Davidson, Russell; MacKinnon, James G.

Working Paper

Improving the Reliability of Bootstrap Tests

Queen's Economics Department Working Paper, No. 995

Provided in Cooperation with:
Queen’s University, Department of Economics (QED)

Suggested Citation: Davidson, Russell; MacKinnon, James G. (2000) : Improving the Reliability of Bootstrap Tests, Queen's Economics Department Working Paper, No. 995, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:
https://hdl.handle.net/10419/189284

Queen's Economics Department Working Paper No. 995

Improving the Reliability of Bootstrap Tests

James MacKinnon
Queen's University

Russel Davidson
GREQAM and Queen's University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

9-2000

Improving the Reliability of Bootstrap Tests

by

Russell Davidson

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13002 Marseille, France

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

russell@ehess.cnrs-mrs.fr

and

James G. MacKinnon

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

jgm@qed.econ.queensu.ca

Abstract

We first propose procedures for estimating the rejection probabilities for bootstrap tests in Monte Carlo experiments without actually computing a bootstrap test for each replication. These procedures are only about twice as expensive (per replication) as estimating rejection probabilities for asymptotic tests. We then propose procedures for computing modified bootstrap P values that will often be more accurate than ordinary ones. These procedures are closely related to the double bootstrap, but they are far less computationally demanding.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada.

Revised, November, 2000.

1. Introduction

In many cases, bootstrap tests are surprisingly easy to perform, and the computational barriers to their routine use are dropping steadily. In general, the simplest approach to bootstrap testing is to calculate bootstrap P values. As we discuss in the next section, the test statistic itself is computed in the usual way, but the P value is computed by comparing the test statistic with the empirical distribution of a number of bootstrap test statistics rather than with the distribution that it follows asymptotically.

Theory suggests that bootstrap tests will generally perform better in finite samples than tests based on asymptotic theory, in the sense that they will commit errors that are of lower order in the sample size n; see, among others, Hall (1992) and Davidson and MacKinnon (1999b). Moreover, there is a growing body of evidence from simulation experiments which indicates that bootstrap tests do indeed yield more reliable inferences than asymptotic tests in a great many cases; relevant papers include Horowitz (1994), Nankervis and Savin (1996), Godfrey (1998), and Davidson and MacKinnon (1999a).

Although bootstrap P values will often be very reliable, this will not be true in every case. For an asymptotic test, one way to check whether it is reliable is simply to use the bootstrap. If the asymptotic and bootstrap P values associated with a given test statistic are similar, we can be fairly confident that the asymptotic one is reasonably accurate. Of course, having gone to the trouble of computing the bootstrap P value, we may well want to use it instead of the asymptotic one.

In a great many cases, however, asymptotic and bootstrap P values are quite different. When this happens, it is almost certain that the asymptotic P value is inaccurate, but we cannot be sure that the bootstrap one is accurate. In this paper, we discuss techniques for computing modified bootstrap P values which will tend to be similar to the ordinary bootstrap P value when the latter is reliable, but which should often be more accurate when it is unreliable. These techniques are closely related to the double bootstrap originally proposed by Beran (1988), but they are far less expensive to compute. In fact, the amount of computational effort beyond that needed to obtain ordinary bootstrap P values is roughly equal to the amount needed to compute the latter in the first place.

In the next section, we rapidly review bootstrap tests and a number of existing results on their properties. Then, in Section 3, we present the basic idea on which the techniques of the paper are based and show how the performance of bootstrap tests may be estimated in simulation experiments that require only twice as much computational effort (per replication) as that needed to estimate the performance of asymptotic tests. In Section 4, we show how modified bootstrap P values may be computed with only twice as much effort as ordinary bootstrap P values. In Section 5, we discuss the double bootstrap. In Section 6, we present some simulation results which illustrate how well the procedures proposed in this paper can work in practice. Finally, in Section 7, we discuss their relation to the sort of Edgeworth expansion often used in the asymptotic theory of the bootstrap.

2. Bootstrap Tests

Beran (1988) showed that bootstrap inference is refined when the quantity bootstrapped is asymptotically pivotal. We formalize the idea of pivotalness by means of a few formal definitions. A *data-generating process*, or *DGP*, is any rule sufficiently specific to allow artificial samples of arbitrary size to be simulated on the computer. Thus all parameter values and all probability distributions must be provided in the specification of a DGP. A *model* is a set of DGPs. Models are usually generated by allowing parameters and probability distributions to vary over admissible sets. A test statistic is a random variable that is a deterministic function of the data generated by a DGP and, possibly, other exogenous variables. A test statistic τ is a *pivot* for a model $\mathbb{M}$ if, for each sample size n, its distribution is independent of the DGP $\mu \in \mathbb{M}$ which generates the data from which τ is calculated. The *asymptotic distribution* of a test statistic τ for a DGP μ is the limit, if it exists, of the distribution of τ under μ as the sample size tends to infinity. The statistic τ is *asymptotically pivotal* for $\mathbb{M}$ if its asymptotic distribution exists for all $\mu \in \mathbb{M}$ and is independent of μ.

In hypothesis testing, the null hypothesis under test is represented by a model, as defined above. A test statistic is said to be pivotal or asymptotically pivotal under the null hypothesis if it is a pivot or an asymptotic pivot for the model that represents the hypothesis. Most test statistics commonly used in econometric practice are asymptotically pivotal under the null hypotheses they test, since asymptotically they have distributions, like standard normal, or chi-squared, that do not depend on unknown parameters. Conventional asymptotic inference is based on these known asymptotic distributions.

If an asymptotic pivot τ is not an exact pivot, its distribution depends on which particular DGP $\mu \in \mathbb{M}$ generates the data used to compute it. In this case, bootstrap inference is no longer exact in general. The bootstrap samples used to estimate the finite-sample distribution of τ are generated by a *bootstrap DGP*, which, although it usually belongs to $\mathbb{M}$, is in general different from the DGP that generated the original data.

It is possible to use the bootstrap either to calculate a critical value for τ or to calculate a P value. In this paper, we prefer the latter approach, as it greatly simplifies the analysis while being theoretically equivalent to an approach based on critical values. Suppose that data are generated by a DGP μ_0 belonging to $\mathbb{M}$, and used to compute a realization $\hat{\tau}$ of the random variable τ. Then, for a test that rejects for large values of the statistic, the P value we would ideally like to compute is

$$p(\hat{\tau}) \equiv \Pr_{\mu_0}(\tau > \hat{\tau}). \tag{1}$$

In practice, (1) cannot be computed, or estimated by simulation, because the DGP μ_0 that generated the observed data is unknown. If τ is an exact pivot, this does not matter, since (1) can be computed using any DGP in $\mathbb{M}$. In this case, $p(\hat{\tau})$ is a drawing from the $U(0,1)$ distribution. If τ is only an asymptotic pivot, the *bootstrap P value* is defined by

$$p^*(\hat{\tau}, \hat{\mu}) \equiv \Pr_{\hat{\mu}}(\tau > \hat{\tau}), \tag{2}$$

where $\hat{\mu}$ is a (random) bootstrap DGP in $\mathbb{M}$, determined in some suitable way from the same data as those used to compute $\hat{\tau}$. We denote by μ^* the random DGP of which $\hat{\mu}$ is a realization.

Let the asymptotic CDF of the asymptotic pivot τ be denoted by F. At nominal level α, an asymptotic test rejects if the asymptotic P value $1 - F(\hat{\tau}) < \alpha$. In order to avoid having to deal with different asymptotic distributions, it is convenient to replace the raw statistic τ by the asymptotic P value $1 - F(\tau)$, of which the asymptotic distribution is always $U(0, 1)$. For the remainder of this section and the next section, τ denotes such an asymptotic P value.

For the sample size of the observed data, the "rejection probability function," or RPF, provides a measure of the true rejection probability of the asymptotic test. This function, which gives the rejection probability under μ of a test at nominal level α, is defined as follows:

$$R(\alpha, \mu) \equiv \Pr_\mu(\tau < \alpha). \tag{3}$$

It is clear that $R(\cdot, \mu)$ is the CDF of τ under μ. The information contained in the function R is also provided by the "critical value function," or CVF, Q, defined implicitly by the equation

$$\Pr_\mu\big(\tau < Q(\alpha, \mu)\big) = \alpha. \tag{4}$$

$Q(\alpha, \mu)$ is just the α quantile of τ under μ. It follows from (3) and (4) that

$$R\big(Q(\alpha, \mu), \mu\big) = \alpha, \quad \text{and, conversely,} \quad Q\big(R(\alpha, \mu), \mu\big) = \alpha, \tag{5}$$

from which it is clear that, for given μ, R and Q are inverse functions.

The bootstrap test rejects at nominal level α if $\tau < Q(\alpha, \mu^*)$, that is, if τ is smaller than the the α quantile of τ under the bootstrap DGP. By acting on both sides with $R(\cdot, \mu^*)$, this condition can also be expressed as

$$R(\tau, \mu^*) < R\big(Q(\alpha, \mu^*), \mu^*\big) = \alpha.$$

This makes it clear that the bootstrap P value is just $R(\tau, \mu^*)$. It follows that, if R actually depends on μ^*, that is, if τ is not an exact pivot, the bootstrap test is not equivalent to the asymptotic test, because the former depends not only on τ, but also on the random μ^*.

In Davidson and MacKinnon (1999b), it is shown that bootstrap tests enjoy a further refinement, over and above that due to the use of an asymptotic pivot, if τ and μ^* are asymptotically independent. In addition, such asymptotic independence makes it possible to obtain an approximate expression for the size distortion of a bootstrap test. Suppose first that τ and μ^* are fully independent under the true DGP μ_0. Then the rejection probability under μ_0 of the bootstrap test at nominal level α is

$$\begin{aligned} \Pr_{\mu_0}\big(\tau < Q(\alpha, \mu^*)\big) &= E_{\mu_0}\Big(\Pr_{\mu_0}\big(\tau < Q(\alpha, \mu^*) \mid \mu^*\big)\Big) \\ &= E_{\mu_0}\Big(R\big(Q(\alpha, \mu^*), \mu_0\big)\Big). \end{aligned} \tag{6}$$

This is an exact result only if τ and μ^* are independent. It is however useful, because it applies approximately if τ and μ^* are only asymptotically independent, and because, as we see in the next section, an approximation to (6) can easily be estimated by simulation.

3. The Basic Idea

The fundamental idea of this paper is based on the fact that we can estimate an approximation to the quantity (6) quite inexpensively. In this section, we discuss how to do so in the context of Monte Carlo experiments. In the next section, we discuss how the same basic idea may be used to calculate modified bootstrap P values. Of course, (6) is exactly valid only if τ and μ^* are independent. However, as we will see in due course, asymptotic independence of τ and μ^*, in a sense we make precise later, is enough for (6) to provide a pretty accurate approximation. The asymptotic independence assumption is not very restrictive. A great many test statistics are asymptotically independent of all parameter estimates under the null hypothesis. This is generally true for extremum estimators where the estimates under the null lie in the interior of the parameter space, including all of the classical test statistics for models estimated by nonlinear least squares and maximum likelihood; see Davidson and MacKinnon (1999b). However, it generally will not be true for inefficient estimators.

