

$$\widehat{\mathrm{RP}}_{\mathrm{A}} \equiv \frac{1}{M} \sum_{m=1}^{M} \mathrm{I}\big(t_m < \hat{Q}^*(\alpha)\big),$$

As a function of α, $\widehat{\mathrm{RP}}_{\mathrm{A}}$ is an estimate of the CDF of the bootstrap P value.

The above estimate is approximate not only because it rests on the assumption of the full independence of t and b, but also because its limit as $B \to \infty$ is not precisely the RP of the bootstrap test. Its limit differs from the RP by an amount of a smaller order of magnitude than the difference between the RP and the nominal level α. But it requires the computation of only $2M$ statistics and M bootstrap DGPs.

Conditional on the bootstrap DGP b, the CDF of τ^* evaluated at x is $R(x, b)$. Therefore, if b is generated by the DGP μ, the unconditional CDF of τ^* is

$$R^*(x, \mu) \equiv \mathrm{E}\big(R(x, \beta(\mu, \omega)\big).$$

We denote the α quantile of the distribution of τ^* under μ by $Q^*(\alpha, \mu)$. In the explicit notation used earlier, since $\tau^* = \tau\big(\beta(\mu, \omega), \omega^*\big)$, we see that

$$R^*(x, \mu) = \int_\Omega \int_\Omega \mathrm{I}\Big(\tau\big(\beta(\omega, \mu), \omega^*\big) < x\Big)\, \mathrm{d}P(\omega^*)\, \mathrm{d}P(\omega),$$

and $Q^*(x, \mu)$ is the inverse with respect to x of $R^*(x, \mu)$.

Multivariate models

Some models have more than one endogenous variable, and so, except in a few cases in which we can legitimately condition on some of them, the bootstrap DGP has to be able to generate all of the endogenous variables simultaneously. This is not at all difficult for models such as vector autoregressive (VAR) models. A typical VAR model can be written as

$$\boldsymbol{Y}_t = \sum_{i=1}^{p} \boldsymbol{Y}_{t-i}\boldsymbol{\Pi}_i + \boldsymbol{X}_t\boldsymbol{B} + \boldsymbol{U}_t, \quad t = p+1, \ldots, n.$$

Here $\boldsymbol{Y}_t$ and $\boldsymbol{U}_t$ are $1 \times m$ vectors, the $\boldsymbol{\Pi}_i$ are all $m \times m$ matrices, $\boldsymbol{X}_t$ is a $1 \times k$ vector, and $\boldsymbol{B}$ is a $k \times m$ matrix. The m elements of $\boldsymbol{Y}_t$ are the endogenous variables for observation t. The elements of $\boldsymbol{X}_t$ are exogenous explanatory variables. The vectors $\boldsymbol{U}_t$ have expectation zero, and are usually assumed to be mutually independent, although correlated among themselves; with covariance matrix $\boldsymbol{\Sigma}$.

Among the hypotheses that can be tested in the context of a VAR model are tests for **Granger causality**. The null hypothesis of these tests is Granger *non-causality*, and it imposes zero restrictions on subsets of the elements of the $\boldsymbol{\Pi}_i$. Unrestricted, our VAR model can be efficiently estimated by least squares applied to each equation separately, with the covariance matrix $\boldsymbol{\Sigma}$ estimated by the empirical covariance matrix of the residuals. Subject to restrictions, the model is usually estimated by maximum likelihood under the assumption that the disturbances are jointly normally distributed.

Bootstrap DGPs can be set up for models that impose varying levels of restrictions. In all cases, the $\boldsymbol{\Pi}_i$ matrices, the $\boldsymbol{\Sigma}$ matrix, and the $\boldsymbol{B}$ matrix, if present, should be set equal to their restricted estimates. In all cases, as well, bootstrap samples should be conditioned on the first p observations from the original sample, unless stationarity is assumed, in which case the first p observations of each bootstrap sample should be drawn from the stationary distribution of p contiguous $m-$vectors $\boldsymbol{Y}_t, \ldots, \boldsymbol{Y}_{t+p-1}$. If normal disturbances are assumed, the bootstrap disturbances can be generated as IID drawings from the multivariate $\mathrm{N}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}})$ distribution – one obtains by Cholesky decomposition an $m \times m$ matrix $\boldsymbol{A}$ such that $\boldsymbol{A}\boldsymbol{A}^\top = \tilde{\boldsymbol{\Sigma}}$, and generates $\boldsymbol{U}_t^*$ as $\boldsymbol{A}\boldsymbol{V}_t^*$, where the m elements of $\boldsymbol{V}_t^*$ are IID standard normal. If it is undesirable to assume normality, then the *vectors* of restricted residuals $\tilde{\boldsymbol{U}}_t$ can be resampled. If it is undesirable even to assume that the $\boldsymbol{U}_t$ are IID, a wild bootstrap can be used in which each of the vectors $\tilde{\boldsymbol{U}}_t$ is multiplied by a scalar s_t^*, with the s_t^* IID drawings from a distribution with expectation 0 and variance 1.

Simultaneous equations

Things are a little more complicated with a **simultaneous-equations model**, in which the endogenous variables for a given observation are determined as the solution of a set of simultaneous equations that also involve exogenous explanatory variables. Lags of the endogenous variables can also appear as explanatory variables; they are said to be **predetermined**. If they are present, the bootstrap DGP must rely on recursive simulation.

A simultaneous-equations model can be written as

$$\boldsymbol{Y}_t\boldsymbol{\Gamma} = \boldsymbol{W}_t\boldsymbol{B} + \boldsymbol{U}_t,$$

with $\boldsymbol{Y}_t$ and $\boldsymbol{U}_t$ $1 \times m$ vectors, $\boldsymbol{W}_t$ a $1 \times k$ vector or exogenous or predetermined explanatory variables, $\boldsymbol{\Gamma}$ an $m \times m$ matrix, and $\boldsymbol{B}$ a $k \times m$ matrix.

The above set of equations is called the **structural form** of the model. The **reduced form** is obtained by solving the equations of the structural form to get

$$\boldsymbol{Y}_t = \boldsymbol{W}_t \boldsymbol{B} \boldsymbol{\Gamma}^{-1} + \boldsymbol{V}_t.$$

The reduced form can be estimated unrestricted, using least squares on each equation of the set of equations

$$\boldsymbol{Y}_t = \boldsymbol{W}_t \boldsymbol{\Pi} + \boldsymbol{V}_t$$

separately, with $\boldsymbol{\Pi}$ a $k \times m$ matrix of parameters. Often, however, the structural form is **overidentified**, meaning that restrictions are imposed on the matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$. This is always the case if the null hypothesis imposes such restrictions. Many techniques exist for the restricted estimation of either one of the equivalent structural and reduced-form models. When conventional asymptotic theory is used, asymptotic efficiency is achieved by two techniques, **three-stage least squares** (3SLS), and **full-information maximum likelihood** (FIML). These standard techniques are presented in most econometrics textbooks.

Bootstrap DGPs should in all cases use efficient restricted estimates of the parameters, obtained by 3SLS or FIML, with a slight preference for FIML, which has higher-order optimality properties not shared by 3SLS. Bootstrap disturbances can be generated from the multivariate normal distribution, or by resampling vectors of restricted residuals, or by a wild bootstrap procedure.

A special case involving weak instruments

A very simple model consists of just two equations,

$$\begin{aligned} \boldsymbol{y}_1 &= \beta \boldsymbol{y}_2 + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}_1, \text{ and} \\ \boldsymbol{y}_2 &= \boldsymbol{W}\boldsymbol{\pi} + \boldsymbol{u}_2. \end{aligned}$$

Here $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are n–vectors of observations on endogenous variables, $\boldsymbol{Z}$ is an $n \times k$ matrix of observations on exogenous variables, and $\boldsymbol{W}$ is an $n \times l$ matrix of instruments such that $\mathcal{S}(\boldsymbol{Z}) \subset \mathcal{S}(\boldsymbol{W})$, where the notation $\mathcal{S}(\boldsymbol{A})$ means the linear span of the columns of the matrix $\boldsymbol{A}$. The disturbances are assumed to be serially uncorrelated and homoskedastic. We assume that $l > k$, so that the model is either exactly identified or, more commonly, overidentified.

The parameters of this model are the scalar β, the k–vector $\boldsymbol{\gamma}$, the l–vector $\boldsymbol{\pi}$, and the 2×2 contemporaneous covariance matrix of the disturbances $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$:

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

We wish to test the hypothesis that $\beta = 0$. There is no loss of generality in considering only this null hypothesis, since we could test the hypothesis that $\beta = \beta_0$ for any nonzero β_0 by replacing the left-hand side variable by $\boldsymbol{y}_1 - \beta_0 \boldsymbol{y}_2$.

Since we are not directly interested in the parameters contained in the $l-$vector $\boldsymbol{\pi}$, we may without loss of generality suppose that $\boldsymbol{W} = [\boldsymbol{Z} \;\; \boldsymbol{W}_1]$, with $\boldsymbol{Z}^\top \boldsymbol{W}_1 = \mathbf{O}$. Notice that $\boldsymbol{W}_1$ can easily be constructed by projecting the columns of $\boldsymbol{W}$ that do not belong to $\mathcal{S}(\boldsymbol{Z})$ off $\boldsymbol{Z}$.

We consider three test statistics: an asymptotic t statistic on which we may base a Wald test, the K statistic of Kleibergen (2002) and Moreira (2001), and a likelihood ratio (LR) statistic. The 2SLS (or IV) estimate $\hat{\beta}$, with instruments the columns of $\boldsymbol{W}$, satisfies the estimating equation

$$\boldsymbol{y}_2^\top \boldsymbol{P}_1 (\boldsymbol{y}_1 - \hat{\beta} \boldsymbol{y}_2) = 0,$$

where $\boldsymbol{P}_1 \equiv \boldsymbol{P}_{\boldsymbol{W}_1}$ is the matrix that projects on to $\mathcal{S}(\boldsymbol{W}_1)$. This follows because $\boldsymbol{Z}^\top \boldsymbol{W}_1 = \mathbf{O}$, but the estimating equation would hold even without this assumption if we define $\boldsymbol{P}_1$ as $\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{Z}}$, where the matrices $\boldsymbol{P}_{\boldsymbol{W}}$ and $\boldsymbol{P}_{\boldsymbol{Z}}$ project orthogonally on to $\mathcal{S}(\boldsymbol{W})$ and $\mathcal{S}(\boldsymbol{Z})$, respectively.

It is not hard to see that the asymptotic t statistic for a test of the hypothesis that $\beta = 0$ is

$$t = \frac{n^{1/2} \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_1}{\|\boldsymbol{P}_1 \boldsymbol{y}_2\| \left\| \boldsymbol{M}_{\boldsymbol{Z}} \left(\boldsymbol{y}_1 - \dfrac{\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_1}{\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2} \boldsymbol{y}_2 \right) \right\|},$$

where $\boldsymbol{M}_{\boldsymbol{Z}} \equiv \mathbf{I} - \boldsymbol{P}_{\boldsymbol{Z}}$. It can be seen that the right-hand side above is homogeneous of degree zero with respect to $\boldsymbol{y}_1$ and also with respect to $\boldsymbol{y}_2$. Consequently, the distribution of the statistic is invariant to the scales of each of the endogenous variables. In addition, the expression is unchanged if $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are replaced by the projections $\boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_1$ and $\boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_2$, since $\boldsymbol{P}_1 \boldsymbol{M}_{\boldsymbol{Z}} = \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{P}_1 = \boldsymbol{P}_1$, given the orthogonality of $\boldsymbol{W}_1$ and $\boldsymbol{Z}$. It follows that, if $\boldsymbol{M}_{\boldsymbol{W}} \equiv \mathbf{I} - \boldsymbol{P}_{\boldsymbol{W}}$, the t statistic depends on the data only through the six quantities

$$\boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1, \quad \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2, \quad \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2, \quad \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1, \quad \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2, \text{ and } \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2;$$

notice that $\boldsymbol{y}_i^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_j = \boldsymbol{y}_i^\top (\boldsymbol{M}_{\boldsymbol{W}} + \boldsymbol{P}_1) \boldsymbol{y}_j$, for $i, j = 1, 2$. The same turns out to be true of the other two statistics we consider, as well as of the celebrated Anderson-Rubin statistic.

In view of the scale invariance that we have established, the contemporaneous covariance matrix of the disturbances $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ can without loss of generality be set equal to

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

with both variances equal to unity. Thus we can represent the disturbances in terms of two independent n–vectors, say $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, of independent standard normal elements, as follows:

$$\boldsymbol{u}_1 = \boldsymbol{v}_1, \quad \boldsymbol{u}_2 = \rho \boldsymbol{v}_1 + r \boldsymbol{v}_2,$$

where $r \equiv (1 - \rho^2)^{1/2}$. We now show that we can write all the test statistics as functions of $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, the exogenous variables, and just three parameters.

