Diagnostics for the Bootstrap and Fast Double Bootstrap

by

Russell Davidson

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

AMSE-GREQAM
Centre de la Vieille Charité
2 Rue de la Charité
13236 Marseille cedex 02, France

russell.davidson@mcgill.ca

Abstract

The bootstrap is typically much less reliable in the context of time-series models with serial correlation of unknown form than it is when regularity conditions for the conventional IID bootstrap, based on resampling, apply. It is therefore useful for practitioners to have available diagnostic techniques capable of evaluating bootstrap performance in specific cases. The techniques suggested in this paper are closely related to the fast double bootstrap, and, although they inevitably rely on simulation, they are not computationally intensive. They can also be used to gauge the performance of the fast double bootstrap itself. Examples of bootstrapping time series are presented which illustrate the diagnostic procedures, and show how the results can cast light on bootstrap performance.

Keywords: Bootstrap, fast double bootstrap, diagnostics for bootstrap, time series, autocorrelation of unknown form

JEL codes: C10, C15, C22, C63

This research was supported by the Canada Research Chair program (Chair in Economics, McGill University) and by a grant from the Fonds de Recherche du Québec - Société et Culture.

November 2014

1. Introduction

While the bootstrap can provide spectacularly reliable inference in many cases, there are others for which results are much less reliable. Intuition can often suggest reasons for this state of affairs, and the asymptotic theory of bootstrap refinements does so as well; see Hall (1992) and Horowitz (1997) among many other relevant references.

It has often been remarked that heavy-tailed distributions give rise to difficulties for the bootstrap; see Davidson (2012) and the discussion of that paper in Schluter (2012). Autocorrelation of unknown form also presents a severe challenge to the bootstrap. So far, no bootstrap has been proposed that, in the presence of autocorrelation of unknown form, can deliver performance comparable to what can be obtained in its absence. Perhaps in consequence, a considerable number of bootstrap methods have been proposed, some a good deal better than others. By far the most popular of these are the various versions of the block bootstrap, which was originally proposed by Künsch (1989). However, it has been seen that the block bootstrap often works poorly, while, in some circumstances, other schemes may work better. These include (versions of) the sieve bootstrap, frequency-domain bootstraps, and the recently-proposed dependent wild bootstrap.

Simulation experiments can of course be used to study the performance of different bootstrap procedures in different circumstances. In this paper, simulation-based diagnostic methods are proposed, intended to determine when a given procedure works well or not, and, if not, provide an analysis of why. Asymptotic theory, including the theory of bootstrap refinements characterised by a rate at which the bootstrap discrepancy tends to zero, is not very useful for this purpose. One obvious reason is that the bootstrap is a finite-sample procedure, not an asymptotic one. To be useful, therefore, a diagnostic technique should be based on finite-sample arguments only.

Despite the rapidly growing power of computing machinery, it would be more useful for practitioners if a diagnostic technique was no more CPU-intensive, or at least very little more intensive, than simply undertaking a bootstrap test or constructing a bootstrap confidence set. The techniques outlined here satisfy that requirement, although simulations are performed that are more CPU-intensive, for the purpose of evaluating the reliability of the diagnostic methods themselves.

The paper is organised as follows. In Section 2, definitions and notation appropriate for theoretical study of the bootstrap are given. The wild bootstrap is presented in Section 3, and its use in the context of a regression model with disturbances that follow an AR(1) process studied. It turns out that the wild bootstrap is capable of giving essentially perfect inference even with very small samples, and so, in Section 4, once the diagnostic methods are explained, they are applied to this setup and the results illustrated graphically. Section 5 looks at an interesting failure, namely the maximum-entropy bootstrap proposed in Vinod (2006). There is nothing wrong, and much right, with the maximum-entropy idea, but its application to time series with autocorrelation of unknown form fails to yield reliable inference. The reason for this disappointing fact is clearly revealed by the diagnostic analysis.

There is a close link between the principle underlying the diagnostics and that underlying the fast double bootstrap of Davidson and MacKinnon (2007). This is brought out in Section 6, where it is seen that, at a cost of some increase in CPU time, the fast double bootstrap itself can be diagnosed. In Section 7, a very simple and special case of a version of the sieve bootstrap is considered. Its performance and that of its fast double counterpart are diagnosed, as well as a procedure combining the sieve bootstrap and the wild bootstrap. Finally, some concluding remarks are presented in Section 8.

2. Definitions and Notation

A model is a collection of data-generating processes (DGPs). If $\mathbb{M}$ denotes a model, it may also represent a hypothesis, namely that the true DGP, μ say, belongs to $\mathbb{M}$. Alternatively, we say that $\mathbb{M}$ is correctly specified.

We almost always want to define a parameter-defining mapping θ, which 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, as, for instance, 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 that θ exists 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 any DGP $\mu \in \mathbb{M}$ is the same for all $\mu \in \mathbb{M}$.

We can denote by $\mathbb{M}_0$ the set of DGPs that represent 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. I assume as usual that $\mathbb{M}_0 \subset \mathbb{M}$.

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

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

Leaving aside questions of just what real-world randomness – if it exists – might be, we can take the probability space Ω to be that of a random number generator. A realisation of the test statistic is written as $\tau(\mu, \omega)$, for some $\mu \in \mathbb{M}$ and $\omega \in \Omega$.

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 leads to rejection when the statistic is too small. Let $R_0 : [0,1] \times \mathbb{M}_0 \to [0,1]$ be the cumulative distribution function (CDF) of τ under any DGP $\mu \in \mathbb{M}_0$:

$$R_0(x, \mu) = P\bigl\{\omega \in \Omega \,|\, \tau(\mu, \omega) \le x\bigr\}. \tag{1}$$

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 by t. Then the ideal P value that would give exact inference is $R_0(t, \mu_0)$. If t is indeed generated by μ_0, $R_0(t, \mu_0)$ is distributed as U(0,1) if the distribution of τ is absolutely continuous with respect to Lebesgue measure – as we assume throughout – but not, in general, if t comes from some other DGP. The quantity $R_0(t, \mu_0)$ 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.

The principle of the bootstrap is that, when we want to use some function or functional of an unknown DGP μ_0, we use the same function or functional of an estimate of μ_0. 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 $R_0(t, b)$, where t and b are observed, or rather can be computed from the observed data. In terms of the two stochastic processes τ and β, the bootstrap P value is another stochastic process:

$$p_1(\mu, \omega) = R_0\bigl(\tau(\mu, \omega), \beta(\mu, \omega)\bigr). \tag{2}$$

Normally, the bootstrap principle must be implemented by a simulation experiment, and so, analogously to (1), we may define

$$\hat{R}_0(x, \mu) = \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}\bigl(\tau(\mu, \omega_j^*) < x\bigr),$$

where the ω_j^* are independent realisations of the random numbers needed to compute the statistic. As the number of bootstrap repetitions $B \to \infty$, $\hat{R}_0(x, \mu)$ tends almost surely to $R_0(x, \mu)$. Accordingly, the bootstrap P value is estimated by $\hat{R}_0(t, b)$.