Under the independence assumption, the size distortion, or, as we prefer to call it, the error in rejection probability (ERP), of the bootstrap test based on τ and μ^* can be written as the expectation (6) minus α:

$$E_{\mu_0}\Big(R\big(Q(\alpha,\mu^*),\mu_0\big)\Big) - \alpha. \tag{7}$$

Now, by (5), $R\big(Q(\alpha,\mu_0),\mu_0\big) = \alpha$. Thus, for given μ_0, expression (7), considered as a function of α, is a bias function. In the spirit of linear bias correction (MacKinnon and Smith, 1998), we approximate $R(Q(\alpha,\mu^*),\mu_0)$ as an affine function of its first, random, argument, and obtain

$$R\big(Q(\alpha,\mu^*),\mu_0\big) \approx \alpha + R_1 \cdot \big(Q(\alpha,\mu^*) - Q(\alpha,\mu_0)\big),$$

where R_1 is the derivative of R with respect to its first argument, evaluated at $Q(\alpha,\mu_0)$ and μ_0. Similarly,

$$R\big(Q(\alpha,\mu_0),\mu^*\big) \approx \alpha + R_1 \cdot \big(Q(\alpha,\mu_0) - Q(\alpha,\mu^*)\big),$$

and so, approximately, (7) is given by

$$\alpha - E_{\mu_0}\Big(R\big(Q(\alpha,\mu_0),\mu^*\big)\Big). \tag{8}$$

Now consider a random variable τ^*, of which a drawing under μ_0 is generated as follows. A sample is drawn from μ_0 and used to compute a drawing $\hat{\mu}$ of the bootstrap

DGP μ^*. Then a sample is drawn from $\hat{\mu}$ and used to compute a bootstrap statistic, which is then the drawing of τ^*. The CDF of τ^*, evaluated at argument α, can be seen to be just $E_{\mu_0}\big(R(\alpha, \mu^*)\big)$. Conditional on $\hat{\mu}$, the probability that $\tau^* < \alpha$ is given by the CDF of the bootstrap statistic under $\hat{\mu}$, that is, $R(\alpha, \hat{\mu})$. The unconditional expectation of this probability is just $E_{\mu_0}\big(R(\alpha, \mu^*)\big)$, as required.

The above construction allows us to evaluate the expectation in (8) by simulation. For each replication, the DGP μ_0 is used to draw realizations of the statistic τ and of the bootstrap DGP μ^*. Next, the realization $\hat{\mu}$ of μ^* is used to draw a realization of τ^*. The quantile $Q(\alpha, \mu_0)$ is then estimated as usual by $\hat{Q}_0(\alpha)$, the α quantile of the drawings of τ, and the expectation of $R\big(Q(\alpha, \mu_0), \mu^*\big)$ by the proportion of the drawings of τ^* that are less than $\hat{Q}_0(\alpha)$. This method was first suggested by Davidson and MacKinnon (1999b). If we perform M replications, the simulation estimate of the rejection probability (RP) of the bootstrap test, and the corresponding simulated ERP, are then given by

$$\widehat{\mathrm{RP}}_2 \equiv 2\alpha - \frac{1}{M}\sum_{m=1}^{M} I\big(\tau_m^* < \hat{Q}_0(\alpha)\big), \text{ and } \widehat{\mathrm{ERP}}_2 \equiv \alpha - \frac{1}{M}\sum_{m=1}^{M} I\big(\tau_m^* < \hat{Q}_0(\alpha)\big). \quad (9)$$

As the notation suggests, this estimator of the rejection probability is not our preferred one. It suffers from two potential disadvantages: It is bounded above by 2α, and it is not guaranteed to be positive.

A somewhat more accurate estimate, which does not suffer from these disadvantages, can be obtained by a slight modification of the above procedure, in which the roles of the distributions of τ and τ^* are interchanged. Drawings of τ and τ^* are made exactly as described above, but then (7) is estimated directly as the proportion of drawings of τ less than $\hat{Q}^*(\alpha)$, the α quantile of τ^*, minus α. This leads to the following simulation estimates of the RP and ERP of the bootstrap test:

$$\widehat{\mathrm{RP}}_1 \equiv \frac{1}{M}\sum_{m=1}^{M} I\big(\tau_m < \hat{Q}^*(\alpha)\big) \text{ and } \widehat{\mathrm{ERP}}_1 \equiv \frac{1}{M}\sum_{m=1}^{M} I\big(\tau_m < \hat{Q}^*(\alpha)\big) - \alpha. \quad (10)$$

Expressions (9) and (10) are equally easy to compute. Since very little extra effort is needed to compute (9) if (10) is already being computed, it probably makes sense to compute both, on the reasonable supposition that substantial differences between the two estimated ERPs may indicate that neither of them is terribly accurate.

In practice, it is not necessary to convert test statistics to approximate P value form in order to estimate ERPs by the above procedures. Drawings of the statistics are obtained in whatever form is most convenient, and then sorted in order from the most extreme values to the least extreme. For each value of α of interest, it is then straightforward to compute the proportion of realizations of the statistic more extreme than the realization of the bootstrap statistic in position α in the sorted list.

For a given number of replications, these procedures require only about twice as much computational effort as performing an experiment for the asymptotic test only. To

investigate the performance of an asymptotic test, we need to compute M test statistics τ_m. To investigate the performance of the bootstrap version of the same test, we just need to compute M additional test statistics τ_m^*. The procedure outlined above results in drawings of τ and τ^* that are asymptotically independent if τ and μ^* are asymptotically independent. Thus the variance of the estimated RP for a bootstrap test with a given actual RP will always be larger than the variance of the estimated RP for an asymptotic test with the same actual RP. For the asymptotic test, the only source of error would be the randomness of the τ_m. For the bootstrap test, there would also be the randomness of the τ_m^*, which causes $\hat{Q}^*(\alpha)$ to be random. This suggests that more replications will be needed to achieve a given level of accuracy.

There is another technique that, in certain cases, will allow one to obtain positively correlated drawings of τ and τ^*, and thus reduce the variance of the estimated RP of the bootstrap test. The procedure works as follows. Once $\hat{\mu}$ has been obtained for each replication, a new set of random numbers, independent of those used to obtain $\hat{\mu}$, is drawn. These are then used to compute the drawings of both τ and τ^*, the former using μ_0, the latter using $\hat{\mu}$. This should result in substantial positive correlation between the drawings of τ and τ^*, which will reduce the variance of the estimated RP. An additional advantage of this method is that τ and μ^* are genuinely, and not just asymptotically, independent.

Although limited simulation results suggest that this method may be attractive in certain cases, we do not develop it any further here, and we did not use it for most of our simulations. One problem with the method is that it necessarily involves up to fifty per cent more computational cost per replication than the simpler one proposed above. In addition, it may not work well for semiparametric bootstrap procedures, where the methods used to draw random samples from μ_0 are not the same as the ones used to draw them from $\hat{\mu}$. For example, in the case of a regression model, random sampling from μ_0 must be done by drawing from a known distribution, while random sampling from $\hat{\mu}$ is be done by resampling from rescaled residuals. Although this method can certainly be adapted to such cases, one might expect the potential gain in efficiency to be modest.

4. Modified Bootstrap *P* Values

The procedures we have discussed so far are useful only in the context of Monte Carlo experiments. But the same basic ideas also lead to procedures for computing bootstrap P values that are, at least potentially, more accurate than the conventional procedure based on the bootstrap P value (2). In this section, we propose two new ways to compute bootstrap P values, based on the estimates $\widehat{\text{RP}}_1$ and $\widehat{\text{RP}}_2$ of (10) and (9).

For each of B bootstrap replications, two different bootstrap statistics are generated. For bootstrap replication j, a bootstrap data set, which we denote by $\boldsymbol{y}_j^*$, is first drawn from the bootstrap DGP $\hat{\mu}$. In exactly the same way as the original data were used to obtain both the realized test statistic $\hat{\tau}$ and the realized bootstrap DGP $\hat{\mu}$, the simulated data $\boldsymbol{y}_j^*$ are used to compute two things: a bootstrap statistic, denoted

by τ_j^*, and a second-level bootstrap DGP, denoted by μ_j^{**}. Next, a further simulated data set, denoted $\boldsymbol{y}_j^{**}$, is drawn using this second-level bootstrap DGP, and a second-level bootstrap test statistic, τ_j^{**}, is computed. This procedure is completely analogous to that of the previous section, in which τ and τ^* are drawn on the basis of a DGP μ_0. Here μ_0 is replaced by $\hat{\mu}$, the ordinary, or single, bootstrap DGP, and the M drawings τ_m and τ_m^* are replaced by the B drawings τ_j^* and τ_j^{**}, respectively.

The ordinary estimated bootstrap P value is

$$\hat{p}^* \equiv \frac{1}{B}\sum_{j=1}^{B} I(\tau_j^* < \hat{\tau}),$$

the simulation estimate of (2) when the test statistic is in approximate P value form. We maintain this convention here, although, in practice, it is unnecessary for the computation of $\hat{p}^*$. Next, we calculate the $\hat{p}^*$ quantile of the τ_j^{**}, denoted by $\hat{Q}^*(\hat{p}^*)$ and defined implicitly by the equation

$$\frac{1}{B}\sum_{j=1}^{B} I\big(\tau_j^{**} < \hat{Q}^*(\hat{p}^*)\big) = \hat{p}^*. \tag{11}$$

Then the **fast double bootstrap** P value (first version), or $\mathbf{FDB}_1$ for short, is

$$\hat{p}_1^{**} = \frac{1}{B}\sum_{j=1}^{B} I\big(\tau_j^* < \hat{Q}^*(\hat{p}^*)\big). \tag{12}$$

Thus, instead of seeing how often the bootstrap test statistics are more extreme than the actual test statistic, we see how often they are more extreme than the $\hat{p}^*$ quantile of the τ_j^{**}.

Suppose, for concreteness, that the τ_j^{**} tend to be more extreme than the τ_j^*. This suggests that the τ_j^* tend to be more extreme than they would be if they were drawn from μ_0 instead of from μ^*. Therefore, the ordinary bootstrap P value will be too big. In this situation, $\hat{Q}(\hat{p}^*)$ will be more extreme than $\hat{\tau}$ itself, and $\hat{p}_1^{**}$ will consequently be smaller than $\hat{p}^*$. Thus it appears that using $\hat{p}_1^{**}$ instead of $\hat{p}^*$ will be a step in the right direction.

The second version of the fast double bootstrap P value, or $\mathbf{FDB}_2$ for short, is calculated as

$$\hat{p}_2^{**} = 2\hat{p}^* - \frac{1}{B}\sum_{j=1}^{B} I(\tau_j^{**} < \hat{\tau}); \tag{13}$$

compare (9). Expression (13) has a slight computational advantage over (12), in that it is not necessary to compute any quantiles. However, it has the disadvantages that $\hat{p}_2^{**}$ could possibly be negative and that $\hat{p}_2^{**}$ cannot be more than twice as great as $\hat{p}^*$. But since it is almost costless to compute FDB_2 if FDB_1 is already being computed, it may be useful to do so as a check on the accuracy of the latter.

5. The Double Bootstrap

The fast double bootstrap procedures proposed in the previous section are so called because of their relationship to the genuine double bootstrap procedure originally suggested by Beran (1988), which is very much more expensive computationally. In the statistical literature, the double bootstrap is normally used for computing confidence intervals. However, a version of it for computing P values is conceptually quite simple and works as follows:

- Obtain B_1 first-level bootstrap samples from the DGP $\hat{\mu}$ in the usual way, and use them to compute bootstrap statistics τ_j^*, $j = 1, \ldots, B_1$, and the ordinary bootstrap P value $\hat{p}^*$.
- For each first-level bootstrap sample j, compute the second-level bootstrap DGP μ_j^{**}, and use it to compute B_2 second-level bootstrap samples, each of which is used to compute a test statistic τ_{jl}^{**}, $l = 1, \ldots, B_2$. These are just like the τ_j^{**} used in the FDB procedures, except that there are B_2 of them for each bootstrap sample.
- For first-level bootstrap sample j, compute the second-level bootstrap P value

$$\hat{p}_j^{**} = \frac{1}{B_2} \sum_{l=1}^{B_2} I(\tau_{jl}^{**} < \tau_j^*). \tag{14}$$

- Finally, compute the double-bootstrap P value

$$\hat{p}^{**} = \frac{1}{B_1} \sum_{j=1}^{B_1} I(\hat{p}_j^{**} \leq \hat{p}^*). \tag{15}$$

  Thus the double-bootstrap P value is equal to the proportion of the second-level bootstrap P values that are more extreme than the first-level bootstrap P value. The inequality in (15) is not strict, because, depending on the values of B_1 and B_2, there may be a substantial number of cases in which $\hat{p}_j^{**} = \hat{p}^*$.