We see that

$$\begin{aligned} \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 &= (\rho \boldsymbol{v}_1 + r \boldsymbol{v}_2)^\top \boldsymbol{M}_{\boldsymbol{W}} (\rho \boldsymbol{v}_1 + r \boldsymbol{v}_2) \\ &= \rho^2 \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1 + r^2 \boldsymbol{v}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2 + 2\rho r \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2, \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2 = \boldsymbol{\pi}_1^\top \boldsymbol{W}_1^\top \boldsymbol{W}_1 \boldsymbol{\pi}_1 &+ 2\boldsymbol{\pi}_1^\top \boldsymbol{W}_1^\top (\rho \boldsymbol{v}_1 + r \boldsymbol{v}_2) \\ &+ \rho^2 \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + r^2 \boldsymbol{v}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_2 + 2\rho r \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2. \end{aligned}$$

Now let $\boldsymbol{W}_1\boldsymbol{\pi}_1 = a\boldsymbol{w}_1$, with $\|\boldsymbol{w}_1\| = 1$. The square of the parameter a is the so-called scalar concentration parameter; see Phillips (1983) and Stock, Wright, and Yogo (2002). Further, let $\boldsymbol{w}_1^\top \boldsymbol{v}_i = x_i$, for $i = 1, 2$. Clearly, x_1 and x_2 are independent standard normal variables. Then

$$\boldsymbol{\pi}_1^\top \boldsymbol{W}_1^\top \boldsymbol{W}_1 \boldsymbol{\pi}_1 = a^2 \text{ and } \boldsymbol{\pi}_1^\top \boldsymbol{W}_1^\top \boldsymbol{v}_i = a x_i, \quad i = 1, 2.$$

We find that

$$\boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2 = a^2 + 2a(\rho x_1 + r x_2) + \rho^2 \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + r^2 \boldsymbol{v}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_2 + 2\rho r \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2.$$

and

$$\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1 = \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1 + 2\beta(\rho \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1 + r \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2) + \beta^2 \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2. \tag{4}$$

Similarly,

$$\begin{aligned} \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_1 &= \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + 2\beta \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + \beta^2 \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2 \\ &= \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + 2\beta(a x_1 + \rho \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + r \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2) + \beta^2 \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2. \end{aligned}$$

Further,

$$\begin{aligned} \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 &= \rho \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1 + r \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2 + \beta \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2, \text{ and} \\ \boldsymbol{y}_1^\top \boldsymbol{P}_1 \boldsymbol{y}_2 &= a x_1 + \rho \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1 + r \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2 + \beta \boldsymbol{y}_2^\top \boldsymbol{P}_1 \boldsymbol{y}_2. \end{aligned}$$

The six quadratic forms can be generated in terms of eight random variables and three parameters. The eight random variables are x_1 and x_2, along with six quadratic forms of the same sort as those above,

$$\boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_1, \quad \boldsymbol{v}_1^\top \boldsymbol{P}_1 \boldsymbol{v}_2, \quad \boldsymbol{v}_2^\top \boldsymbol{P}_1 \boldsymbol{v}_2, \quad \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_1, \quad \boldsymbol{v}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2, \text{ and } \boldsymbol{v}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{v}_2,$$

and the three parameters are a, ρ, and β. Under the null hypothesis, of course, $\beta = 0$. Since $\boldsymbol{P}_1 \boldsymbol{M}_{\boldsymbol{W}} = \mathbf{O}$, the first three variables are independent of the last three.

It is not hard, at least under the assumption of Gaussian disturbances, to find the distributions of the eight variables, and then to simulate them directly, without any need to generate actual bootstrap samples. Even if the disturbances are not assumed to be Gaussian, we can generate bootstrap disturbances by resampling, and then use them to generate the eight random variables, again with no need to generate bootstrap samples or to estimate anything.

Several bootstrap DGPs

An obvious but important point is that the bootstrap DGP must be able to handle both of the endogenous variables, that is, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. A straightforward, conventional approach is to estimate the parameters β, $\boldsymbol{\gamma}$, $\boldsymbol{\pi}$, σ_1, σ_2, and ρ of the original model, and then to generate simulated data using these equations with the estimated parameters.

However, the conventional approach estimates more parameters than it needs to. In fact, only a and ρ need to be estimated. In order to estimate a, we may use an estimate of $\boldsymbol{\pi}$ with an appropriate scaling factor to take account of the fact that a is defined for DGPs with unit disturbance variances.

We investigate five different ways of estimating the parameters ρ and a. All can be written as

$$\ddot{\rho} = \frac{\ddot{\boldsymbol{u}}_1^\top \ddot{\boldsymbol{u}}_2}{(\ddot{\boldsymbol{u}}_1^\top \ddot{\boldsymbol{u}}_1 \ddot{\boldsymbol{u}}_2^\top \ddot{\boldsymbol{u}}_2)^{1/2}}, \text{ and}$$

$$\ddot{a} = \sqrt{n\, \ddot{\boldsymbol{\pi}}_1^\top \boldsymbol{W}_1^\top \boldsymbol{W}_1 \ddot{\boldsymbol{\pi}}_1 / \ddot{\boldsymbol{u}}_2^\top \ddot{\boldsymbol{u}}_2}.$$

Different bootstrap DGPs use various estimates of $\boldsymbol{\pi}_1$ and various residual vectors. The issue is to what extent Golden Rule 2 is respected, as we will see that the performance of the bootstrap varies enormously in consequence.

The simplest way to estimate ρ and a is probably to use the restricted residuals

$$\tilde{\boldsymbol{u}}_1 = \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_1 = \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1 + \boldsymbol{P}_1 \boldsymbol{y}_1,$$

which, in the case of the simple model, are just equal to $\boldsymbol{y}_1$, along with the OLS estimates $\hat{\boldsymbol{\pi}}_1$ and OLS residuals $\hat{\boldsymbol{u}}_2$ from the reduced-form equation. We call this widely-used method the RI bootstrap, for "Restricted, Inefficient". It can be expected to work better than the pairs bootstrap, and better than other parametric procedures that do not impose the null hypothesis.

As the name implies, the problem with the RI bootstrap is that $\hat{\boldsymbol{\pi}}_1$ is not an efficient estimator. Efficient estimates $\tilde{\boldsymbol{\pi}}_1$ can be obtained by running the artificial regression

$$\boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_2 = \boldsymbol{W}_1 \boldsymbol{\pi}_1 + \delta \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_1 + \text{residuals}. \tag{5}$$

It can be shown that these estimates are asymptotically equivalent to the ones that would be obtained by using 3SLS or FIML. The estimated vector of disturbances from equation (5) is not the vector of OLS residuals but rather the vector $\tilde{\boldsymbol{u}}_2 = \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_2 - \boldsymbol{W}_1 \tilde{\boldsymbol{\pi}}_1$.

Instead of equation (5), it may be more convenient to run the regression

$$\boldsymbol{y}_2 = \boldsymbol{W}_1 \boldsymbol{\pi}_1 + \boldsymbol{Z} \boldsymbol{\pi}_2 + \delta \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_1 + \text{residuals}.$$

This is just the reduced form equation augmented by the residuals from restricted estimation of the structural equation. We call the bootstrap that uses $\tilde{\boldsymbol{u}}_1$, $\tilde{\boldsymbol{\pi}}_1$, and $\tilde{\boldsymbol{u}}_2$ the RE bootstrap, for "Restricted, Efficient".

Two other bootstrap methods do not impose the restriction that $\beta = 0$ when estimating ρ and a. For the purposes of testing, it is a bad idea not to impose this restriction, as argued in Davidson and MacKinnon (1999). However, it is quite inconvenient to impose restrictions when constructing bootstrap confidence intervals, and, since confidence intervals are implicitly obtained by inverting tests, it is of interest to see how much harm is done by not imposing the restriction.

The UI bootstrap, for "Unrestricted, Inefficient", uses the unrestricted residuals $\hat{\boldsymbol{u}}_1$ from IV estimation of the structural equation, along with the estimates $\hat{\boldsymbol{\pi}}_1$ and residuals $\hat{\boldsymbol{u}}_2$ from OLS estimation of the reduced-form equation. The UE bootstrap, for "Unrestricted, Efficient", also uses $\hat{\boldsymbol{u}}_1$, but the other quantities come from the artificial regression

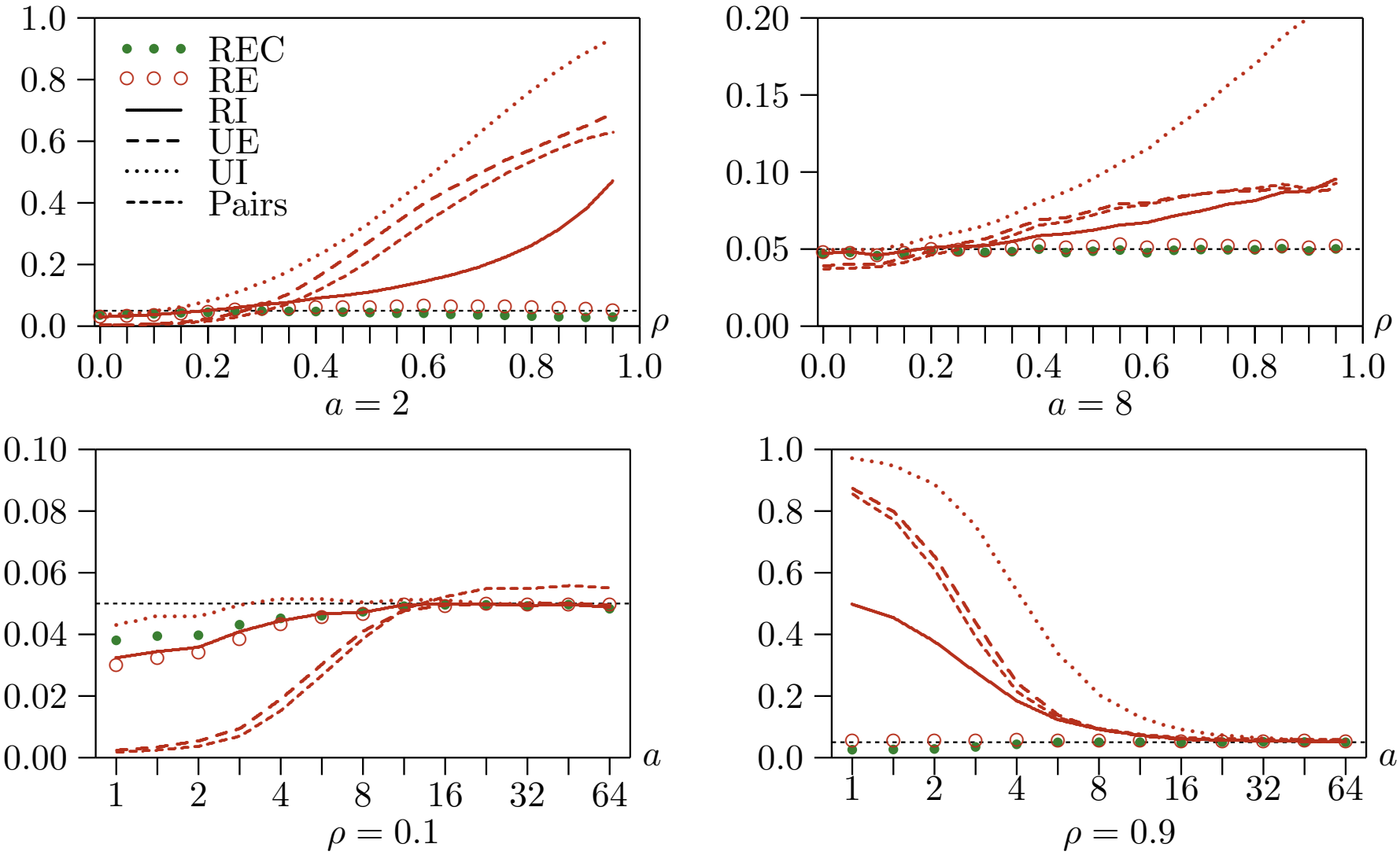
$$\boldsymbol{M_Z y}_2 = \boldsymbol{W}_1 \boldsymbol{\pi}_1 + \delta \hat{\boldsymbol{u}}_1 + \text{residuals},$$

which is similar to regression (5).

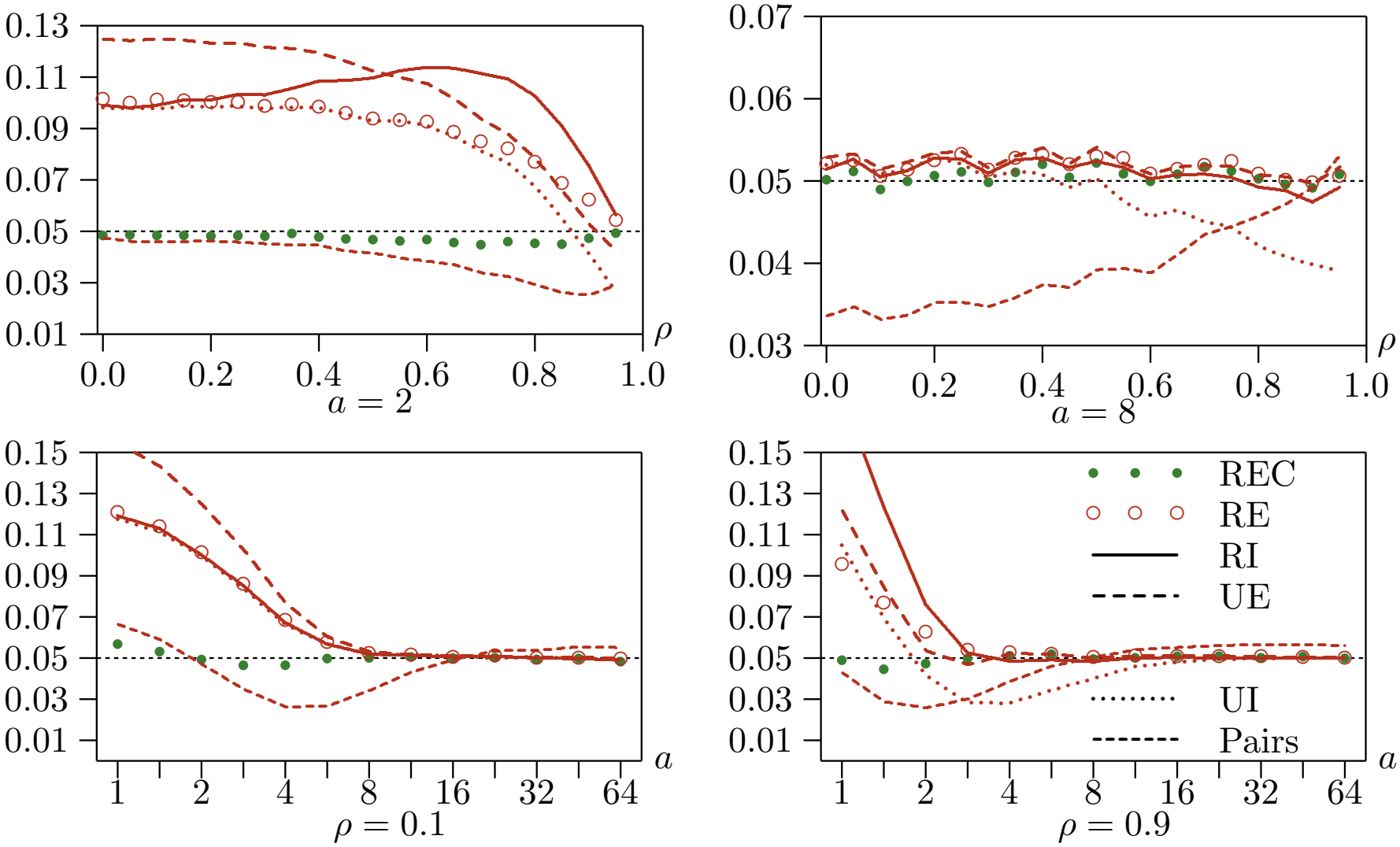
The weak-instrument asymptotic construction shows that $\tilde{a}^2$ is biased and inconsistent, the bias being equal to $r^2(l-k)$. It seems plausible, therefore, that the bias-corrected estimator

$$\tilde{a}^2_{\text{BC}} \equiv \max\bigl(0, \tilde{a}^2 - (l-k)(1-\tilde{\rho}^2)\bigr)$$

may be better for the purposes of defining the bootstrap DGP. Thus we consider a fifth bootstrap method, REC, for "Restricted, Efficient, Corrected." It differs from RE in that it uses $\tilde{a}_{\text{BC}}$ instead of $\tilde{a}$.