Since by absolute continuity R_0 is a continuous function, it follows that p_1 also has an absolutely continuous distribution. We denote the continuous CDF of $p_1(\mu, \omega)$ by $R_1(\cdot, \mu)$. This CDF can also be estimated by simulation, but that is very computationally intensive. The double bootstrap uses this approach, using the bootstrap principle by replacing the unknown true DGP μ by the bootstrap DGP b. An ideal double bootstrap P value that would give exact inference is $R_1\bigl(p_1(\mu, \omega), \mu\bigr)$, which is distributed as U(0,1). The double bootstrap P value is, analogously, $R_1\bigl(R_0(t, b), b\bigr)$.

3. The Wild Bootstrap

Models that incorporate heteroskedasticity can be bootstrapped effectively by use of the wild bootstrap. Early references to this procedure include Wu (1986), Liu (1988), and Mammen (1993). For the linear regression

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \tag{3}$$

the wild bootstrap DGP can be written as

$$\boldsymbol{y}^* = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{u}^*,$$

where, as usual, stars denote simulated quantities, and $\tilde{\boldsymbol{\beta}}$ is a vector of restricted estimates that satisfy the possibly nonlinear null hypothesis under test. The bootstrap disturbances are defined by $u_t^* = |\hat{u}_t| s_t^*$, where $\hat{u}_t$ is the residual for observation t obtained by estimating the restricted model, and the s_t^* are IID drawings from a distribution such that $\mathrm{E}(s_t^*) = 0$, $\mathrm{Var}(s_t^*) = 1$.

Davidson and Flachaire (2008) recommend the Rademacher distribution, defined as follows:

$$s_t^* = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2, \end{cases} \tag{4}$$

for the wild bootstrap. When the Rademacher distribution is used, the covariance structure of the squared bootstrap disturbances is the same as that of the squared residuals from the original sample. This is because the squared bootstrap disturbances are always just the squared residuals, so that any relationship among the squared residuals, like that given by any GARCH model, is preserved unchanged by the Rademacher wild bootstrap.

In order to study the consequences of this fact for a simple GARCH model, a simulation experiment was conducted for the model

$$y_t = a + \rho y_{t-1} + u_t, \tag{5}$$

where u_t are GARCH(1,1) disturbances, defined by the recurrence relation

$$\begin{aligned} \sigma_t^2 &= \alpha + (\delta + \gamma\varepsilon_{t-1}^2)\sigma_{t-1}^2; \\ u_t &= \sigma_t \varepsilon_t, \end{aligned} \tag{6}$$

with the ε_t standard normal white noise, and the recurrence initialised by $\sigma_1^2 = \alpha/(1 - \gamma - \delta)$, which is the unconditional stationary expectation of the process. The parameters of the DGP used in the experiment were $a = 1.5$, $y_0 = 0$, $\alpha = 1$, $\gamma = 0.4$, and $\delta = 0.45$, with sample sizes $n = 10, 30, 50$, and $\rho = 0.3, 0.5, 0.7$ and 0.9. In order to test the hypothesis that $\rho = \rho_0$, the test statistic used was

$$\tau = \frac{\hat{\rho} - \rho_0}{\hat{\sigma}_\rho},$$

where $\hat{\rho}$ is the OLS estimate from (5), run over observations 2 to n. The standard error $\hat{\sigma}_\rho$ was obtained by use of the HC_2 variant of the Eicker-White HCCME; see White (1980) and Eicker (1963).

The bootstrap DGP is determined by first running the constrained regression

$$y_t - \rho_0 y_{t-1} = a + u_t, \quad t = 2, \ldots, n,$$

in order to obtain the estimate $\tilde{a}$, and the constrained residuals $\tilde{u}_t$, $t = 2, \ldots n$. A bootstrap sample is defined by

$$y_1^* = y_1 \quad \text{and} \quad y_t^* = \tilde{a} + \rho_0 y_{t-1}^* + s_t^* \tilde{u}_t, \quad t = 2, \ldots, n,$$

where the s_t^* are IID realisations from the Rademacher distribution. The bootstrap statistics are

$$\tau_j^* = \frac{\hat{\rho}^* - \rho_0}{\hat{\sigma}_\rho^*}, \quad j = 1, \ldots, B$$

with $\hat{\rho}^*$ and $\hat{\sigma}_\rho^*$ defined as the bootstrap counterparts of $\hat{\rho}$ and $\hat{\sigma}_\rho$ respectively. The bootstrap P value is the proportion of the τ_j^* that are more extreme than τ. The performance of the bootstrap test, as revealed by experiments with $N = 100{,}000$ replications with $B = 199$ bootstrap samples for each, is excellent. This will be seen in the context of the diagnostic procedures presented in the next section, in preference to presenting results here in tabular form.

Davidson and Flachaire (2008) show that there is a special setup where the wild bootstrap can deliver perfect inference. If one wishes to test the hypothesis that the entire vector $\boldsymbol{\beta}$ in the linear regression (3) is zero when the disturbances $\boldsymbol{u}$ may be heteroskedastic, the obvious test statistic is

$$\tau \equiv \boldsymbol{y}^\top \boldsymbol{X} (\boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \tag{7}$$

where the dependent variable $\boldsymbol{y}$ is the vector of restricted residuals under the null hypothesis, and $\hat{\boldsymbol{\Omega}}$ is one of the inconsistent estimates of the covariance matrix of the disturbances used in the HCCME. When the Rademacher distribution (4) is used, the wild bootstrap P value is uniformly distributed under the null up to discreteness due to a finite sample size.

In what follows, simulation results are presented for a variety of different bootstrap procedures that are found in the literature, with a setup similar to the above. The model is a linear regression with disturbances that are possibly serially correlated as well as heteroskedastic, with null hypothesis that all the regression parameters are zero, and a test statistic with the form of (7), but with a HAC covariance matrix estimator instead of the HCCME. It serves as a useful test bed, as it allows us to compare the performance of these bootstrap tests with the perfect inference obtainable with only heteroskedasticity and the wild bootstrap.

4. Diagnostic Procedures

A standard way of evaluating bootstrap performance by simulation is to graph the P value and P value discrepancy plots for a test based on the bootstrap P value. The former is just a plot of the CDF of this P value; the latter a plot of the CDF minus its argument. Perfect inference appears as a P value plot that coincides with the diagonal of the unit square, or a P value discrepancy plot that coincides with the horizontal axis, because there is perfect inference when the P value is uniformly distributed on the [0,1] interval.

If the bootstrap discrepancy, that is, the ordinate of the P value discrepancy plot, is acceptably small, there is no need to look further. But, if not, it is useful to see why, and it is for this purpose that we may use the procedures of this section. A simulation experiment that provides the information for a P value plot also provides the information needed for these. Suppose that there are N replications in the experiment, with B bootstrap repetitions for each replication. The data for each replication are generated using a DGP denoted by μ, which satisfies the null hypothesis that $\boldsymbol{\beta} = \mathbf{0}$, and with a chosen specification of the joint distribution of the random elements needed to generate a bootstrap sample. Let τ_j, $j = 1, \ldots, N$, be the IID realisations of the test statistic (7), and let τ_j^*, $j = 1, \ldots, N$, be a single bootstrap statistic taken from the B bootstrap statistics computed for replication j, the first perhaps, or the last, or one chosen at random.