The advantage of this procedure, relative to the new ones proposed in the previous section, is that it does not require any sort of independence between the bootstrap DGP and the test statistic. But this comes at an enormous computational cost. For each of B_1 bootstrap samples, we need to compute $B_2 + 1$ test statistics. Thus the total number of test statistics that must be computed is $1 + B_1 + B_1 B_2$. Even if B_2 is somewhat smaller than B_1, as is often recommended, this will be vastly more expensive than computing $1 + 2B$ test statistics, for reasonable values of $B \approx B_1$ and $B_2 \leq B_1$. For example, if $B = B_1 = 999$ and $B_2 = 199$, the FDB procedures will require the computation of 1999 test statistics, while the genuine double bootstrap will require the computation of no less than 199,801 of them.

6. Simulation Evidence

In this section, we present results from a number of simulation experiments. Our objective here is to see whether the procedures proposed in this paper can work well enough to be useful in practice. First of all, we want to see whether $\widehat{\text{RP}}_1$ and $\widehat{\text{RP}}_2$ provide good approximations to the actual rejection probabilities for ordinary bootstrap tests based on $\hat{p}^*$. Secondly, we want to see whether the FDB_1 and FDB_2 procedures can yield bootstrap tests with smaller errors in rejection probability than tests based on $\hat{p}^*$.

In many cases, ordinary bootstrap tests appear to work so well that there is no point using anything more complicated. Even when they do not, they at least tend to work so well that an extraordinarily large number of replications would be needed in order to show conclusively that anything else works better. This made it somewhat difficult to find a good example to focus on. We have chosen to study the OPG version of the LM test for omitted variables in the probit model. This test has noticeably worse finite-sample properties than other tests of the same hypothesis, such as the LR test and the efficient score version of the LM test; see Davidson and MacKinnon (1984). Therefore, it is rarely used in practice. However, its poor finite-sample performance, and the dependence of that performance on parameter values, makes it a good example for the study of alternative bootstrap procedures.

The probit model we study can be written as

$$E(y_t \mid \boldsymbol{X}_t) = \Phi(\boldsymbol{X}_{1t}\boldsymbol{\beta}_1 + \boldsymbol{X}_{2t}\boldsymbol{\beta}_2), \tag{16}$$

where y_t is a binary dependent variable that can equal either 0 or 1, $\Phi(\cdot)$ is the cumulative standard normal distribution function, $\boldsymbol{X}_t = [\boldsymbol{X}_{1t} \;\; \boldsymbol{X}_{2t}]$ is a $1 \times k$ vector of exogenous variables, with $k = k_1 + k_2$, $\boldsymbol{\beta}_1$ is a k_1–vector, and $\boldsymbol{\beta}_2$ is a k_2–vector. The null hypothesis is that $\boldsymbol{\beta}_2 = \mathbf{0}$. The OPG test statistic is the explained sum of squares from a regression of an n–vector of 1s on the derivatives of the contributions to the loglikelihood with respect to each of the parameters, where those derivatives are evaluated at the ML estimates under the null hypothesis. See Davidson and MacKinnon (1984) for more details.

Our experimental results pertain to four different cases. In Case 1, $k_1 = 2$, $k_2 = 6$, $\beta_1 = 0$, and $\beta_2 = 1$. The number of restrictions, k_2, is relatively large because the finite-sample performance of the test becomes worse as k_2 increases, and preliminary experiments revealed that the finite-sample performance of the bootstrap test does likewise. In Case 2, $k_1 = 2$, $k_2 = 6$, $\beta_1 = 1$, and $\beta_2 = 2$. In Case 3, $k_1 = 4$, $k_2 = 6$, $\beta_1 = 0$, and $\beta_2 = \beta_3 = \beta_4 = 1$. Finally, in Case 4, $k_1 = 4$, $k_2 = 10$, $\beta_1 = 0$, and $\beta_2 = \beta_3 = \beta_4 = 1$. In all cases, the first of the exogenous variables was a constant, and the others were independent drawings from the $N(0, 1)$ distribution.

For each case, we performed 171 experiments, with 100,000 replications, for all sample sizes between 30 and 200. The exogenous variables, other than the constant, were redrawn for each replication, so as to avoid undue dependence on the design matrix. In these experiments, rejection frequencies of the asymptotic test and the approximate

rejection probabilities of the bootstrap test were estimated. In addition, we performed 18 much more expensive experiments, for $n = 30, 40, 50, \ldots, 200$, also with 100,000 replications, in which we estimated the actual performance of the bootstrap test and the two FDB tests using $B = 399$. The results of these experiments are presented graphically in Figures 1 through 5.

Figure 1 shows the rejection frequencies for the asymptotic test. It always overrejects severely, with Case 4 being the worst and Case 1 the least bad. For several of the cases, the overrejection initially becomes worse as the sample size increases. Eventually, as expected, the overrejection gradually diminishes as the sample size increases, although it remains quite substantial at $n = 200$.

Figures 2 through 5 pertain to Cases 1 through 4. Each of these figures shows the rejection frequencies for the three bootstrap tests, along with the approximate rejection probabilities given by $\widehat{\text{RP}}_1$ and $\widehat{\text{RP}}_2$. A number of empirical regularities are apparent in these figures:

- Compared with the asymptotic test, the bootstrap test always performs remarkably well. However, it overrejects for very small sample sizes (except, perhaps, in Case 1) and then underrejects for a range of somewhat larger sample sizes.
- Both $\widehat{\text{RP}}_1$ and $\widehat{\text{RP}}_2$ provide good approximations to the actual rejection frequencies of the bootstrap test. When they differ appreciably, as they do most noticeably in Figures 4 and 5, $\widehat{\text{RP}}_1$ always provides a somewhat better approximation than does $\widehat{\text{RP}}_2$. However, both approximations seem to make systematic errors for certain ranges of sample sizes.
- The rejection frequencies for FDB_2 are always greater than for FDB_1, and they are always too large.
- FDB_1 usually, but not always, outperforms the ordinary bootstrap test. This is most noticeable for the sample sizes where the ordinary bootstrap test systematically underrejects. For very small sample sizes, where the ordinary bootstrap test sometimes overrejects quite severely, FDB_1 tends to underrreject.

In Figures 6 and 7, the simulated errors in rejection probability, $\widehat{\text{ERP}}_1$ and $\widehat{\text{ERP}}_2$, are plotted as functions of α. These figures are essentially P value discrepancy plots in the sense of Davidson and MacKinnon (1998). The plots are for sample sizes $n = 50$ and $n = 200$ for Cases 1 (Figure 6) and 4 (Figure 7). Although the ERP for the nominal .05 level seems reasonably typical of the rest of the distribution in Case 1, things are quite different for Case 4 with $n = 50$. There, the overrejection rapidly changes to underrejection for larger nominal levels.

In Section 3, we discussed the relationship between the variance of $\widehat{\text{RP}}_1$ and the variance of the estimated rejection probability for the asymptotic test. In order to investigate this relationship, we regressed both estimates of bootstrap rejection probability errors on a number of powers of $n^{-1/2}$ (with no constant, since asymptotically there is no error) for each of the four cases. The standard errors of the preferred regressions for Cases 1 through 4 were 0.000925, 0.000955, 0.000875, and 0.000921. These are substantially larger than the theoretical value of 0.000689 that applies to an estimate

of $\alpha = .05$ based on 100,000 replications. For the $\widehat{\text{RP}}_2$ estimates, the standard errors were 0.001023, 0.000909, 0.000950, and 0.000996, respectively. Thus it appears that, as expected, the sampling variability of the approximate bootstrap estimators is moderately larger than the sampling variability of ordinary rejection frequencies.

The tendency of the ordinary bootstrap test, and of FDB_2, to overreject severely in very small samples in Cases 3, 4, and 5 has a simple explanation. In these cases, ML estimation of the null model not infrequently achieves a perfect fit. When this happens, the test statistic is equal to zero. As is well known, probit models tend to fit too well in small samples. Therefore, the slope coefficients used to generate the bootstrap samples tend to be larger than the ones used to generate the original samples; see MacKinnon and Smith (1998). This means that perfect fits are achieved more often for the bootstrap samples than they are for the original samples. In consequence, there are fewer large values of the τ_j^* than there are of the τ_j, and the bootstrap P values are therefore biased downwards. The problem tends to go away rapidly as n increases.

It would have been computationally infeasible to compute genuine double bootstrap P values in most of our experiments. However, we did perform a few experiments in which these were calculated. Even though these had only 10,000 replications, and we set $B_1 = B_2 = 199$, they were far more expensive than any of the other experiments. Results are presented in Table 1. This table provides no evidence to suggest that double bootstrap P values are any more accurate than FDB_1 P values. In Case 2 with 50 observations, where perfect fits occur with some frequency, the double bootstrap actually performs substantially less well than does FDB_1. In the other five cases, bearing in mind that the standard errors of the estimated rejection frequencies are roughly 0.0022, there is little to choose between them.

7. Relation to Edgeworth Expansions

Much of the asymptotic theory of the bootstrap is based on Edgeworth expansions of statistics that are asymptotically distributed as $N(0, 1)$. The classic reference for this approach is Hall (1992), and the methods and assumptions used in this section are similar to those in Hall (1988). This theory most often serves to determine the order of the ERP of a bootstrap test as a negative power of the sample size n. At a certain algebraic cost, approximate expressions for the ERP can also be obtained. However, these approximate expressions are rarely used either to estimate the ERPs of specific bootstrap tests or to improve bootstrap P values in the manner of the FDB procedures. This is most likely due to the poor quality of the Edgeworth approximation, especially in the tails of distributions.

Nonetheless, it is interesting to see to what extent Edgeworth expansions give results in accord with the simulation results of the previous section. It turns out that the difference between expressions (7) and (8) for the bootstrap ERP is very generally of order no greater than $O(n^{-2})$. When the bootstrapped statistic τ and the bootstrap DGP μ^* are independent, the same is true of the difference between the true rejection probability, that is, the first term in (7), and RP_1, by which we mean the theoretical

expression to which $\widehat{\mathrm{RP}}_1$ converges when M, the number of replications, tends to infinity. It is thus not surprising that, in our simulation experiments, $\widehat{\mathrm{RP}}_1$ and $\widehat{\mathrm{RP}}_2$ are very similar to each other and to the true bootstrap rejection probability. Just as the bootstrap itself obviates the need to perform analytic Edgeworth expansions, and provides a much more reliable approximation than an Edgeworth expansion truncated after a few terms, so do $\widehat{\mathrm{ERP}}_1$ and $\widehat{\mathrm{ERP}}_2$ obviate the need for an Edgeworth expansion of a bootstrap ERP. They provide what, on the basis of the simulations of this paper, seems to be a very reliable approximation to the bootstrap ERP.