Rejection frequencies for bootstrap Wald (t) tests for $l - k = 9$, $n = 100$



Rejection frequencies for bootstrap LR tests for $l - k = 9$, $n = 100$

Nonlinear models

Bootstrapping is often seen as a very computationally intensive procedure, although with the hardware and software available at the time of writing, this is seldom a serious problem in applied work. Models that require nonlinear estimation can be an exception to this statement, because the algorithms used in nonlinear estimation may fail to converge after a small number of iterations. If this happens while estimating a model with real data, the problem is not related to bootstrapping, but arises rather from the relation between the model and the data. The problem for bootstrapping occurs when an estimation procedure that works with the original data does not work with one or more of the bootstrap samples.

In principle nonlinear estimation should be easier in the bootstrap context than otherwise. One knows the true bootstrap DGP, and can use the true parameters for that DGP as the starting point for the iterative procedure used to implement the nonlinear estimation. In those cases in which it is necessary to estimate two models, one restricted, the other unrestricted, one can use the estimates from the restricted model, say, as the starting point for the unrestricted estimation, thus making use of properties specific to a particular bootstrap sample.

When any nonlinear procedure is repeated thousands of times, it seems that anything that can go wrong will go wrong at least once. Most of the time, the arguments of the previous paragraph apply, but not always. Any iterative procedure can go into an infinite loop if it does not converge, with all sorts of undesirable consequences. It is therefore good practice to set a quite modest upper limit to the number of iterations permitted for each bootstrap sample.

In many cases, an upper limit of just 3 or 4 iterations can be justified theoretically. Asymptotic theory can usually provide a *rate of convergence*, with respect to the sample size, of the bootstrap discrepancy to zero. It can also provide the rate of convergence of Newton's method, or a quasi-Newton method, used by the estimation algorithm. If the bootstrap discrepancy goes to zero as $n^{-3/2}$ say, then there is little point in seeking numerical accuracy with a better rate of convergence. With most quasi-Newton methods, the Gauss-Newton algorithm for instance, each iteration reduces the distance between the current parameters of the algorithm and those to which the algorithm will converge (assuming that it does converge) by a factor of $n^{-1/2}$. Normally, we can initialise the algorithm with parameters that differ from those at convergence by an amount of order $n^{-1/2}$. After three iterations, the difference is of order only n^{-2}, a lower order than that of the bootstrap discrepancy. The same order of accuracy is thus achieved on average as would be attainable if the iterations continued until convergence by some stricter criterion. Since bootstrap inference is based on an average over the bootstrap repetitions, this is enough for most purposes.

Bootstrapping LM tests

For a classical LM test, the criterion function is the loglikelihood function, and the test statistic is based on the estimates obtained by maximising it subject to the restrictions of the null hypothesis. However, the test statistic is expressed in terms of the gradient and Hessian of the loglikelihood of the full unrestricted model. One form of the statistic is

$$LM = -\boldsymbol{g}^\top(\tilde{\boldsymbol{\theta}})\boldsymbol{H}^{-1}(\tilde{\boldsymbol{\theta}})\boldsymbol{g}(\tilde{\boldsymbol{\theta}}),$$

where $\boldsymbol{g}(\tilde{\boldsymbol{\theta}})$ and $\boldsymbol{H}(\tilde{\boldsymbol{\theta}})$ are, respectively, the gradient and Hessian of the unrestricted loglikelihood, evaluated at the restricted estimates $\tilde{\boldsymbol{\theta}}$.

We propose to replace the nonlinear estimation by a predetermined, finite, usually small, number of Newton or quasi-Newton steps, starting from the estimates given by the real data. It is usually possible to determine an integer l such that the rejection probability for the bootstrap test at nominal level α differs from α by an amount that is $O(n^{-l/2})$; typically, $l = 3$ or $l = 4$. This being so, the same order of accuracy will be achieved even if there is an error that is $O_p(n^{-l/2})$ in the computation of the bootstrap P values.

The true value of the parameters for the bootstrap DGP is $\tilde{\boldsymbol{\theta}}$. If we denote the fully nonlinear estimates from a bootstrap sample by $\tilde{\boldsymbol{\theta}}^*$, then, by construction, we have that $\tilde{\boldsymbol{\theta}}^* - \tilde{\boldsymbol{\theta}} = O_p(n^{-1/2})$. Thus $\tilde{\boldsymbol{\theta}}$ is a suitable starting point for Newton's method or a quasi-Newton method applied to the restricted model. If the exact Hessian is used, then the successive estimates $\tilde{\boldsymbol{\theta}}^*_{(i)}$, $i = 0, 1, 2, \ldots$, satisfy

$$\tilde{\boldsymbol{\theta}}^*_{(i)} - \tilde{\boldsymbol{\theta}}^* = O_p(n^{-2^{i-1}}).$$

If an approximate Hessian is used, they instead satisfy

$$\tilde{\boldsymbol{\theta}}^*_{(i)} - \tilde{\boldsymbol{\theta}}^* = O_p(n^{-(i+1)/2}).$$

The successive approximations to the LM statistic are defined by

$$LM(\tilde{\boldsymbol{\theta}}^*_{(i)}) \equiv -\boldsymbol{g}^\top(\tilde{\boldsymbol{\theta}}^*_{(i)})\boldsymbol{H}^{-1}(\tilde{\boldsymbol{\theta}}^*_{(i)})\boldsymbol{g}(\tilde{\boldsymbol{\theta}}^*_{(i)}),$$

where the functions $\boldsymbol{g}$ and $\boldsymbol{H}$ are the same as the ones used to compute the actual test statistic.

At this point, a little care is necessary. The successive approximations can be written as

$$\left(n^{-1/2}\boldsymbol{g}^\top(\tilde{\boldsymbol{\theta}}^*_{(i)})\right)\left(-n^{-1}\boldsymbol{H}(\tilde{\boldsymbol{\theta}}^*_{(i)})\right)^{-1}\left(n^{-1/2}\boldsymbol{g}(\tilde{\boldsymbol{\theta}}^*_{(i)})\right),$$

where each factor in parentheses is $O_p(1)$. We have

$$n^{-1}\boldsymbol{H}(\tilde{\boldsymbol{\theta}}^*_{(i)}) = n^{-1}\boldsymbol{H}(\tilde{\boldsymbol{\theta}})+O_p(\tilde{\boldsymbol{\theta}}^*_{(i)}-\tilde{\boldsymbol{\theta}}), \text{ and } n^{-1/2}\boldsymbol{g}(\tilde{\boldsymbol{\theta}}^*_{(i)}) = n^{-1/2}\boldsymbol{g}(\tilde{\boldsymbol{\theta}})+n^{1/2}O_p(\tilde{\boldsymbol{\theta}}^*_{(i)}-\tilde{\boldsymbol{\theta}}).$$

Note that $n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}) = O_p(1)$ whenever $n^{1/2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) = O_p(1)$, where $\hat{\boldsymbol{\theta}}$ maximises the unrestricted loglikelihood, so that $\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$. We find that

$$LM(\tilde{\boldsymbol{\theta}}^*_{(i)}) = LM(\tilde{\boldsymbol{\theta}}) + n^{1/2}O_p(\tilde{\boldsymbol{\theta}}^*_{(i)} - \tilde{\boldsymbol{\theta}}).$$

This is the key result for LM tests.

For Newton's method, the difference between $LM(\tilde{\boldsymbol{\theta}}^*_{(i)})$ and $LM(\tilde{\boldsymbol{\theta}})$ is of order $n^{-(2^i-1)/2}$. For the quasi-Newton case, it is of order $n^{-i/2}$. After just one iteration, when $i = 1$, the difference is of order $n^{-1/2}$ in both cases. Subsequently, the difference diminishes much more rapidly for Newton's method than for quasi-Newton methods. In general, if bootstrap P values are in error at order $n^{-l/2}$, and we are using Newton's method with an exact Hessian, the number of steps needed to achieve at least the same order of accuracy as the bootstrap, m, should be chosen so that $2^m - 1 \geq l$. Thus, for $l = 3$, the smallest suitable m is 2, and for $l = 4$, it is 3. If we are using a quasi-Newton method, we simply need to choose m so that $m \geq l$.

Bootstrapping LR tests

Likelihood Ratio tests are particularly expensive to bootstrap, because two nonlinear optimisations must normally be performed for each bootstrap sample. However, both of these can be replaced by a small number of Newton steps, starting from the restricted estimates. Under the null, this is done exactly as above for an LM test. Under the alternative, the only difference is that the gradient and Hessian correspond to the unrestricted model, and thus involve derivatives with respect to all components of $\boldsymbol{\theta}$. The starting point for the unrestricted model may well be the same as for the restricted model, but it is probably preferable to start from the endpoint of the restricted iterations. This endpoint contains possibly relevant information about the current bootstrap sample, and the difference between it and the unrestricted bootstrap fully nonlinear estimate $\hat{\boldsymbol{\theta}}^*$ is $O_p(n^{-1/2})$, as required.

For each bootstrap sample, we can compute a bootstrap LR statistic. The true value of this statistic is $2\big(\ell(\hat{\boldsymbol{\theta}}^*) - \ell(\tilde{\boldsymbol{\theta}}^*)\big)$. Consider replacing $\hat{\boldsymbol{\theta}}^*$ by an approximation $\acute{\boldsymbol{\theta}}$ such that $\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^* = O_p(n^{-1/2})$. Since $\boldsymbol{g}(\hat{\boldsymbol{\theta}}^*) = \mathbf{0}$ by the first-order conditions for maximising $\ell(\boldsymbol{\theta})$, a Taylor expansion gives

$$\ell(\hat{\boldsymbol{\theta}}^*) - \ell(\acute{\boldsymbol{\theta}}) = -\frac{1}{2}\big(\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^*\big)^\top \boldsymbol{H}(\bar{\boldsymbol{\theta}})\big(\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^*\big),$$

where $\bar{\boldsymbol{\theta}}$ is a convex combination of $\acute{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^*$. Since $\boldsymbol{H}$ is $O_p(n)$, it follows that

$$\ell(\hat{\boldsymbol{\theta}}^*) - \ell(\acute{\boldsymbol{\theta}}) = nO_p\big((\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^*)^2\big).$$

The above result is true for both the restricted and unrestricted loglikelihoods, and is therefore true as well for the LR statistic.

We see that, if Newton's method is used,

$$\ell(\hat{\boldsymbol{\theta}}^*) - \ell(\hat{\boldsymbol{\theta}}^*_{(i)}) = O_p(n^{-2^i+1}),$$

while, if a quasi-Newton method is used,

$$\ell(\hat{\boldsymbol{\theta}}^*) - \ell(\hat{\boldsymbol{\theta}}^*_{(i)}) = O_p(n^{-i}).$$

These results imply that, for both Newton and quasi-Newton methods when $l = 3$ and $l = 4$, the minimum number of steps m for computing $\hat{\boldsymbol{\theta}}^*_{(m)}$ and $\tilde{\boldsymbol{\theta}}^*_{(m)}$ needed to ensure that the error in the LR statistic is at most of order $n^{-l/2}$ is just 2.

Heteroskedasticity

All the bootstrap DGPs that we have looked at so far are based on models where either the observations are IID, or else some set of quantities that can be estimated from the data, like the disturbances of a regression model, are IID. But if the disturbances of a regression are heteroskedastic, with an unknown pattern of heteroskedasticity, there is nothing that is even approximately IID. There exist of course test statistics robust to heteroskedasticity of unknown form, based on one of the numerous variants of the Eicker-White Heteroskedasticity Consistent Covariance Matrix Estimator (HCCME). Use of an HCCME gives rise to statistics that are approximately pivotal for models that admit heteroskedasticity of unknown form.

