# 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, not only in the time-series context.

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. I am grateful to two anonymous referees, seminar participants at Emory University, the Universities of Aarhus, York, and Durham for helpful comments, and especially to James MacKinnon.

March 2016

## 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); see also Hall, Horowitz, and Jing (1995), Lahiri (1999) and (2003). 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 (see for instance Bühlmann (1997) and (2002)), frequency-domain bootstraps (Kreiss and E. Paparoditis (2003), Kirch and Politis (2011)), and the recently-proposed dependent wild bootstrap, Shao (2010).

Simulation experiments can of course be used to study the performance of different bootstrap procedures in different circumstances. When the bootstrap is used for hypothesis testing, this is most readily done by looking at the bootstrap discrepancy, defined as the difference between the actual rejection frequency of the bootstrap test and the nominal significance level at which the test is performed. 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. A paper that does not seem to be very widely known that proposes such a technique is Beran (1997). The technique, although implemented with finite samples, is justified by sophisticated asymptotic arguments. Beran's paper is not oriented towards econometrics, but contains useful ideas that are developed differently here.

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 a GARCH(1,1) process studied. It turns out that the wild bootstrap is capable of giving essentially perfect inference even with very small samples. The ideas behind the fast double bootstrap (FDB) of Davidson and MacKinnon (2007) are laid out in Section 4, and used to motivate a method for estimating the bootstrap discrepancy that is much less computationally intensive than a direct simulation-based approach. This leads in Section 5 to the development of the diagnostic techniques that constitute the main contribution of this paper. In Section 6 the GARCH model of Section 3 is revisited, and the diagnostic techniques applied to it. It is seen that very good bootstrap performance can be diagnosed as well as poor performance. In Section 7, a version of the block bootstrap is diagnosed. There, it is seen to perform rather poorly, while the FDB, which can also be studied using the diagnostics of this paper, yields only a slight improvement. Another technique sometimes used with time series, the sieve bootstrap, is the topic of Section 8. It can perform reasonably well in favourable circumstances, and has the advantage that it can be combined with the wild bootstrap to take account of possible heteroskedasticity. Finally, some concluding remarks are presented in Section 9.

## 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, $\mu$ say, belongs to $\mathbb{M}$. Alternatively, we say that $\mathbb{M}$ is correctly specified.

We almost always want to define a parameter-defining mapping $\theta$, which maps the model $\mathbb{M}$ into a parameter space $\Theta$, 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 $\theta_\mu$, is the parameter vector that corresponds to $\mu$. Sometimes the mapping $\theta$ is one-one, as, for instance, with models estimated by maximum likelihood. More often, $\theta$ is many-one, so that a given parameter vector does not uniquely specify a DGP. Supposing that $\theta$ 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 $\mu$, then a statistic $\tau(\boldsymbol{y})$ is a realisation of a random variable $\tau$ of which the distribution is determined by $\mu$. A statistic $\tau$ 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 $\tau$. Unless $\tau$ 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 $\tau$ 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 single probability space $(\Omega, \mathcal{F}, P)$, with just one probability measure, $P$. Then we treat the test statistic $\tau$ as a stochastic process with the set $\mathbb{M}$ as index set. 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 $\Omega$ 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 $\tau$ 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 $\tau$ under any DGP $\mu \in \mathbb{M}_0$:

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

Suppose that we have a statistic computed from a data set that may or may not have been generated by a DGP $\mu \in \mathbb{M}_0$. Denote this statistic by $t$. Then the ideal $P$ value that would give exact inference is $R_0(t, \mu)$. If $t$ is indeed generated by $\mu$, $R_0(t, \mu)$ is distributed as U(0,1) if the distribution of $\tau$ 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 $R_0(t, \mu)$ is available by simulation only if $\tau$ is a pivot with respect to $\mathbb{M}_0$, since then we need not know the precise DGP $\mu$. 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 $\mu$, we use the same function or functional of an estimate of $\mu$. Analogously to the stochastic process $\tau$, we define the DGP-valued process

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

The estimate of $\mu$, which we call the bootstrap DGP, is $\beta(\mu, \omega)$, where $\omega$ 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 $\tau$ and $\beta$, the bootstrap $P$ value is another stochastic process:

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

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

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

where the $\omega_j^*$ 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 $\mu$ by the bootstrap DGP $b$. An ideal double bootstrap $P$ value that would give exact inference is $R_1\big(p_1(\mu, \omega), \mu\big)$, which is distributed as U(0,1). The double bootstrap $P$ value is, analogously, $R_1\big(R_0(t, b), b\big)$.

## 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

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

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

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

with the $\varepsilon_t$ standard normal white noise, and the recurrence initialised by $\sigma_1^2 = \lambda/(1-\gamma-\delta)$, which is the unconditional stationary expectation of the process. The parameters of the DGP used in the experiment were $a = 1.5$, $y_0 = 0$, $\lambda = 1$, $\gamma = 0.4$, and $\delta = 0.45$, with very small sample size $n = 10$, and $\rho = 0.3$. 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 heteroskedasticity-consistent covariance matrix estimator (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

$$y_1^* = y_1 \quad \text{and} \quad y_t^* = \tilde{a} + \rho_0 y_{t-1}^* + s_t^* \tilde{u}_t, \quad t = 2, \ldots, n, \tag{7}$$

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 $\tau_j^*$ that are more extreme than $\tau$. The performance of the bootstrap test, as revealed by experiments with $N = 100{,}000$ replications with $B = 399$ bootstrap samples for each, is excellent. Rather than presenting the simulation results here in tabular form, they will be seen in Section 6 in the context of the diagnostic procedures presented in Section 5.

