A Discrete Model for Bootstrap Iteration

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

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

Abstract

In an attempt to free bootstrap theory from the shackles of asymptotic considerations, this paper studies the possibility of justifying, or validating, the bootstrap, not by letting the sample size tend to infinity, but by considering the sequence of bootstrap P values obtained by iterating the bootstrap. The main idea of the paper is that, if this sequence converges to a random variable that follows the uniform $\mathrm{U}(0,1)$ distribution, then the bootstrap is valid. The idea is studied by making the model under test discrete and finite, so that it is characterised by a finite three-dimensional array of probabilities. This device, when available, renders bootstrap iteration to any desired order feasible. It is used for studying a unit-root test for a process driven by a stationary MA(1) process, where it is known that the unit-root test, even when bootstrapped, becomes quite unreliable when the MA(1) parameter is in the vicinity of -1. Iteration of the bootstrap P value to convergence achieves reliable inference except for a parameter value very close to -1. The paper then endeavours to see these specific results in a wider context, and tries to cast new light on where bootstrap theory may be going.

Keywords: Bootstrap, bootstrap iteration

JEL codes: C10, C12, C15

This research was supported by the Canada Research Chair program (Chair in Economics, McGill University), and by grants from the Fonds de Recherche du Québec: Société et Culture. I am grateful for discussions with, and comments from, Andrew Chesher, James Davidson, and Joel Horowitz. This version of the paper has substantially benefited from the comments of three anonymous referees. The usual disclaimer applies.

August 2016

1. Introduction, Background, and Motivation

The statistical method called the bootstrap makes no use of any asymptotic considerations, but current bootstrap theory relies on them. I find this an extremely undesirable state of affairs, if for no other reason than that the choice of an asymptotic construction is inevitably somewhat arbitrary. This paper tries to take a first step towards remedying the sitution.

Asymptotic arguments rely on sequences of random variables that converge, in probability or in distribution, to some desirable limit. Since bootstrap inference is exact only in rare instances, any justification of it must also rely on some sequence of random variables with a desirable limit.

The approach of this paper involves the convergence or otherwise of a sequence of bootstrap P values for testing the correct specification of an econometric (or statistical) model, obtained by iterating the bootstrap. Computing resources are nowadays such that the first iterate, called the double bootstrap, is in many cases feasible; when it is, it is presumed that it can provide more reliable statistical inference than either conventional asymptotics or the single uniterated bootstrap.

Bootstrap iteration first appeared in the literature in the 1980s, with one approach due to Hall (1986), another to Beran (1987) and (1988). Shortly after, Hall and Martin (1988) set out a unified approach that subsumes Hall's earlier one and that of Beran. In these papers, bootstrap iteration was justified on the strength of asymptotic expansions in negative powers of the sample size, which show that errors in inference are of progressively lower order as the order of iteration increases. Among other relevant papers from around this time, see also Martin (1990). Bootstrap iteration is discussed in Hall's (1992) influential book, in which Edgeworth expansions play a vital theoretical role.

The bootstrap can provide asymptotic refinements over leading-order asymptotic theory, as brought out clearly in Horowitz (2001). These refinements have sometimes been described as "Edgeworth corrections". Indeed, the bootstrap was thought of as having many of the same properties as an Edgeworth expansion, and, on this view, iterated bootstraps are akin to higher-order Edgeworth expansions.

Bootstrap iteration is very costly in terms of the required computing effort. Various attempts have therefore been made to alleviate this cost. An early one, making use of analytical approximations, is contained in DiCiccio, Martin, and Young (1992), where Edgeworth expansions are not used, although the method is justified by asymptotic arguments. The approach taken in Lee and Young (1995) makes somewhat different use of Edgeworth expansions as a way to avoid the cost of a double bootstrap. Chan and Lee (2001) derive an algorithm for infinitely iterated bootstrap bias correction, by considering bootstrap iteration as a Markov provess. Lee and Young (2003) develop methods of weighted bootstrap iteration, shown to be asymptotically equivalent to two consecutive conventional bootstrap iterations. In Davidson and MacKinnon (2007), the so-called fast double bootstrap is described; it had been used in various studies before that.

The constant theme throughout this literature is that bootstrap iteration leads to more reliable inference. In a wide variety of circumstances, it has been possible to demonstrate that asymptotic refinements accrue with bootstrap iteration. It is certainly tempting, therefore, to suppose that the only barrier to completely reliable inference *via* bootstrap iteration is computational infeasibility. This paper attempts to show, in the finite-sample context, and with no use of asymptotic arguments, that bootstrap iteration can indeed improve reliability of inference.

In the next Section, I introduce the definitions and notation needed to formulate the problem. In Section 3, the discretisation procedure is laid out, and the mechanics of bootstrap iteration developed. The notion of the *bootstrap discrepancy* is considered in Section 4, and some simple conclusions drawn. Then, in Section 5, a specific model is treated, and analysed numerically. The model under test is characterised by a unit-root process, obtained by cumulating a stationary MA(1) process, and the specification is tested by an augmented Dickey-Fuller (ADF) test. The model is completely parametric, with one single parameter, namely the MA(1) parameter. It is well known that, when this parameter is close to -1, the ADF test becomes thoroughly unreliable. A parametric bootstrap suggests itself as a way of improving reliability. The numerical study shows that it does so except for a parameter in a very small neighbourhood of -1, if the sequence of iterated P values is followed until convergence. This occurs for the 60^{th} bootstrap iteration, an order of iteration completely inconceivable in normal circumstances. Potential loss of power due to use of an iterated bootstrap P value is considered in Section 6, and shown to be unimportant unless the hypothesis under test is substantially false. The results obtained for this example are discussed in a more general context in Section 7, and some concluding remarks are offered in Section 8.

2. Framework

The focus in this paper is on bootstrap P values as the tool of inference. Reliability of inference is measured by means of the *bootstrap discrepancy*, by which is meant the difference between the rejection probability (RP) of a bootstrap test and the nominal significance level. For level α and a bootstrap P value, P say, rejection is the event $P < \alpha$. Inference is called reliable, or exact, if the probability that $P < \alpha$ is equal to α.

A *model* is defined as a set of data-generating processes (DGPs), a DGP being, for present purposes, a unique recipe for simulation. A model $\mathbb{M}$ may correspond to a *hypothesis*, namely that $\mathbb{M}$, as a set, contains all DGPs that satisfy the hypothesis. In order to test the hypothesis, a test statistic, τ, is used. For notational convenience, we may suppose that the range of τ is the $[0, 1]$ interval rather than the whole real line, and that the statistic takes the form of an approximate P value, which thus leads to rejection of the hypothesis characterised by $\mathbb{M}$ when the statistic is too small.

Let $R_0 : [0, 1] \times \mathbb{M} \to [0, 1]$ be the cumulative distribution function (CDF) of τ under any DGP $\mu \in \mathbb{M}$:

$$R_0(\alpha, \mu) = \Pr{}_\mu(\tau \leq \alpha).$$

Thus $R_0(\alpha, \mu)$ is the RP, at nominal level α, of a test based on τ.

Assumption 1:

> The distributions characterised by the CDFs $R_0(\cdot, \mu)$, $\mu \in \mathbb{M}$, are absolutely continuous on $[0, 1]$.

The statistic τ is said to be a *pivot*, or be *pivotal*, if $R_0(\alpha, \mu)$ is equal to some function $R(\alpha)$ for all $\mu \in \mathbb{M}$, and for all $\alpha \in [0, 1]$. In this case, the random variable $R(\tau)$ follows the $\mathrm{U}(0, 1)$ distribution under all DGPs $\mu \in \mathbb{M}$, and so permits exact inference.

If τ is not pivotal, exact inference based on τ would be possible, using the statistic $R_0(\tau, \mu)$, if μ were known. In practice, of course, μ is unknown, and exact inference is no longer possible. The principle of the bootstrap is that, when we want to use some function or functional of an unknown DGP μ, we use an estimate in place of μ. This estimated DGP is the *bootstrap DGP*, denoted by β. Let a realisation of τ be denoted as t, and a realisation of β, obtained from the same data set, as b. Then the bootstrap statistic is $R_0(t, b)$. If indeed the bootstrap DGP β is a good estimator of the true DGP μ, then $R_0(t, b)$ is a realisation of the random variable $R_0(\tau, \beta)$, which can be expected to follow the $\mathrm{U}(0, 1)$ distribution approximately under the null. Note that it is necessary that $b \in \mathbb{M}$, according to the first Golden Rule for bootstrapping; see, for instance, Davidson (2007)).