The Edgeworth expansion of the CDF $F(x)$ of an asymptotically $N(0,1)$ statistic τ can be written as

$$F(x) = \Phi(x) - n^{-1/2}\phi(x)\sum_{i=1}^{\infty}\lambda_i He_{i-1}(x). \tag{17}$$

Here $\Phi(\cdot)$ and $\phi(\cdot)$ are, respectively, the CDF and the density of the $N(0,1)$ distribution, $He_i(\cdot)$ is the Hermite polynomial of degree i (see for instance Abramowitz and Stegun (1965), Chapter 22 for details of these polynomials), and the λ_i are coefficients that are at most of order unity. Thus (17) implicitly supposes that the rate of convergence of the statistic to the asymptotic $N(0,1)$ distribution is at least as fast as $n^{-1/2}$. The expansion as written in (17) is more properly referred to as the Gram-Charlier series, but, unless truncated, the Edgeworth and Gram-Charlier series are equivalent. As we truncate everything of order lower than some fixed negative power of n, we obtain true Edgeworth series. The relation between the Hermite polynomials and the derivatives of the density $\phi(\cdot)$ is

$$\phi^{(i)}(x) = (-1)^i He_i(x)\phi(x), \tag{18}$$

and they can be defined by the following recursion, easily derived from (18),

$$He_0(x) = 1; \qquad He_{i+1}(x) = x\,He_i(x) - He_i'(x).$$

The λ_i can be related to the (uncentered) moments μ_i of the statistic by means of the equation

$$\lambda_j = \frac{n^{1/2}}{j!}E\big(He_j(\tau)\big), \tag{19}$$

so that, for the first few values of j, $\lambda_1 = n^{1/2}\mu_1$, $\lambda_2 = n^{1/2}(\mu_2 - 1)/2$, $\lambda_3 = n^{1/2}(\mu_3 - 3\mu_1)/6$, $\lambda_4 = n^{1/2}(\mu_4 - 6\mu_2 + 3)/24$, and so on. The formula (19) follows directly from differentiating (17) with respect to x to find the density, and then using the orthogonality of the Hermite polynomials with the standard normal density as weighting function. It is easily checked that the λ_i vanish if they are defined using the moments of the standard normal distribution. In practice, the sum over i in (17) is truncated after a finite, usually small, number of terms, because, as i increases, the λ_i are of the order of progressively higher negative powers of n as $n \to \infty$. In many

cases, only λ_1 and λ_3 are $O(1)$. The λ_i can quite generally be expressed in terms of the cumulants κ_j of τ, and, for $j \geq 4$, $\kappa_j = O(n^{(j-2)/2})$.

The formula (17) can be thought of as an expansion of the RPF (3) for an asymptotically $N(0,1)$ rather than an asymptotically $U(0,1)$ statistic. In order to facilitate comparison of the results here with those of Sections 2 and 3, we consider tests that reject in the *left*-hand tail of the $N(0,1)$ distribution. In this way, we can retain the signs of the inequalities used earlier. There is, of course, no conceptual difficulty in considering tests that reject to the right or in both tails of the distribution. With this convention, if a DGP μ yields the values λ_i for the statistic τ, the quantity $R(\alpha, \mu)$ of (3) is given by (17) evaluated at $x = z_\alpha$, where z_α is the α quantile of the $N(0,1)$ distribution. In general, we write $\boldsymbol{\lambda}(\mu)$ to denote the, in principle infinite, sequence of λ_i, $i = 1, \ldots, \infty$ for statistic τ generated by the DGP μ; instead of $R(\alpha, \mu)$ we write $R(\alpha, \boldsymbol{\lambda}(\mu))$; and we have explicitly from (17) that, for a given sequence $\boldsymbol{\lambda}$,

$$R(\alpha, \boldsymbol{\lambda}) = \alpha - n^{-1/2}\phi(z_\alpha) \sum_{i=1}^{\infty} \lambda_i He_{i-1}(z_\alpha). \tag{20}$$

For the distribution characterized by the expansion (17), let the α quantile be denoted, by a slight extension of the notation of Sections 2 and 3, by $Q(\alpha, \boldsymbol{\lambda})$. This quantile can, if needed, be expanded in a Cornish-Fisher expansion that is the inverse of the Edgeworth expansion (17). The relations (5) must be modified with our new definitions of R and Q, as follows:

$$R\big(\Phi(Q(\alpha, \boldsymbol{\lambda})), \boldsymbol{\lambda}\big) = \alpha \quad \text{and} \quad \Phi\big(Q\big(R(\alpha, \boldsymbol{\lambda}), \boldsymbol{\lambda}\big)\big) = \alpha. \tag{21}$$

Suppose next that the statistic τ is bootstrapped. If the true DGP is as usual denoted by μ_0, then we write $\boldsymbol{\lambda}_0 \equiv \boldsymbol{\lambda}(\mu_0)$, with elements λ_i^0. In many circumstances, in particular if the statistic is a smooth function of sample moments, or if a parametric bootstrap based on root-n consistent parameter estimates is used, the bootstrap DGP will generate bootstrap statistics whose CDF can be expanded as in (17), with an infinite sequence $\boldsymbol{\lambda}^* \equiv \boldsymbol{\lambda}(\mu^*)$ of coefficients λ_i^* such that $\lambda_i^* = \lambda_i^0 + O(n^{-1/2})$. The λ_i^* are of course random, and in the circumstances we are discussing, they are such that $E(\lambda_i^* - \lambda_i^0) = O(n^{-1})$; see Hall (1988). In some cases, and for larger values of i, $\lambda_i^* - \lambda_i^0$ is of lower order than $O(n^{-1/2})$, and it will often vanish for some i. On the other hand, some bootstrap procedures may lead to differences of order 1 between λ_i^* and λ_i^0. The following theory does not apply in such cases.

When it does apply, we may write $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}_0 + n^{-1/2}\boldsymbol{l}$, where the elements of the sequence $\boldsymbol{l}$ are of order at most unity. Let us now fix the true DGP μ_0, and, for this DGP, define another sequence $\boldsymbol{\nu}$ by the element-wise rule $\nu_i = n^{1/2}E_{\mu_0}(l_i)$, so that the ν_i are of order no greater than 1. We define q_α^* by the relation $Q(\alpha, \boldsymbol{\lambda}^*) \equiv Q(\alpha, \boldsymbol{\lambda}_0) + n^{-1}q_\alpha^*$, and, similarly, we define q_α by $Q(\alpha, \boldsymbol{\lambda}_0 + n^{-1}\boldsymbol{\nu}) = Q(\alpha, \boldsymbol{\lambda}_0) + n^{-1}q_\alpha$. Clearly q_α^* is random but q_α is not. Finally, we make the definition $\gamma_\alpha = q_\alpha^* - q_\alpha$. The following lemma establishes the orders of the quantities we have just defined.

Lemma 1

Under the conditions of this section on the orders of $\boldsymbol{\lambda}^* - \boldsymbol{\lambda}_0$ and its expectation, $q^*_\alpha = O(1)$, $q_\alpha = O(n^{-1/2})$, $\gamma_\alpha = O(1)$, and $E(\gamma_\alpha) = O(n^{-1})$.

All proofs are found in the Appendix.

It will often be necessary to consider the difference between values of the function R for the same first argument but different second arguments. In general, we have:

$$R(\alpha, \boldsymbol{\lambda}) - R(\alpha, \boldsymbol{\lambda}') = n^{-1/2}\phi(z_\alpha)\sum_i (\lambda'_i - \lambda_i) He_{i-1}(z_\alpha). \tag{22}$$

It is thus convenient to make the following definition:

$$r(x, \boldsymbol{\lambda}) = \phi(x)\sum_i \lambda_i He_{i-1}(x), \tag{23}$$

in which we can exploit the linearity of R with respect to its second argument.

Our first result serves to express the ERP of the bootstrap test in terms of the quantities we have defined.

Theorem 1

Consider a bootstrap test based on an asymptotically standard normal statistic τ of which the distribution has an Edgeworth expansion that we can characterize by $R(\alpha, \boldsymbol{\lambda}(\mu))$ when τ is generated by the DGP μ. For a DGP μ_0, let $\boldsymbol{\lambda}_0 \equiv \boldsymbol{\lambda}(\mu_0)$, and let the bootstrap DGP μ^* be characterized by a sequence $\boldsymbol{\lambda}^* \equiv \boldsymbol{\lambda}(\mu^*)$ that satisfies the conditions of this section on the orders of $\boldsymbol{\lambda}^* - \boldsymbol{\lambda}_0$ and its expectation. Then the rejection probability error of the bootstrap test at nominal level α can be written as

$$n^{-1} r\big(Q_\alpha + n^{-1}q_\alpha, n^{-1/2}\boldsymbol{\nu} - \boldsymbol{\eta}_\alpha\big), \tag{24}$$

where $Q_\alpha \equiv Q(\alpha, \boldsymbol{\lambda}_0)$, and the sequence $\boldsymbol{\eta}_\alpha$, which is at most of order unity as $n \to \infty$, is such that the distribution of $\tau - n^{-1}\gamma_\alpha$ is characterized by the sequence $\boldsymbol{\lambda}_0 + n^{-1/2}\boldsymbol{\eta}_\alpha$.

Remarks: Expression (24) makes it clear that, under the conditions of the theorem, the ERP of the bootstrap test is of order n^{-1} at most, in accord with the results of Beran (1988).

Since r is linear with respect to its second argument, the ERP (24) can also be written as $-n^{-1}r(Q_\alpha + n^{-1}q_\alpha, \boldsymbol{\eta}_\alpha) + n^{-3/2}r(Q_\alpha + n^{-1}q_\alpha, \boldsymbol{\nu})$. Note that this expression is complete: It is not just an approximate expression correct through order $n^{-3/2}$.

If $\boldsymbol{\eta}_\alpha = O(n^{-1/2})$, then the entire bootstrap ERP is of order $n^{-3/2}$ at most. This will turn out to be the case whenever τ and μ^* are asymptotically independent.

For some purposes, the dependence of the sequence $\boldsymbol{\eta}_\alpha$ on α is inconvenient. Fortunately, there exists another representation of the ERP in which $\boldsymbol{\eta}_\alpha$ is replaced by a sequence that does not depend on α.

Corollary 1

There exists a sequence $\boldsymbol{\theta}$, of order at most unity as $n \to \infty$, independent of α, such that the ERP (24) can be written as

$$n^{-1} r(z_\alpha, n^{-1/2}\boldsymbol{\nu} - \boldsymbol{\theta}). \tag{25}$$

Remark: The bootstrap P value is expressed as $R(\Phi(\tau), \lambda^*)$ in the notation of this section. By Theorem 1, the probability that this P value is less than α is α plus the ERP (24), or, equivalently, (25). Thus the Edgeworth expansion of the distribution of the bootstrap P value is

$$\begin{aligned} \Pr(p^* < \alpha) &= \Phi(z_\alpha) + n^{-1} \sum_i (n^{-1/2}\nu_i - \theta_i) He_{i-1}(z_\alpha) \\ &= R(\alpha, n^{-1}\boldsymbol{\nu} - n^{-1/2}\boldsymbol{\theta}). \end{aligned} \tag{26}$$

In the next theorem, we obtain expressions, similar to (24), for the approximations to the bootstrap ERP.

Theorem 2

Under the conditions of Theorem 1, the approximate rejection probability errors ERP_2 and ERP_1, estimated by (9) and (10) respectively, are equal to

$$n^{-3/2} r(Q_\alpha, \boldsymbol{\nu}) \quad \text{and} \quad n^{-3/2} r(Q_\alpha + n^{-1} q_\alpha, \boldsymbol{\nu}). \tag{27}$$

Remarks: The results of Theorem 2 do not depend on whether τ and μ^* are asymptotically independent. Expressions (27) are just algebraic consequences of the definitions of the approximate ERPs that follow from the statistical properties of τ and μ^* separately, and not from their joint distribution.

Whether either of the expressions (27) is a reasonable approximation to the true bootstrap ERP (24) depends on the impact of the sequence $\boldsymbol{\eta}_\alpha$ on (24). In any event, it is clear why, in favorable cases, ERP_1 is a better approximation than ERP_2, since the former, like (24), evaluates r at $Q_\alpha + n^{-1} q_\alpha$ rather than at Q_α.

In the next theorem, we obtain results on the order of magnitude of the elements of $\boldsymbol{\eta}_\alpha$.

Theorem 3

Under the conditions of Theorem 1, the first component of $\boldsymbol{\eta}_\alpha$, η_1^α, is of order at most n^{-1}. When τ and μ^* are independent, all components of $\boldsymbol{\eta}_\alpha$ except the first two are $O(n^{-3/2})$, these two components being $O(n^{-1})$. When τ and μ^* are asymptotically independent, in the precise sense that, for all i, the covariance of l_i and τ is of order at most $n^{-1/2}$, all components other than the first may be of order $n^{-1/2}$. When τ and μ^* are not asymptotically independent, however, all components of $\boldsymbol{\eta}_\alpha$ except the first may be of order 1.