The next step is to graph kernel-density estimates of the distribution of the statistic τ and that of the bootstrap statistic τ^*. If these are not similar, then clearly the bootstrap DGP fails to mimic the true DGP at all well. Bootstrap failure is then a consequence of this fact. Another diagnostic is based on running an OLS regression of the τ_j^* on a constant and the τ_j. Suppose that this regression reveals that, for the data generated on one replication, τ is strongly positively correlated with τ^*, and suppose without loss of generality that τ itself is in nominal P value form, so that the rejection region is on the left. The bootstrap P value for replication j is

$$P = \frac{1}{B} \sum_{i=1}^{B} \mathrm{I}(\tau_{ji}^* < \tau_j), \tag{8}$$

where the τ_{ji}^*, $i = 1, \ldots, B$, are the bootstrap statistics computed for replication j, and $\mathrm{I}(\cdot)$ is the indicator function. The positive correlation then implies that, if τ_j is small, then the τ_{ji}^* tend to be small as well. It follows from (8) that the P value is greater than it would be in the absence of the correlation, and that the bootstrap tests under-rejects. Similarly, on the right-hand side of the distribution of the P value, there is more probability mass than there would be with no or smaller correlation. A similar argument shows that, *mutatis mutandis*, a negative correlation leads to over-rejection.

The presence or otherwise of a significant correlation is related to the extent of bootstrap refinements. An argument borrowed from Davidson and MacKinnon (2006) can help shed light on this point. The argument assumes that the distribution of τ is absolutely continuous for any DGP that satisfies the null hypothesis. Under DGP μ, the CDF of τ, which is supposed to be in approximate P value form, is denoted by $R_0(\cdot, \mu)$, and the inverse quantile function by $Q_0(\cdot, \mu)$. A bootstrap test based on τ rejects at nominal level α if $\tau < Q_0(\alpha, \mu^*)$, where μ^* denotes the bootstrap DGP, or, equivalently, if $R_0(\tau, \mu^*) < \alpha$.

Let the random variable p be defined as $p = R_0(\tau, \mu)$. Since $R_0(\cdot, \mu)$ is the CDF of τ under μ, p is distributed as $\mathrm{U}(0, 1)$. Further, for a given α, define the random variable q as $q = R_0\big(Q_0(\alpha, \mu^*), \mu\big) - \alpha$, so that q is just the difference in the rejection probabilities under μ according to whether the bootstrap critical value or the true critical value for μ is used. These variables allow another representation of the rejection event: $p < \alpha + q$.

Let $F(q \,|\, p)$ denote the CDF of q conditional on p. The rejection probability (RP) of the bootstrap test at nominal significance level α under μ is then

$$\Pr{}_{\mu}(p < \alpha + q) = \mathrm{E}_{\mu}\big(\mathrm{I}(q > p - \alpha) \,|\, p\big) = \mathrm{E}_{\mu}\big(1 - F(p - \alpha \,|\, p)\big) = 1 - \int_0^1 F(p - \alpha \,|\, p)\,\mathrm{d}p.$$

On integrating by parts and changing variables, we find that the RP of the bootstrap test is

$$\int_0^1 p\,\mathrm{d}F(p - \alpha \,|\, p) = \int_{-\alpha}^{1-\alpha} (x + \alpha)\,\mathrm{d}F(x \,|\, \alpha + x) = \alpha + \int_{-\alpha}^{1-\alpha} x\,\mathrm{d}F(x \,|\, \alpha + x). \tag{9}$$

The integral in the rightmost expression above is the bootstrap discrepancy.

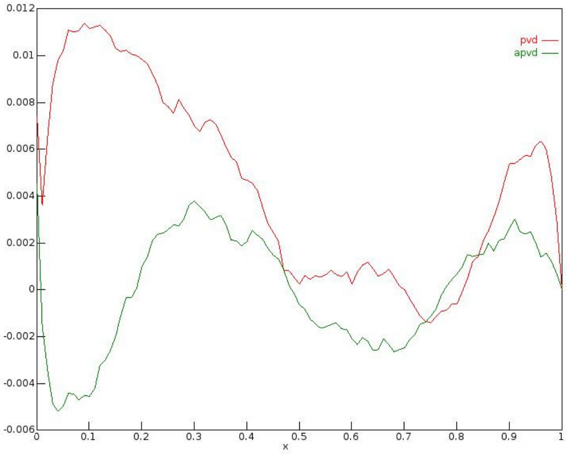
If we use an asymptotic construction such that μ^* converges to μ as $n \to \infty$, then q tends to zero asymptotically, the conditional CDF corresponds to a degenerate distribution at zero, and the bootstrap discrepancy vanishes. The usual criterion for the (asymptotic) validity of the bootstrap is that this happens for all $\alpha \in [0, 1]$.

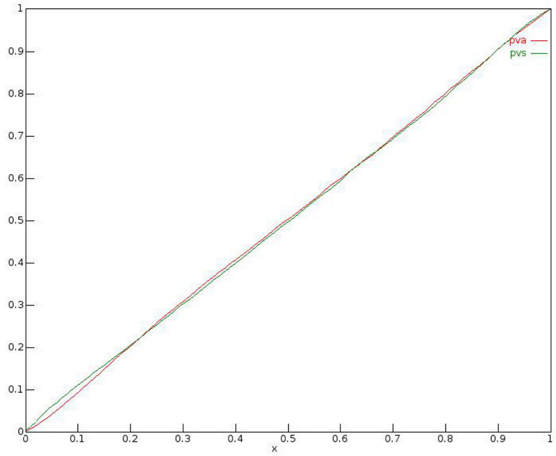
The bootstrap discrepancy in (9) can be interpreted as the expectation of q conditional on the bootstrap P value being equal to α, that is, being at the margin between rejection and non-rejection at level α. The random variable p is random through the statistic τ, while q is random only through the bootstrap DGP μ^*. If p and q were independent, then the τ_j and the τ_j^* of the simulation experiment would also be independent, and so uncorrelated. Independence is unlikely to hold exactly in finite samples, but it often holds asymptotically, and so presumably approximately, in finite samples.

When p and q are approximately independent, the conditional expectation of q is close to the unconditional expectation, which is not in general zero. Conventional bootstrap refinements arise when the unconditional expectation tends to zero sufficiently fast as the sample size grows. The conditional expectation can be expected to tend to zero more slowly than the unconditional expectation, except when there is near-independence, in which case there is a further refinement; see Davidson and MacKinnon (1999). The comparison of the densities of τ and τ^* reveals a non-zero unconditional expectation of q in the form of a shift of the densities, while a significant correlation reveals a failure of the condition for the second refinement.

The Model with GARCH(1,1) disturbances

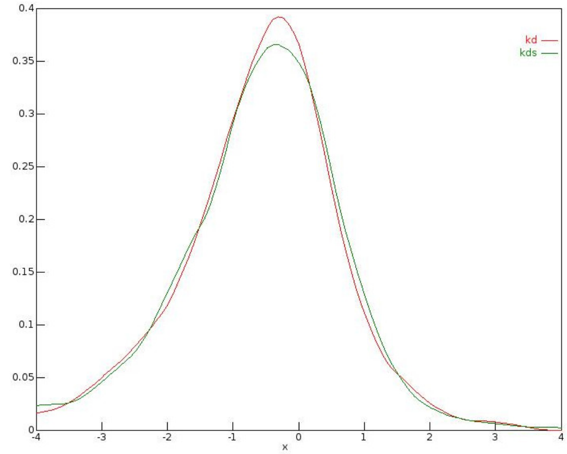
As a first example of the diagnostic tests, results are given here for the test of $\rho = \rho_0$ in the model specified by (5) and (6), with $\rho = 0.3$ and $n = 10$. First, the P value and P value discrepancy plots. They appear below. Since the test has only one degree of freedom, it was possible to look separately at a one-tailed test that rejects to the right and the two-tailed test. The curves in red are for a two-tailed test; those in green for a one-tailed test that rejects to the right. It can be seen that use of a two-tailed test confers no significant advantage.