## 4. 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 $\mu$ 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

$$p_2(\mu,\omega) = R_1\big(p_1(\mu,\omega),\beta(\mu,\omega)\big). \tag{8}$$

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,

$$R_1(\alpha,\mu) = P\big\{\omega \in \Omega \,|\, p_1(\mu,\omega) < \alpha\big\} = \mathrm{E}\big[\mathrm{I}\big(R_0(\tau(\mu,\omega),\beta(\mu,\omega)) < \alpha\big)\big], \tag{9}$$

where $\mathrm{I}(\cdot)$ is the indicator function, with value 1 when its argument is true, and 0 otherwise. 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 (9) 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

$$\begin{aligned} R_1(\alpha,\mu) &= \mathrm{E}\Big[\mathrm{E}\big[\mathrm{I}\big(\tau(\mu,\omega) < Q_0(\alpha,\beta(\mu,\omega))\big) \,|\, \beta(\mu,\omega)\big]\Big] \\ &\approx \mathrm{E}\big[R_0\big(Q_0(\alpha,\beta(\mu,\omega)),\mu\big)\big] \end{aligned} \tag{10}$$

Define the stochastic process

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

where $\Omega_1$ and $\Omega_2$ 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 $\mu$. We denote the CDF of $\tau^1$ under $\mu$ by $R^1(\cdot,\mu)$. Thus

$$\begin{aligned} R^1(\alpha,\mu) &= \Pr\big\{(\omega_1,\omega_2) \in \Omega_1 \times \Omega_2 \,|\, \tau\big(\beta(\mu,\omega_1),\omega_2\big) < \alpha\big\} \\ &= \mathrm{E}\big[\mathrm{I}\big(\tau(\beta(\mu,\omega_1),\omega_2) < \alpha\big)\big] \\ &= \mathrm{E}\Big[\mathrm{E}\big[\mathrm{I}\big(\tau(\beta(\mu,\omega_1),\omega_2) < \alpha\big) \,|\, \mathcal{F}_1\big]\Big] \\ &= \mathrm{E}\big[R_0\big(\alpha,\beta(\mu,\omega_1)\big)\big]. \end{aligned} \tag{11}$$

Here $\mathcal{F}_1$ denotes the sigma-algebra generated by functions of $\omega_1$.

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

$$\mathrm{E}\big[R_0\big(Q_0(\alpha,\beta(\mu,\omega)),\mu\big)\big] \approx R_0\big(Q^1(\alpha,\mu),\mu\big), \tag{12}$$

where $Q^1(\cdot,\mu)$ is the quantile function inverse to the CDF $R^1(\cdot,\mu)$. See Davidson and MacKinnon (2007) for the reasoning that leads to this approximation.

On putting the two approximations, (10) and (12), together, we obtain

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

The fast double bootstrap substitutes $R_1^f$ for $R_1$ in the double bootstrap $P$ value (8). The FDB $P$ value is therefore

$$p_2^f(\mu,\omega) = R_1^f\big(p_1(\mu,\omega),\beta(\mu,\omega)\big) = R_0\big(Q^1(p_1(\mu,\omega),\beta(\mu,\omega)),\beta(\mu,\omega)\big). \tag{13}$$

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 in Figure 1 perform this comparison for the wild bootstrap applied to the GARCH model of Section 3. 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.

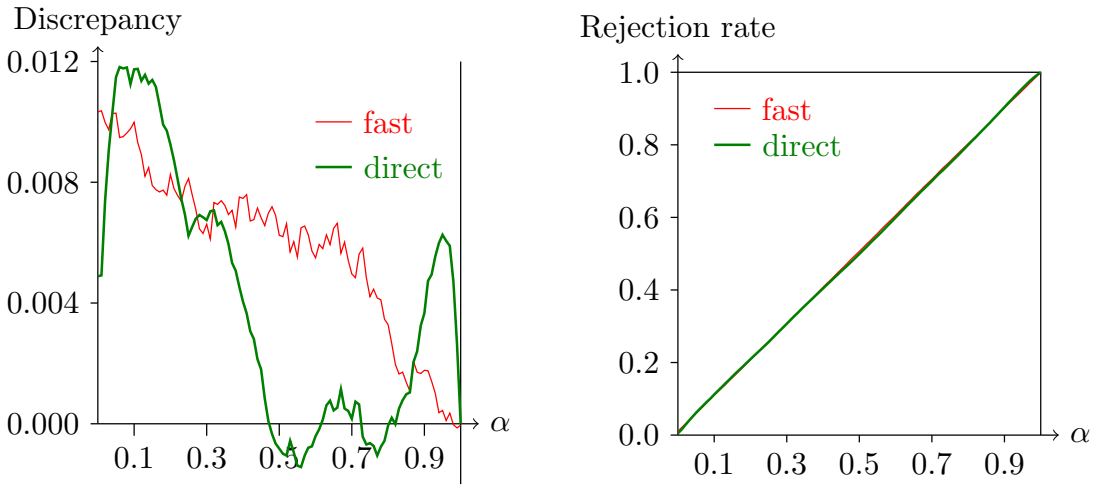


**Figure 1: Direct and fast estimates of the bootstrap discrepancy**

## 5. 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; see Davidson and MacKinnon (1998). 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.

A simulation experiment that provides the information for a $P$ value plot for a bootstrap test can be described as follows. For each of $N$ replications in the experiment, there are $B$ bootstrap repetitions for each replication. The data for each replication $i$, $i = 1, \ldots, n$, are generated using a DGP denoted by $\mu$, which satisfies the null hypothesis. These data are used to generate the realisation $\tau_i$ of the statistic $\tau$, and the realisation of the corresponding bootstrap DGP, which is then used to generate $B$ bootstrap statistics $\tau^*_{ij}$, $j = 1, \ldots, B$. The bootstrap $P$ value $P^*_i$ for replication $i$ is the proportion of the $\tau^*_{ij}$ that are more extreme than $\tau_i$. The graph of the empirical distribution function of the $P^*_i$ is then a simulation-based estimate of the $P$ value plot.

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. For replication $i$, $i = 1, \ldots, n$, let $\tau_i$ be as above a realisation of the test statistic, and let $\tau^*_i$ be a *single* bootstrap statistic generated by the bootstrap DGP for replication $i$.