For bootstrapping, it is very easy to satisfy Golden Rule 1, since either a parametric bootstrap or a resampling bootstrap of the sort we have described belongs to a null hypothesis that, since it allows heteroskedasticity, must also allow the special case of homoskedasticity. But Golden Rule 2 poses a more severe challenge.

The pairs bootstrap

The first suggestion for bootstrapping models with heteroskedasticity bears a variety of names: among them the (y, X) bootstrap or the pairs bootstrap. The approach was proposed in Freedman (1981). Instead of resampling the dependent variable, or residuals, possibly centred or rescaled, one resamples **pairs** consisting of an observation of the dependent variable along with the set of explanatory variables for that same observation. One selects an index s at random from the set $1, \ldots, n$, and then an observation of a bootstrap sample is the pair $(y_s, \boldsymbol{X}_s)$, where $\boldsymbol{X}_s$ is a row vector of all the explanatory variables for observation s.

This bootstrap implicitly assumes that the pairs $(y_t, \boldsymbol{X}_t)$ are IID under the null hypothesis. Although this is still a restrictive assumption, ruling out any form of dependence among observations, it does allow for any sort of heteroskedasticity of y_t conditional of $\boldsymbol{X}_t$. The objects resampled are IID drawings from the *joint* distribution of y_t and $\boldsymbol{X}_t$.

Suppose that the regression model itself is written as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad t = 1, \ldots, n,$$

with $\boldsymbol{X}_t$ a $1 \times k$ vector and $\boldsymbol{\beta}$ a $k \times 1$ vector of parameters. The disturbances u_t are allowed to be heteroskedastic, but must have an expectation of 0 conditional on the explanatory variables. Thus $\mathrm{E}(y_t|\boldsymbol{X}_t) = \boldsymbol{X}_t\boldsymbol{\beta}_0$ if $\boldsymbol{\beta}_0$ is the parameter vector for the true DGP. Let us consider a null hypothesis according to which a subvector of $\boldsymbol{\beta}$, $\boldsymbol{\beta}_2$ say, is zero. This null hypothesis is not satisfied by the pairs bootstrap DGP. In order to respect Golden Rule 1, therefore, we must modify either the null hypothesis to be tested in the bootstrap samples, or the bootstrap DGP itself.

In the empirical joint distribution of the pairs $(y_t, \boldsymbol{X}_t)$, the expectation of the first element y conditional on the second element $\boldsymbol{X}$ is defined only if $\boldsymbol{X} = \boldsymbol{X}_t$ for some $t = 1, \ldots, n$. Then $\mathrm{E}(y|\boldsymbol{X} = \boldsymbol{X}_t) = y_t$. This result does not help determine what the true value of $\boldsymbol{\beta}$, or of $\boldsymbol{\beta}_2$, might be for the bootstrap DGP. Given this, what is usually done is to use the OLS estimate $\hat{\boldsymbol{\beta}}_2$ as true for the bootstrap DGP, and so to test the hypothesis that $\boldsymbol{\beta}_2 = \hat{\boldsymbol{\beta}}_2$ when computing the bootstrap statistics.

In Flachaire (1999), the bootstrap DGP is changed. It now resamples pairs $(\hat{u}_t, \boldsymbol{X}_t)$, where the $\hat{u}_t$ are the OLS residuals from estimation of the *unrestricted* model, possibly rescaled in various ways. Then, if s is an integer drawn at random from the set $1, \ldots, n$, y_t^* is generated by

$$y_t^* = \boldsymbol{X}_{s1}\tilde{\boldsymbol{\beta}}_1 + \hat{u}_s,$$

where $\boldsymbol{\beta}_1$ contains the elements of $\boldsymbol{\beta}$ that are not in $\boldsymbol{\beta}_2$, and $\tilde{\boldsymbol{\beta}}_1$ is the *restricted* OLS estimate. Similarly, $\boldsymbol{X}_{s1}$ contains the elements of $\boldsymbol{X}_s$ of which the coefficients are elements of $\boldsymbol{\beta}_1$. By construction, the vector of the $\hat{u}_t$ is orthogonal to all of the vectors containing the observations of the explanatory variables. Thus in the empirical joint distribution of the pairs $(\hat{u}_t, \boldsymbol{X}_t)$, the first element, $\hat{u}$, is uncorrelated with the second element, $\boldsymbol{X}$. However any relation between the variance of $\hat{u}$ and the explanatory variables is preserved, as with Freedman's pairs bootstrap. In addition, the new bootstrap DGP now satisfies the null hypothesis as originally formulated.

The wild bootstrap

The null model on which any form of pairs bootstrap is based posits the joint distribution of the dependent variable y and the explanatory variables. If it is assumed that the explanatory variables are exogenous, conventional practice is to compute statistics, and their distributions, conditional on them. One way in which this can be done is to use the so-called **wild bootstrap**; see Wu (1986) Liu (1988), Mammen (1993), and Davidson and Flachaire (2008).

For a regression model, the wild bootstrap DGP takes the form

$$y_t^* = \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} + s_t^*\tilde{u}_t$$

where $\tilde{\boldsymbol{\beta}}$ is as usual the restricted least-squares estimate of the regression parameters, and the $\tilde{u}_t$ are the restricted least-squares residuals. Notice that no resampling takes place here; both the explanatory variables and the residual for bootstrap observation t come from observation t of the original sample. The new random elements introduced are the s_t^*, which are IID drawings from a distribution with expectation 0 and variance 1.

The bootstrap DGP satisfies Golden Rule 1 easily: since s_t^* and $\tilde{u}_t$ are independent, the latter having been generated by the real DGP and the former by the random number generator, the expectation of the bootstrap disturbance $s_t^*\tilde{u}_t$ is 0. Conditional on the residual $\tilde{u}_t$, the variance of $s_t^*\tilde{u}_t$ is $\tilde{u}_t^2$. If the residual is accepted as a proxy for the unobserved disturbance u_t, then the unconditional expectation of $\tilde{u}_t^2$ is the true variance of u_t, and this fact goes a long way towards satisfying Golden Rule 2. The HCCME uses exactly the same strategy to estimate the latent variances.

For a long time, the most commonly used distribution for the s_t^* was the following two-point distribution,

$$s_t^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}), \\ (\sqrt{5}+1)/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}), \end{cases}$$

which was suggested by Mammen because, with it, $\mathrm{E}((s_t^*)^3) = 1$. If the true disturbances, and also the explanatory variables, are skewed, Edgeworth expansions suggest that this last property is desirable. (But Edgeworth expansions are not valid with the discrete distribution of the wild bootstrap disturbances.) A simpler two-point distribution is the **Rademacher distribution**

$$s_t^* = \begin{cases} -1 & \text{with probability } \frac{1}{2}, \\ 1 & \text{with probability } \frac{1}{2}. \end{cases}$$

Davidson and Flachaire propose this simpler distribution, which leaves the absolute value of each residual unchanged in the bootstrap DGP, while assigning it an arbitrary sign. They show by means of simulation experiments that their choice often leads to more reliable bootstrap inference than other choices.

The wild bootstrap can be generalised quite easily to the IV case studied earlier. The idea of the wild bootstrap is to use for the bootstrap disturbance(s) associated with the i^{th} observation the actual residual(s) for that observation, possibly transformed in some way, and multiplied by a random variable, independent of the data, with mean 0 and variance 1. We propose the **wild restricted efficient residual bootstrap**, or **WRE bootstrap**. The DGP has a structural and a reduced form equation as before, with

$$\begin{bmatrix} \tilde{u}_{1i}^* \\ \tilde{u}_{2i}^* \end{bmatrix} = \begin{bmatrix} \big(n/(n-k)\big)^{1/2} \tilde{u}_{1i} v_i^* \\ \big(n/(n-l)\big)^{1/2} \tilde{u}_{2i} v_i^* \end{bmatrix},$$

where v_i^* is a random variable that has mean 0 and variance 1. Notice that both rescaled residuals are multiplied by the *same* value of v_i^*. This preserves the correlation between the two disturbances, at least when they are symmetrically distributed. Using the Rademacher distribution imposes symmetry on the bivariate distribution of the bootstrap disturbances, and this may affect the correlation when they are not actually symmetric.

There is a good deal of evidence that the wild bootstrap works reasonably well for univariate regression models, even when there is quite severe heteroskedasticity. See, among others, Gonçalves and Kilian (2004) and MacKinnon (2006). Although the wild bootstrap cannot be expected to work quite as well as a comparable residual bootstrap method when the disturbances are actually homoskedastic, the cost of insuring against heteroskedasticity generally seems to be very small.

Bootstrap DGPs for Dependent Data

The bootstrap DGPs that we have discussed so far are not valid when applied to models with dependent disturbances having an unknown pattern of dependence. For such models, we wish to specify a bootstrap DGP which generates correlated disturbances that exhibit approximately the same pattern of dependence as the real disturbances, even though we do not know the process that actually generated them. There are two main approaches, neither of which is entirely satisfactory in all cases.

The first approach is a semiparametric one called the **sieve bootstrap**. It is based on the fact that any linear, invertible time-series process can be approximated by an AR(∞) process. The idea is to estimate a stationary AR(p) process and use this estimated process, perhaps together with resampled residuals from the estimation of the AR(p) model, to generate bootstrap samples. For example, suppose we are concerned with a static linear regression model, but the covariance matrix $\boldsymbol{\Omega}$ is no longer assumed to be diagonal. Instead, it is assumed that $\boldsymbol{\Omega}$ can be well approximated by the covariance matrix of a stationary AR(p) process, which implies that the diagonal elements are all the same.

In this case, the first step is to estimate the regression model, possibly after imposing restrictions on it, so as to generate a parameter vector $\hat{\boldsymbol{\beta}}$ and a vector of residuals $\hat{\boldsymbol{u}}$ with typical element $\hat{u}_t$. The next step is to estimate the AR(p) model

$$\hat{u}_t = \sum_{i=1}^{p} \rho_i \hat{u}_{t-i} + \varepsilon_t$$

for $t = p+1, \ldots, n$. In theory, the order p of this model should increase at a certain rate as the sample size increases. In practice, p is most likely to be determined either by using an information criterion like the AIC or by sequential testing. Care should be taken to ensure that the estimated model is stationary. This may require the use of full maximum likelihood, rather than least squares.

Estimation of the AR(p) model yields residuals and an estimate $\hat{\sigma}_\varepsilon^2$ of the variance of the ε_t, as well as the estimates $\hat{\rho}_i$. We may use these to set up a variety of possible bootstrap DGPs, all of which take the form

$$y_t^* = \boldsymbol{X}_t \hat{\boldsymbol{\beta}} + u_t^*.$$

There are two choices to be made, namely, the choice of parameter estimates $\hat{\boldsymbol{\beta}}$ and the generating process for the bootstrap disturbances u_t^*. One choice for $\hat{\boldsymbol{\beta}}$ is just the OLS estimates. But these estimates, although consistent, are not efficient if $\boldsymbol{\Omega}$ is not a scalar matrix. We might therefore prefer to use feasible GLS estimates. An estimate $\hat{\boldsymbol{\Omega}}$ of the covariance matrix can be obtained by solving the Yule-Walker equations, using the $\hat{\rho}_i$ in order to obtain estimates of the autocovariances of the AR(p) process. Then a Cholesky decomposition of $\hat{\boldsymbol{\Omega}}^{-1}$ provides the feasible GLS transformation to be applied to the dependent variable $\boldsymbol{y}$ and the explanatory variables $\boldsymbol{X}$ in order to compute feasible GLS estimates of $\boldsymbol{\beta}$, restricted as required by the null hypothesis under test.

For observations after the first p, the bootstrap disturbances are generated as follows:

$$u_t^* = \sum_{i=1}^{p} \hat{\rho}_i u_{t-i}^* + \varepsilon_t^*, \quad t = p+1, \ldots, n,$$

where the ε_t^* can either be drawn from the $\mathrm{N}(0, \hat{\sigma}_\varepsilon^2)$ distribution for a parametric bootstrap or resampled from the residuals $\hat{\varepsilon}_t$, preferably rescaled by the factor $\sqrt{n/(n-p)}$. First, of course, we must generate the first p bootstrap disturbances, the u_t^*, for $t = 1, \ldots, p$.

One way to do so is just to set $u_t^* = \hat{u}_t$ for the first p observations of each bootstrap sample. This is analogous to what we proposed for the bootstrap DGP used in conjunction with a dynamic model: We initialise with fixed starting values given by the real data. Unless we are sure that the AR(p) process is really stationary, rather than just being characterised by values of the ρ_i that correspond to a stationary covariance matrix, this is the only appropriate procedure.

If we are happy to impose full stationarity on the bootstrap DGP, then we may draw the first p values of the u_t^* from the p–variate stationary distribution. This is easy to do if we have solved the Yule-Walker equations for the first p autocovariances, provided that we assume normality. If normality is an uncomfortably strong assumption, then we can initialise the recurrence in any way we please and then generate a reasonably large number (say 200) of bootstrap disturbances recursively, using resampled rescaled values of the $\hat{\varepsilon}_t$ for the ε_t^*. We then throw away all but the last p of these disturbances and use those for initialisation. In this way, we approximate a stationary process with the correct estimated stationary covariance matrix, but with no assumption of normality.

The sieve bootstrap method has been used to improve the finite-sample properties of unit root tests by Park (2003) and Chang and Park (2003), but it has not yet been widely used in econometrics. The fact that it does not allow for heteroskedasticity is a limitation. Moreover, AR(p) processes do not provide good approximations to every time-series process that might arise in practice. An example for which the approximation is exceedingly poor is an MA(1) process with a parameter close to -1. The sieve bootstrap cannot be expected to work well in such cases. For more detailed treatments, see Bühlmann (1997, 2002), Choi and Hall (2000), and Park (2002).