Only in exceptional, and usually trivial, cases can $R_0(t, b)$ be evaluated analytically, and so the usual practice is to estimate it by simulation. One obtains some large number, B say, of *bootstrap statistics*, τ_j^*, $j = 1, \ldots, B$, by simulation under the bootstrap DGP b, and then estimates $R_0(t, b)$ by

$$\hat{R}_0(t, b) = \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}(\tau_j^* < t),$$

where $\mathrm{I}(\cdot)$ is the indicator function, equal to 1 if its Boolean argument is true, and to 0 otherwise. By the law of large numbers, as $B \to \infty$, $\hat{R}_0(t, b)$ tends almost surely to the "ideal" bootstrap P value $R_0(t, b)$.

Let p_1 denote the ideal bootstrap P value, which, expressed as a random variable, is $R_0(\tau, \beta)$, and let $R_1(\alpha, \mu)$ be the CDF of p_1 under a DGP $\mu \in \mathbb{M}$. Observe that, since $b \in \mathbb{M}$, by Assumption 1, the distribution of p_1 is absolutely continuous on $[0, 1]$. Consequently, the random variable $R_1(p_1, \mu)$ follows the $\mathrm{U}(0, 1)$ distribution. As μ is unknown, the *double bootstrap* P value is defined as $p_2 \equiv R_1(p_1, \beta)$. Estimating p_2 by simulation is much more costly than estimating p_1.

Iterated bootstrap P values are then defined by the following recursive scheme:

$$\begin{aligned} R_r(\alpha, \mu) &= \Pr{}_\mu(p_r \leq \alpha), \\ p_{r+1} &= R_r(p_r, \beta), \end{aligned} \tag{1}$$

where the recurrence is initialised by the definition $p_0 = \tau$. Thus p_{r+1} is the bootstrap P value obtained by bootstrapping the r^{th} order P value p_r. It estimates the probability mass in the distribution of the r^{th} order P value to the left of its realisation.

In order for bootstrap iteration to be useful in the sense of permitting exact inference, it is necessary for the sequence $\{p_r\}$ of iterated bootstrap P values to converge as $r \to \infty$, and for the limit of the sequence to be distributed as $\mathrm{U}(0,1)$. Note that it is possible to have convergence to different distributions for different $\mu \in \mathbb{M}$, and for these, or some of them, to be quite different from $\mathrm{U}(0,1)$. For instance, with a resampling bootstrap, since with high probability each resample does not contain some of the observations of the original sample, repeated iteration leads almost surely to iterated bootstrap DGPs that have only one observation of the original sample, repeated as many times as the original sample has observations. This is one reason for our having limited attention until now to absolutely continuous distributions. As a resampling bootstrap DGP has a discrete distribution, and since, by the first Golden Rule, it must belong to the null model $\mathbb{M}$, we have ruled it out by Assumption 1.

This research project aims at sorting this out, and finding sufficient conditions for convergence. Necessary conditions would be even better, if we achieve sufficiency. It would then be possible to free the bootstrap from asymptotic theory in order to say what bootstrap procedures are valid and what not. The new criterion will be convergence of the sequence of iterated bootstrap P values to a known distribution.

3. Making Things Discrete

It's not obvious where to start in the study of convergence. But it is sometimes helpful, at least for intuition, if a discrete model is considered. This means that it is useless to aim for the continuous $\mathrm{U}(0,1)$ distribution, but that need not be harmful. Various wild bootstraps exist, and although they are in many ways similar to resampling bootstraps, they do not suffer from the problem that the bootstrap DGP becomes degenerate if the bootstrap is iterated. The P value remains discrete, however. In the case of a wild bootstrap that assigns random signs to N residuals, there are at most 2^N possible bootstrap samples, and so at most 2^N possible bootstrap statistics. But if the bootstrap statistic is distributed uniformly over these 2^N values, that is quite enough for reliable inference, even although one cannot achieve exact inference for a test at any arbitrary level in $[0,1]$.

In what follows, I assume that the statistic, in approximate P value form, can take on only the values π_i, $i = 0, 1, \ldots, n$, with

$$0 = \pi_0 < \pi_1 < \pi_2 < \ldots < \pi_{n-1} < \pi_n = 1. \tag{2}$$

Here n is not the sample size. Further, I assume that there are only m possible DGPs in the model that represents the null hypothesis. Thus the outcome space on which the random variables τ, the test statistic, and β, the bootstrap DGP, are defined consists of just $m(n+1)$ points, labelled by two integer coordinates (i, j), $i = 0, 1, \ldots, n$, $j = 1, \ldots, m$.

The model is then completely characterised by the probabilities p_{kij}, $k, j = 1, \ldots, m$, $i = 0, 1, \ldots, n$, where, under the DGP indexed by k,

$$p_{kij} = \Pr{}_k\big[\tau = \pi_i \text{ and } \beta = j\big]. \tag{3}$$

It follows for all $k = 1, \ldots, m$, that

$$\sum_{i=0}^{n} \sum_{j=1}^{m} p_{kij} = 1.$$

Make the definitions

$$a_{kij} = \sum_{l=0}^{i-1} p_{klj} \qquad \text{and} \qquad A_{ki} = \sum_{j=1}^{m} a_{kij} \quad i = 0, \ldots, n+1. \tag{4}$$

Thus a_{kij} is the probability under DGP k that τ is less than π_i and that $b = j$, while A_{ki} is the marginal probability under k that $\tau < \pi_i$. The CDF of τ can therefore be written as

$$R_0(\alpha, k) = A_{k,\lfloor \alpha n \rfloor + 1}.$$

Note that $a_{k0j} = A_{k0} = 0$ for all $k, j = 1, \ldots, m$. Further, $A_{k(n+1)} = 1$ and $a_{k(n+1)j}$ is the marginal probability under k that $\beta = j$, for all $k = 1, \ldots, m$.

Consider next the bootstrap P value p_1 under DGP k with realisation (i, j). It is the probability mass under the bootstrap DGP j of a value of τ less than π_i. It follows that

$$p_1 = A_{ji}, \tag{5}$$

In order to iterate the bootstrap, we need the CDF of p_1 under the DGP k. It can be written as

$$R_1(\alpha, k) = \Pr{}_k(p_1 < \alpha) = \sum_{i=0}^{n} \sum_{j=1}^{m} p_{kij} \mathrm{I}(A_{ji} < \alpha), \tag{6}$$

where $\mathrm{I}(\cdot)$ is an indicator function. Now let $q_j^1(\alpha)$ be defined by

$$q_j^1(\alpha) = \max_{i=0,\ldots,n+1} \left\{ i \;\middle|\; A_{ji} \leq \alpha \right\}$$

In this discrete context, $q_j^1(\alpha)/n$ can be interpreted analogously to an α-quantile of the distribution of the statistic τ with DGP j. In fact, for $\alpha = A_{ji}$, it is easy to see that $q_j^1(A_{ji}) = i$. Since the a_{jil} are increasing in i, so too are the A_{ji}, and we see that

$$A_{ji} < \alpha \quad \text{for all } i < q_j^1(\alpha) \text{ and } \quad A_{ji} \geq \alpha \quad \text{for all } i \geq q_j^1(\alpha).$$

Thus the event $\{A_{ji} < \alpha\}$ is equivalent to the inequality $i < q_j^1(\alpha)$.

This allows us to compute $R_1(\alpha, k)$. We find from (6) that

$$R_1(\alpha, k) = \sum_{j=1}^{m} \sum_{i=0}^{q_j^1(\alpha)-1} p_{kij} = \sum_{j=1}^{m} a_{kq_j^1(\alpha)j}. \tag{7}$$

We may think of the indices k (a DGP), i (a statistic), and j (a DGP) as forming a three-dimensional array, with k varying in the horizontal direction and i in the vertical direction of a flat two-dimensional surface, and j varying perpendicular to that surface. Fixing k in $R_1(\alpha, k)$ puts us on a sheet defined by the vertical i-direction and the perpendicular j-direction. The sum in the right-hand expression above takes us on a trip across this sheet, where, as we vary j, the i-coordinate traces out the values $q_j^1(\alpha)$.

Next, we wish to compute the double bootstrap P value p_2. By definition, $p_2 = R_1(p_1, \beta)$, so that, for a realisation (i, j),

$$p_2 = R_1(A_{ji}, j) = \sum_{l=1}^{m} a_{jq_l^1(A_{ji})l} \tag{8}$$

where the first equality follows because $p_1 = A_{ji}$, and $\beta = j$, and the second follows from the definition (7). Make the definition

$$A_{ji}^2 = \sum_{l=1}^{m} a_{jq_l^1(A_{ji})l}, \tag{9}$$

where the superscript 2 is an index, not an exponent. Then the double bootstrap P value p_2 is A_{ji}^2; compare with the relation (5).