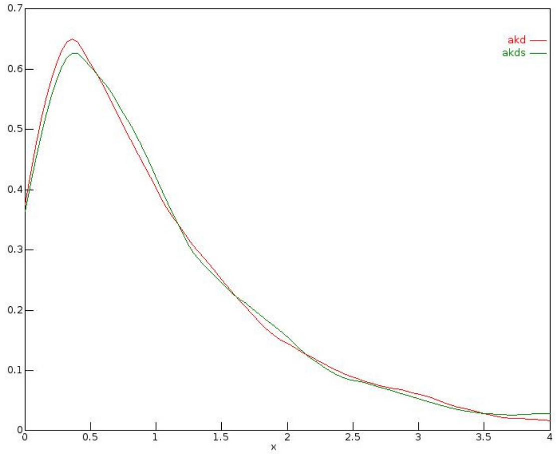




It is reasonable to claim that the discrepancy is acceptably small, even though it does not seem to be exactly zero. For $\alpha = 0.05$, its simulated value is 0.010 for the one-tailed test, and -0.005 for the two-tailed test.

Next the results of the diagnostic procedure. Below are plotted the kernel density estimates of the distributions of the statistic and the bootstrap statistic for both cases.





For the one-tailed test, the regression of the bootstrap statistic τ^* on a constant and τ gives (standard errors in parentheses)

$$\underset{(0.005)}{\tau^* = -0.640} + \underset{(0.003)}{0.0003\tau}, \qquad \text{centred } R^2 = 7 \times 10^{-8}$$

so that the constant is highly significant, but the coefficient of τ is completely insignificant. For the two-tailed test, the result is

$$\tau^* = \underset{(0.005)}{1.023} + \underset{(0.003)}{0.044\tau}, \qquad \text{centred } R^2 = 0.002$$

Here, both estimated coefficients are significant, although the overall fit of the regression is very slight. The negative constant for the one-tailed test means that the distribution of the bootstrap statistic is to the left of that of τ, leading to over-rejection since the test rejects to the right. Similarly the positive constant for the two-tailed test explains the under-rejection for interesting values of α.

5. An Interesting Failure: the Maximum-Entropy Bootstrap

The principle of maximum entropy was propounded by Jaynes (1957) as an interpretation of statistical mechanics that treats the problems of thermodynamics as problems of statistical inference on the basis of extremely limited information. One application of the principle was proposed by Theil and Laitinen (1980), for the estimation, from a random IID sample, of the density of the underlying distribution, under the assumption that the distribution is continuous and is almost everywhere differentiable. For a brief discussion of the method, see the more accessible Fiebig Denzil and Theil (1982). For a sample of size n, with order statistics $x_{(i)}$, $i = 1, \ldots, n$, the estimated distribution has, except in the tails, a continuous piecewise linear CDF that assigns probability mass $1/n$ to each interval $I_i \equiv [(x_{(i-1)} + x_{(i)}/2, (x_{(i)} + x_{(i+1)})/2]$, for $i = 2, \ldots, n-1$. The distribution is exponential in the tails, defined as the intervals I_1 from $-\infty$ to $(x_{(1)} + x_{(2)})/2$, and I_n from $(x_{(n-1)} + x_{n)})/2$ to $+\infty$. Each of the infinite intervals receives a probability mass of $1/n$, and the lower interval is constructed to have an expectation of $0.75x_{(1)} + 0.25x_{(2)}$, the upper an expectation of $0.25x_{(n-1)} + 0.75x_{(n)}$.

This way of estimating a distribution was picked by Vinod (2006), who bases a technique for bootstrapping time series on it. He modifies the procedure described above so as to allow for the possibility of a bounded rather than an infinite support, but I cannot follow the details of his discussion. Aside from this, his method proceeds as follows:

1. Define an $n \times 2$ sorting matrix S_1 and place the index set $T_0 = \{1, 2, \ldots, n\}$ in the first column and the observed time series x_t in the second column.
2. Sort the matrix S_1 with respect to the numbers in its second column while carrying along the numbers in the first column. This yields the order statistics

$x_{(i)}$ in the second column and a vector $I_{\rm rev}$ of sorted T_0 in the first column. From the $x_{(i)}$ construct the intervals I_i defined above.

3. Denote by $\hat{F}$ the CDF of the maximum-entropy distribution defined above. Generate n random numbers p_i, $i = 1, \ldots, n$ distributed uniformly on $[0, 1]$. Obtain a resample x_i^* as the p_i quantiles of $\hat{F}$, $i = 1, \ldots, n$.

5. Define another $n \times 2$ sorting matrix S_2. Sort the x_i^* in increasing order and place the result in column 1 of S_2. Place the vector $I_{\rm rev}$ in column 2.

6. Sort the S_2 matrix with respect to the second column to restore the order $\{1, 2, \ldots, n\}$ there. Redefine the x_i^* as the elements of the jointly sorted column 1 of S_2.

The idea is clearly to preserve as much of the correlation structure of the original series as possible. It is a pity that Vinod went on directly to apply his method to real data, as it turns out that altogether too many of the specific properties of the original series are retained in each bootstrap sample, so that there is not enough variability in the bootstrap DGP.

I have documented this method in full because, although it does not work, it shows up a number of interesting things. First, resampling from the continuous distribution $\hat{F}$ can very well be employed instead of resampling from the discrete empirical distribution. Rescaling, and other operations that specify higher moments, can easily be incorporated into the maximum entropy algorithm. Although in most cases one may expect there to be little difference relative to conventional resampling, there are situations in which it may be necessary to impose the continuity of the bootstrap distribution.

The other reason for my dwelling on this method is that the diagnostic procedures show clearly what is wrong with it. Consider the following model, which I will use as a test case for this and other bootstrapping methods.