The second principal method of dealing with dependent data is the **block bootstrap**, which was originally proposed by Künsch (1989). This method is much more widely used than the sieve bootstrap. The idea is to divide the quantities that are being resampled, which might be either rescaled residuals or $[\boldsymbol{y}, \boldsymbol{X}]$ pairs, into blocks of l consecutive observations, and then resample the blocks. The blocks may be either overlapping or non-overlapping. In either case, the choice of block length, l, is evidently very important. If l is small, the bootstrap samples cannot possibly mimic the patterns of dependence in the original data, because these patterns are broken whenever one block ends and the next begins. However, if l is large, the bootstrap samples will tend to be excessively influenced by the random characteristics of the actual sample.

For the block bootstrap to work asymptotically, the block length must increase as the sample size n increases, but at a slower rate, which varies depending on what the bootstrap samples are to be used for. In some common cases, l should be proportional to $n^{1/3}$, but with a factor of proportionality that is, in practice, unknown. Unless the sample size is very large, it is generally impossible to find a value of l for which the bootstrap DGP provides a really good approximation to the unknown true DGP.

A variation of the block bootstrap is the **stationary bootstrap** proposed by Politis and Romano (1994), in which the block length is random rather than fixed. This procedure is commonly used in practice. However, Lahiri (1999) provides both theoretical arguments and limited simulation evidence which suggest that fixed block lengths are better than variable ones and that overlapping blocks are better than non-overlapping ones. Thus, at the present time, the procedure of choice appears to be the **moving-block bootstrap**, in which there are $n-l+1$ blocks, the first containing observations 1 through l, the second containing observations 2 through $l+1$, and the last containing observations $n-l+1$ through n.

It is possible to use block bootstrap methods with dynamic models. Let

$$\boldsymbol{Z}_t \equiv [y_t, y_{t-1}, \boldsymbol{X}_t].$$

For this model, we could construct $n-l+1$ overlapping blocks

$$\boldsymbol{Z}_1 \dots \boldsymbol{Z}_l, \;\; \boldsymbol{Z}_2 \dots \boldsymbol{Z}_{l+1}, \dots\dots, \boldsymbol{Z}_{n-l+1} \dots \boldsymbol{Z}_n$$

and resample from them. This is the moving-block analog of the pairs bootstrap. When there are no exogenous variables and several lagged values of the dependent variable, the $\boldsymbol{Z}_t$ are themselves blocks of observations. Therefore, this method is sometimes referred to as the **block-of-blocks bootstrap**. Notice that, when the block size is 1, the block-of-blocks bootstrap is simply the pairs bootstrap adapted to dynamic models, as in Gonçalves and Kilian (2004).

Block bootstrap methods are conceptually simple. However, there are many different versions, most of which we have not discussed, and theoretical analysis of their properties tends to require advanced techniques. The biggest problem with block bootstrap methods is that they often do not work very well. We have already provided an intuitive explanation of why this is the case. From a theoretical perspective, the problem is that, even when the block bootstrap offers higher-order accuracy than asymptotic methods, it often does so to only a modest extent. The improvement is always of higher order in the independent case, where blocks should be of length 1, than in the dependent case, where the block size must be greater than 1 and must increase at an optimal rate with the sample size. See Hall, Horowitz, and Jing (1995) and Andrews (2002, 2004), among others.

There are several valuable, recent surveys of bootstrap methods for time-series data. These include Bühlmann (2002), Politis (2003), and Härdle, Horowitz, and Kreiss (2003). Surveys that are older or deal with methods for time-series data in less depth include Li and Maddala (1996), Davison and Hinkley (1997, Chapter 8), Berkowitz and Kilian (2000), Horowitz (2001), and Horowitz (2003).

Bootstrapping the Bootstrap Discrepancy

Any procedure that gives an estimate of the rejection probability of a bootstrap test, or of the CDF of the bootstrap P value, allows one to compute a corrected P value. Two techniques sometimes used to obtain a corrected bootstrap P value are the double bootstrap, as originally proposed by Beran (1988), and the fast double bootstrap proposed by Davidson and MacKinnon (2007).

The double bootstrap

An estimate of the bootstrap RP or the bootstrap discrepancy is specific to the DGP that generates the data. Thus what is in fact done by all techniques that aim to correct a bootstrap P value is to *bootstrap* the estimate of the bootstrap RP, in the sense that the bootstrap DGP itself is used to estimate the bootstrap discrepancy. This can be seen for the ordinary double bootstrap as follows.

The brute force method described earlier for estimating the RP of the bootstrap test is employed, but with the (first-level) bootstrap DGP b in place of μ. The first step is to compute the usual (estimated) bootstrap P value $\hat{p}$, using B_1 bootstrap samples generated from the bootstrap DGP b. Now one wants an estimate of the actual RP of a bootstrap test at nominal level $\hat{p}$. This estimated RP is the double bootstrap P value, $\hat{p}_2$. Thus we set $\mu = b$, $M = B_1$, and $B = B_2$ in the brute-force algorithm described in the previous section. The computation of $\hat{p}$ has already provided us with B_1 statistics τ_j^*, $j = 1, \ldots, B_1$, corresponding to the t_m of the algorithm.

For each of these, we compute the (double) bootstrap DGP β_j^* realised jointly with τ_j^*. Then β_j^* is used to generate B_2 second-level statistics, which we denote by τ_{jl}^{**}, $l = 1, \ldots, B_2$; these correspond to the τ_{mj}^* of the algorithm. The second-level bootstrap P value is then computed as

$$\hat{p}_j^* = \frac{1}{B_2} \sum_{l=1}^{B_2} \mathrm{I}(\tau_{jl}^{**} < \tau_j^*).$$

The estimate of the bootstrap RP at nominal level $\hat{p}$ is then the proportion of the $\hat{p}_j^*$ that are less than $\hat{p}$:

$$\hat{p}_2 = \frac{1}{B_1} \sum_{j=1}^{B_1} \mathrm{I}\big(\hat{p}_j^* \leq \hat{p}\big).$$

The inequality above is not strict, because there may well be cases for which $\hat{p}_j^* = \hat{p}$. For this reason, it is desirable that $B_2 \neq B_1$. The whole procedure requires the computation of $B_1(B_2 + 1) + 1$ statistics and $B_1 + 1$ bootstrap DGPs.

Recall that $R_1(x, \mu)$ is our notation for the CDF of the first-level bootstrap P value. The ideal double bootstrap P value is thus

$$p_2(\mu, \omega) \equiv R_1\big(p(\mu, \omega), \beta(\mu, \omega)\big) = R_1\big(R(\tau(\mu, \omega), \beta(\mu, \omega)), \beta(\mu, \omega)\big).$$

The fast double bootstrap

The so-called fast double bootstrap (FDB) of Davidson and MacKinnon (2007) is much less computationally demanding than the double bootstrap, being based on the fast approximation of the previous section. Like the double bootstrap, the FDB begins by computing the usual bootstrap P value $\hat{p}$. In order to obtain the estimate of the RP of the bootstrap test at nominal level $\hat{p}$, we use the algorithm of the fast approximation with $M = B$ and $\mu = b$. For each of the B samples drawn from β, we obtain the ordinary bootstrap statistic τ_j^*, $j = 1, \dots, B$, and the double bootstrap DGP β_j^*, exactly as with the double bootstrap. *One* statistic τ_j^{**} is then generated by β_j^*. The $\hat{p}$ quantile of the τ_j^{**}, say $Q^{**}(\hat{p})$, is then computed. Of course, for finite B, there is a range of values that can be considered to be the relevant quantile, and we must choose one of them somewhat arbitrarily. The FDB P value is then

$$\hat{p}_{\text{FDB}} = \frac{1}{B} \sum_{j=1}^{B} \text{I}\big(\tau_j^* < Q^{**}(\hat{p})\big).$$

To obtain it, we must compute $2B + 1$ statistics and $B + 1$ bootstrap DGPs.

The ideal fast double bootstrap P value is

$$p_{\mathrm{FDB}}(\mu,\omega) = \Pr\Big(\tau(\beta(\mu,\omega),\omega^*) < Q^*\big(p(\mu,\omega),\beta(\mu,\omega)\big) \;\Big|\; \omega\Big).$$

where $Q^*(x,\mu)$ is the quantile function corresponding to the CDF $R^*(x,\mu)$, so that $Q^*(x,\beta(\mu,\omega))$ corresponds to the distribution $R^*(x,b)$.

More explicitly, we have that

$$\begin{aligned} p_{\mathrm{FDB}}(\mu,\omega) &= \int_\Omega \mathrm{I}\Big(\tau\big(\beta(\mu,\omega),\omega^*\big) < Q^*\big(p(\mu,\omega),\beta(\mu,\omega)\big)\Big)\,\mathrm{d}P(\omega^*) \\ &= R\Big(Q^*\big(p(\mu,\omega),\beta(\mu,\omega)\big),\beta(\mu,\omega)\Big) \\ &= R\Big(Q^*\big(R(\tau(\mu,\omega),\beta(\mu,\omega)),\beta(\mu,\omega)\big),\beta(\mu,\omega)\Big). \end{aligned}$$

Generalising this to a fast triple bootstrap is not routine!

Bootstrap Iteration

We need a slight extension of our notation in order to be able to discuss double, triple, *etc.*, bootstraps. The original statistic in approximate P value form, $t = \tau(\mu, \omega)$, is also denoted by $p_0(\mu, \omega)$, and its CDF by $R_0(x, \mu)$. The bootstrap P value, denoted till now as $p(\mu, \omega)$, becomes $p_1(\mu, \omega)$, and its CDF, as before, by $R_1(x, \mu)$.

Recall that $p_1(\mu, \omega) = R_0(p_0(\mu, \omega), \beta(\mu, \omega))$. The double bootstrap P value was denoted as $p_2(\mu, \omega)$, and defined as $R_1(p_1(\mu, \omega), \beta(\mu, \omega))$. Let the CDF of $p_2(\mu, \omega)$ be $R_2(x, \mu)$. Then we have the iterative scheme: for $k = 0, 1, 2, \ldots$, we define

$$\begin{aligned} R_k(x, \mu) &= P\big\{\omega \in \Omega \,|\, p_k(\mu, \omega) \leq x\big\}, \\ p_{k+1}(\mu, \omega) &= R_k\big(p_k(\mu, \omega), \beta(\mu, \omega)\big), \end{aligned}$$

where we initialise the recurrence by the definition $p_0(\mu, \omega) = \tau(\mu, \omega)$. Thus $p_{k+1}(\mu, \omega)$ is the bootstrap P value obtained by bootstrapping the k^{th} order P value $p_k(\mu, \omega)$. It estimates the probability mass in the distribution of the k^{th} order P value to the left of its realisation.

The next step of the iteration gives

$$\begin{aligned} R_2(x,\mu) &= P\big\{\omega \in \Omega \,|\, p_2(\mu,\omega) \leq x\big\} \text{ and} \\ p_3(\mu,\omega) &= R_2(p_2(\mu,\omega),\beta(\mu,\omega)) = R_2(R_1(p_1(\mu,\omega),\beta(\mu,\omega)),\beta(\mu,\omega)) \\ &= R_2(R_1(R_0(p_0(\mu,\omega),\beta(\mu,\omega)),\beta(\mu,\omega)),\beta(\mu,\omega)). \end{aligned}$$

Estimating this is very computationally challenging.

The fast double bootstrap treats $\tau(\mu,\omega)$ and $\beta(\mu,\omega)$ as though they were independent. This gives

$$R_1(x,\mu) = \mathrm{E}\Big[\mathrm{E}\big[\mathrm{I}\big(\tau(\mu,\omega) < Q_0(x,\beta(\mu,\omega)\big) \,|\, \beta(\mu,\omega)\big]\Big] = \mathrm{E}\big[R_0\big(Q_0(x,\beta(\mu,\omega)),\mu\big)\big]$$

Of course in general this is just an approximation.