Remarks: Without the asymptotic independence of τ and μ^*, the bootstrap ERP (24) is in general of order n^{-1}, the highest order allowed by the results of Beran (1988).

If τ and μ^* are independent, the contribution to the bootstrap ERP made by $\boldsymbol{\eta}_\alpha$ is of order at most n^{-2}. Thus the ERP (24) is of order at most $n^{-3/2}$, and and the approximate expressions in (27) coincide with (24) and each other at least through order $n^{-3/2}$.

If τ and μ^* are only asymptotically independent, which is of course more frequent in practice than full independence, then the contribution to the ERP from $\boldsymbol{\eta}_\alpha$ may be of order $n^{-3/2}$. In that case, (24) is, as in the case of independence, of order at most $n^{-3/2}$, but the approximations in (27) may differ from (24) already at this order. It is beyond the scope of this paper to investigate this issue in greater detail, but in many cases of asymptotic independence it can be seen that the $O(n^{-3/2})$ contribution from $\boldsymbol{\eta}_\alpha$ is in fact of lower order. Exceptions in the framework of regression models include cases in which both regressors and error terms are skewed, or in which there is no constant term in the regression; see Hall (1992) for more details.

Finally, in the next theorem, we present results on the orders of the ERPs of tests based on the FDB P values and the genuine double bootstrap P value.

Theorem 4

Under the conditions of Theorem 1, the ERP of either one of the FDB methods for nominal level α is

$$-n^{-1}r\big(Q_\alpha, \boldsymbol{\eta}_\alpha + O(n^{-1})\big) + O(n^{-5/2}). \tag{28}$$

For the full double bootstrap method, there exists $\boldsymbol{\zeta}_\alpha$ such that the ERP is

$$n^{-3/2}r\big(Q(\alpha, n^{-1}\boldsymbol{\nu} - n^{-1/2}\boldsymbol{\theta}), \boldsymbol{\zeta}_\alpha\big),$$

where $\boldsymbol{\theta}$ is the sequence defined in Corollary 1, and $\boldsymbol{\zeta}_\alpha$ is of order 1 in general, but of order $n^{-1/2}$ if τ and μ^* are asymptotically independent.

Remarks: Without the asymptotic independence of τ and μ^*, the ERP of the FDB methods can be of the same order as that of the ordinary bootstrap test. Whether there is any advantage to using an FDB method is therefore not revealed by this sort of analysis.

With asymptotic independence, the ERP of the FDB methods is of order at most $n^{-3/2}$, and with full independence, at most n^{-2}. Thus, with only asymptotic independence, the ERP may be of the same order as that of the ordinary bootstrap test. However, the FDB methods knock out the term in (24) proportional to $\boldsymbol{\nu}$.

Without asymptotic independence, the double bootstrap ERP is in general of the same order as that of FDB with asymptotic independence. It too benefits from an added refinement with asymptotic independence.

The analysis of this section applies only to one-tailed tests based on statistics that are asymptotically distributed as $N(0,1)$. In particular, we have not studied, as does Hall (1988), refinements that may arise when two-tailed tests are used. Nevertheless, the simulation results of Section 6 make it clear that similar considerations apply to statistics in asymptotically χ^2 form. It cannot be too much emphasized that analyses based on Edgeworth expansions are at best indicative of the actual behavior of test statistics and bootstrap P values in finite samples. Terms of order $n^{-3/2}$ may be completely dominated by terms of order n^{-2} with certain configurations of the coefficients in the expansion. Further, since it is rare that Hermite polynomials of degree greater than 4 appear in truncated expansions, it is clear that the first four polynomials do not provide anything like enough functional flexibility to capture the actual distributions of many statistics. With these caveats, however, we have obtained in this section some intuition as to why the approximate bootstrap ERPs and the FDB procedures, in appropriate circumstances, work so well in practice.

8. Conclusion

In this paper, we have proposed two different, but closely related, techniques to solve two rather different problems. The first problem is the high cost of Monte Carlo experiments that involve bootstrap tests. Our procedures make it possible to study the finite-sample performance of bootstrap tests for only about three or four times the computational cost of studying asymptotic tests. In contrast, with the standard approach, the cost is generally hundreds of times as great.

The second problem is the errors in rejection probability that sometimes occur for bootstrap tests. Our FDB_1 procedure seems to reduce these errors quite substantially in some cases. Of course, bootstrap tests often work so well that there is no point trying to reduce the ERP any further. In many cases, the principal reason for calculating the FDB_1 P value is simply to verify that it is similar to the ordinary bootstrap P value. When that is the case, the investigator can feel very confident that both of them are providing good estimates of the true P value.

In both cases, the method based on the estimate $\widehat{\text{RP}}_1$ of (10) seems to work better than the other method, based on $\widehat{\text{RP}}_2$ of (9). The second method is therefore recommended primarily as a way of checking that the first one is giving reasonable results. If the two methods yield sharply different results, neither should be trusted.

References

Abramowitz, M., and I. A. Stegun (1965). *Handbook of Mathematical Functions*, New York, Dover.

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

Davidson, R., and J. G. MacKinnon (1984). "Convenient specification tests for logit and probit models," *Journal of Econometrics*, **25**, 241–262.

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 (1999a). "Bootstrap testing in nonlinear models," *International Economic Review*, **40**, 487–508.

Davidson, R., and J. G. MacKinnon (1999b). "The size distortion of bootstrap tests," *Econometric Theory*, **15**, 361–376.

Davidson, R., and J. G. MacKinnon (2000). "Bootstrap tests: How many bootstraps?" *Econometric Reviews*, **19**, 55–68

Godfrey, L. G. (1998). "Tests of non-nested regression models: Some results on small sample behaviour and the bootstrap," *Journal of Econometrics*, **84**, 59–74.

Hall, P. (1988). "On symmetric bootstrap confidence intervals," *Journal of the Royal Statistical Society*, Series B, **50**, 35–45.

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

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

MacKinnon, J. G. and A. A. Smith, Jr. (1998). "Approximate bias correction in econometrics," *Journal of Econometrics*, **85**, 205–230.

Nankervis, J. C. and N. E. Savin (1996). "The level and power of the bootstrap t test in the AR(1) model with trend," *Journal of Business and Economic Statistics*, **14**, 161–168.

Appendix

Proof of Lemma 1: The Cornish-Fisher expansion of $Q(\alpha, \boldsymbol{\lambda})$ can be written schematically as follows:

$$\begin{aligned} Q(\alpha, \boldsymbol{\lambda}) = z_\alpha &+ n^{-1/2} \sum_i \lambda_i He_{i-1}(z_\alpha) + n^{-1} \sum_i \sum_j \lambda_i \lambda_j h_{ij}(z_\alpha) \\ &+ n^{-3/2} \sum_i \sum_j \sum_k \lambda_i \lambda_j \lambda_k h_{ijk}(z_\alpha) + \ldots, \end{aligned} \tag{29}$$

where z_α is the α quantile of the $N(0,1)$ distribution, and where the h_{ij}, h_{ijk}, and so on, are polynomials, of which we will not need the explicit form, that can be expressed in terms of the Hermite polynomials. It is, however, easy to see that the term of order $n^{-1/2}$ in (29) is the one given there.

By definition, $q^*_\alpha = n\big(Q(\alpha, \boldsymbol{\lambda}_0 + n^{-1/2}\boldsymbol{l}) - Q(\alpha, \boldsymbol{\lambda}_0)\big)$. Using (29), we obtain

$$q^*_\alpha = \sum_i l_i He_{i-1}(z_\alpha) + n^{-1/2} \sum_i \sum_j (\lambda^0_i l_j + l_i \lambda^0_j) h_{ij}(z_\alpha) + O(n^{-1}).$$

The right-hand side of this equation is plainly of order 1. Similarly

$$\begin{aligned} q_\alpha &= n\big(Q(\alpha, \boldsymbol{\lambda}_0 + n^{-1}\boldsymbol{\nu}) - Q(\alpha, \boldsymbol{\lambda}_0)\big) \\ &= n^{-1/2} \sum_i \nu_i He_{i-1}(z_\alpha) + O(n^{-1}), \end{aligned} \tag{30}$$

which is of order $n^{-1/2}$. Thus we see that, with accuracy through order $n^{-1/2}$,

$$\gamma_\alpha = \sum_i (l_i - n^{-1/2}\nu_i) He_{i-1}(z_\alpha) + n^{-1/2} \sum_i \sum_j (l_i \lambda^0_j + \lambda^0_i l_j) h_{ij}(z_\alpha). \tag{31}$$

Thus $\gamma_\alpha = O(1)$, and the expectation of γ_α is zero through order $n^{-1/2}$ at least, and so is in general of order no greater than n^{-1}. ∎

Proof of Theorem 1: The event that leads to rejection by the bootstrap test at nominal level α is $\tau < Q(\alpha, \boldsymbol{\lambda}^*)$. Since $Q(\alpha, \boldsymbol{\lambda}^*) = Q_\alpha + n^{-1} q^*_\alpha = Q_\alpha + n^{-1}(q_\alpha + \gamma_\alpha)$, this is equivalent to the event $\tau - n^{-1}\gamma_\alpha < Q_\alpha + n^{-1} q_\alpha$, where all random quantities are on the left-hand side.

The sequence $\boldsymbol{\eta}_\alpha$ was defined to be such that the Edgeworth expansion of the distribution of $\tau - n^{-1}\gamma_\alpha$ is characterized by the sequence $\boldsymbol{\lambda}_0 + n^{-1/2}\boldsymbol{\eta}_\alpha$. By Lemma 1, $\gamma_\alpha = O(1)$, and so $\boldsymbol{\eta}_\alpha$ is $O(1)$ at most.

The rejection probability of the bootstrap test at nominal level α is the probability that the random variable $\tau - n^{-1}\gamma_\alpha$ is less than the critical value $Q_\alpha + n^{-1} q_\alpha$. This probability is

$$R\big(\Phi(Q_\alpha + n^{-1} q_\alpha), \boldsymbol{\lambda}_0 + n^{-1/2}\boldsymbol{\eta}_\alpha\big). \tag{32}$$

By the definition of q_α and the relation (21), we have that

$$R\big(\Phi(Q_\alpha + n^{-1}q_\alpha), \boldsymbol{\lambda}_0 + n^{-1}\boldsymbol{\nu}\big) = \alpha. \tag{33}$$

It follows that the ERP of the bootstrap test at nominal level α is just the difference between (32) and (33). Using (22) and (23), we see that the ERP of the bootstrap test becomes

$$n^{-1}r\big(Q_\alpha + n^{-1}q_\alpha, n^{-1/2}\boldsymbol{\nu} - \boldsymbol{\eta}_\alpha\big). \tag{34}$$

This completes the proof, since (34) is just expression (24) in the statement of the theorem. ∎

Proof of Theorem 2: In Section 3, it was shown that the distribution of the statistic τ^* under the DGP μ_0 was given by $E_{\mu_0}\big(R(\alpha, \mu^*)\big)$, or, in the notation of Section 7, $E_{\mu_0}\big(R(\alpha, \boldsymbol{\lambda}^*)\big)$. Thus expression (8), to which (9) tends as the number of simulations becomes large, is given by this expression evaluated at $\alpha = Q(\alpha, \mu_0)$, in the notation of Section 3, and $\alpha = \Phi(Q(\alpha, \boldsymbol{\lambda}_0))$, in that of Section 7. Thus (8) becomes $E_{\mu_0}\big(R(\Phi(Q(\alpha, \boldsymbol{\lambda}_0)), \boldsymbol{\lambda}^*)\big) = E_{\mu_0}\big(R(\Phi(Q_\alpha), \boldsymbol{\lambda}_0 + n^{-1/2}\boldsymbol{l})\big)$. Since R is linear with respect to its second (sequence) argument, this expectation is just $R\big(\Phi(Q_\alpha), \boldsymbol{\lambda}_0 + n^{-1}\boldsymbol{\nu}\big)$. Since $R\big(\Phi(Q_\alpha), \boldsymbol{\lambda}_0\big) = \alpha$, it follows that (8) is

$$R\big(\Phi(Q_\alpha), \boldsymbol{\lambda}_0\big) - R\big(\Phi(Q_\alpha), \boldsymbol{\lambda} + n^{-1}\boldsymbol{\nu}\big) = n^{-3/2}r(Q_\alpha, \boldsymbol{\nu}).$$

This demonstrates the assertion of the Theorem for (9) and $\widehat{\mathrm{ERP}}_2$.