The recursive scheme for obtaining the sequence of iterated bootstrap P values is given in the following Theorem.

Theorem 1

Let the finite, discrete, model $\mathbb{M}$ consist of m DGPs, labelled by $j = 1, \ldots, m$. Each DGP in $\mathbb{M}$ can generate realisations of the pair of random variables (τ, β), where the statistic τ takes on the values π_i, $i = 0, 1, \ldots, n$ satisfying (2), and the bootstrap DGP β takes on values $j = 1, \ldots, m$. Under the DGP labelled by k, the probability of the realisation (i, j) is p_{kij}. Define the cumulative probabilities a_{kij} and A_{ki} by the relations (4). Then the sequence $\{p_r\}$, of iterated bootstrap P values for realisation (i, j) is given by the recurrence:

$$p_r = A_{ji}^r, \tag{10}$$

$$A_{ji}^{r+1} = \sum_{l=1}^{m} a_{jq_l^r(A_{ji}^r)l}, \tag{11}$$

$$q_l^r(\alpha) = \max_{i=1,\ldots,n+1} \{i \mid A_{li}^r \leq \alpha\}, \tag{12}$$

initialised by the definitions

$$A_{ji}^1 = A_{ji}; \quad q_l^1(\alpha) = \max_{i=0,1,\ldots,n+1} \{i \mid A_{ji} \leq \alpha\}.$$

Proof:

By (5), the ordinary bootstrap P value p_1 is given by A_{ji} for realisation (i, j). By (11) with $r = 1$ and the given initialisation, the double bootstrap P value p_2 is

$$A_{ji}^2 = \sum_{l=1}^{m} a_{jq_l^1(A_{ji}^1)l},$$

in agreement with (9). Note that the A_{ji}^1 are strictly increasing in i for all $j = 1, \ldots, m$, given that the inequalities in (2) are strict. Suppose now that (10) holds for some given $r \geq 2$, and that the A_{ji}^r are strictly increasing in i for all $j = 1, \ldots, m$. It follows from (11) that the A_{ji}^{r+1} are also strictly increasing in i. With p_r defined by (10), the CDF of p_r under DGP k is

$$R_r(\alpha, k) = \sum_{i=0}^{n} \sum_{j=1}^{m} p_{kij} \mathrm{I}(A_{ji}^r \leq \alpha). \tag{13}$$

From (12), it can be seen that the inequality $A_{ji}^r < \alpha$ is equivalent to $i < q_j^r(\alpha)$, and so

$$R_r(\alpha, j) = \sum_{l=1}^{m} \sum_{i=0}^{q_l^r(\alpha)-1} p_{jil} = \sum_{l=1}^{m} a_{jq_l^r(\alpha)l}. \tag{14}$$

By the general definition (1) of iterated bootstrap P values, it can be seen that, for realisation (i, j),

$$p_{r+1} = R_r(p_r, j) = \sum_{l=1}^{m} a_{jq_l^r(A_{ji}^r)l} = A_{ji}^{r+1}, \tag{15}$$

where the last equality follows from (11). Thus (10) holds for $r + 1$, and the theorem is proved. ■

The question of the convergence of the sequence $\{p_r\}$ as $r \to \infty$ is the subject of the next theorem.

Theorem 2

> The recurrence (10), (11), and (12) defines the sequence $\{A_{ji}^r\}$ as a Markov process. The sequence converges for some finite r, or else it is locked in a limit cycle.

Proof:

Inasmuch as the probabilities a_{jil} are given once and for all in the definition of the probabilistic structure of the model, it follows from (11) that the transition from the A_{ji}^r to the A_{ji}^{r+1} is completely determined by the three-dimensional array of integers $q_l^r(A_{ji}^r)$, $l, j = 1, \ldots, m$, $i = 0, 1, \ldots, n$. By (12) with $\alpha = A_{li}^r$, these integers in turn are completely

determined by the A^r_{li}. This shows that the sequence $\{A^r_{ji}\}$ is a Markov process, as also the sequence $\{q^r_l(A_{ji})\}$.

Since the values of each integer $q^r_l(A_{ji})$ are limited to the finite set $\{0, 1, \ldots, n, n+1\}$, there must be at least one configuration of these integers that occurs infinitely often in the sequence. If the first occurrence is at iteration r, then the successor configuration for $r+1$ may be the same as for r, in which case the sequence converges at iteration r, or else there is a limit cycle constituted by the set of configurations from iteration r on until the next occurrence of the configuration at iteration r. ■

Remarks:

The limit cycle possibility seems unlikely to arise with realistic models, for which the statistic τ is an approximate pivot, and the bootstrap DGP β under a DGP k is in some sense close to k. Indeed, I have been unable to construct even a pathological example that leads to a limit cycle.

Although one of the outcomes allowed by Theorem 2 must certainly occur for some finite r, convergence to this outcome need not be quick, given the number of possible configurations of the $q^r_l(A_{ji})$, namely $m^2(n+2)$.

The mapping, independent of r, which takes the A^r_{ji} to their successors A^{r+1}_{ji}, is of course very nonlinear, and is not a contraction mapping, as can be seen by the following argument. For some $l = 1, \ldots, m$ and some iteration r, consider the $(n+2) \times m$ matrix with element (i, j) given by $q^r_l(A^r_{ji})$. All elements of the row with $i = 0$ are necessarily equal to 0, since $A^r_{j0} = 0$ and all elements of the row with $i = n+1$ are equal to $n+1$, since $A^r_{j(n+1)} = 1$. If all other elements of a column j are equal to 0 at iteration r, they remain equal to 0 at all successive iterations, and similarly, if all elements except the first are equal to 1, they remain so subsequently. Thus any configuration in which every column either has all elements except the last equal to 0, or else has all elements except the first equal to 1 is a fixed point of the mapping. Since a contraction mapping has a unique fixed point, the mapping defined by the recurrence cannot be a contraction mapping.

4. Bootstrap Discrepancies

In general, the bootstrap discrepancy is defined, for a given DGP μ in a model $\mathbb{M}$, and for a given significance level α, as the difference between the rejection probability of the bootstrap test under the DGP μ and level α and α itself. It is therefore given by $R_1(\alpha, \mu) - \alpha$, where as before $R_1(\cdot, \mu)$ is the CDF of the bootstrap P value p_1 under μ. For iterated bootstraps, the definition is the same: at level r, the order-r discrepancy is $R_r(\alpha, \mu) - \alpha$. The aim in bootstrapping, and in iterated bootstrapping, is to minimise the bootstrap discrepancy. If it is zero, the bootstrap permits exact inference.

There are two trivial cases in which the discrepancy is zero, and the iterated bootstrap P values coincide with the single bootstrap P value. The first case arises when the statistic τ is a pivot, the second when the bootstrap DGP coincides with the true DGP. It is a useful

sanity check to ensure that these properties hold for the discrete model of the preceding section.

First, if τ is a pivot, the A_{ki} do not depend on k; recall that A_{ki} is the probability under k that $\tau < \pi_i$. We may express this by writing $A_{ki} = A_{\cdot i}$. This means that $q_j^1(\alpha)$ depends only on α, not on j; we write $q_j^1(\alpha) = q_\cdot^1(\alpha)$. We see that, in this case,

$$R_1(\alpha, k) = \sum_{j=1}^{m} a_{kq_\cdot^1(\alpha)j} = A_{\cdot q_\cdot^1(\alpha)}, \tag{16}$$

and this does not depend on k. Therefore the bootstrap P value is also a pivot, with the same distribution whatever the DGP k, and the probability that it is less than α is the probability that $i < q_\cdot^1(\alpha)$, which would be exactly equal to α without the discreteness of the problem, and is indeed equal to α for $\alpha = A_{\cdot i}$, since $q_\cdot^1(A_{\cdot i}) = i$, and the probability that $\tau < \pi_i$ is $A_{\cdot i} = \alpha$. For these values of α, therefore, the bootstrap discrepancy is zero.