Define the stochastic process τ^1 (new notation for the old τ^*) by the formula

$$\tau^1(\mu, \omega_1, \omega_2) = \tau\big(\beta(\mu, \omega_1), \omega_2\big),$$

ω_1 and ω_2 being independent. Thus $\tau^1(\mu, \omega_1, \omega_2)$ can be thought of as a realisation of the bootstrap statistic when the underlying DGP is μ. We denote the CDF of τ^1 under μ by $R^1(\cdot, \mu)$. In this notation, we saw earlier that

$$\begin{aligned} R^1(x, \mu) &= \mathrm{E}\big[\mathrm{I}\big(\tau(\beta(\mu, \omega_1), \omega_2) < x\big)\big] = \mathrm{E}\Big[\mathrm{E}\big[\mathrm{I}\big(\tau(\beta(\mu, \omega_1), \omega_2) < x\big) \,|\, \omega_1\big]\Big] \\ &= \mathrm{E}\big[R_0\big(x, \beta(\mu, \omega_1)\big)\big]. \end{aligned}$$

Let $Q^1(\cdot, \mu)$ be the quantile function inverse to the CDF $R^1(\cdot, \mu)$. The second approximation underlying the FDB can now be stated as follows:

$$\mathrm{E}\big[R_0\big(Q_0(x, \beta(\mu, \omega)), \mu\big)\big] \approx R_0\big(Q^1(x, \mu), \mu\big),$$

On putting the two approximations together, we obtain

$$R_1(x, \mu) \approx R_0\big(Q^1(x, \mu), \mu\big) \equiv R_1^f(x, \mu).$$

In order to study the distribution of the FDB P value, we wish to evaluate the expression

$$\mathrm{E}\Big[\mathrm{I}\Big(R_0\big(Q^1\big(R_0(\tau(\mu,\omega),\beta(\mu,\omega)),\beta(\mu,\omega)\big),\beta(\mu,\omega)\big)<\alpha\Big)\Big],$$

which is the probability, under the DGP μ, that the FDB P value is less than α. The inequality that is the argument of the indicator above is equivalent to several other inequalities, as follows:

$$\begin{aligned}
& Q^1\big(R_0(\tau(\mu,\omega),\beta(\mu,\omega)),\beta(\mu,\omega)\big) < Q_0(\alpha,\beta(\mu,\omega)) \\
\iff & R_0(\tau(\mu,\omega),\beta(\mu,\omega)) < R^1\big(Q_0(\alpha,\beta(\mu,\omega)),\beta(\mu,\omega)\big) \\
\iff & \tau(\mu,\omega) < Q_0\big(R^1\big(Q_0(\alpha,\beta(\mu,\omega)),\beta(\mu,\omega)\big),\beta(\mu,\omega)\big).
\end{aligned}$$

At this point, we can again invoke an approximation that would be exact if $\tau(\mu,\omega)$ and $\beta(\mu,\omega)$ were independent. The final inequality above separates $\tau(\mu,\omega)$ from $\beta(\mu,\omega)$ on the left- and right-hand sides respectively, and so the expectation of the indicator of that inequality is approximated by

$$\mathrm{E}\Big[R_0\big(Q_0\big(R^1(Q_0(\alpha,\beta(\mu,\omega)),\beta(\mu,\omega)\big),\mu\big)\Big].$$

We can make a further approximation in the spirit of the second of the approximations that lead to the FDB. The approximation can be written as

$$\begin{aligned} &\mathrm{E}\Big[R_0\big(Q_0\big(R^1(Q_0(\alpha,\beta(\mu,\omega)),\beta(\mu,\omega)),\beta(\mu,\omega)\big),\mu\big)\Big] \\ \approx\ &\mathrm{E}\big[R_0\big(Q_0\big(R^1(Q^1(\alpha,\mu),\beta(\mu,\omega)),\beta(\mu,\omega)\big),\mu\big)\big] \end{aligned}$$

Now define the random variable

$$\tau^2(\mu,\omega_1,\omega_2,\omega_3) = \tau(\beta(\beta(\mu,\omega_1),\omega_2),\omega_3),$$

which can be thought of as a realisation of the second-order bootstrap statistic. The CDF of τ^2 under μ, denoted by $R^2(\cdot,\mu)$ is given by

$$\begin{aligned} R^2(\alpha,\mu) &= \mathrm{E}\big[\mathrm{E}\big[\mathrm{I}\big(\tau(\beta(\beta(\mu,\omega_1),\omega_2),\omega_3)<\alpha\big)\,|\,\omega_1,\omega_2\big]\big] \\ &= \mathrm{E}\big[R_0\big(\alpha,\beta(\beta(\mu,\omega_1),\omega_2)\big)\big] \\ &= \mathrm{E}\big[R^1\big(\alpha,\beta(\mu,\omega_1)\big)\big], \end{aligned}$$

where the last equality follows from the definition of R^1.

Now, an argument based on this last result shows that the CDF of the FDB P value is approximately

$$\mathrm{E}\big[R_0\big(Q_0\big(R^2(Q^1(\alpha,\mu),\mu),\beta(\mu,\omega)\big),\mu\big)\big].$$

Finally, another approximation we made earlier shows show that this last expression is, approximately,

$$R_2^f(\alpha,\mu) \equiv R_0\big(Q^1\big(R^2(Q^1(\alpha,\mu),\mu),\mu\big),\mu\big).$$

The point of all these approximations is to replace the argument $\beta(\mu,\omega)$ by μ itself, which allows us to avoid inner loops in the calculations. Estimation of $R_2^f(\alpha,\mu)$ by simulation, for given α and μ, can be done using the following algorithm.

Algorithm FastR2:

1. For each $i = 1, \ldots, N$:
   (i) Generate an independent realisation ω_{i1} from $(\Omega, \mathcal{F}, P)$;
   (ii) Compute a statistic $t_i = \tau(\mu, \omega_{i1})$ and corresponding bootstrap DGP $b_{i1} = \beta(\mu, \omega_{1i})$;
   (iii) Generate a second independent realisation ω_{i2}, a realisation $t_i^1 = \tau(b_{i1}, \omega_{i2})$ of τ^1, and corresponding bootstrap DGP $b_{i2} = \beta(b_{i1}, \omega_{i2})$;
   (iv) Generate a third independent realisation ω_{i3} and subsequently a realisation $t_i^2 = \tau(b_{i2}, \omega_{i3})$ of τ^2.
2. Sort the t_i^1 in increasing order, and form an estimate $\hat{Q}^1(x, \mu)$ as the order statistic of rank $\lceil xN \rceil$.
3. Estimate $R^2\big(Q^1(x, \mu), \mu\big)$ by the proportion of the t_i^2 that are less than $\hat{Q}^1(x, \mu)$. Denote the estimate by $\hat{r}_2$.
4. Estimate $Q^1\big(R^2\big(Q^1(x, \mu), \mu\big), \mu\big)$ as the order statistic of the t_i^1 of rank $\lceil \hat{r}_2 N \rceil$. Denote the estimate by $\hat{q}_1$.
5. Finally, estimate $R_2^f(x, \mu)$ as the proportion of the t_i that are smaller than $\hat{q}_1$.

The theoretical FDB P value is the approximation evaluated with x set equal to the first-level bootstrap P value, and μ replaced by the bootstrap DGP. The theoretical fast triple bootstrap (FTB) P value is formed analogously from $R_2^f(x, \mu)$ by setting x equal to the FDB P value, and again replacing μ by the (first-level) bootstrap DGP, according to the bootstrap principle. The result is

$$p_3^f(\mu, \omega) \equiv R_0\big(Q^1\big(R^2(Q^1(p_2^f(\mu, \omega), \beta(\mu, \omega)), \beta(\mu, \omega)), \beta(\mu, \omega)\big), \beta(\mu, \omega)\big),$$

The simulation estimate, which must be expressed as a function of the observed statistic t and bootstrap DGP b, is

$$\hat{p}_3^f(t, b) = \hat{R}_0\big(\hat{Q}^1\big(\hat{R}^2(\hat{Q}^1(\hat{p}_2^f(t, b), b), b), b\big), b\big).$$

Here is the algorithm for the FTB P value.

Algorithm FTB:

1. From the data set under analysis, compute the realised statistic t and the bootstrap DGP b.
2. Draw B bootstrap samples and compute B bootstrap statistics $t_j^* = \tau(\beta, \omega_j^*)$, $j = 1, \ldots, B$, and B iterated bootstrap DGPs $b_j^* = \beta(b, \omega_j^*)$.
3. Compute B second-level bootstrap statistics $t_j^{1*} = \tau(b_j^*, \omega_j^{**})$, and sort them in increasing order. At the same time, compute the corresponding second-level bootstrap DGPs $b_j^{**} = \beta(b_j^*, \omega_j^{**})$.
4. Compute B third-level bootstrap statistics $t_j^{2*} = \tau(b_j^{**}, \omega_j^{***})$.
5. Compute the estimated first-level bootstrap P value $\hat{p}_1(t, b)$, as the proportion of the t_j^* smaller than t.
6. Obtain the estimate $\hat{Q}^{1*} \equiv \hat{Q}^1\big(\hat{p}_1(t, b), b\big)$ as the order statistic of the t_j^{1*} of rank $\lceil B\hat{p}_1(t, b) \rceil$.
7. Compute the estimated FDB P value $\hat{p}_2^f(t, b)$ as the proportion of the t_j^* smaller than $\hat{Q}^{1*}$.

8. Compute $\hat{Q}^{1**} \equiv \hat{Q}^1\big(\hat{p}_2^f(t,b), b\big)$ as the order statistic of the t_j^{1*} that is of rank $\lceil B\hat{p}_2^f(t,b)\rceil$.
9. Compute $\hat{R}^{2*} \equiv \hat{R}^2\big(\hat{Q}^1(\hat{p}_2^f(t,b), b), b\big)$ as the proportion of the t_j^{2*} smaller than $\hat{Q}^{1**}$.
10. Compute $\hat{Q}^{1***} \equiv \hat{Q}^1\big(\hat{R}^2\big(\hat{Q}^1(\hat{p}_2^f(t,b), b), b\big), b\big)$ as the order statistic of the t_j^{1*} of rank $\lceil r\hat{R}^{2*}\rceil$.
11. Compute $\hat{p}_3^f(t,b)$ as the proportion of the t_j^* smaller than $\hat{Q}^{1***}$.

Although this looks complicated, it is in fact easy to program, and, as we will see in the simulation experiments, it can give rise to a useful improvement in the reliability of inference.

The ideas that lead to the FDB and FTB P values can obviously be extended to higher orders. For the FDB, we approximate the distribution of the first-level bootstrap P value $p_1(\mu,\omega)$, and evaluate it at the computed first-level P value $p_1(t,b)$ and the bootstrap DGP b. For the FTB, we approximate the distribution of the FDB P value $p_2^f(\mu,\omega)$ and evaluate it at the computed FDB P value $p_2^f(t,b)$ and b. For a fast quadruple bootstrap, we wish to approximate the distribution of the FTB P value $p_3^f(\mu,\omega)$ and evaluate it at the computed FTB P value $p_3^f(t,b)$ and b. And so on.

Illustrations with simulation experiments

Testing for a unit root

The model studied in this section may be summarised as follows:

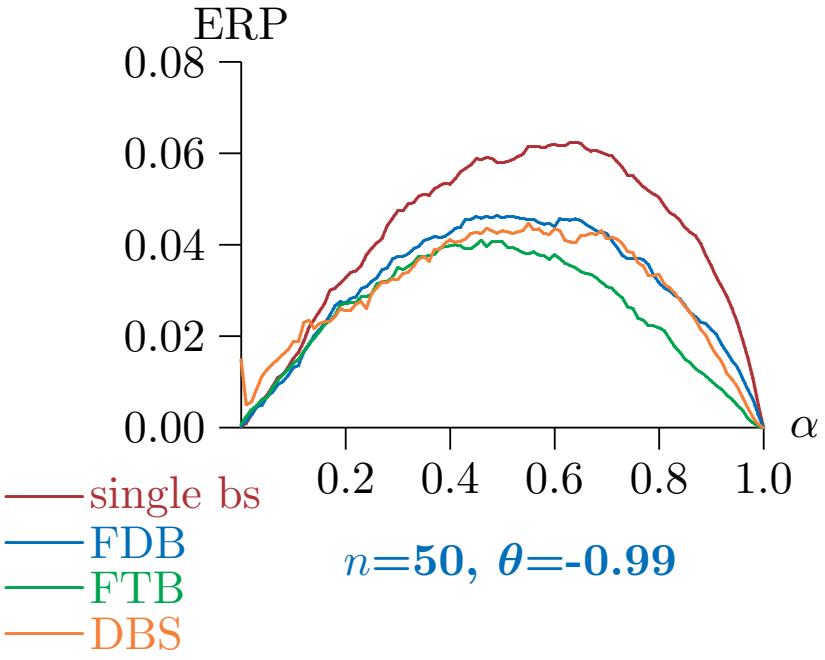
$$y_t = \rho y_{t-1} + v_t \tag{6}$$

$$v_t = u_t + \theta u_{t-1}, \qquad u_t \sim \text{NID}(0, \sigma^2), \qquad t = 1, \ldots, n. \tag{7}$$

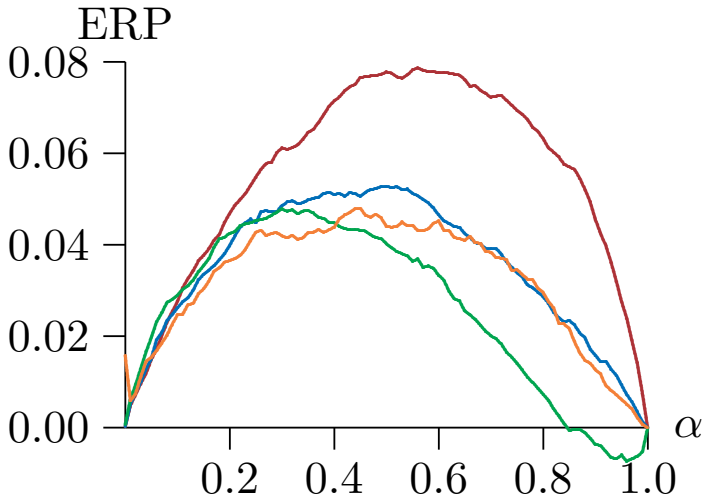
The observed series is y_t, and the null hypothesis of a unit root sets $\rho = 1$. Under that hypothesis, $v_t = \Delta y_t$, where Δ is the first-difference operator. We may write (7) in vector notation using the lag operator L, as follows:

$$\boldsymbol{v} = (1 + \theta \text{L})\boldsymbol{u}, \quad \text{or} \quad \boldsymbol{v} = R(\text{L})\boldsymbol{v} + \boldsymbol{u},$$

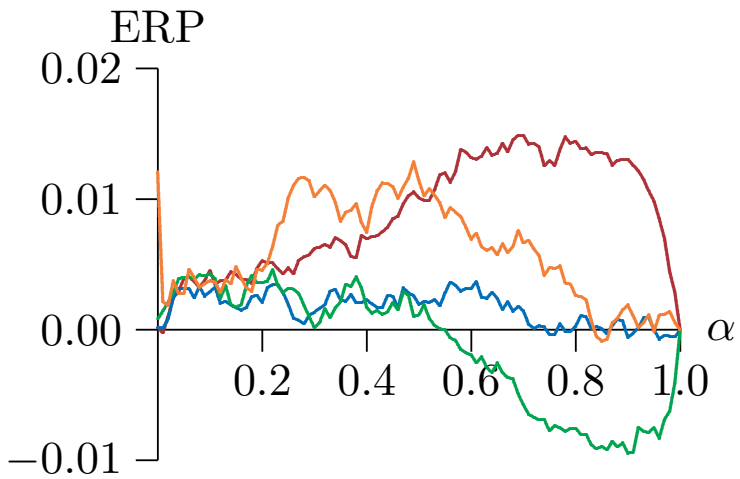
where we define $R(\text{L}) = \theta(1+\theta\text{L})^{-1}\text{L}$. The parameter θ is an MA parameter affecting the innovations to the unit-root process. If $\theta = -1$, then the MA component cancels out the unit root to leave a white-noise process. Thus near $\theta = -1$, we can expect serious size distortions of any unit root test.