The next step is to graph kernel-density estimates of the distribution of the statistic $\tau$ and that of the bootstrap statistic $\tau^*$. If these are not similar, then clearly the bootstrap DGP fails to mimic the true DGP at all well. Poor bootstrap performance is then a consequence of this fact. Alternatively, graphs of the empirical distributions of the simulated $\tau$ and $\tau^*$ may be compared. Another diagnostic is based on running an OLS regression of the $\tau^*_i$ on a constant and the $\tau_i$. Suppose without loss of generality that $\tau$ itself is in nominal $P$ value form, so that the rejection region is on the left. A significant constant in the regression indicates that the distribution of $\tau^*$ is shifted relative to that of $\tau$, and this typically causes under-rejection in one tail and over-rejection in the other. Next, suppose that this regression reveals that, for the data generated on any one replication, $\tau$ is strongly positively correlated with $\tau^*$. The positive correlation then implies that, if $\tau$ is small, then $\tau^*$ tends to be small as well. It follows 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. In Davidson and MacKinnon (1999), it is shown that if the test statistic $\tau(\mu, \omega)$ and the bootstrap DGP $\beta(\mu, \omega)$ are asymptotically independent, then there is a refinement, in the sense that the bootstrap discrepancy tends to zero faster

as the sample size tends to infinity than it would in the presence of an asymptotic correlation. In the finite-sample context, a significant correlation between $\tau$ and $\tau^*$ shows that the statistic and the bootstrap DGP are not independent, giving rise to over- or under-rejection, as suggested by the asymptotic theory. For the bootstrap to work well, the distribution of the bootstrap statistic $\tau^*$, conditional on the data that led to $\tau$, should be close to the distribution of $\tau$ itself. A correlation between $\tau$ and $\tau^*$ suggests that, even if the unconditional distributions of $\tau$ and $\tau^*$ are close, the conditional distribution of $\tau^*$ may be rather different from the unconditional distribution of $\tau$.

## 6. 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), parametrised as described there, with $\rho = 0.3$ and $n = 10$. First, the $P$ value and $P$ value discrepancy plots. They appear in Figure 2. 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.

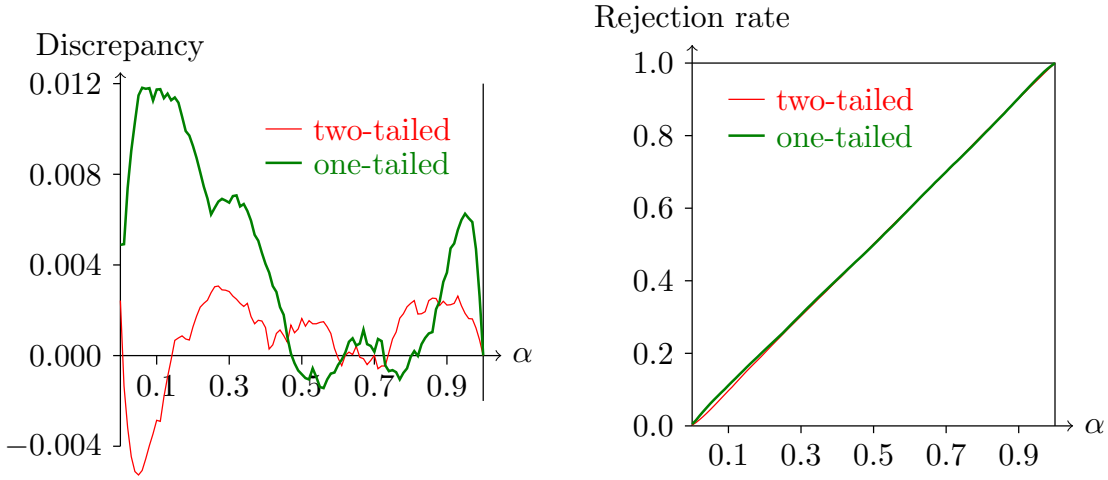


**Figure 2: $P$ value and $P$ value discrepancy plots, GARCH model, wild bootstrap**

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 0.011 for the one-tailed test, and -0.005 for the two-tailed test. The experiment that led to these results, with $N = 100{,}000$ replications with $B = 399$ bootstrap repetitions each, used just under 100 minutes of computer time on a computer with Intel Xeon processors at 2.3 GHz.*

Next the results of the diagnostic procedure. In Figure 3 are plotted the kernel density estimates of the distributions of the statistic and the bootstrap statistic for both cases.

* The actual running time was only between two and two and a half minutes, since 50 cores were used in parallel.

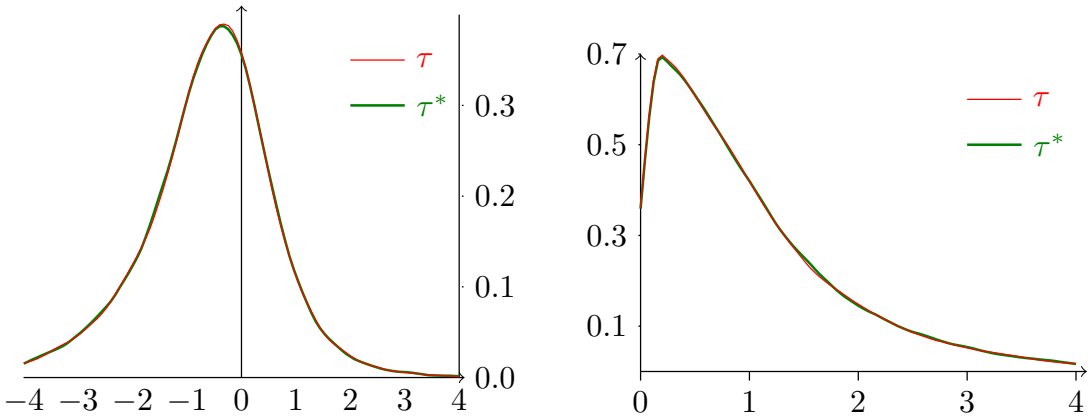


**Figure 3: Diagnostic comparison of densities, GARCH model, wild bootstrap**

The information above can also be represented by the plots of the distribution functions of the estimates, seen in Figure 4..