The expressions in (10) make use of the α quantile of the distribution of τ^*, which is given by $Q(\alpha, \boldsymbol{\lambda}_0 + n^{-1}\boldsymbol{\nu}) = Q_\alpha + n^{-1}q_\alpha$. Thus it is clear that RP_1, estimated by the $\widehat{\mathrm{RP}}_1$ in (10), tends to the probability that τ, generated by μ_0, is less than $Q_\alpha + n^{-1}q_\alpha$. This is equal to $R\big(\Phi(Q_\alpha + n^{-1}q_\alpha), \boldsymbol{\lambda}_0\big)$. Since, by the definition of q_α, $R\big(\Phi(Q_\alpha + n^{-1}q_\alpha), \boldsymbol{\lambda}_0 + n^{-1}\boldsymbol{\nu}\big) = \alpha$, it follows at once that the ERP estimate in (10) tends to $n^{-3/2}r(Q_\alpha + n^{-1}q_\alpha, \boldsymbol{\nu})$. ∎

Proof of Theorem 3: By the results following (19) and derived from it, we see that

$$\lambda_1^0 + n^{-1/2}\eta_1^\alpha = n^{1/2}E_{\mu_0}(\tau - n^{-1}\gamma_\alpha).$$

By the same token, $n^{1/2}E_{\mu_0}(\tau) = \lambda_1^0$, and so $\eta_1^\alpha = -E(\gamma_\alpha) = O(n^{-1})$, by Lemma 1. This result holds generally, regardless of the nature of the joint distribution of τ and μ^*.

For the second component, we have that

$$\begin{aligned}
\lambda_2^0 + n^{-1/2}\eta_2^\alpha &= \frac{1}{2}n^{1/2}E\big((\tau - n^{-1}\gamma_\alpha)^2 - 1\big) \\
&= \frac{1}{2}n^{1/2}\big(E(\tau^2 - 1) + E(-2n^{-1}\tau\gamma_\alpha + n^{-2}\gamma_\alpha^2)\big) \\
&= \lambda_2^0 - n^{-1/2}E(\tau\gamma_\alpha - n^{-1}\gamma_\alpha^2/2),
\end{aligned}$$

whence $\eta_2^\alpha = -E(\tau\gamma_\alpha - n^{-1}\gamma_\alpha^2/2)$. If τ and μ^* are independent, it follows that τ and γ_α are independent, and in that case $E(\tau\gamma_\alpha) = E(\tau)E(\gamma_\alpha) = O(n^{-3/2})$. Since the second

term in η_2^α is $O(n^{-1})$, this implies that $\eta_2^\alpha = O(n^{-1})$. If τ and γ_α are not independent, then in general the first term of η_2^α is $O(1)$, since both τ and γ_α are $O(1)$ in general. For the case of asymptotic independence, note that by (31) γ_α is a deterministic function of the l_i. If, therefore, τ and the l_i are asymptotically independent in the sense of the statement of the theorem, $E(\tau\gamma_\alpha) = E(\tau)E(\gamma_\alpha) + O(n^{-1/2}) = O(n^{-1/2})$. In this case, then, $\eta_2^\alpha = O(n^{-1/2})$.

For all other components, we note that, from (19) with $j > 2$,

$$\lambda_j^0 + n^{-1/2}\eta_j^\alpha = n^{1/2}\frac{1}{j!}E\big(He_j(\tau - n^{-1}\gamma_\alpha)\big) \text{ and } \lambda_j^0 = n^{1/2}\frac{1}{j!}E\big(He_j(\tau)\big).$$

Thus

$$\eta_j^\alpha = \frac{n}{j!}E\big(He_j(\tau - n^{-1}\gamma + \alpha) - He_j(\tau)\big).$$

Now, since $He_j(x)$ is a polynomial of degree j in x, we have that

$$He_j(\tau - n^{-1}\gamma_\alpha) - He_j(\tau) = \sum_{i=1}^{j}\frac{1}{i!}(-1)^i n^{-i}\gamma_\alpha^i He_j^{(i)}(\tau).$$

One of the properties of the Hermite polynomials (see, for instance, Abramowitz and Stegun (1965), p. 783) is that $He_j'(x) = j\,He_{j-1}(x)$, from which it follows that the i^{th} derivative, for $i \le j$, is

$$He_j^{(i)}(x) = \frac{j!}{(j-i)!}He_{j-i}(x).$$

Thus we find that

$$\eta_j^\alpha = \frac{1}{j!}\sum_{i=1}^{j} n^{-(i-1)}(-1)^i\binom{j}{i}E\big(\gamma_\alpha^i He_{j-i}(\tau)\big). \tag{35}$$

In general, the first term on the right-hand side of (35) is $O(1)$. In the case of full independence of τ and γ_α, we have $E(\gamma_\alpha He_{j-1}(\tau)) = E(\gamma_\alpha)E(He_{j-1}(\tau)) = O(n^{-3/2})$, and $E(\gamma_\alpha^2 He_{j-2}(\tau)) = O(n^{-1/2})$. Thus the first two terms in (35) are of order $n^{-3/2}$, and the others are of order no greater than n^{-2}. With only asymptotic independence, the first term may be of order $n^{-1/2}$, the second of order n^{-1}, and all others of order no greater than n^{-2}. We have now proved all the assertions in the statement of the theorem. ∎

Proof of Corollary 1: It is clear from (35) and the expressions in the proof of Theorem 3 for η_1^α and η_2^α that all elements of the sequence $\boldsymbol{\eta}_\alpha$ depend on α only through γ_α. In turn, it is clear from (31) that γ_α is a sum of terms each of which is the product of a random variable independent of α and a polynomial in z_α. Thus, for all j, η_j^α is a sum of terms each of which is the product of an expectation independent of α and a polynomial in z_α.

Any polynomial can be expressed as a linear combination of the Hermite polynomials. Therefore, it follows that

$$r(z_\alpha, \boldsymbol{\eta}_\alpha) \equiv \phi(z_\alpha) \sum_i \eta_i^\alpha He_{i-1}(z_\alpha)$$

can be expressed, by expanding the η_i^α as described above, multiplying the polynomials in z_α by the $He_{i-1}(z_\alpha)$, and rearranging, as an expression of the form

$$\phi(z_\alpha) \sum_i \theta_i He_{i-1}(z_\alpha) = r(z_\alpha, \boldsymbol{\theta}), \tag{36}$$

where the order as $n \to \infty$ of θ_i cannot be greater than the maximal order of the η_i^α. We can see from (29) and (30) that $Q_\alpha + n^{-1}q_\alpha$ is equal to z_α plus a sum of terms of the same type as those in $\boldsymbol{\eta}_\alpha$. Since the derivative of $\phi(x)He_{i-1}(x)$ is $-\phi(x)He_i(x)$, a Taylor expansion shows that $r(Q_\alpha + n^{-1}q_\alpha, \boldsymbol{\theta})$ can also be expressed in the form (36), with a suitable redefinition of $\boldsymbol{\theta}$, in which each element is changed by a quantity of order at most $n^{-1/2}$. Since r is linear with respect to its second argument, the conclusion of the corollary follows. ∎

Proof of Theorem 4: Conditional on τ and μ^*, the properties of the τ_j^* and the τ_j^{**} in (11), (12), and (13) are the same as those of τ and τ^* in the proofs of the earlier theorems, except that μ_0 is replaced by the realization $\hat{\mu}$ of μ^*. Thus the FDB_2 P value p_2^{**} to which (13) tends as $B \to \infty$ is the random variable

$$2p^* - R\big(\Phi(\tau), \boldsymbol{\lambda}^* + n^{-1}\boldsymbol{\nu}^*\big), \tag{37}$$

where $p^* \equiv R\big(\Phi(\tau), \boldsymbol{\lambda}^*\big)$ is the ordinary bootstrap P value, and $\boldsymbol{\nu}^*$, the bootstrap version of $\boldsymbol{\nu}$, is nonrandom conditional on μ^*. It is easy to see that (37) becomes

$$R\big(\Phi(\tau), \boldsymbol{\lambda}^*\big) + n^{-3/2} r(\tau, \boldsymbol{\nu}^*) = R\big(\Phi(\tau), \boldsymbol{\lambda}^* - n^{-1}\boldsymbol{\nu}^*\big), \tag{38}$$

by (23).

The quantity $\hat{Q}^*(\hat{p}^*)$ in (11) is an estimate of the p^* quantile of the distribution characterized by the sequence $\boldsymbol{\lambda}^* + n^{-1}\boldsymbol{\nu}^*$. Since the τ^* are generated by μ^*, their distribution is characterized by $\boldsymbol{\lambda}^*$. Thus, as $B \to \infty$, (11) tends to

$$R\big(\Phi(Q(p^*, \boldsymbol{\lambda}^* + n^{-1}\boldsymbol{\nu}^*)), \boldsymbol{\lambda}^*\big) = p^* + n^{-3/2} r\big(Q(p^*, \boldsymbol{\lambda}^* + n^{-1}\boldsymbol{\nu}^*), \boldsymbol{\nu}^*\big), \tag{39}$$

as in the proof of Theorem 2. From (29), it can be seen that $Q(p^*, \boldsymbol{\lambda}^* + n^{-1}\boldsymbol{\nu}^*) = Q(p^*, \boldsymbol{\lambda}^*) + O(n^{-3/2})$. Further,

$$Q(p^*, \boldsymbol{\lambda}^*) = Q\big(R(\Phi(\tau), \boldsymbol{\lambda}^*), \boldsymbol{\lambda}^*\big) = \tau,$$

by (21). Thus the P value (39) becomes

$$R\big(\Phi(\tau), \boldsymbol{\lambda}^*\big) + n^{-3/2} r(\tau, \boldsymbol{\nu}^*) + O(n^{-3}),$$

which is the same as (38) except for the $O(n^{-3})$ term. We may therefore treat both FDB P values as the same up to any order in which we are interested.

The P value (38) can now be considered as a random variable, being a deterministic function of τ, $\boldsymbol{\lambda}^*$, and $\boldsymbol{\nu}^*$. The probability, under the DGP μ_0, that (38) is less than α is then just the probability that $\tau < Q(\alpha, \boldsymbol{\lambda}^* - n^{-1}\boldsymbol{\nu}^*)$, and this probability can be evaluated using the method of the proof of Theorem 1. Since $\boldsymbol{\nu}^*$ is a bootstrap estimate of $\boldsymbol{\nu}$, in exactly the same sense that $\boldsymbol{\lambda}^*$ is a bootstrap estimate of $\boldsymbol{\lambda}_0$, we have that $\boldsymbol{\nu}^* = \boldsymbol{\nu} + n^{-1/2}\boldsymbol{r}$, with $\boldsymbol{r} = O(1)$, and $E(\boldsymbol{r}) = n^{-1/2}\boldsymbol{\rho}$, with $\boldsymbol{\rho} = O(1)$. We obtain that

$$\boldsymbol{\lambda}^* - n^{-1}\boldsymbol{\nu}^* = \boldsymbol{\lambda}_0 + n^{-1/2}\boldsymbol{l} - n^{-1}\boldsymbol{\nu} - n^{-3/2}\boldsymbol{r}.$$

Thus the sequence $\boldsymbol{\nu}$ in the proof of Theorem 1 is to be replaced here by $nE(n^{-1/2}\boldsymbol{l} - n^{-1}\boldsymbol{\nu} - n^{-3/2}\boldsymbol{r}) = n^{-1}\boldsymbol{\rho}$. By (30) therefore, q_α is replaced by a quantity of order n^{-1}. Similarly, we see that q^*_α in Theorem 1 is changed here only by a quantity of order n^{-1}. Thus γ_α is changed only at order n^{-1}, and consequently also $\boldsymbol{\eta}_\alpha$. Making these changes in (34) gives (28).