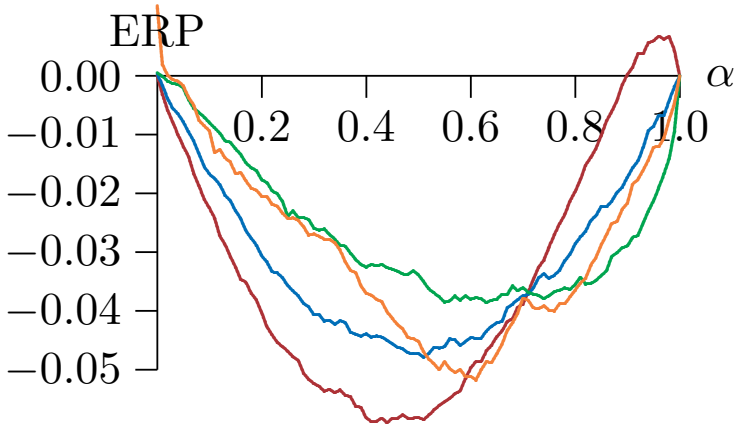
n=**50,** θ=**-0.99**



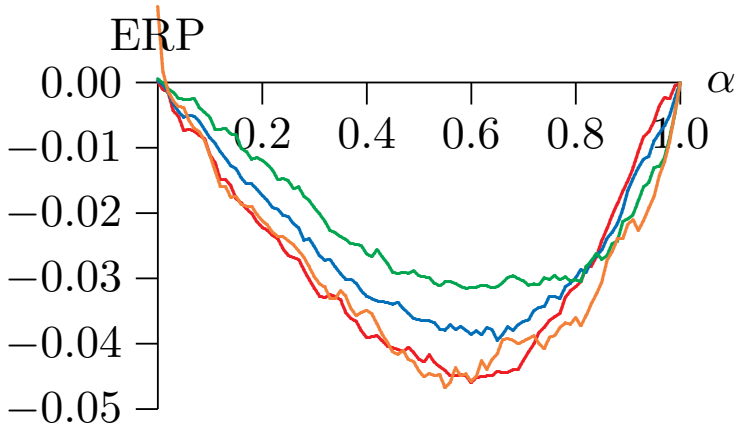
n=**100,** θ=**-0.99**



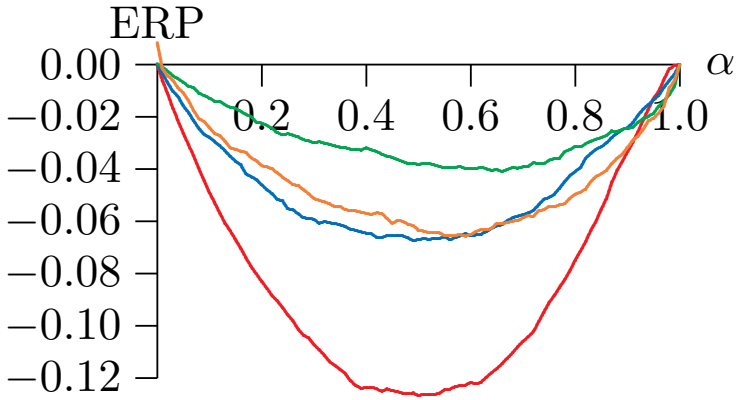
n=**50,** θ=**-0.95**



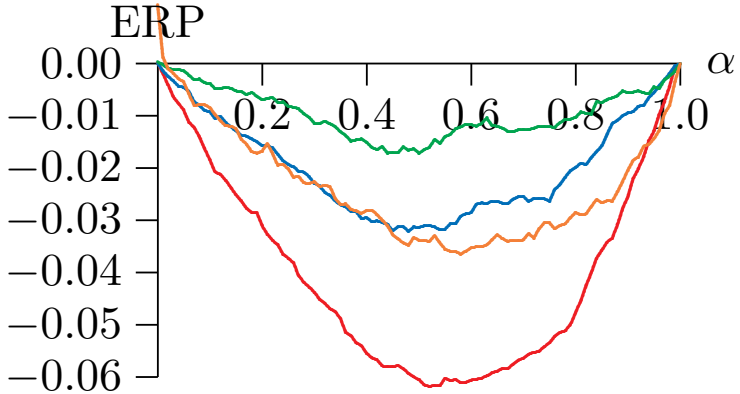
n=**100,** θ=**-0.95**



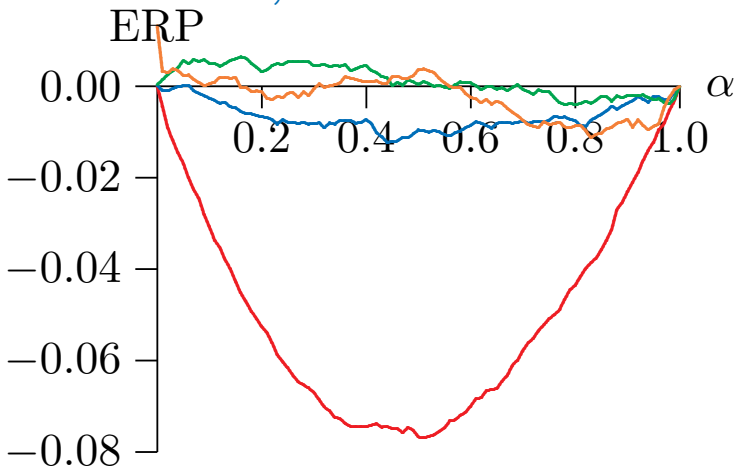
n=**50,** θ=**-0.90**



n=**100,** θ=**-0.90**



n=**50,** θ=**-0.80**



n=**100,** θ=**-0.80**

A test for ARCH

Consider the linear regression model

$$\begin{aligned} y_t &= \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t = \sigma_t \varepsilon_t, \quad t = 1, \ldots, n, \\ \sigma_t^2 &= \sigma^2 + \gamma u_{t-1}^2 + \delta \sigma_{t-1}^2, \quad \varepsilon_t \sim \text{IID}(0, 1). \end{aligned} \tag{8}$$

The disturbances of this model follow a GARCH(1,1) process. The easiest way to test the null hypothesis that the u_t are IID in the model (8) is to run the regression

$$\hat{u}_t^2 = b_0 + b_1 \hat{u}_{t-1}^2 + \text{residual}, \tag{9}$$

where $\hat{u}_t$ is the t^{th} residual from an OLS regression of y_t on $\boldsymbol{X}_t$. The null hypothesis that $\gamma = \delta = 0$ can be tested by testing the hypothesis that $b_1 = 0$. Besides the ordinary t statistic for b_1, a commonly used statistic is n times the centred R^2 of the regression, which has a limiting asymptotic distribution of χ_1^2 under the null hypothesis.

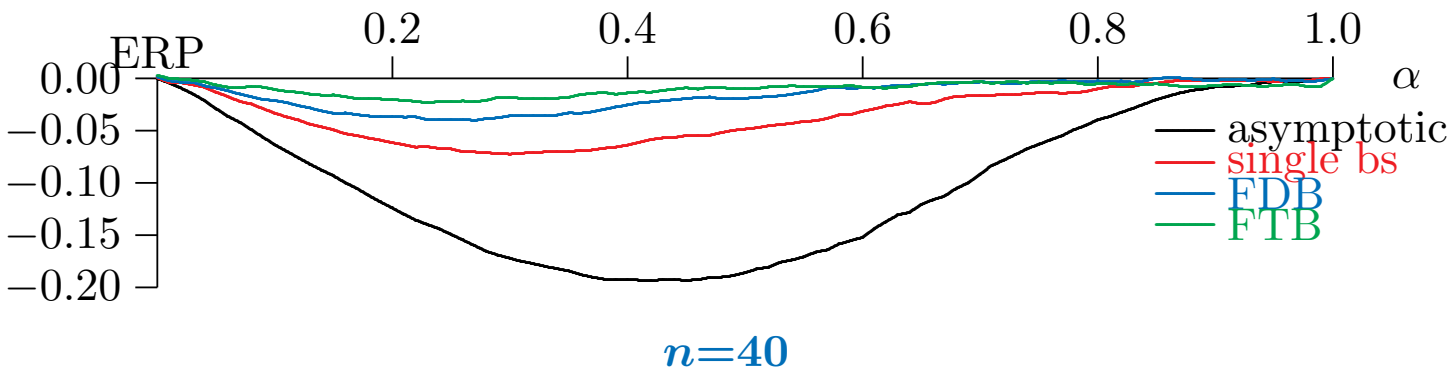
Since in general one is unwilling to make any restrictive assumptions about the distribution of the ε_t, a resampling bootstrap seems the best choice. It is of course of interest to see to what extent the theory of fast iterated bootstraps can be used effectively with resampling.

Without loss of generality, we set $\boldsymbol{\beta} = \mathbf{0}$ and $\sigma^2 = 1$ in the bootstrap DGP, since the test statistic is invariant to changes in the values of these parameters. The invariance means that we can use as bootstrap DGP the following:

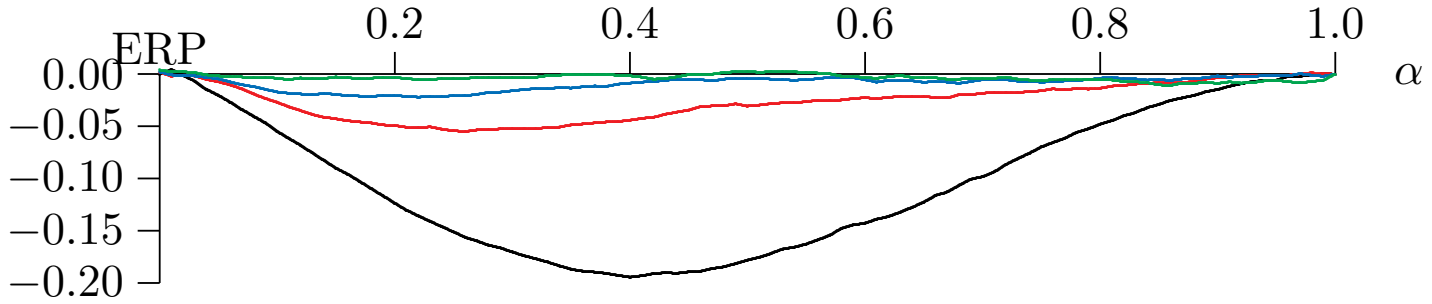
$$y_t^* = u_t^*, \quad u_t^* \sim \text{EDF}(y_t),$$

where the notation EDF (for "empirical distribution function") means simply that the bootstrap data are resampled from the original data. For iterated bootstraps, y_t^{**} is resampled from the y_t^*, and y_t^{***} is resampled from the y_t^{**}.

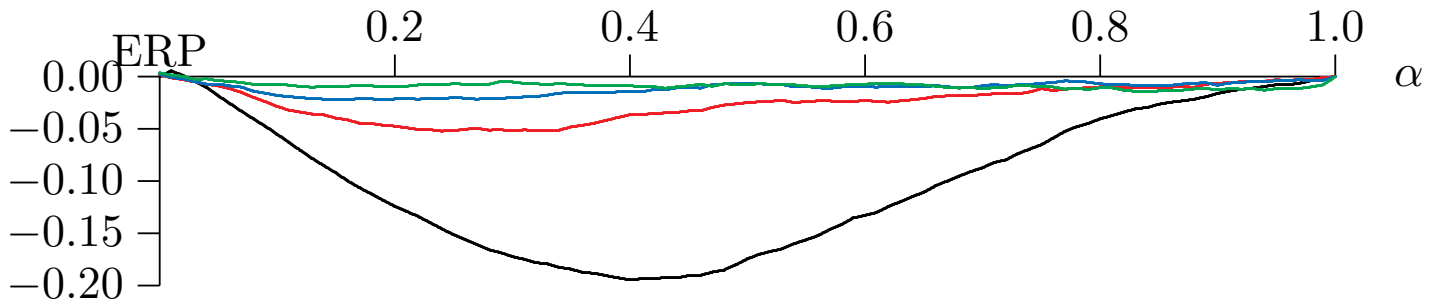
The next page shows results under the null, and after that the following page shows results under the alternative, with $\alpha = 1$, $\gamma = \delta = 0.3$.