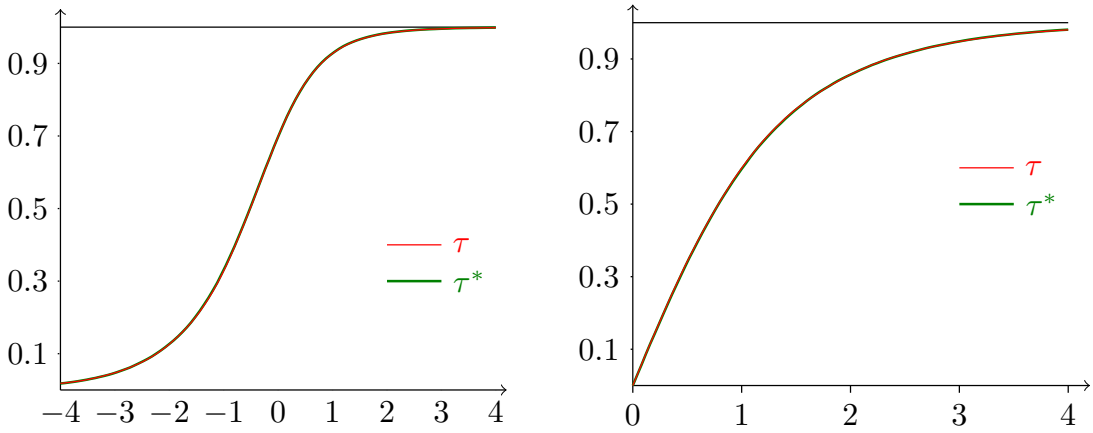


**Figure 4: Comparison of EDFs, GARCH model, wild bootstrap**

The computing time needed for these diagnostics was about two and a quarter minutes.

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

$$\underset{(0.005)}{\tau^* = -0.637} + \underset{(0.003)}{0.0015\tau}, \qquad \text{centred } R^2 = 0.000002$$

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

$$\underset{(0.005)}{\tau^* = 1.024} + \underset{(0.003)}{0.044\tau}, \qquad \text{centred } R^2 = 0.002$$

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 $\tau$, leading to the over-rejection seen in the $P$ value discrepancy plot, since the test rejects to the right.

Similarly the positive constant for the two-tailed test explains the under-rejection for interesting values of $\alpha$.

For a second example of the diagnostics, consider testing a hypothesis for the same model as above, but with a larger value of $\rho$, equal to 0.9, and using a conventional IID resampling bootstrap instead of the wild bootstrap. That is, a bootstrap sample is generated, not as in (7), but by

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

where the $u_t^*$ are resampled from the constrained residuals $\tilde{u}_t$, $t = 2, \ldots, n$. We may reasonably expect considerably worse bootstrap performance in this case. The expectation is borne out, as shown in the (costly) $P$ value and $P$ value discrepancy plots in Figure 5.

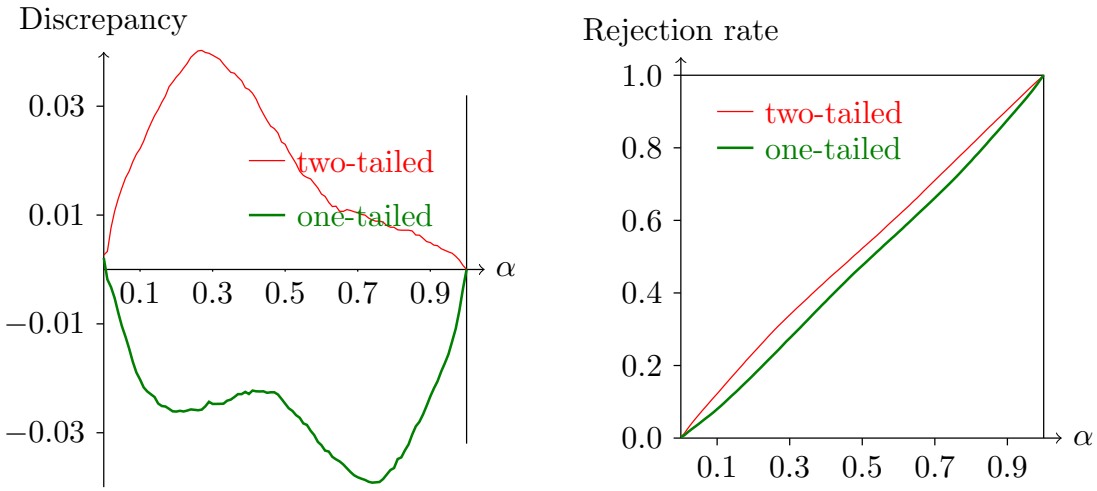


**Figure 5: *P* plots, GARCH model, resampling bootstrap**

Now the diagnostics: there is no need for more than the kernel density plots.

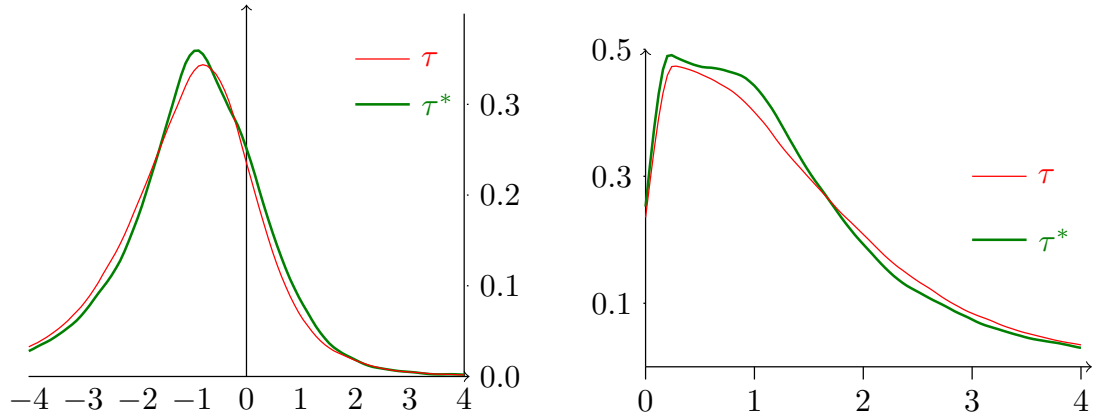


**Figure 6: Density comparison, GARCH model, resampling bootstrap**

For both tests, it is clear that the two distributions differ considerably. For the one-tailed test, the regression results are:

$$\underset{(0.006)}{\tau^* = -1.127} + \underset{(0.003)}{0.062\tau}, \qquad \text{centred } R^2 = 0.003$$

and, for the two-tailed test:

$$\underset{(0.007)}{\tau^* = 1.448} + \underset{(0.004)}{0.027\tau}, \qquad \text{centred } R^2 = 0.0006$$

The very significant estimated constants also show that the two distributions differ. The slope coefficients, although small, are also highly significant, and indicate that there is a positive correlation between the statistic and the bootstrap counterpart.