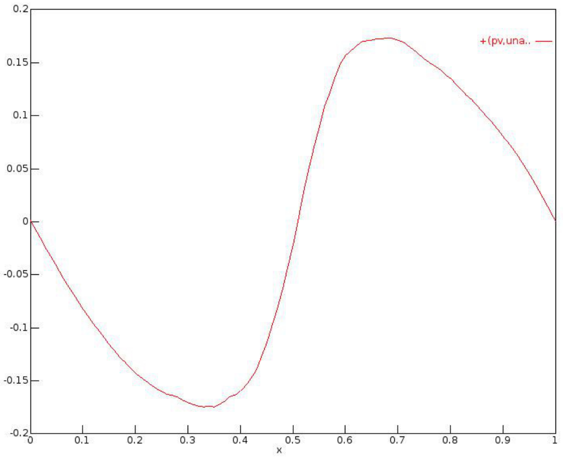
$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \quad u_t = \rho u_{t-1} + v_t. \tag{10}$$

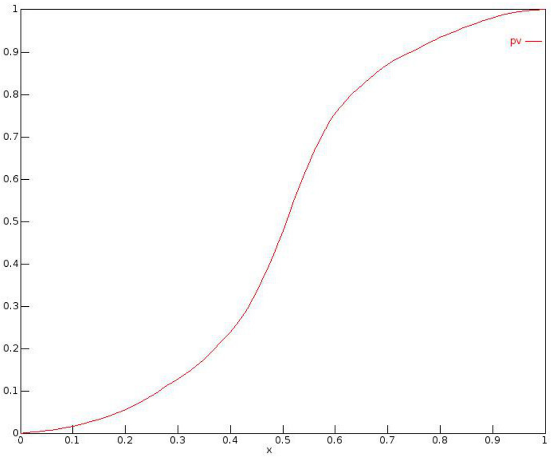
The regressor matrix $\boldsymbol{X}$ includes a constant and three other variables, constructed so that they are serially correlated with autocorrelation coefficient ρ_1. The disturbances follow an AR(1) process. The null hypothesis is that the full coefficient vector $\boldsymbol{\beta} = \boldsymbol{0}$; just as in the case of the exact result with the wild bootstrap with heteroskedasticity only. The test statistic is the asymptotic chi-squared statistic, with four degrees of freedom:

$$\tau = \boldsymbol{y}^\top \boldsymbol{X} (\boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \tag{11}$$

where $\hat{\boldsymbol{\Omega}}$ is the well-known Newey-West HAC covariance matrix estimator based on the Bartlett kernel; see Newey and West (1987).

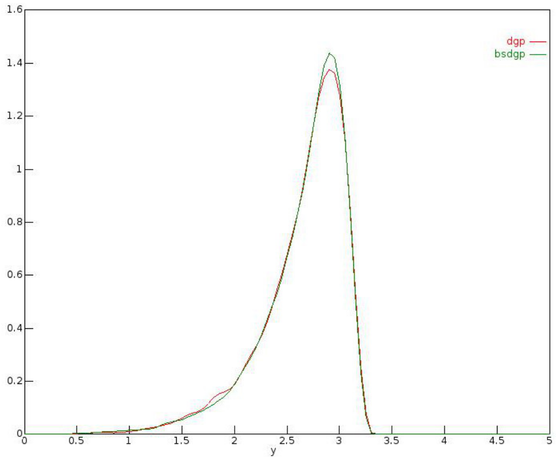
Below are the P value discrepancy and P value plots for $n = 50$, $\rho = 0.9$, $\rho_1 = 0.8$, and a lag-truncation parameter $p = 20$ for $\hat{\boldsymbol{\Omega}}$. There are $9,999$ replications with 399 bootstrap repetitions each.





It is quite clear that something is badly wrong! There is severe under-rejection for small α, and equally severe over-rejection for large α. There are at least two possible reasons for this. The first is that, if the distribution of the bootstrap statistic is on average more dispersed than that of the statistic itself, then the mass in the bootstrap distribution to the right of τ is too great for large values of τ, so that the P value is too small, leading to over-rejection, and it is too small when τ is small, so that the the P value is too great, leading to under-rejection. A second possible explanation is that, for each replication, the bootstrap statistics are strongly positively correlated with τ. In that event, when τ is large, the bootstrap distribution is shifted right, and conversely.

Below is presented the kernel density plots of the statistics τ and τ^*; for τ in red, for τ^* in green. The distributions are clearly almost identical, thus ruling out the first possible explanation.



The regression of τ^* on τ gave (standard errors in parentheses):

$$\underset{(0.016)}{\tau^* = 0.508} + \underset{(0.006)}{0.810\tau}, \qquad \text{centred } R^2 = 0.662$$

Both coefficients are highly significant. Thus this is clear evidence of the second possible explanation: τ^* is indeed strongly positively correlated with τ. What this shows is that the attempt to make the bootstrapped time series mimic the real series is too successful, and so there is too little variation in the bootstrap distribution.

6. The Fast Approximation

The idea behind the diagnostic procedures discussed here is closely related to the fast double bootstrap (FDB) of Davidson and MacKinnon (2007). It is convenient at this point to review the FDB.

As a stochastic process, the bootstrap P value can be written as $p_1(\mu,\omega)$, as in (2). The double bootstrap bootstraps this bootstrap P value, as follows: If $R_1(\cdot,\mu)$ is the CDF of $p_1(\mu,\omega)$, then the random variable $R_1\big(p_1(\mu,\omega),\mu\big)$ follows the U(0,1) distribution. Since μ is unknown in practice, the double bootstrap P value follows the bootstrap principle by replacing it by the bootstrap DGP, $\beta(\mu,\omega)$. We define the stochastic process

$$p_2(\mu,\omega) = R_1\big(p_1(\mu,\omega),\beta(\mu,\omega)\big). \tag{12}$$

Of course it is computationally expensive to estimate the CDF R_1 by simulation, as it involves two nested loops.

Davidson and MacKinnon (2007) suggested a much less expensive way of estimating R_1, based on two approximations. The first arises by treating the random variables $\tau(\mu,\omega)$ and $\beta(\mu,\omega)$, for any $\mu \in \mathbb{M}_0$, as independent. Of course, this independence does not hold except in special circumstances, but it holds asymptotically in many commonly encountered situations. By definition,

$$R_1(\alpha,\mu) = P\big\{\omega\in\Omega \,|\, p_1(\mu,\omega) < \alpha\big\} = \mathrm{E}\Big[\mathrm{I}\big(R_0(\tau(\mu,\omega),\beta(\mu,\omega)) < \alpha\big)\Big]. \tag{13}$$

Let $Q_0(\cdot,\mu)$ be the quantile function corresponding to the distribution $R_0(\cdot,\mu)$. Since R_0 is absolutely continuous, we have

$$R_0\big(Q_0(\alpha,\mu),\mu\big) = \alpha = Q_0\big(R_0(\alpha,\mu),\mu\big).$$

Use of this relation between R_0 and Q_0 lets us write (13) as

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

If $\tau(\mu,\omega)$ and $\beta(\mu,\omega)$ are treated as though they were independent, then we have

$$\begin{aligned} R_1(\alpha,\mu) &= \mathrm{E}\Big[\mathrm{E}\big[\mathrm{I}\big(\tau(\mu,\omega) < Q_0(\alpha,\beta(\mu,\omega))\big) \,|\, \beta(\mu,\omega)\big]\Big] \\ &\approx \mathrm{E}\big[R_0\big(Q_0(\alpha,\beta(\mu,\omega)),\mu\big)\big] \end{aligned} \tag{14}$$

Define the stochastic process

$$\tau^1 \;:\; \mathbb{M} \times (\Omega_1 \times \Omega_2) \to \mathbb{R},$$

where Ω_1 and Ω_2 are two copies of the outcome space, by the formula

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

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 $R^1(\cdot, \mu)$. Thus

$$\begin{aligned} R^1(\alpha, \mu) &= \Pr\big\{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 \,|\, \tau\big(\beta(\mu, \omega_1), \omega_2\big) < \alpha\big\} \\ &= \mathrm{E}\big[\mathrm{I}\big(\tau(\beta(\mu, \omega_1), \omega_2) < \alpha\big)\big] \\ &= \mathrm{E}\Big[\mathrm{E}\big[\mathrm{I}\big(\tau(\beta(\mu, \omega_1), \omega_2) < \alpha\big) \,|\, \mathcal{F}_1\big]\Big] \\ &= \mathrm{E}\big[R_0\big(\alpha, \beta(\mu, \omega_1)\big)\big]. \end{aligned} \tag{15}$$

Here $\mathcal{F}_1$ denotes the sigma-algebra generated by functions of ω_1.