For the full double bootstrap, note that the $\hat{p}^{**}_j$ in (14) have the same properties, conditional on τ and μ^*, as the ordinary bootstrap P value p^* under μ_0, of which the distribution is given by (26). Thus the distribution of the p^{**}_j is given by (26) with $\boldsymbol{\nu}$ and $\boldsymbol{\theta}$ replaced by $\boldsymbol{\nu}^*$ and $\boldsymbol{\theta}^*$, their bootstrap counterparts. As $B \to \infty$, therefore, the double bootstrap P value tends to this distribution evaluated at the single bootstrap P value, p^*, so that the ideal double bootstrap P value is

$$p^{**} = R(p^*, n^{-1}\boldsymbol{\nu}^* - n^{-1/2}\boldsymbol{\theta}^*).$$

The double bootstrap test rejects at nominal level α if $p^{**} < \alpha$, that is, if

$$\Phi^{-1}(p^*) < Q(\alpha, n^{-1}\boldsymbol{\nu}^* - n^{-1/2}\boldsymbol{\theta}^*). \tag{40}$$

The rest of the proof closely mirrors the proof of Theorem 1. As with $\boldsymbol{\nu}^*$, we may write $\boldsymbol{\theta}^* = \boldsymbol{\theta} + n^{-1/2}\boldsymbol{h}$, with $E(\boldsymbol{h}) = O(n^{-1/2})$. Then we implicitly define δ_α by the relation

$$Q(\alpha, n^{-1}\boldsymbol{\nu}^* - n^{-1/2}\boldsymbol{\theta}^*) = Q(\alpha, n^{-1}\boldsymbol{\nu} - n^{-1/2}\boldsymbol{\theta}) - n^{-3/2}\delta_\alpha,$$

where, by use of (29), we see that $\delta_\alpha = \sum_i h_i He_{i-1}(z_\alpha) + O(n^{-1/2})$. Inequality (40) can now be written as $\Phi^{-1}(p^*) + n^{-3/2}\delta_\alpha < Q(\alpha, n^{-1}\boldsymbol{\nu} - n^{-1/2}\boldsymbol{\theta})$. Let the distribution of the left-hand side of this inequality be characterized by $n^{-1}\boldsymbol{\nu} - n^{-1/2}\boldsymbol{\theta} - n^{-1}\boldsymbol{\zeta}_\alpha$, where $\boldsymbol{\zeta}_\alpha$ is of order at most unity. As can easily be checked, it is at most of order $n^{-1/2}$ in the case of asymptotic independence of τ and μ^*. Thus the probability that (40) holds is

$$R\big(\Phi(Q(\alpha, n^{-1}\boldsymbol{\nu} - n^{-1/2}\boldsymbol{\theta})), n^{-1}\boldsymbol{\nu} - n^{-1/2}\boldsymbol{\theta} - n^{-1}\boldsymbol{\zeta}_\alpha\big),$$

and so the ERP is $n^{-3/2} r\big(Q(\alpha, n^{-1}\boldsymbol{\nu} - n^{-1/2}\boldsymbol{\theta}), \boldsymbol{\zeta}_\alpha\big)$, as required. ∎

Table 1. Rejection frequencies at .05 level, 10,000 replications

	n	B	Bootstrap	FDB_1	FDB_2	Double Bootstrap
Case 1	50	199	0.04030	0.04930	0.05870	0.04660
Case 1	80	199	0.04490	0.05120	0.05950	0.04870
Case 1	120	199	0.05050	0.05400	0.06180	0.05210
Case 2	50	199	0.05790	0.04630	0.05900	0.03420
Case 2	80	199	0.04940	0.05090	0.05920	0.04970
Case 2	120	199	0.04600	0.04980	0.05660	0.04700