## 7. The Block Bootstrap

Davidson and Flachaire (2008) show that there is a special setup where the wild bootstrap can deliver perfect inference. Unlike the GARCH example just considered, the setup here is of a static linear regression model. If one wishes to test the hypothesis that the entire vector $\boldsymbol{\beta}$ in the static linear regression $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ is zero when the disturbances $\boldsymbol{u}$ may be heteroskedastic, the obvious test statistic is

$$\tau \equiv \boldsymbol{y}^\top \boldsymbol{X}(\boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \tag{14}$$

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 different versions of either the HCCME or a heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimator. When the Rademacher distribution (4) is used, the wild bootstrap $P$ value, in the presence of heteroskedasticity, is uniformly distributed under the null up to discreteness due to a finite sample size. In this and the next section, simulation results are presented for two 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 (14), 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.

As remarked in the Introduction, when it is suspected that the disturbances are not necessarily white noise, the commonest way to proceed is to use some version of the block bootstrap. A recent paper by Nordman and Lahiri (2012) reviews many of these and compares their performance. Here, the one chosen as an illustration is not the

best to emerge from these comparisons, so that mediocre performance can be readily detected by the diagnostics of this paper.

The model is a static linear regression with a constant and three non-constant regressors, all themselves serially correlated with an autoregressive parameter of 0.8. The disturbances are an AR(1) process, with an autoregressive parameter of 0.9. The hypothesis to be tested is that the complete vector $\boldsymbol{\beta}$ of regression coefficients is zero. The sample size is $n = 64$. The test statistic takes the form (14), and makes use of a Newey-West HAC covariance estimator – see Newey and West (1987) – with lag truncation parameter of $p = 16$.

The bootstrap procedure is the basic moving-block bootstrap with overlapping blocks. The block length was chosen to be $b = 24$, after some experimentation with different choices of $p$ and $b$ showed that these choices were far from the worst, although performance was similar for many other choices in the neighbourhood of this one. In Figure 7 are shown, on the left, the $P$ value discrepancy plot obtained using the (costly) direct method, and on the right the (cheap) diagnostic comparison of the densities of the statistics $\tau$ and $\tau^*$. The distortion is clearly considerable.

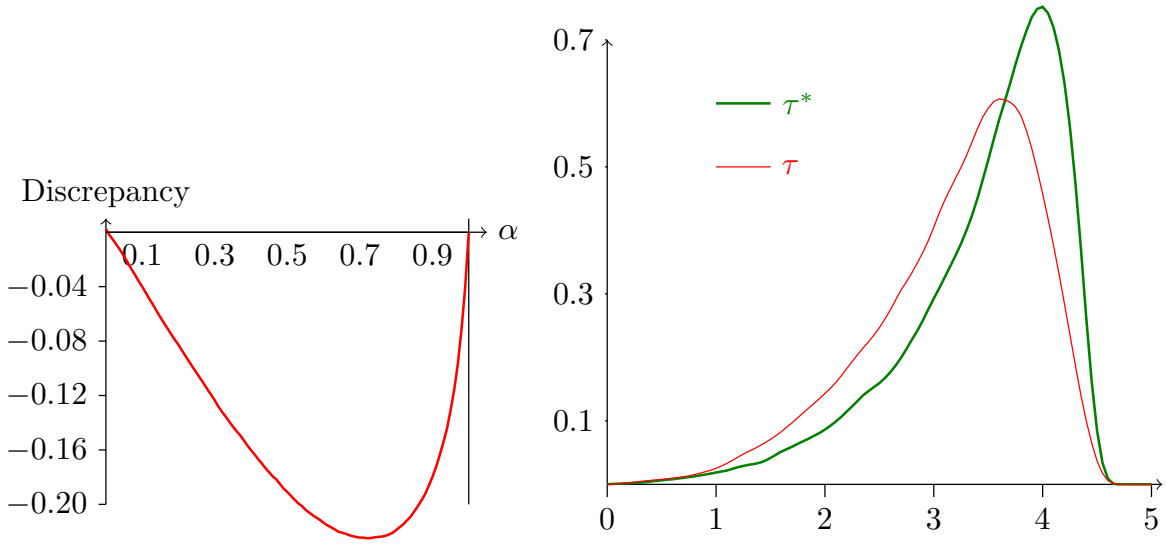


**Figure 7: $P$ value discrepancy plot; density comparison, block bootstrap**

The results of the regression diagnostic are as follows:

$$\underset{}{\tau^* = 2.854 + 0.192\tau,} \qquad \text{centred } R^2 = 0.040$$
$$(0.010) \quad (0.003)$$

The constant shows what is already seen in the density plot, namely that the distribution of $\tau^*$ is shifted to the right relative to that of $\tau$. Since the statistic is asymptotically chi-squared, it rejects to the right, and so the shift tends to make $P$ values larger than they would ideally be. The resulting tendency to under-reject is reinforced by the very significant positive correlation between the two statistics.

Figure 8 compares the $P$ value and $P$ value discrepancies as computed directly with the computation by the fast method.