The second approximation underlying the fast method can now be stated as follows:

$$\mathrm{E}\big[R_0\big(Q_0(\alpha, \beta(\mu, \omega)), \mu\big)\big] \approx R_0\big(Q^1(\alpha, \mu), \mu\big), \tag{16}$$

where $Q^1(\cdot, \mu)$ is the quantile function inverse to the CDF $R^1(\cdot, \mu)$. Since by definition $R^1\big(Q^1(\alpha, \mu), \mu\big) = \alpha$, it follows from (15) that

$$\mathrm{E}\big[R_0\big(Q^1(\alpha, \mu), \beta(\mu, \omega)\big)\big] = \alpha. \tag{17}$$

In order to motivate the approximation (16), we follow Davidson and MacKinnon (2007), and suppose that, for any DGP $\mu \in \mathbb{M}_0$ and for all $\alpha \in [0, 1]$, $R_0(\alpha, \mu) - \alpha$ is small in some appropriate sense. In other words, suppose that τ is expressed as an approximate P value, and is approximately pivotal with respect to $\mathbb{M}_0$. Next, assume that R_0 is not only continuous but also continuously differentiable with respect to its first argument α for all $\mu \in \mathbb{M}_0$. Thus the statistic τ has a continuous density for all $\mu \in \mathbb{M}_0$. Finally, we assume that $R_0'(\alpha, \mu) - 1$, where R_0' denotes the derivative of R_0 with respect to its first argument, is small in the same sense as that in which $R_0(\alpha, \mu) - \alpha$ is small.

The assumption about the derivative R_0' implies that $Q_0(\alpha, \mu) - \alpha$ is small for $\mu \in \mathbb{M}_0$. The definition (15) implies that $R^1(\alpha, \mu) - \alpha$ is small, and so also $Q^1(\alpha, \mu) - \alpha$. Now (17) can be written as

$$\mathrm{E}\big[R_0\big(Q^1(\alpha, \mu), \beta(\mu, \omega)\big) - R_0\big(Q_0(\alpha, \beta(\mu, \omega), \beta(\mu, \omega)\big)\big] = 0,$$

and our assumption about the derivative of R_0, along with Taylor's Theorem, lets us rewrite this equation as

$$\mathrm{E}\big[(1 + \eta_1)\big(Q^1(\alpha, \mu) - Q_0(\alpha, \beta(\mu, \omega)\big)\big] = 0, \tag{18}$$

where the random variable η_1 is small. Further applications of our smallness assumptions give us

$$Q^1(\alpha,\mu) - Q_0(\alpha,\beta(\mu,\omega)) = \alpha - \alpha + \eta_2$$

where η_2 is another small random variable. Thus (18) becomes

$$\mathrm{E}\bigl[Q_0(\alpha,\beta(\mu,\omega))\bigr] = Q^1(\alpha,\mu) + \mathrm{E}(\eta_1\eta_2), \tag{19}$$

so that the expectation of $Q_0\bigl(\alpha,\beta(\mu,\omega)\bigr)$ is equal to $Q^1(\alpha,\mu)$ up to an error of the second order of small quantities.

The difference between the left- and right-hand sides of (16) is

$$\begin{aligned} &\mathrm{E}\bigl[R_0\bigl(Q_0(\alpha,\beta(\mu,\omega)),\mu\bigr) - R_0\bigl(Q^1(\alpha,\mu),\mu\bigr)\bigr] \\ &= \mathrm{E}\bigl[(1+\eta_3)\bigl(Q_0(\alpha,\beta(\mu,\omega)) - Q^1(\alpha,\mu)\bigr)\bigr], \end{aligned}$$

where η_3 is small. By (19) the last expression above is a sum of products of two small quantities, thus justifying the approximation (16).

On putting the two approximations, (14) and (16), together, we obtain

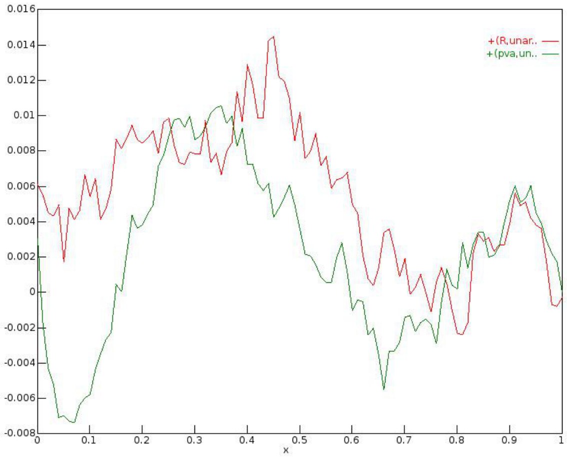
$$R_1(\alpha,\mu) \approx R_0\bigl(Q^1(\alpha,\mu),\mu\bigr) \equiv R_1^f(\alpha,\mu).$$

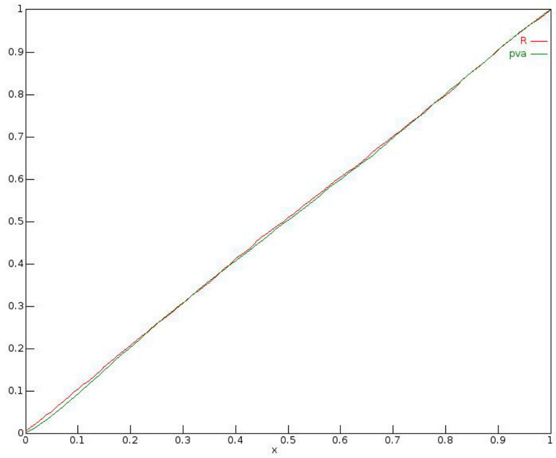
The fast double bootstrap substitutes R_1^f for R_1 in the double bootstrap P value (12). The FDB P value is therefore

$$p_2^f(\mu,\omega) = R_1^f\bigl(p_1(\mu,\omega),\beta(\mu,\omega)\bigr) = R_0\bigl(Q^1(p_1(\mu,\omega),\beta(\mu,\omega)),\beta(\mu,\omega)\bigr). \tag{20}$$

Estimating it by simulation involves only one loop.

A way to see to what extent the FDB may help improve reliability is to compare the (estimated) distribution of the bootstrap P value and the fast approximation to that distribution. The graphs below perform this comparison for the wild bootstrap applied to the model of Section 4. On the right are plotted the distribution estimated directly (in green) and the fast approximation (in red). On the left the same thing, but in deviations from the uniform distribution.





Such differences as are visible are clearly within the simulation noise of the experiment.