For realisation (i, j), the bootstrap P value is $A_{\cdot i}$. The double bootstrap P value is, from (8), $A_{ji}^2 = R_1(A_{\cdot i}, \cdot)$, and, from (16), this is $A_{\cdot q_\cdot^1(A_{\cdot i})}$. Now

$$q_\cdot^1(A_{\cdot i}) = \max_j \{j \mid A_{\cdot j} \leq A_{\cdot i}\},$$

and the value of the right-hand side is clearly just i. Thus the double bootstrap P value is $A_{\cdot i}$, the same as the single bootstrap P value, and also therefore independent of j.

The other special case arises when the bootstrap DGP always coincides with the true DGP. If so, then, for all admissible k, i, and j, we have $a_{kij} = A_{ki}\delta_{kj}$, where δ_{kj} is the Kronecker delta. The factor of A_{ki} on the right-hand side of this equation is justified, because

$$A_{ki} \equiv \sum_{j=1}^{m} a_{kij} = A_{ki} \sum_{j=1}^{m} \delta_{kj} = A_{ki}.$$

From (7) we have

$$R_1(\alpha, k) = \sum_{j=1}^{m} a_{kq_j^1(\alpha)j} = \sum_{j=1}^{m} \delta_{kj} A_{kq_j^1(\alpha)} = A_{kq_k^1(\alpha)}.$$

In particular, for $\alpha = A_{ki}$, $i = 0, \ldots, n$, since $q_k^1(A_{ki}) = i$, $R_1(A_{ki}, k) = A_{ki}$, which implies that for these values of α, the bootstrap discrepancy is zero. The double bootstrap P value is A_{ki}^2, which is

$$A_{ki}^2 = \sum_{j=1}^{m} a_{kq_j^1(A_{ki})j} = A_{kq_k^1(A_{ki})} = A_{ki}.$$

Thus, as in the case of a pivotal statistic, so too here the bootstrap discrepancy is zero for $\alpha = A_{ki}$, and the double bootstrap P value is identical to the single bootstrap P value, as are all higher-order P values.

When the discrepancy is not zero, it can be examined graphically using a P value plot or a P value discrepancy plot; see Davidson and MacKinnon (1998). Without bootstrapping, the former plot under DGP k of the discrete model is the locus of points (π_i, A_{ki}), since A_{ki} is the probability that $\tau < \pi_i$, that is, the actual RP at significance level π_i. The latter plot is the locus of points $(\pi_i, A_{ki} - \pi_i)$. Exact inference corresponds to the former tracing out the 45° line across the unit square, or the latter tracing out the horizontal axis.

Theorem 3

Under DGP k of the discrete model, the P value plot for the order r bootstrap contains the points (A^r_{ki}, A^{r+1}_{ki}), $i = 0, 1, \ldots, n$, and the P value discrepancy plot the points $(A^r_{ki}, A^{r+1}_{ki} - A^r_{ki})$. If the recurrence defined by (11) and (12) converges at order r, the bootstrap discrepancy is zero for all significance levels A^r_{ki}.

Proof:

The CDF under DGP k of the order r bootstrap P value p_r is the function $R_r(\cdot, k)$, given in (13). Thus the probability that $p_r < A^r_{ki}$ is $R_r(A^r_{ki}, k)$, which, by (15), is equal to A^{r+1}_{ki}. Hence the point (A^r_{ki}, A^{r+1}_{ki}) lies on the P value plot for order r, and the point $(A^r_{ki}, A^{r+1}_{ki} - A^r_{ki})$ lies on the P value discrepancy plot.

If the recurrence defined by (11) and (12) converges at order r under DGP k, then $A^{r+1}_{ki} = A^r_{ki}$ for all $i = 0, 1, \ldots, n$. Hence the bootstrap discrepancy at level A^r_{ki} is $A^{r+1}_{ki} - A^r_{ki} = 0$. ■

Remark:

It is important to note here that Theorem 3 does *not* guarantee that exact inference is available with the fully iterated bootstrap for an arbitrarily chosen significance level. In the example of the next section, it is seen that, in some cases, the levels at which the bootstrap discrepancy is zero are roughly uniformly spread out over the unit interval, so that a level chosen arbitrarily will be close to one with zero discrepancy. In other cases, however, levels with zero discrepancy are bunched up near zero or one, with the result that the discrepancy at some conventional level may be far from zero.

5. An Example

Whereas it is easy to make the set of P values discrete, for instance by choosing $n = 100$ and letting a P value take on values $0.00, 0.01, 0.02, \ldots, 0.99, 1.00$ only, there are not very many examples for which it is obvious how to make the null model discrete. An exception to this, which we study in this section, is the model treated in Davidson (2010). The model can be expressed as

$$y_t = y_{t-1} + u_t, \quad u_t = v_t + \theta v_{t-1}, \quad t = 1, \ldots, T \tag{17}$$

where the v_t are Gaussian white noise with variance 1, so that y_t is a unit-root process, obtained by cumulating the MA(1) process u_t with MA parameter θ. Since (17) takes the

form of a recurrence relation, it must be initialised. In order to make u_t stationary, v_0 is set to a standard normal variable, independent of the v_t, $t = 1, 2, \ldots, T$.

The model is tested by a unit-root test, for which the test statistic is the τ_c version of the augmented Dickey-Fuller (ADF) test, proposed by Dickey and Fuller (1979), and justified asymptotically under much less restrictive assumptions by Said and Dickey (1984) and Phillips and Perron (1988). The statistics are computed using the ADF testing regression

$$\Delta y_t = \beta_0 + \beta_1 y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta y_{t-1} + \text{residual},$$

where p is a lag truncation parameter, usually chosen in a data-determined way on the basis of some information criterion. When this regression is run by ordinary least squares, the τ_c statistic is the conventional t statistic for the hypothesis that $\beta_1 = 0$. Under the null hypothesis that the series y_t has a unit root, this statistic has a well-known but nonstandard asymptotic distribution.

With all of these assumptions, the model (17) is purely parametric, with only one parameter, namely θ. It is therefore easy to make the model discrete, by choosing a grid of values for θ. Since the model is parametric, the bootstrap is also parametric, and is characterised by a value of θ. In our discrete model, these values are restricted to the chosen grid. The parameter θ has to be estimated under the null hypothesis, as discussed in Davidson (2007) – the second "Golden Rule of Bootstrapping". In his earlier study, Davidson used a nonlinear least-squares (NLS) procedure that gave results comparable to those of maximum likelihood.

Testing for a unit root in a series obtained by summing a stationary MA(1) process with a parameter close to -1 leads to serious size distortions under the null, on account of the near cancellation of the unit root by the MA component in the driving stationary series u_t. Davidson (2010) found that, for a sample size of 100, the best results in terms of minimising size distortion under the null were obtained with a lag truncation parameter of $p = 12$. Since we wish to study the effect of bootstrap iteration, it is very desirable to consider a setup in which the size distortion is non-negligible, and so the neighbourhood of $\theta = -1$ is suitable in this regard. Further, as θ approaches -1 from above, the distortion changes sign from negative to positive.

In Figure 1 are plotted the 0.01, 0.05, and 0.10 quantiles of the distribution of the asymptotic P value obtained from the raw τ_c statistic by use of the program described in MacKinnon (1994) and MacKinnon (1996), plotted as a function of θ from -0.4 to -0.99, and for a sample size $T = 100$.