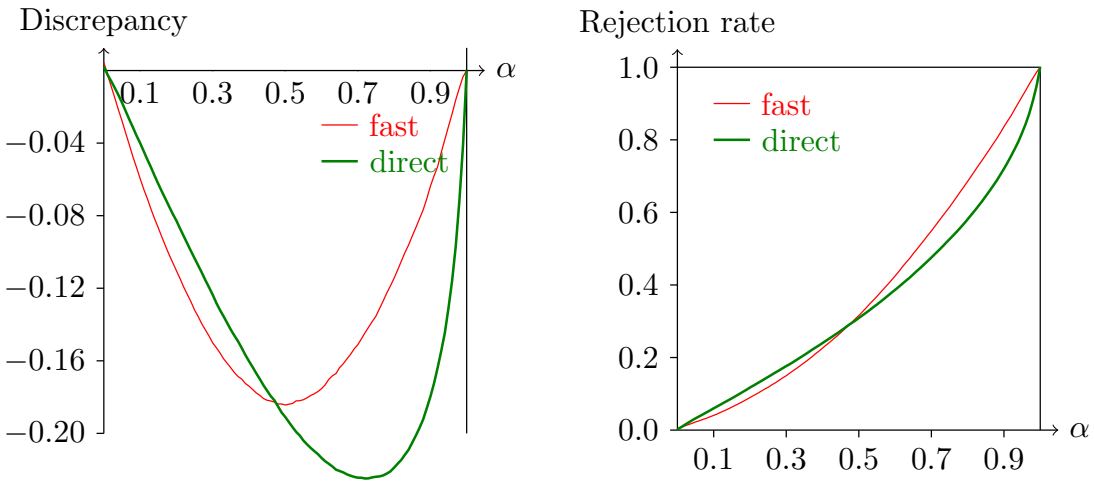


**Figure 8: Fast and direct computations of the *P* value; block bootstrap**

It can be seen that the fast method underestimates the under-rejection for large significance levels, but overestimates it somewhat for conventional levels. It is thus interesting to see whether use of the FDB improves inference. To this end, the simulation experiment was extended to compute FDB $P$ values. For each replication, a realisation of the statistic $\tau$ was obtained, and a realisation of the bootstrap DGP $\mu^*$. Then $B$ first-level bootstrap statistics, $\tau_j^*$, $j = 1, \ldots, B$, were generated using the realisation of $\mu^*$, along with a second-level bootstrap DGP $\mu_j^{**}$, using which the second-level statistic $\tau_j^{**}$ was generated. The FDB bootstrap $P$ value was then computed as an estimate of the theoretical formula (13): the function $R_0$ estimated as the empirical distribution of the $\tau_j^*$, and the quantile function $Q^1$ as an empirical quantile of the $\tau_j^{**}$.

Figure 9 repeats the plots in Figure 7, with the corresponding graphs for the FDB overlaid.

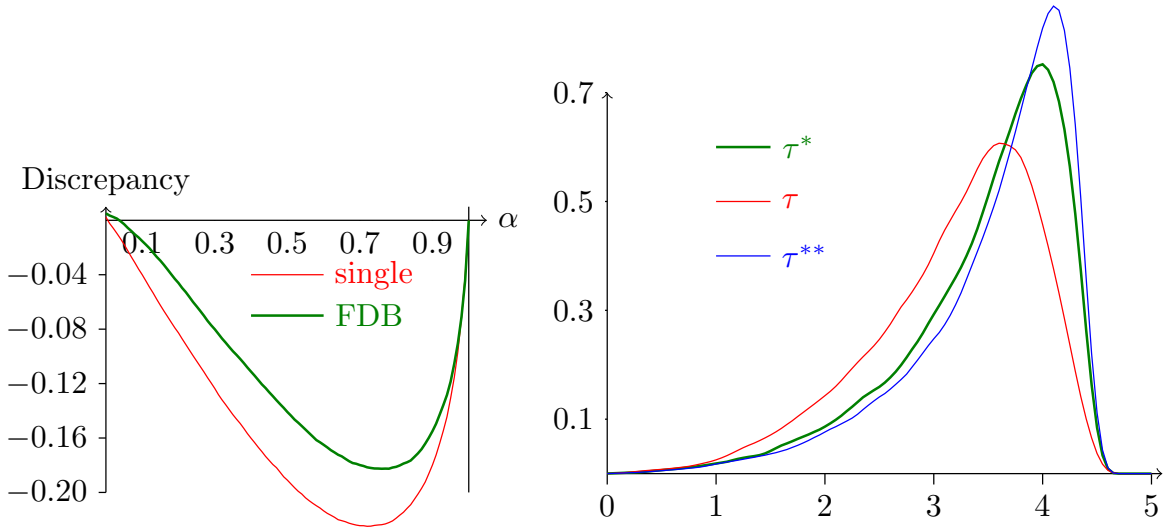


**Figure 9: Plots for single bootstrap and FDB; block bootstrap**

It can be seen that the FDB does yield some slight improvement, only slight, presumably, because the shift of the distribution of $\tau^*$ to that of $\tau^{**}$ is less than the shift

from $\tau$ to $\tau^*$. If $\tau^{**}$ is regressed on a constant and $\tau$, we obtain

$$\underset{(0.010)}{\tau^{**} = 3.005} + \underset{(0.003)}{0.171\tau}, \qquad \text{centred } R^2 = 0.030$$

The estimated constant shows the greater shift from $\tau$ to $\tau^{**}$, but it seems that $\tau^{**}$ is less correlated with $\tau$ than is $\tau^*$.

## 8. 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($\infty$) 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 once again concerned with the static linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, where it is assumed that the covariance matrix $\boldsymbol{\Omega}$ of the disturbance vector $\boldsymbol{u}$ can be well approximated by the covariance matrix of a stationary AR($p$) process, which implies that the diagonal elements are all the same.

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

$$\hat{u}_t = \sum_{i=1}^{p} \rho_i \hat{u}_{t-i} + \varepsilon_t \tag{15}$$

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 (15), rather than least squares.

Estimation of (15) yields residuals and an estimate $\hat{\sigma}_\varepsilon^2$ of the variance of the $\varepsilon_t$, 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), or (3) subject to the restrictions of the null hypothesis. 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 obtained under the assumption that the AR($p$) model is correctly specified.