7. The Sieve Bootstrap

The sieve bootstrap most commonly used with time series when there is serial correlation of unknown form 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) process, to generate bootstrap samples. For example, suppose we are concerned with the static linear regression model (3), 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

$$\hat{u}_t = \sum_{i=1}^{p} \rho_i \hat{u}_{t-i} + \varepsilon_t \tag{21}$$

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 probably be taken to ensure that the estimated model is stationary. This may require the use of full maximum likelihood to estimate (21), rather than least squares.

Estimation of (21) 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 from running (3). 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:

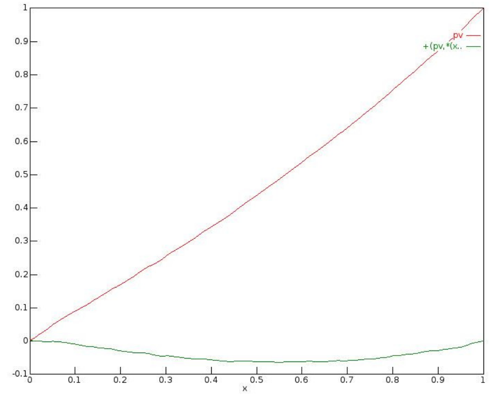
$$u_t^* = \sum_{i=1}^{p} \hat{\rho}_i u_{t-i}^* + \varepsilon_t^*, \quad t = p+1, \ldots, n, \tag{22}$$

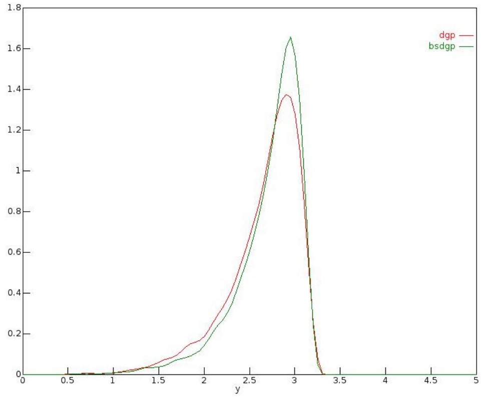
where the ε_t^* can either be drawn from the $\mathrm{N}(0, \hat{\sigma}_\varepsilon^2)$ distribution for a parametric bootstrap or resampled from the residuals $\hat{\varepsilon}_t$ from the estimation of (21), preferably rescaled by the factor $\sqrt{n/(n-p)}$. Before we can use (22), of course, we 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. We initialize (22) 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 characterized 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 initialize (22) 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 to initialize (22). In this way, we approximate a stationary process with the correct estimated stationary covariance matrix, but with no assumption of normality.

We again consider our test-bed case, with model (10) and test statistic (11). As the disturbances are AR(1), we set $p = 1$ in the AR(p) model estimated with the OLS residuals. It would have been possible, and, arguably, better to let P be chosen in a data-driven way. Otherwise, the setup is identical to that used with the maximum-entropy bootstrap. The graphs following show the P value and P value discrepancy plots, on the left, and the kernel density plots on the right.



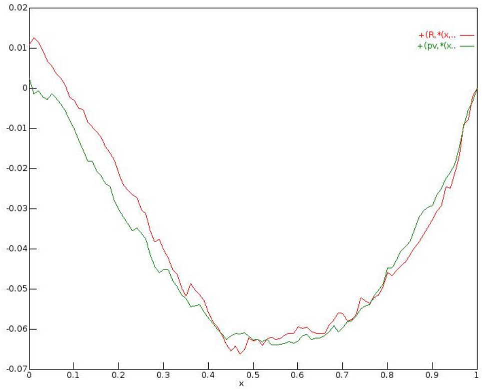


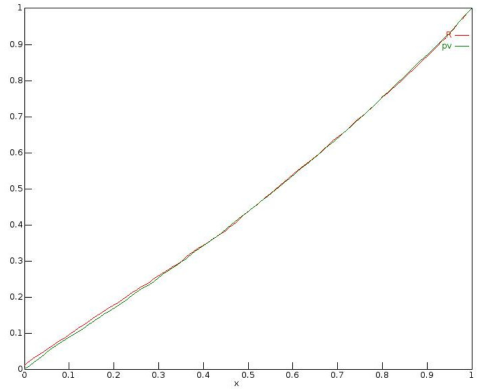
Overall, the bootstrap test performs well, although there is significant distortion in the middle of the distribution of the P value. In the left-hand tail, on the other hand, there is very little. The distributions of τ and τ^* are very similar.

The regression of τ^* on τ gave (OLS standard errors in parentheses):

$$\underset{}{\tau^*} = \underset{(0.025)}{2.55} + \underset{(0.009)}{0.06\tau}, \qquad \text{centred } R^2 = 0.004$$

The next graphs show the comparison between the P value plot as estimated directly by simulation (in green) and as estimated using the fast method (in red) on the right, and that for the P value discrepancy plot, on the left.



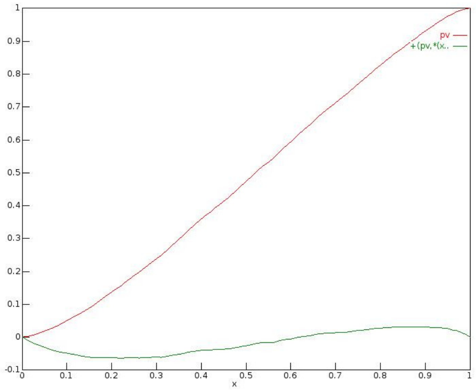


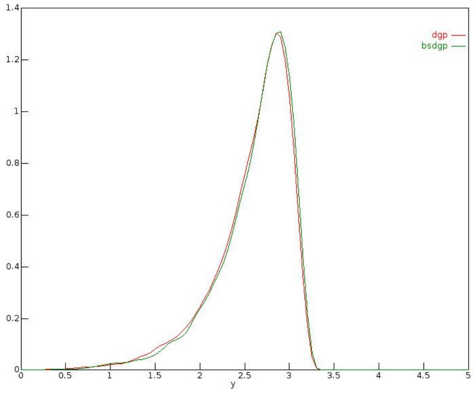
It is clear that the fast method gives results very close indeed to those obtained by direct simulation, and this suggests that the FDB could improve performance substantially. This is also supported by the insignificant correlation between τ and τ^*, and the fact that there is no visible shift in the density plot for τ and that for τ^*, although the significant positive constant in the regression shows that τ^* tends to be greater than τ, which accounts for the under-rejection in the middle of the distribution.

The sieve bootstrap as described so far takes no account of heteroskedasticity. It is interesting, therefore, to see whether it performs well when combined with the wild bootstrap. For that purpose, equation (22) is replaced by

$$u_t^* = \sum_{i=1}^{p} \hat{\rho}_i u_{t-i}^* + s_t^* \hat{\varepsilon}_t,$$

where $\hat{\varepsilon}_t$ is the residual from (21), and the s_t^* are IID drawings from the Rademacher distribution. Below are shown results of the diagnostic procedure for a simulation experiment in which the disturbances are scaled by one of the regressors.