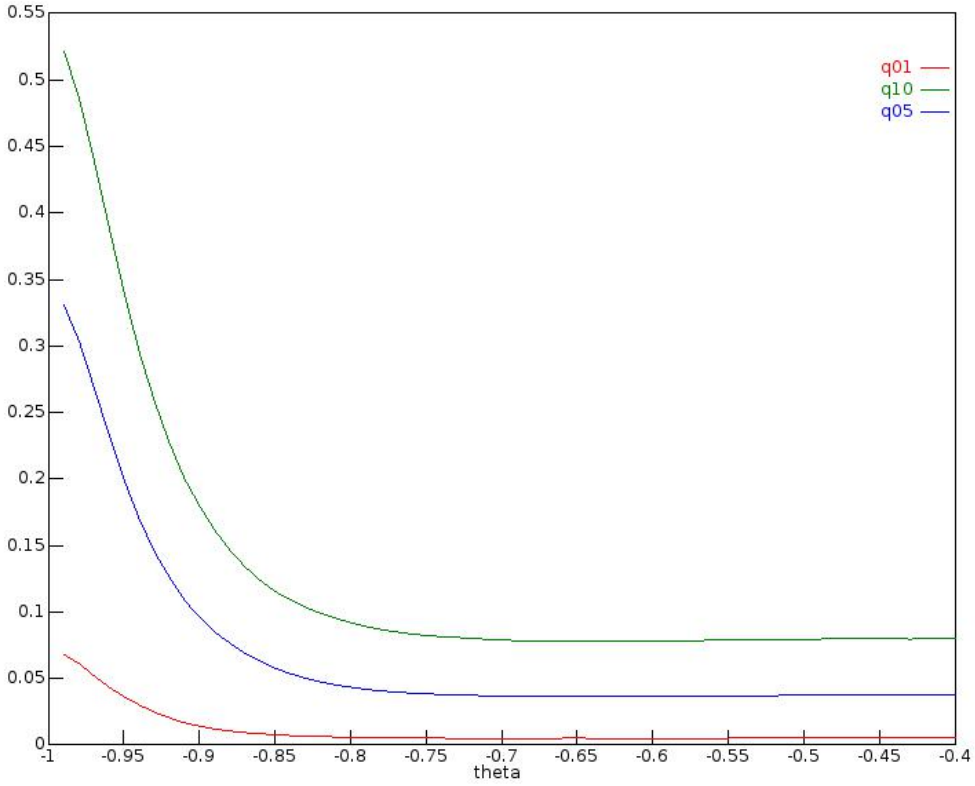

Figure 1

It can be see that, for values of θ from -0.4 to around -0.7, the curves are nearly flat, with P values close to nominal. Thus the distribution does not change much over this range. From -0.7 to -0.99, however, the distribution is very sensitive indeed to the value of θ, and this implies difficulties for the bootstrap if data are generated by a DGP in this region. The greatest size distortions are to be expected around $\theta = -0.90$, since it is in that neighbourhood that the curves bend most. Closer to $\theta = -1$, although the curves are very far from flat, they follow more or less straight lines. According to the analysis in Davidson and MacKinnon (1999), this fact helps to reduce distortion more than in regions in which the graphs are curved.

The discrete model is as follows:

- The number of possible test statistics, $n + 1$, is given by $n = 100$, with probabilities $\pi_i = i/100$, $i = 0, \dots, n$, forming the grid $0.00, 0.01, 0.02, \dots, 0.99, 1.00$.
- The τ_c test statistic is converted into an asymptotic P value by use of the program described in MacKinnon (1994) and MacKinnon (1996). The P value is rounded to the closest point on the grid defined above.
- The discrete model $\mathbb{M}$ is defined, with $m = 60$ points, for the DGPs (17) with parameters θ on the grid $-0.40, -0.41, -0.42, \dots, -0.98, -0.99$. The value of $\theta = -1$ must of course be excluded, as belonging to the alternative hypothesis by which y_t is an I(0) process rather than to the null of an I(1) process.

- The parameter θ which defines the bootstrap DGP is estimated by the NLS procedure mentioned above. Values greater than -0.40 are rounded down to this value; values between -0.99 and -1 are rounded up to -0.99; intermediate values are rounded to the nearest point on the grid.

For the iteration procedure defined in Section 3 to be implemented, it was necessary to estimate by simulation the $60 \times 101 \times 60 = 363{,}600$ probabilities p_{kij} defined in (3). To this end, 60 separate simulation experiments were undertaken, for each of the 60 DGPs of the discrete model, on a machine with 64 cores. There were 100,000 replications for each of the 60 experiments, and the whole set of experiments took about an hour and a quarter of computer time. When all of the p_{kij} had been estimated in this way, the probabilities A_{ki} defined in (4) were constructed, and then the successive iterates A^r_{ki} computed using (11) and (12). Convergence of the recurrence was achieved after 60 iterations. Since one iteration of the bootstrap gives what we call the double bootstrap, it is the 61-tuple bootstrap which can be identified with the infinitely iterated bootstrap. As expected on the basis of the evidence in Figure 1, convergence was slowest for the DGPs with θ around -0.95.

Selected results

The iterative procedure yields so many results, many of them rather uninteresting, that I limit myself here to a selection of more interesting things that emerge from the computations.

In Figures 2, 3, and 4, are shown results similar to some results in Davidson (2010). For $\theta = -0.90$, -0.95, and -0.99, P value discrepancy plots are given for bootstrap iterations, starting with the ordinary (single) bootstrap, through the quintuple bootstrap. In order to give a complete picture, the abscissa α in these graphs takes the values $i/100$, $i = 0, \dots, 100$, and the ordinate is computed as $R_r(\alpha, j) - \alpha$, with R_r given by (14), for the values of j that correspond to the given values of θ. If a graph hits the horizontal axis for some α, this means that that iteration of the bootstrap gives exact inference for level α. Rather than studying the discrepancy for conventional significance levels only, this graphical approach allows us to see the extent to which the iterated bootstrap P values converge to a discrete distribution close to the continuous $\mathrm{U}(0, 1)$ distribution. Although this may be of limited practical importance, it is of essential theoretical interest.

For $\theta = -0.90$, the discrepancy plots appear to converge monotonically to the horizontal axis, implying that the quintuple bootstrap and higher iterations give exact inference for a rather dense set of levels, up to the errors induced by the discretisation and the simulation errors in estimating the probabilities p_{kij}. However, for $\theta = -0.95$ things look rather different. The results are plainly contaminated by more noise than for $\theta = -0.90$, and it is not clear whether there really is convergence. For $\theta = -0.99$, the switch from under- to over-rejection is evident. Here, it is possible to believe that the discrepancy plots are converging to the horizontal axis, but that has not happened even for the quintuple bootstrap.