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

$$u_t^* = \sum_{i=1}^{p} \hat{\rho}_i u_{t-i}^* + \varepsilon_t^*, \quad t = p+1, \ldots, n, \tag{16}$$

where the $\varepsilon_t^*$ 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 (15), preferably rescaled by the factor $\sqrt{n/(n-p)}$. In order to initialise the recurrence (16), the simplest way is just to set $u_t^* = \hat{u}_t$ for the first $p$ observations of each bootstrap sample, thereby conditioning on the initial observations of the real data. Unless we are sure that the AR($p$) process is really stationary, rather than just being characterized by values of the $\rho_i$ 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 (16) 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 $\varepsilon_t^*$. We then throw away all but the last $p$ of these disturbances and use those to initialize (16). In this way, we approximate a stationary process with the correct estimated stationary covariance matrix, but with no assumption of normality.

We now consider the same setup as in the previous section, with the statistic (14) used to test the hypothesis that the full vector $\boldsymbol{\beta}$ of regression parameters is zero. As the disturbances are in fact 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. Figure 10 shows the $P$ value and $P$ value discrepancy plots, on the left, and the kernel density plots on the right.

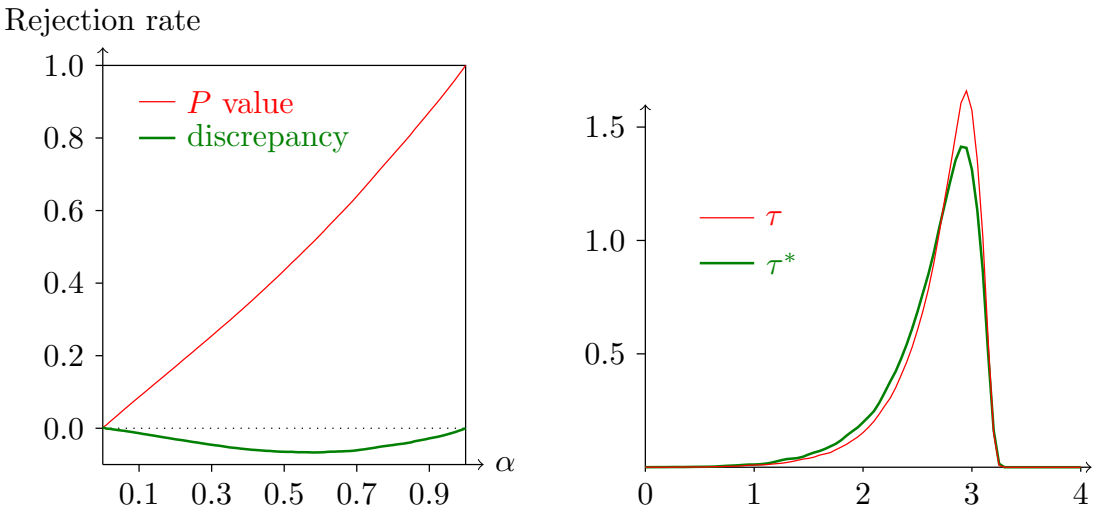


**Figure 10: $P$ values and density diagnostic; sieve bootstrap**

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 (unconditional) distributions of $\tau$ and $\tau^*$ are very similar.

The regression of $\tau^*$ on $\tau$ gave (OLS standard errors in parentheses):

$$\begin{array}{rcl} \tau^* = & 2.57 & + \quad 0.05\tau, \qquad \text{centred } R^2 = 0.003 \\ & (0.008) & \quad (0.003) \end{array}$$

Figure 11 shows 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.

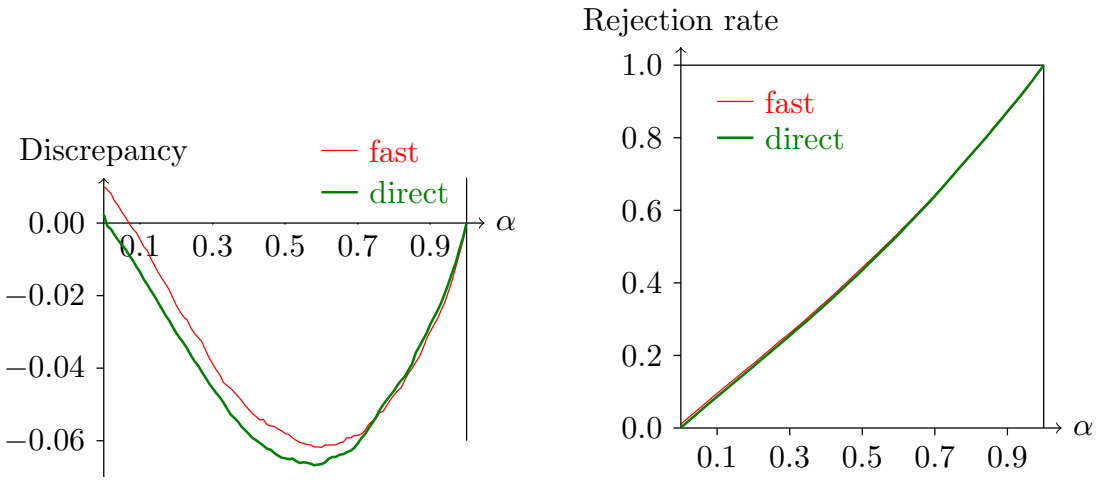


**Figure 11: Fast and direct computations of the $P$ value; sieve bootstrap**

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

Figure 12 shows on the left 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 $\tau$ (red), $\tau^*$ (green), and $\tau^{**}$ (blue). All three are very similar, but the densities of $\tau^*$ and $\tau^{**}$ are closer than is either of them to the density of $\tau$. This fact is probably the explanation of why the FDB does not do a better job.