ERP
0.00
−0.05
−0.10
−0.15
−0.20
0.2
0.4
0.6
0.8
1.0
α
asymptotic
single bs
EDB
FTB

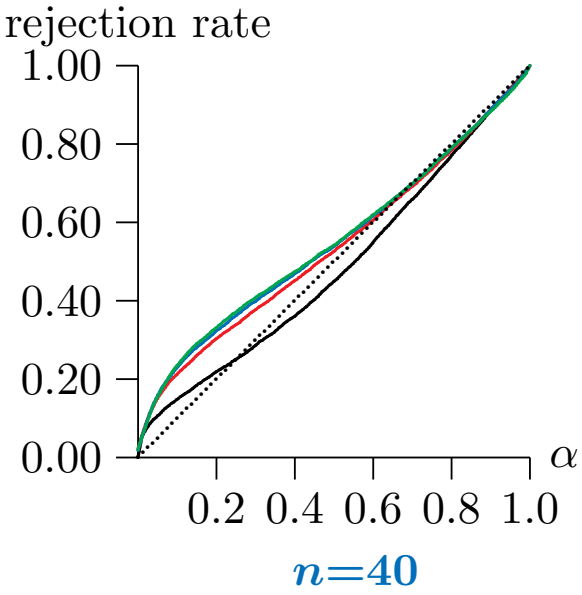
n=40


ERP
0.00
−0.05
−0.10
−0.15
−0.20
0.2
0.4
0.6
0.8
1.0
α

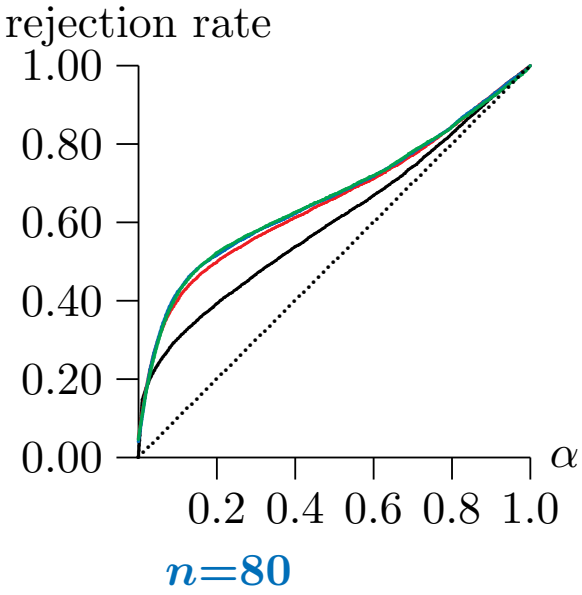
n=80


ERP
0.00
−0.05
−0.10
−0.15
−0.20
0.2
0.4
0.6
0.8
1.0
α

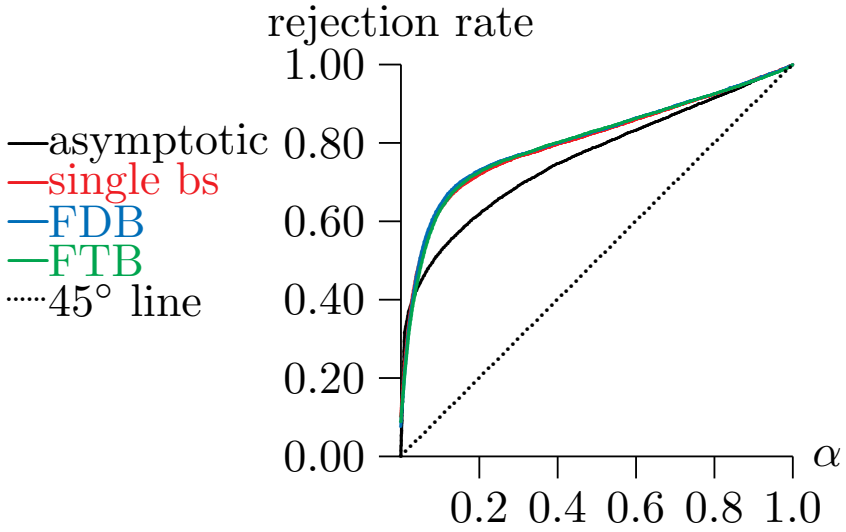
n=160


rejection rate
1.00
0.80
0.60
0.40
0.20
0.00
0.2 0.4 0.6 0.8 1.0
α

n=40


rejection rate
1.00
0.80
0.60
0.40
0.20
0.00
0.2 0.4 0.6 0.8 1.0
α

n=80


rejection rate
asymptotic
single bs
FDB
FTB
45° line
1.00
0.80
0.60
0.40
0.20
0.00
0.2 0.4 0.6 0.8 1.0
α

n=160

A test for serial correlation

Another of the examples of the good performance of the FDB found in Davidson and MacKinnon (2007) is given by the Durbin-Godfrey test for serial correlation of the disturbances in a linear regression model. The model that serves as the alternative hypothesis for the test is the linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \gamma y_{t-1} + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma^2), \quad t = 1, \ldots, n, \tag{10}$$

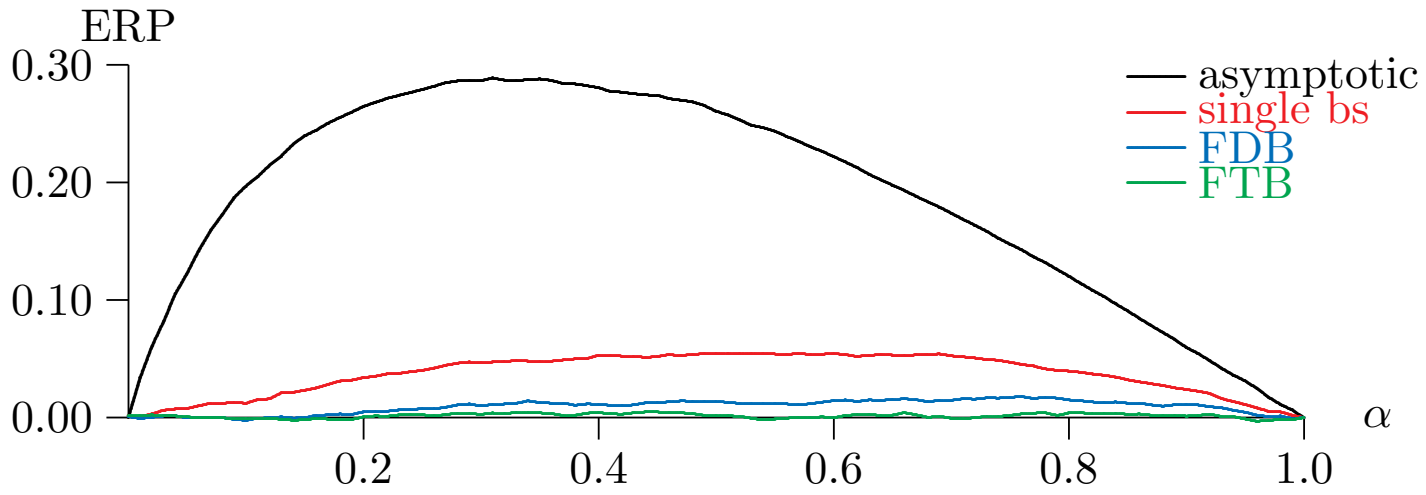
where $\boldsymbol{X}_t$ is a $1 \times k$ vector of observations on exogenous variables. The null hypothesis is that $\rho = 0$. Let the OLS residuals from running regression (10) be denoted $\hat{u}_t$. Then the Durbin-Godfrey (DG) test statistic is the t statistic for $\hat{u}_{t-1}$ in a regression of y_t on $\boldsymbol{X}_t$, y_{t-1}, and $\hat{u}_{t-1}$. It is asymptotically distributed as $\text{N}(0, 1)$ under the null hypothesis.

For the bootstrap DGP, from running regression (10), we obtain estimates $\tilde{\boldsymbol{\beta}}$, $\tilde{\gamma}$, as well as the residuals $\tilde{u}_t$. The semiparametric bootstrap DGP can be written as

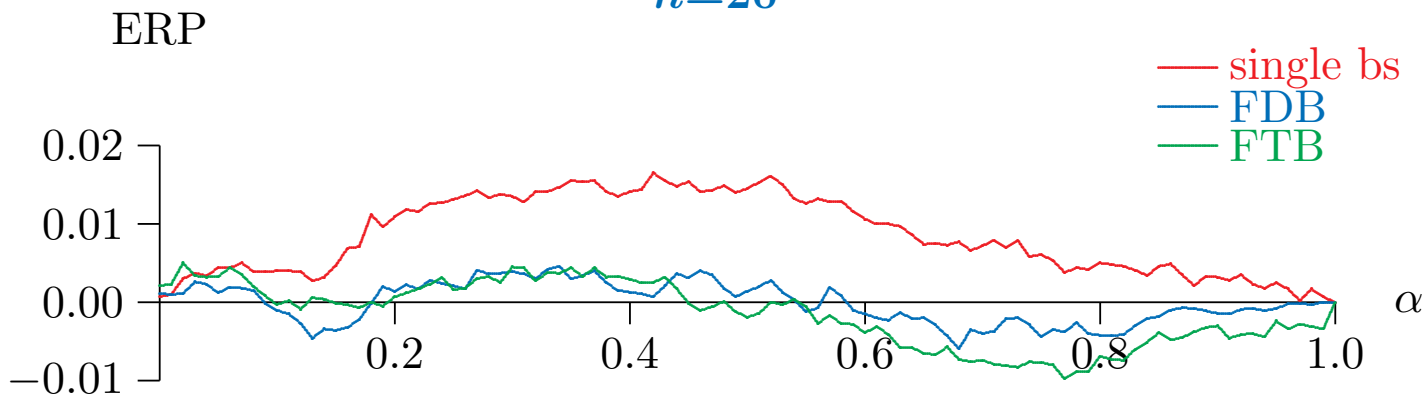
$$y_t^* = \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} + \tilde{\gamma} y_{t-1}^* + u_t^*. \tag{11}$$

The u_t^* are obtained by resampling the rescaled residuals $(n/(n-k-1))^{1/2}\tilde{u}_t$. The initial value y_0^* is set equal to the actual pre-sample value y_0.

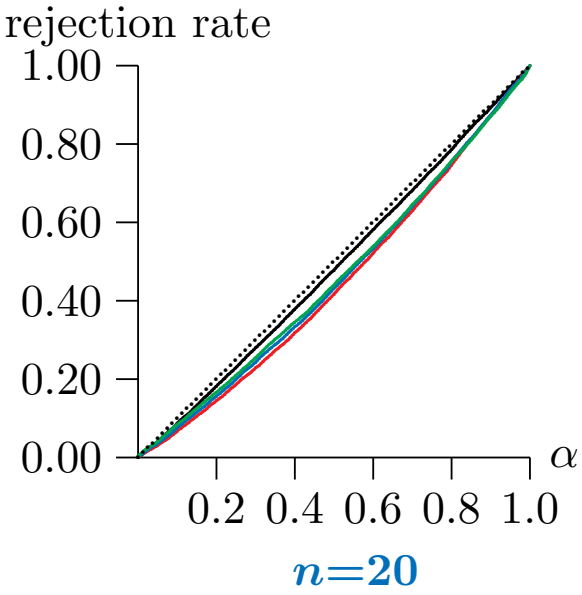
We again give results for both size and power.


ERP
0.30
0.20
0.10
0.00
0.2
0.4
0.6
0.8
1.0
α
asymptotic
single bs
FDB
FTB

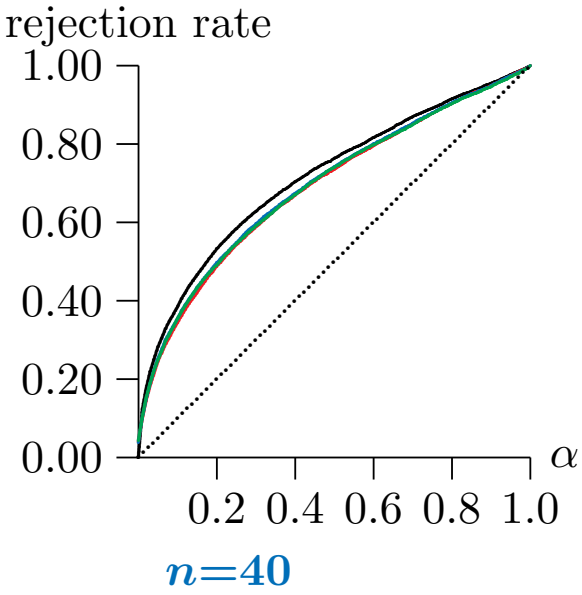
n=20


ERP
0.02
0.01
0.00
−0.01
0.2
0.4
0.6
0.8
1.0
α
single bs
FDB
FTB

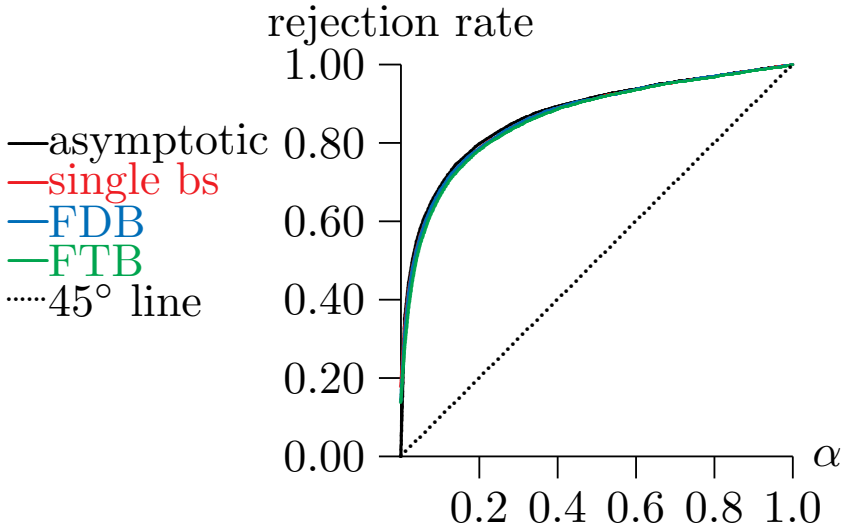
n=40


rejection rate
1.00
0.80
0.60
0.40
0.20
0.00
0.2 0.4 0.6 0.8 1.0
α

n=20


rejection rate
1.00
0.80
0.60
0.40
0.20
0.00
0.2 0.4 0.6 0.8 1.0
α

n=40


rejection rate
asymptotic
single bs
FDB
FTB
45° line
1.00
0.80
0.60
0.40
0.20
0.00
0.2 0.4 0.6 0.8 1.0
α

n=56