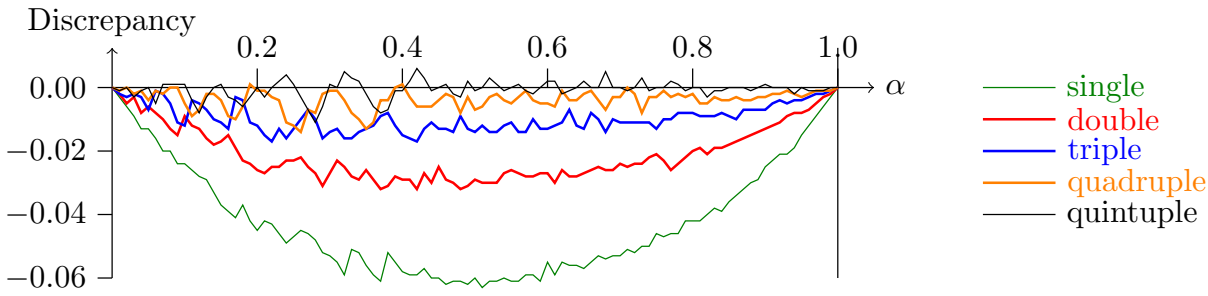

Figure 2: $T = 100$, $\theta = -0.90$

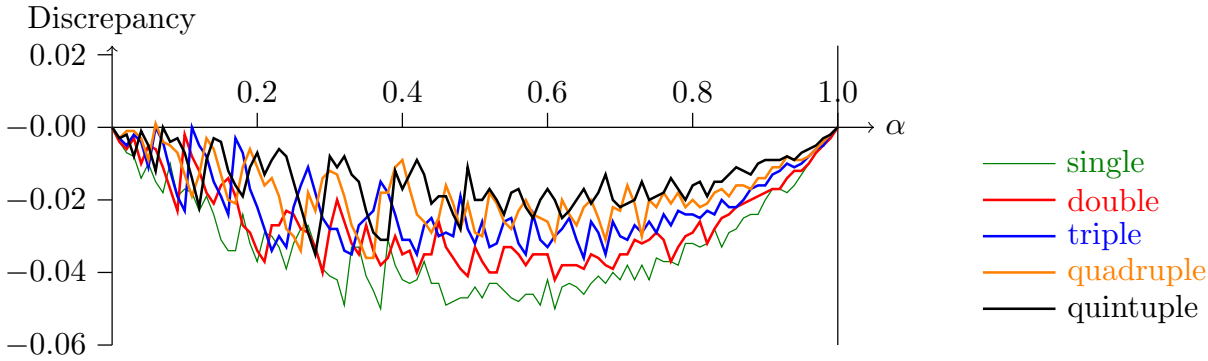

Figure 3: $T = 100$, $\theta = -0.95$

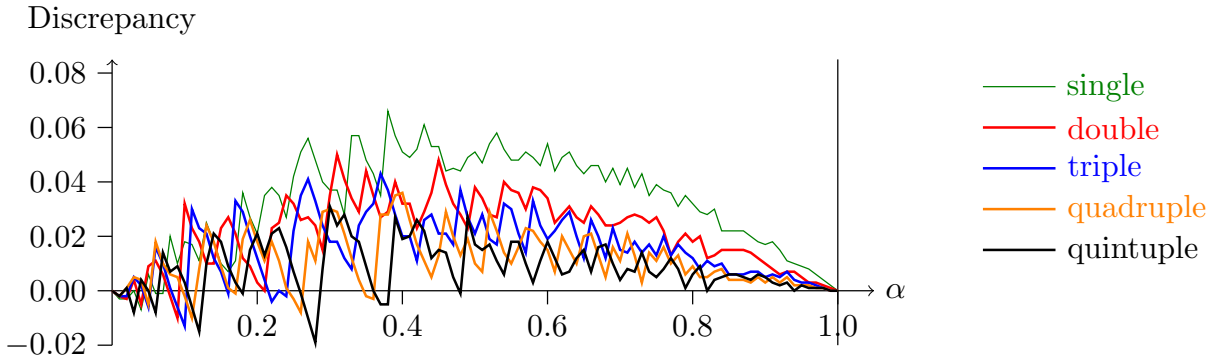

Figure 4: $T = 100$, $\theta = -0.99$

It can be remarked here that, for values of θ greater than -0.90, convergence to the horizontal axis is essentially complete by the quadruple or quintuple bootstrap. One example is presented in Figure 5, for $\theta = -0.70$. Notice the scale of the vertical axis. Evidently, the distortion even of the single bootstrap is not very great.

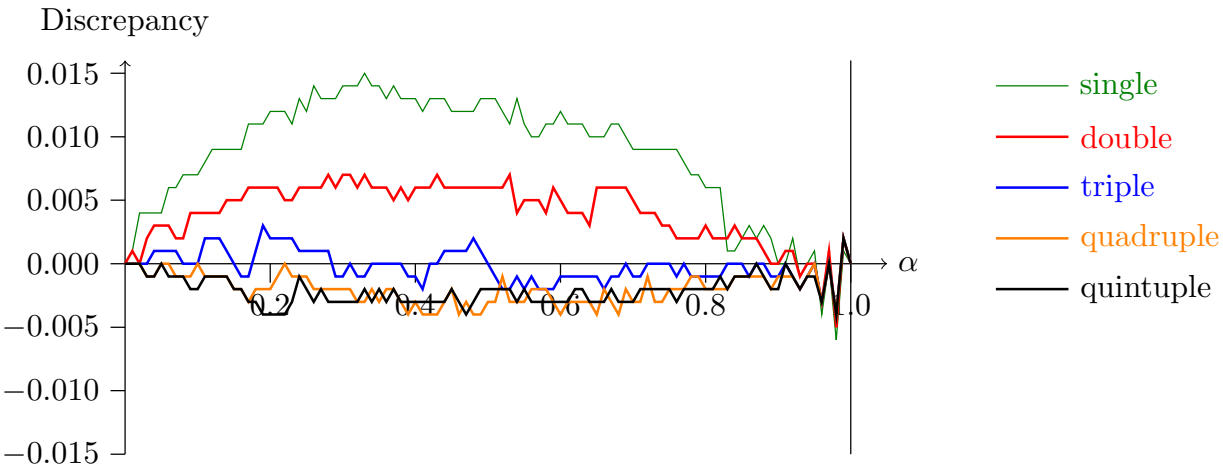

Figure 5: $T = 100$, θ = -0.70

After the recurrence has converged, we can look at the discrepancy plot for the infinitely iterated, or 61-tuple, bootstrap. In Figure 6, these discrepancy plots are shown for θ equal to -0.70, -0.90, -0.95, and -0.99. Because the recurrence converged, the discrepancy in all cases must be exactly equal to zero for the nominal levels in the relevant rows of the matrix we may write as $\boldsymbol{A}^{\infty}$, with element (i, j) equal to A^{60}_{ji}. For θ between -0.95 and -0.99, a great many of these levels are in the neighbourhood of 1. For $\alpha = 1$, of course, inference is trivially exact. All of the curves cross the horizontal axis several times, and are almost coincident with it for values of θ not in the close neighbourhood of -1. The fact that the discrepancies are not zero everywhere is of course a consequence of the discrete nature of the model. However, even for $\theta = -0.90$, and more so for $\theta = -0.70$, the discrepancy is everywhere very small indeed.

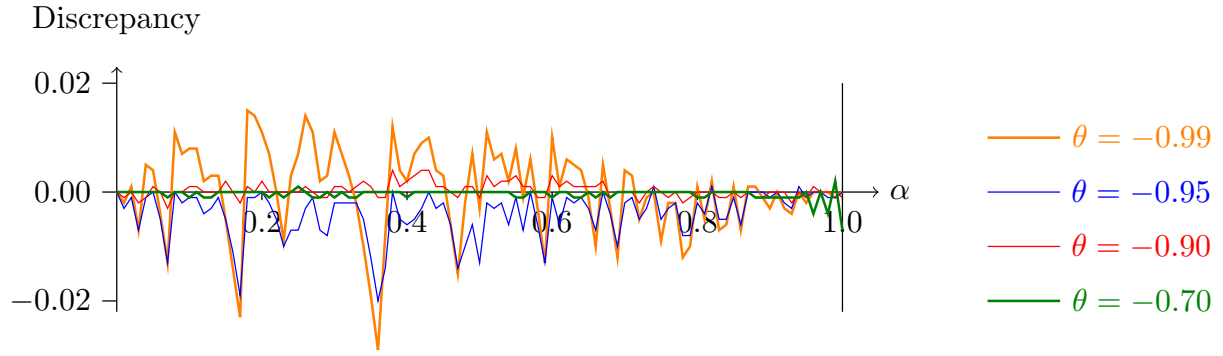

Figure 6: $T = 100$, fully iterated bootstraps

A referee suggested that, as the approach of this paper is in no way asymptotic, it would be interesting to see how it performs for a much smaller sample size. A second set of experiments was therefore undertaken with $T = 20$. Convergence of the iterated bootstrap P values was much slower, requiring 660 iterations. The P value discrepancy plots for the fully iterated bootstrap are shown in Figure 7 for $\theta = -0.7, -0.9, -0.95$, and -0.99. The distortions are, if anything, smaller than those for $T = 100$.

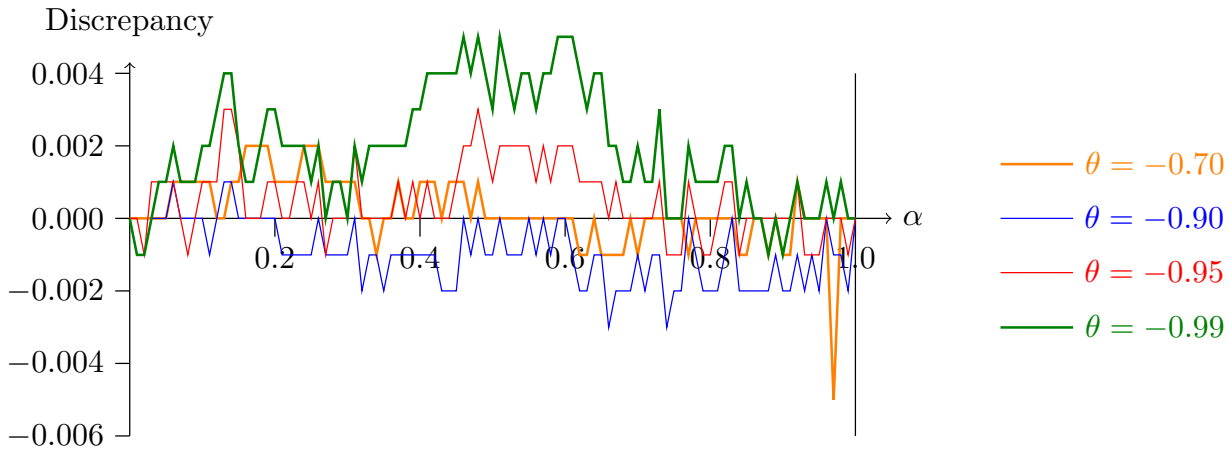

Figure 7: $T = 20$, fully iterated bootstraps

6. Power

Any time that a procedure seems to give better control of Type I error, it is advisable to see to what extent this has been bought at a cost in terms of power. However, power cannot be unambiguously defined when a non-pivotal test statistic is used, since, without some means of size or level adjustment, the raw RP of the test under a DGP not in the null hypothesis cannot be interpreted as power. See Horowitz and Savin (2000) for a full discussion of this point; they recommend the raw RP of a well-constructed bootstrap test as the best that can be done. That is what is done in this section using a fully iterated bootstrap P value for the test statistic.