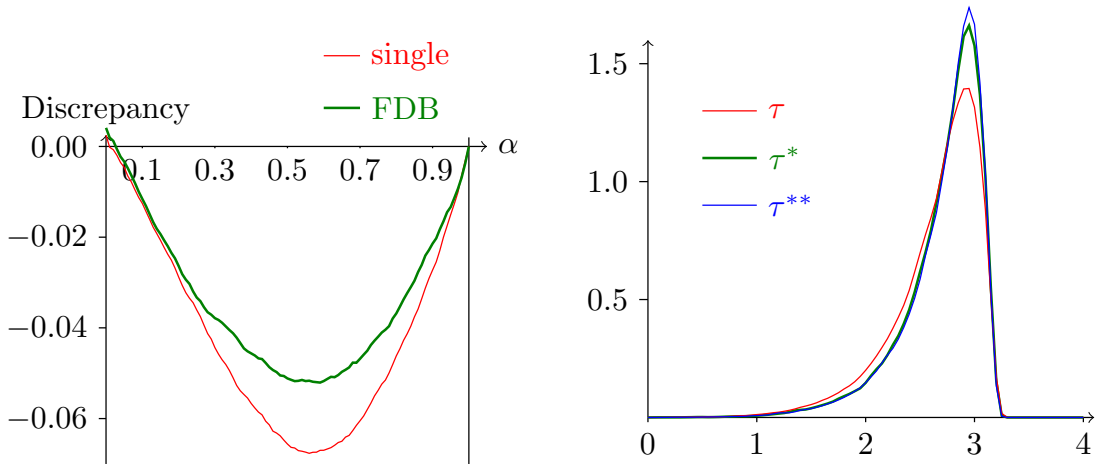


**Figure 12: Plots for single bootstrap and FDB; sieve bootstrap**

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

$$\begin{array}{rcll} \tau^* &=& 2.57 + 0.05\tau, & \text{centred } R^2 = 0.003 \\ & & (0.008)\quad (0.003) & \\ \tau^{**} &=& 2.50 + 0.08\tau^*, & \text{centred } R^2 = 0.006 \\ & & (0.009)\quad (0.003) & \\ \tau^{**} &=& 2.60 + 0.04\tau, & \text{centred } R^2 = 0.002 \\ & & (0.008)\quad (0.003) & \end{array}$$

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.

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 (16) 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 (15), and the $s_t^*$ are IID drawings from the Rademacher distribution. In Figure 13 are shown results of the diagnostic procedure for a simulation experiment in which the disturbances are scaled by one of the regressors.

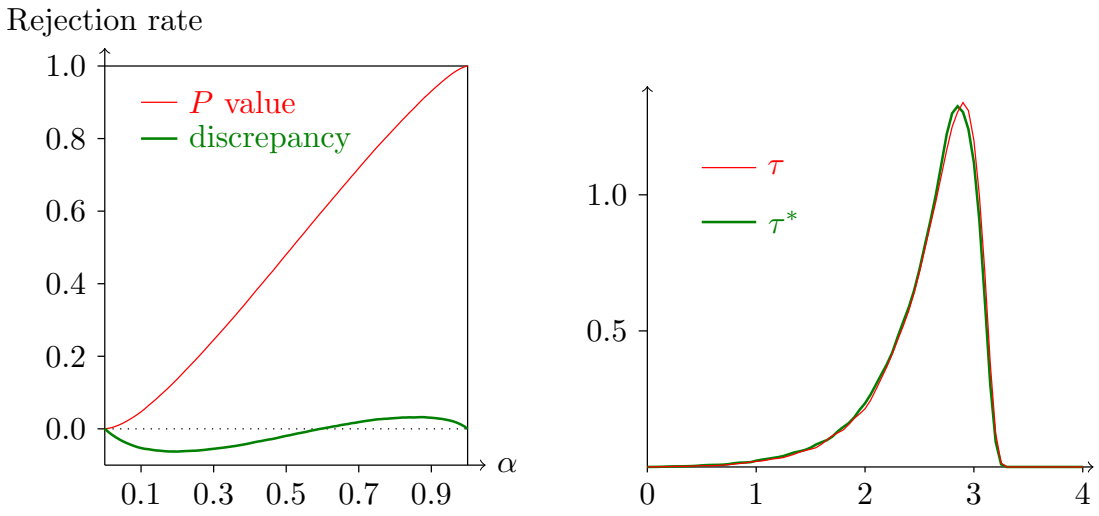


**Figure 13: Results with heteroskedasticity; sieve and wild bootstrap**

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

$$\underset{}{\tau^* =} \underset{(0.008)}{2.02} + \underset{(0.003)}{0.23\tau}, \qquad \text{centred } R^2 = 0.054$$

This time, there is significant correlation, and consequent under-rejection in the left-hand part of the distribution, although the distributions of $\tau$ and $\tau^*$ are at least as similar as in the homoskedastic case.

## 9. 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, and to what extent the distributions of this quantity and its bootstrap counterpart are shifted relative to each other. 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.

The methods of the paper can, of course, be applied equally well in circumstances that do not necessarily involve time series, but in which the bootstrap seems not to work very well.

## References


Beran, R. (1997) "Diagnosing Bootstrap Success", *Annals of the Institute of Statistical Mathematics*, **49**, 1–24.

Bühlmann, P. (1997). "Sieve bootstrap for time series", *Bernoulli* 3, 123–148.

Bühlmann, P. (2002). "Bootstraps for time series", *Statistical Science* 17, 52–72.

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 (1998). "Graphical Methods for Investigating the Size and Power of Hypothesis Tests," *The Manchester School*, **66**, 1-26.

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

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.

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

Hall, P., J. L. Horowitz, and B.-Y. Jing (1995). "On blocking rules for the bootstrap with dependent data", *Biometrika* 82, 561–574.

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.

Kirch, C. and D. N. Politis (2011). "TFT Bootstrap: Resampling time series in the frequency domain to obtain replicates in the time domain", *Annals of Statistics* 39, 1427–1470.

Kreiss, J. P. and E. Paparoditis (2003). "Autoregressive-aided periodogram bootstrap for time series", *Annals of Statistics* 31, 1923-1955.

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

Lahiri, S. N. (1999). "Theoretical comparisons of block bootstrap methods", *Annals of Statistics* 27, 386–404.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*, New York: Springer.

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.

Nordman, D. J. and S. N. Lahiri (2012). "Block Bootstraps for Time Series with Fixed Regressors", *Journal of the American Statistical Association*, 107, 233–46.

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

Shao, X. (2010). "The Dependent Wild Bootstrap", *Journal of the American Statistical Association* 105, 218–235.

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.