Heteroskedasticity, it appears, can be handled by the wild bootstrap in this context as well. However, the regression of τ^* on τ gave:

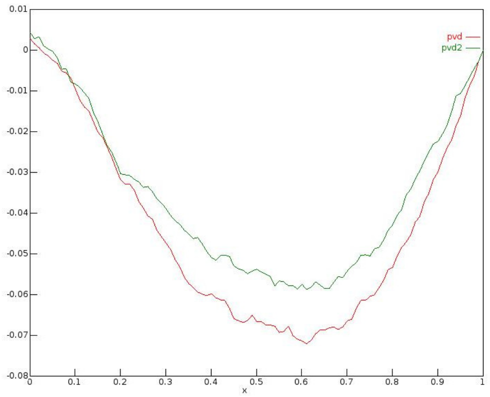
$$\underset{}{\tau^*} = \underset{(0.025)}{2.00} + \underset{(0.010)}{0.24\tau}, \qquad \text{centred } R^2 = 0.058$$

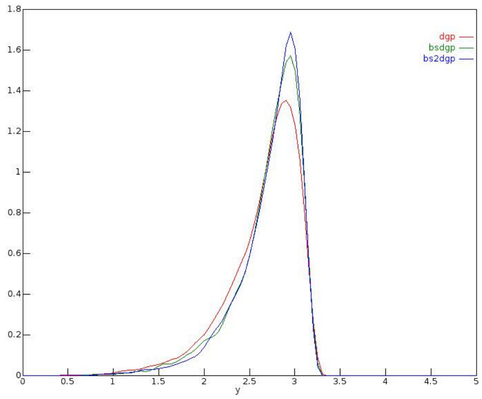
This time, there is significant correlation, although the distributions of τ and τ^* are at least as similar as in the homoskedastic case.

Performance of the FDB

Since the diagnostic test for the test-bed model suggested that use of the FDB might improve the reliability of the bootstrap, the simulation experiment was extended to compute FDB P values. For each replication, a realisation of the statistic τ of (11) was obtained, and a realisation of the bootstrap DGP μ^*. Then B first-level bootstrap statistics, τ^*_j, $j = 1, \ldots, B$, were generated using the realisation of μ^*, along with a second-level bootstrap DGP μ^{**}_j, using which the second-level statistic τ^{**}_j was generated. The FDB bootstrap value was then computed as an estimate of the theoretical formula (20): the function R_0 estimated as the empirical distribution of the τ^*_j, and the quantile function Q^1 as an empirical quantile of the τ^{**}_j.

Below on the left is a comparison of the P value discrepancy plots for the single bootstrap (in red) and the FDB (in green). There is a slight improvement, but it is not very impressive. On the right are the kernel density plots for τ (red), τ^* (green), and τ^{**} (blue). All three are very similar, but the densities of τ^* and τ^{**} are closer than is either of them to the density of τ. This fact is probably the explanation of why the FDB does not do a better job.





The three statistics are at most very weakly correlated, as seen in the regression results:

$$\begin{array}{llll}
\tau^* = & 2.55 & + \;\; 0.25\tau, & \qquad \text{centred } R^2 = 0.004 \\
 & (0.025) & \;\;\;(0.009) & \\
\tau^{**} = & 2.51 & + \; 0.075\tau^*, & \qquad \text{centred } R^2 = 0.006 \\
 & (0.026) & \;\;\;(0.010) & \\
\tau^{**} = & 2.58 & + \;\; 0.051\tau, & \qquad \text{centred } R^2 = 0.003 \\
 & (0.024) & \;\;\;(0.009) &
\end{array}$$

The significant constants, though, are indicators of shifts in the distributions that are not visible in the kernel density plots, and contribute to the under-rejection by both single and fast double bootstrap P values.

8. Concluding Remarks

The diagnostic techniques proposed in this paper do not rely in any way on asymptotic analysis. Although they require a simulation experiment for their implementation, this experiment is hardly more costly than undertaking bootstrap inference in the first place. Results of the experiment can be presented graphically, and can often be interpreted very easily. A simple OLS regression constitutes the other part of the diagnosis. It measures to what extent the quantity being bootstrapped is correlated with its bootstrap counterpart. Significant correlation not only takes away the possibility of an asymptotic refinement, but also degrades bootstrap performance, as shown by a finite-sample analysis.

Since bootstrapping time series is an endeavour fraught with peril, the examples for which the diagnostic techniques are applied in this paper all involve time series. In some cases, the bootstrap method is parametric; in others it is intended to be robust to autocorrelation of unknown form. Such robustness can be difficult to obtain, and the reasons for this in the particular cases studied here are revealed by the diagnostic analysis.

References

Davidson, R. (2012). “Statistical Inference in the Presence of Heavy Tails”, *Econometrics Journal*, **15**, C31–C53.

Davidson, R. and E. Flachaire (2008). “The wild bootstrap, tamed at last”, *Journal of Econometrics*, 146, 162–9.

Davidson, R. and J. G. MacKinnon (1999). “The Size Distortion of Bootstrap Tests,” *Econometric Theory*, **15**, 361-376.

Davidson, R. and J. G. MacKinnon (2006). “The power of bootstrap and asymptotic tests”, *Journal of Econometrics*, **133**, 421–441.

Davidson, R. and J. G. MacKinnon (2007). “Improving the Reliability of Bootstrap Tests with the Fast Double Bootstrap”, *Computational Statistics and Data Analysis*, **51**, 3259–3281.

Eicker, F. (1963). “Asymptotic normality and consistency of the least squares estimators for families of linear regressions,” *The Annals of Mathematical Statistics*, 34, 447–456.

Fiebig Denzil, G. and H. Theil (1982). Comment, *Econometric Reviews* 1:2, 263–269

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

Horowitz, J. L. (1997). “Bootstrap methods in econometrics: Theory and numerical performance,” in D. M. Kreps and K. F. Wallis (ed.), Advances in Economics and Econometrics: Theory and Applications: Seventh World Congress, Cambridge, Cambridge University Press.

Jaynes. E. T. (1957). “Information Theory and Statistical Mechanics”, *Physical Review* 106, 620–630.

Künsch, H. R. (1989). “The jackknife and the bootstrap for general stationary observations”, *Annals of Statistics* 17, 1217–1241.

Liu, R. Y. (1988). “Bootstrap procedures under some non-I.I.D. models”, *Annals of Statistics* 16, 1696–1708.

Mammen, E. (1993). “Bootstrap and wild bootstrap for high dimensional linear models”, *Annals of Statistics* 21, 255–285.

Newey, W. K. and K. D. West (1987). “A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix”, *Econometrica* 55, 703–8.

Schluter, C. (2012) “Discussion of S. G. Donald *et al* and R. Davidson”, *Econometrics Journal*, **12**, C54-C57

Theil, H. and K. Laitinen (1980). “Singular moment matrices in applied econometrics”, in *Mu1tivariate Analysis - V* (P.R. Krishnaiah, Ed.). Amsterdam: North-Holland Pub1ishing Co., 629–649.

Vinod, H.D. (2006). “Maximum entropy ensembles for time series inference in economics”, *Journal of Asian Economics* 17, 955–978.

White, H. (1980). “A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity,” *Econometrica*, 48, 817–838.

Wu, C. F. J. (1986). “Jackknife, bootstrap and other resampling methods in regression analysis”, *Annals of Statistics* 14, 1261–1295.