Four DGPs were investigated, characterised as follows:

$$y_t = \rho y_{t-1} + u_t, \quad u_t = v_t + \theta v_{t-1}, t = 1, \dots, 100,$$

where $\theta = -0.7$ for all four, and the values of the AR parameter ρ, successively farther from the null-hypothesis value of 1, were 0.99, 0.95, 0.90, and 0.85. The value of θ was chosen in order to avoid the near cancellation that arises if ρ is close to $-\theta$.

As with the DGPs in the null model, an experiment involving 100,000 replications was carried out for each of the four DGPs (in parallel), in order to estimate the probabilities of obtaining each of the $101 \times 60 = 6060$ realisations (i, j), $i = 0, 1 \dots, 100$, $j = 1, \dots, 60$, allowed by the model. For each realisation (i, j), the corresponding fully iterated bootstrap P value, A^{60}_{ji} was found. The probability under any of the four non-null DGPs of realisations (i, j) such that $A^{60}_{ji} < \alpha$ is the RP of the fully iterated bootstrap test at level α, which can then properly be interpreted as the power.

Graphs of the size-power tradeoff, as proposed in Davidson and MacKinnon (1998), are shown in Figure 8, with significance level α on the horizontal axis, and the RP at that level on the vertical axis.

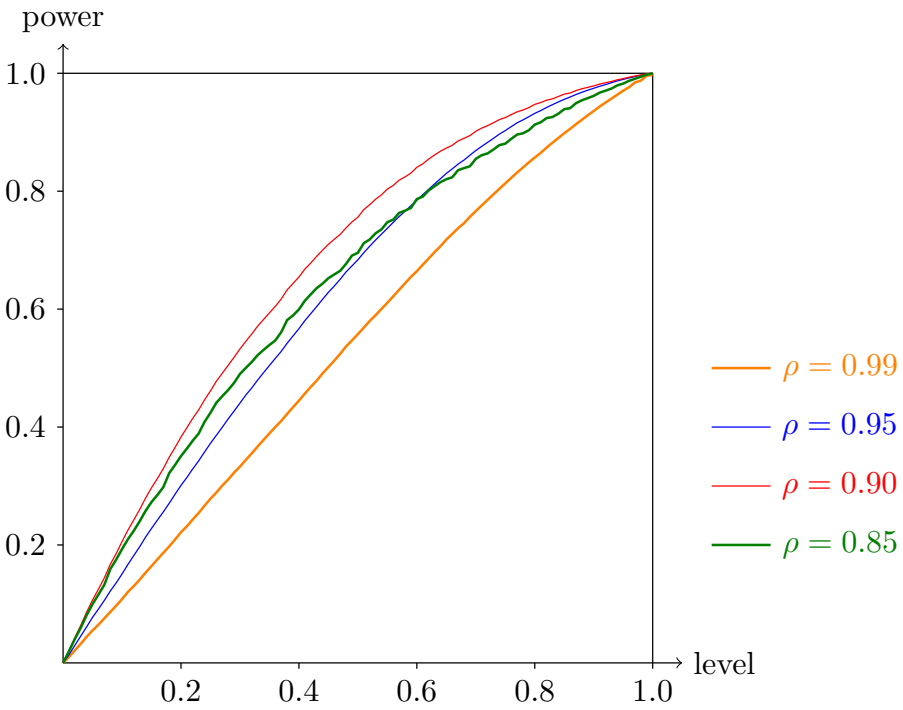

Figure 8: Size-power tradeoff; iterated bootstrap, θ=-0.7

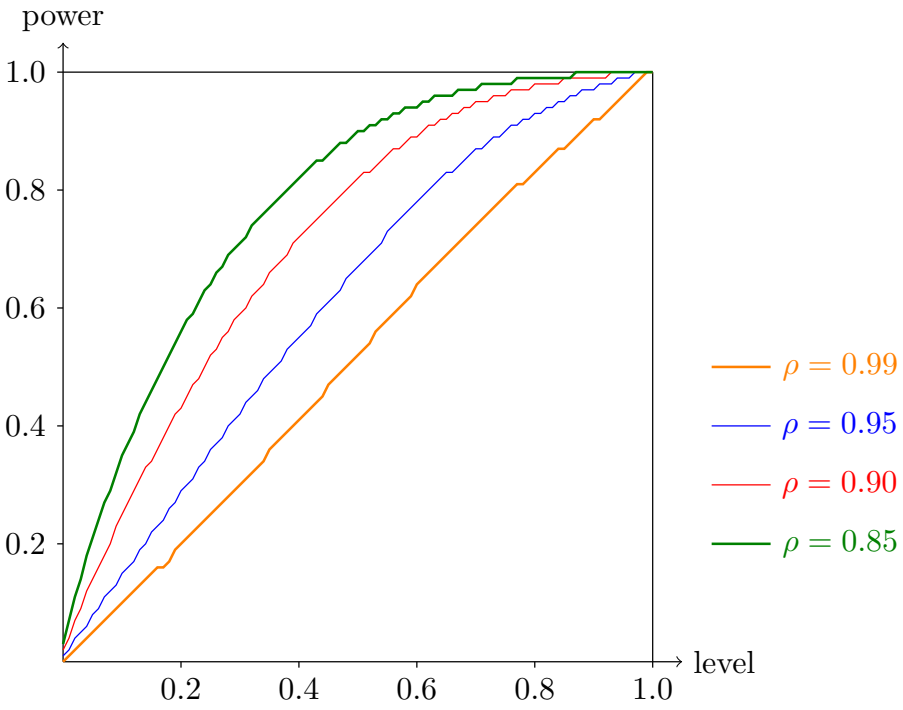

Figure 9: Size-power tradeoff; asymptotic test, θ=-0.7

For comparison purposes, Figure 9 shows the apparent size-power tradeoff if the test is based on the asymptotic P value τ. For $\theta = -0.7$, Figure 1 suggests that the distribution of this asymptotic P value should be close to nominal under the null hypothesis, implying that a comparison of the curves in Figures 8 and 9 is not unreasonable. It is clear that

there is no serious power loss occasioned by use of the iterated bootstrap P value, except for $\rho = 0.85$. This last feature is accounted for by noting that estimating θ under the assumption of a unit root when in fact ρ is not far removed from $-\theta$ biases the estimate in the direction of -1. Again according to Figure 1, the critical values of the bootstrap distribution are larger than for the true DGP, leading to a loss of power. This presumably could be alleviated by estimating θ under the alternative, although this would normally be bad practice.

7. Discussion

The study in this paper raises a number of issues. One obvious one is to what extent the use of a discrete model is feasible in practice, with the sort of bootstrap commonly used, rather than one governed by a single scalar parameter. Two parameters are probably feasible, but with any more one swiftly runs afoul of the curse of dimensionality.

It is clear that the most important advantage of this discrete approach is that it eliminates conventional bootstrapping based on a simulation experiment. In exchange, it is necessary to conduct the probably costly simulation experiment needed in order to estimate the p_{kij} of (3). However, once that experiment is carried out, it serves as a fixed overhead for arbitrary levels of bootstrap iteration. It thus becomes feasible to examine the convergence or otherwise of the sequence of iterated bootstrap P values.

In the discrete case, Theorem 2 tells us that the sequence of iterated bootstrap P values converges except in the unlikely case of a limit cycle. However, this need not imply that exact inference is possible for significance levels of interest, since Theorem 3 makes it clear that the fully iterated bootstrap discrepancy is exactly zero only for a specific set of significance levels. Figure 6 makes it clear that, for $\theta = -0.99$, there is non-negligible distortion away from these levels.

It is quite possible to concoct pathological examples in which, after convergence, the only levels capable of giving exact inference are zero and one; hardly a desirable state of affairs. One such example is given by a model with only two DGPs, and a statistic the distribution of which under one of these strictly stochastically dominates the distribution under the other. If the bootstrap DGP is always, deterministically, just one of the two, then, under that one, as we saw in Section 4, convergence is immediate and allows exact inference. Under the other DGP, numbered 2 say, however, convergence is eventually achieved, with all of the A_{2i} equal to 0 or 1, except for the first or the last, depending on the direction of the dominance.

What we saw for $\theta = -0.99$ is far from being as extreme as the above example, but it brings out another point, namely that, for one and the same model, for some DGPs we may have convergence, in the discrete case, of the sort we observed for θ greater than -0.90, while, for others, it may be more like what we saw with $\theta = -0.99$. In the continuous case, convergence to a uniform limit may occur for some DGPs in the model, but not for others. If there is such convergence for all DGPs in the null model, then the infinitely iterated bootstrap P value is a pivot with respect to that model.

The most frequently used sort of bootstrap is a resampling bootstrap. This implies that the bootstrap distribution is discrete, although it is usually presumed that the underlying distribution or distributions are continuous. It is in any case impossible to iterate a resampling bootstrap indefinitely, because each resample has fewer separate objects than the sample from which the resample is drawn. If one iterates a resampling bootstrap, eventually there will be only one element of the original sample left, which one presumably being randomly selected with equal probabilities for all the elements. "Convergence" in such a case would be meaningless.

If we abstract from the simulation noise in the estimation of the p_{kij}, the discrete model is quite nonrandom. We are, in effect, working simultaneously with every point in the outcome space. Convergence, therefore, is to be understood in the ordinary sense of convergence of a sequence of real numbers. In the continuous case, of course, we have to speak of stochastic convergence, which may perhaps be almost sure, or in probability. If this discrete approach were to be used with real data, it would be necessary to use these data to compute realisations of the quantity being bootstrapped and of the bootstrap DGP, and then to discretise them according to the plan of discretisation in use. If the realised quantity is indexed by i and the realised bootstrap DGP by j, then for the r-tuple bootstrap, the bootstrap P value is A^r_{ji}.

Many questions remain regarding the numerical stability of the discrete model. It would for instance be useful to examine to what extent simulation error in the preliminary estimation of the p_{kij} is propagated through the iterations; I plan to study this in future work. Another point that will bear future investigation is to what extent coarseness or fineness of the discretisation matters, both for the statistics (the P values) and for the DGPs of the model. If relatively coarse discretisation yields satisfactory results, this would be of enormous importance for any practical use of this approach.

It would be immensely useful to find ways of discretising the set of bootstrap DGPs used in situations that are not purely parametric. While it is easy enough to replace the use of a discrete empirical distribution for resampling by a continuous version, thus avoiding the problem inherent to iterating a conventional resampling bootstrap, it is not obvious how to make discrete the set of bootstrap DGPs that would be obtained in this way. I conjecture that, when bootstrapping an approximately pivotal statistic, it may be possible to cover the set of bootstrap DGPs rather coarsely and still achieve satisfactory results. Although how best to do so remains to be seen, it is not at all necessary to adopt the strategy of the example in Section 5, and to use an equally spaced grid of either P values or the parameter(s) that define the bootstrap DGP, and this suggests that it may be advantageous to make use of adaptive sparse grids, of the sort proposed in Brumm and Scheidegger (2015); see also the many relevant references in that paper.

The double bootstrap was introduced by Beran in two papers, Beran (1987) and Beran (1988), in which he refers to "pre-pivoting", meaning making some quantity more close to being pivotal for a model by bootstrapping it, and then bootstrapping the result. This interpretation clearly applies to higher orders of bootstrap iteration. In some sense, the iterative procedure, if it converges appropriately, serves to project the original statistic

into a space of pivotal statistics. It would be desirable to formalise this intuition. It is also necessary to see to what extent this "projection" may adversely affect the power of a test. Of course, power is not uniquely defined when a non-pivotal statistic is used, but, as remarked previously, if an iterated bootstrap P value follows the uniform $\mathrm{U}(0,1)$ distribution, it is by definition a pivot.

In discussing bootstrap "validity", it is conventional to make use of an appropriate asymptotic construction in order to show that the limiting distribution of the quantity considered is the same as the limiting distribution of its bootstrap counterpart, conditional on the realised data. This is of course a very weak requirement. A somewhat better justification for the bootstrap comes from any refinements that can be demonstrated by an asymptotic argument, as in Hall's (1992) book, where he uses Edgeworth expansion. But the bootstrap is not intrinsically an asymptotic procedure; rather, current bootstrap *theory* relies on asymptotic arguments. It seems to me that convergence of the sequence of iterated bootstrap P values to the uniform distribution is a much richer and more satisfactory means of justifying or validating the bootstrap. No asymptotic argument is involved, so that the potential arbitrariness of the choice of an asymptotic construction is avoided. To the extent that the approach of this paper can be made operational for problems of interest, the approach carries its validity along with it.

Further, the new proposed criterion for validity is by no means equivalent to asymptotic validity. An example of this is when a regression model, the disturbances of which are not necessarily Gaussian, is bootstrapped using a bootstrap DGP that imposes Gaussianity. Under very weak conditions on the asymptotic construction, this bootstrap is asymptotically valid. But it certainly is not, by the criterion of convergence of iterated P values to $\mathrm{U}(0,1)$, for any DGPs in the model the disturbances of which are in fact not Gaussian.

8. Concluding Remarks

In this paper, I have tried to take a step towards realising my ambition of freeing bootstrap theory from the use of asymptotic methods. The main idea is that the bootstrap can be justified – or not, as the case may be – by the convergence of the iterations of the bootstrap. This idea is developed here by making it tractable to study the bootstrap to any order of iteration, by means of a discretisation procedure that makes the model under test represented by a finite three-dimensional array of probabilities.

Much work remains to be done if the approach of the paper is to be useful, either in practice, or in the further development of bootstrap theory. One question seems particularly urgent: can we find sufficient conditions for convergence of the sequence of iterated P values to the uniform $\mathrm{U}(0,1)$ distribution in the continuous case, conditions that can be checked without excessive trouble in situations of practical interest? This, and other related questions, will be pursued in future work.

References

Beran, R. (1987). "Prepivoting to Reduce Level Error of Confidence Sets", *Biometrika* **74**, 457–68.

Beran, R. (1988). "Prepivoting test statistics: A bootstrap view of asymptotic refinements", *Journal of the American Statistical Association*, **83**, 687–97.

Brumm, J. and S. Scheidegger (2015). "Using Adaptive Sparse Grids to Solve High-Dimensional Dynamic Models".
Available at SSRN: http://ssrn.com/abstract=2349281 or
http://dx.doi.org/10.2139/ssrn.2349281

Chan, K. Y. F., and S. M. S. Lee (2001). "An exact iterated bootstrap algorithm for small-sample bias reduction", *Computational Statistics and Data Analysis*, **36**, 1–13.

Davidson, R. (2007). "Bootstrapping Econometric Models", *Quantile*, **3**, 13–36 (in Russian). English version available from the author's website:
http://russell-davidson.arts.mcgill.ca/articles/

Davidson, R. (2010). "Size Distortion of Bootstrap Tests: Application to a Unit Root Test", *Review of Economic Analysis*, **2**, 169–93.

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**, 361-76.

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.

DiCiccio, T. J., M. A. Martin, and G. A. Young (1992). "Analytical approximations for iterated bootstrap confidence intervals", *Statistics and Computing*, **2**, 161–71.

Dickey, D. A., and W. A. Fuller (1979). "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, **74**, 427–31.

Hall, P. (1986). "On the Bootstrap and Confidence Intervals", *The Annals of Statistics*, **14**, 1431–52.

Hall, P., and M. A. Martin (1988). "On Bootstrap Resampling and Iteration", *Biometrika*, **75**, 661–71.

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

Horowitz, J. L. (2001). "The Bootstrap", in *Handbook of Econometrics*, Volume 5, eds J. J Heckman and E Leamer, Elsevier.

Horowitz, J. L., and N. E. Savin (2000). "Empirically relevant critical values for hypothesis tests", *Journal of Econometrics*, 95, 375–389.

Lee, S. M. S., and G. A. Young (1995). "Asymptotic Iterated Bootstrap Confidence Intervals", *The Annals of Statistics*, **23**, 1301–30.

Lee, S. M. S., and G. A. Young (2003). "Prepivoting by weighted bootstrap iteration", *Biometrika*, **90**, 393–410.

MacKinnon, J. G. (1994). "Approximate asymptotic distribution functions for unit root and cointegration tests," *Journal of Business and Economic Statistics*, **12**, 167–76.

MacKinnon, J. G. (1996). "Numerical distribution functions for unit root and cointegration tests", *Journal of Applied Econometrics*, **11**, 601–18.

Martin, M. A. (1990). "On bootstrap iteration for coverage correction in confidence intervals", *Journal of the American Statistical Association*, **85**, 1105–18.

Phillips, P. C. B., and P. Perron (1988). "Testing for a unit root in time series regression," *Biometrika*, **75**, 335–46.

Said, E. S., and D. A. Dickey (1984). "Testing for unit roots in autoregressive-moving average models of unknown order," *Biometrika*, **71**, 599–607.