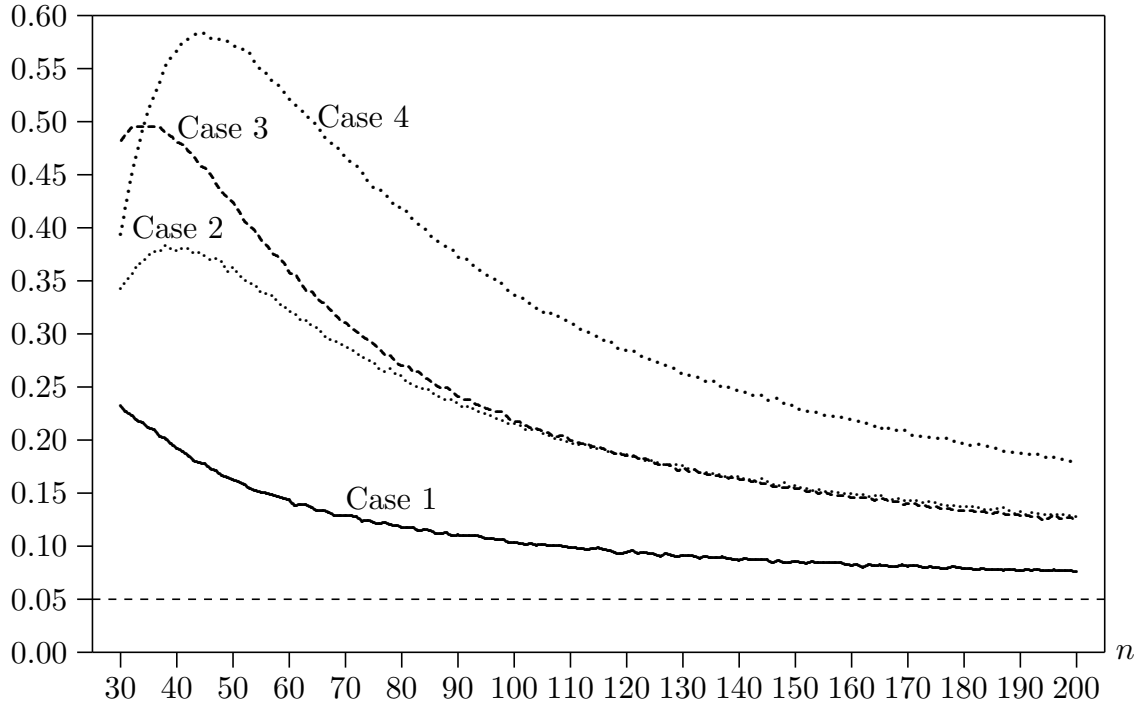

Figure 1. Rejection Frequencies at Nominal .05 Level, Asymptotic Tests

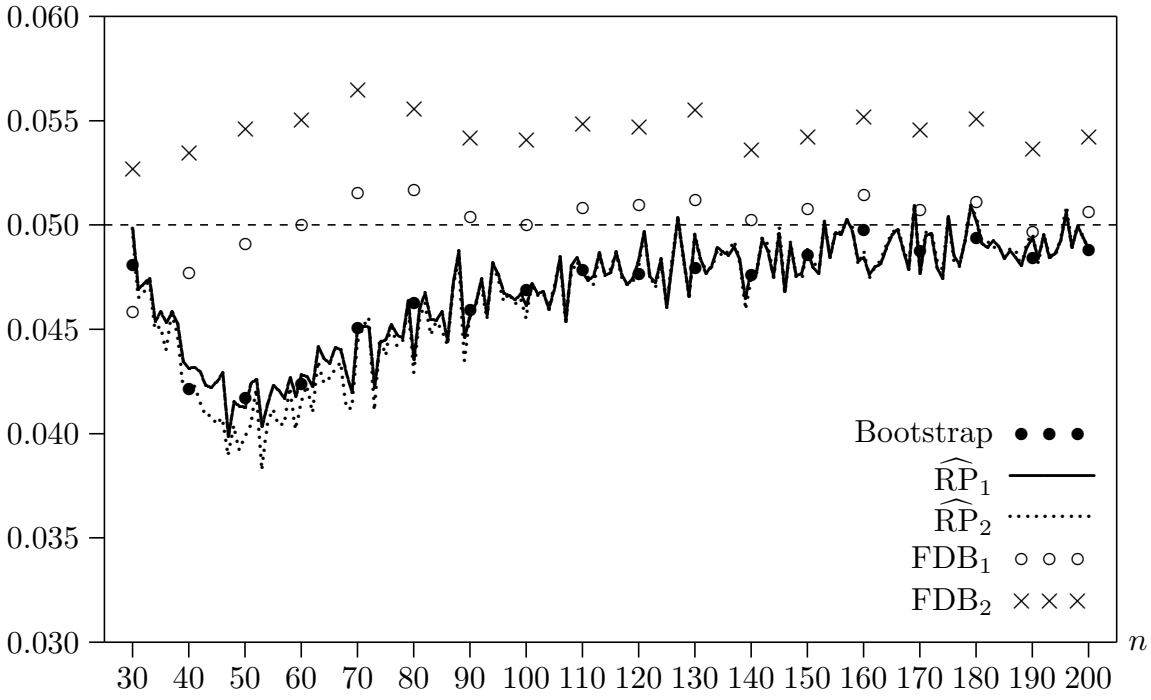

Figure 2. Rejection Frequencies at Nominal .05 Level, Case 1

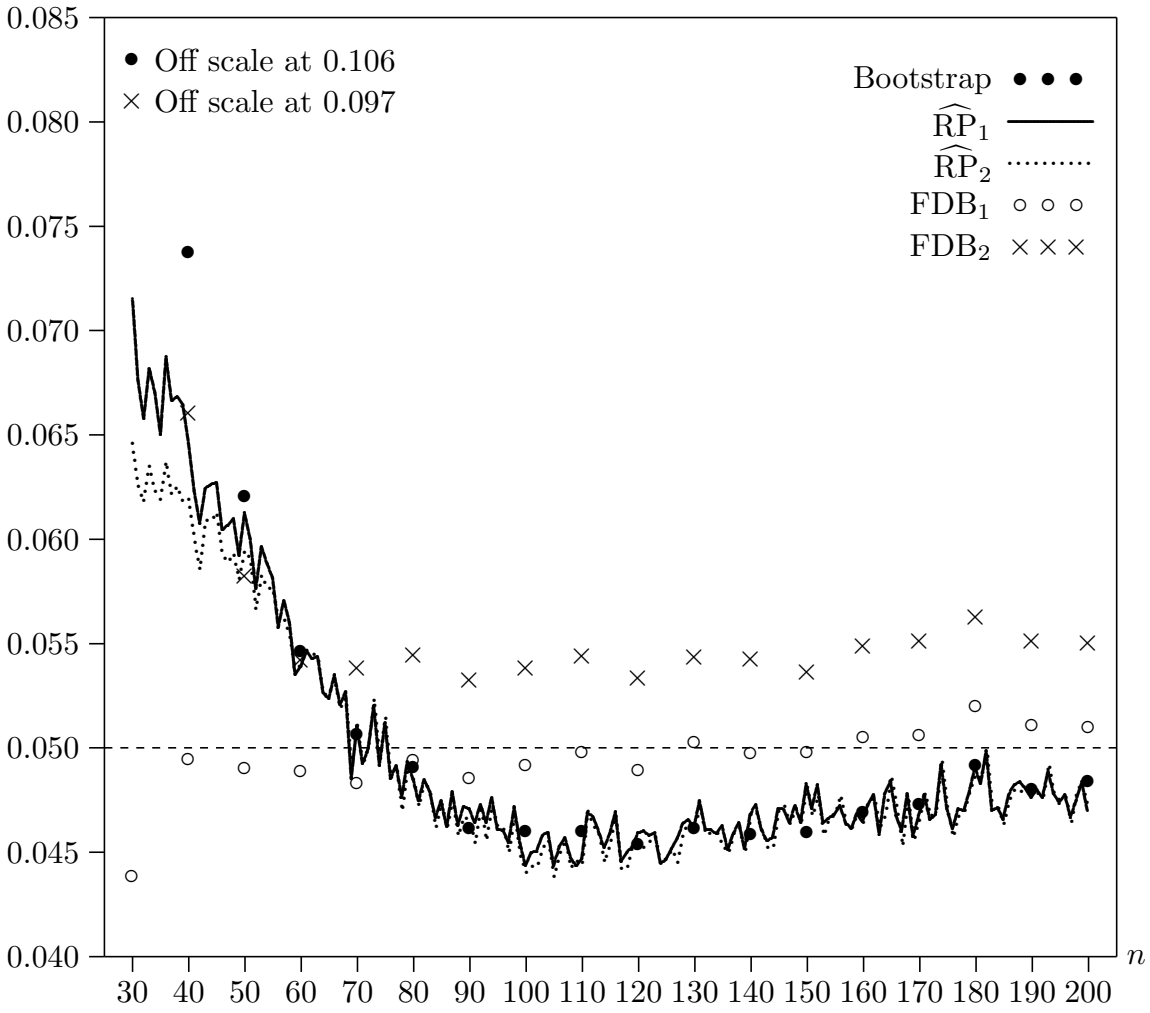

Figure 3. Rejection Frequencies at Nominal .05 Level, Case 2

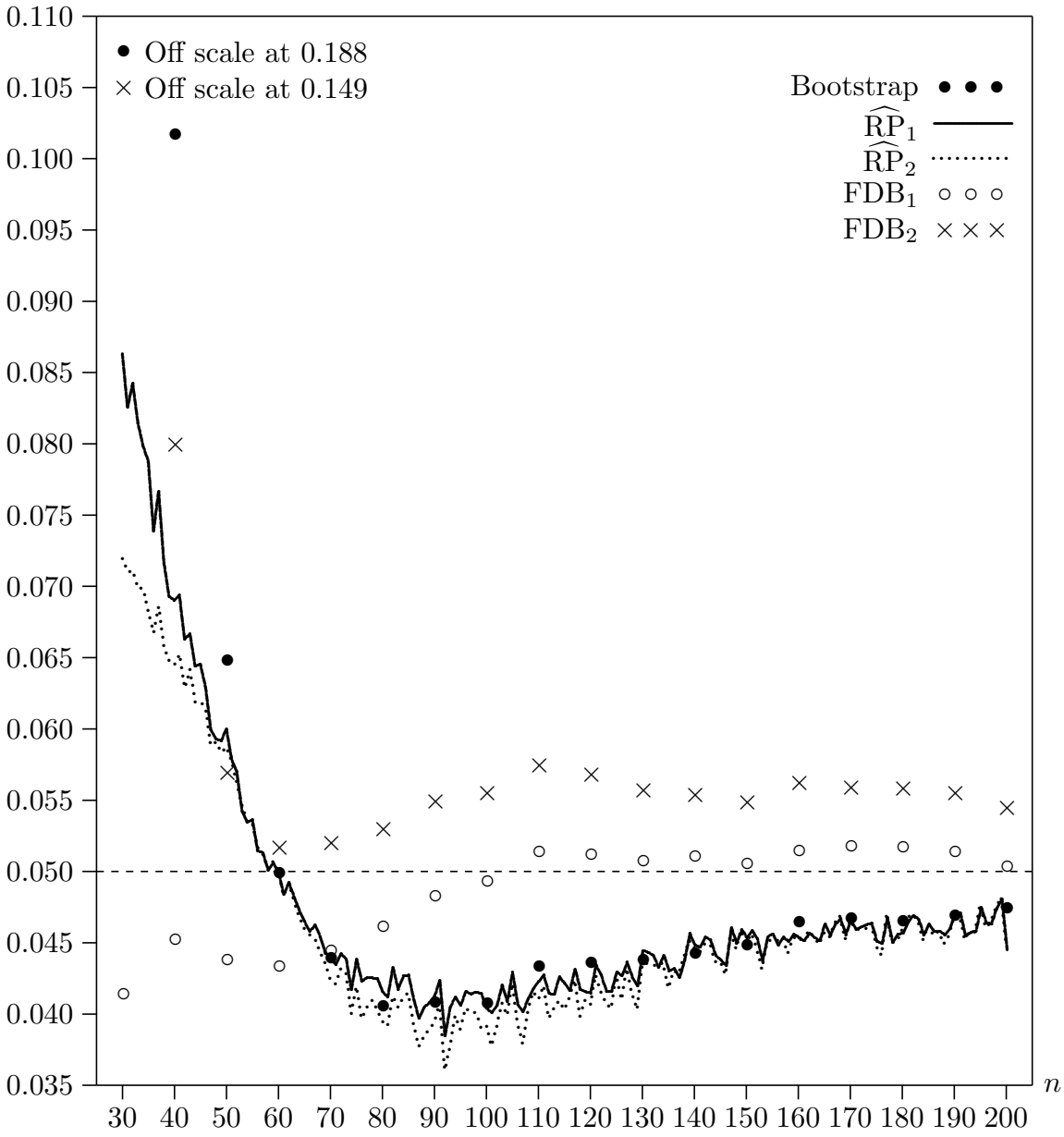

Figure 4. Rejection Frequencies at Nominal .05 Level, Case 3

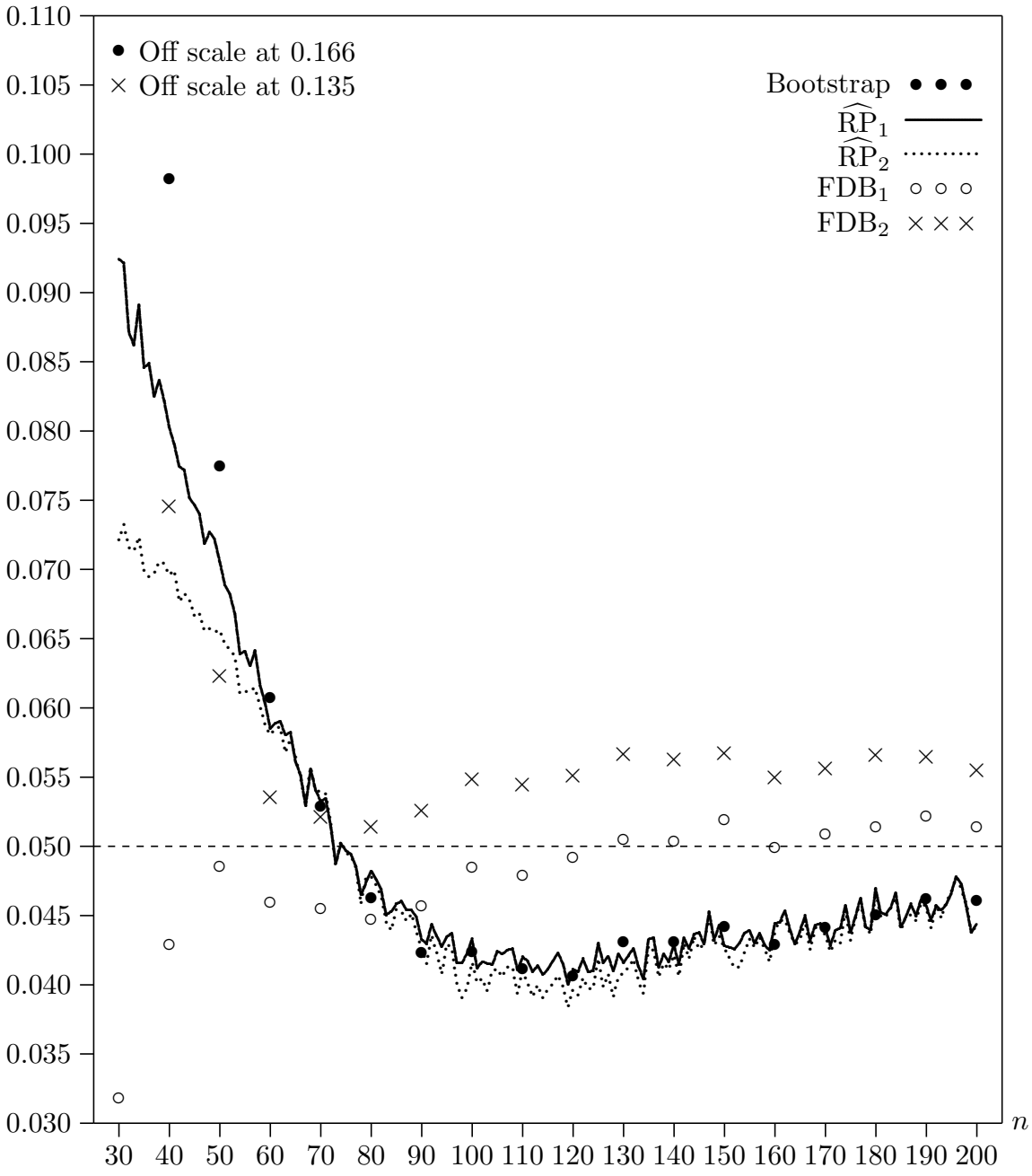

Figure 5. Rejection Frequencies at Nominal .05 Level, Case 4

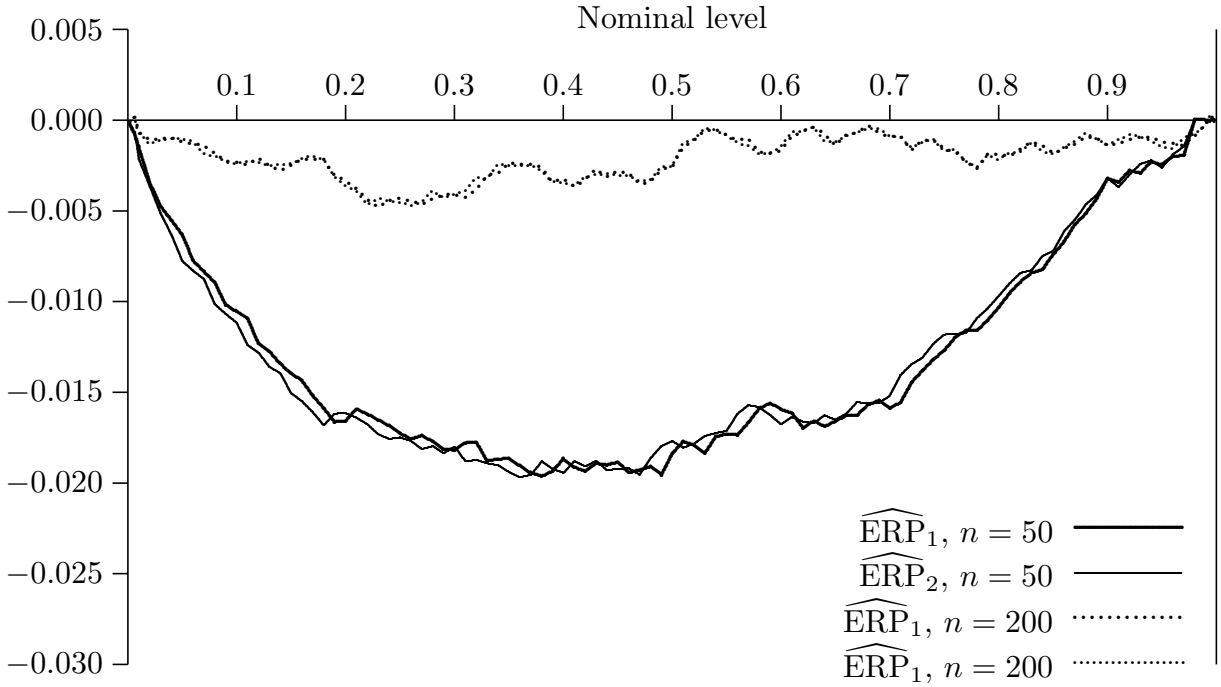

Figure 6. *P* value discrepancy plots, Case 1

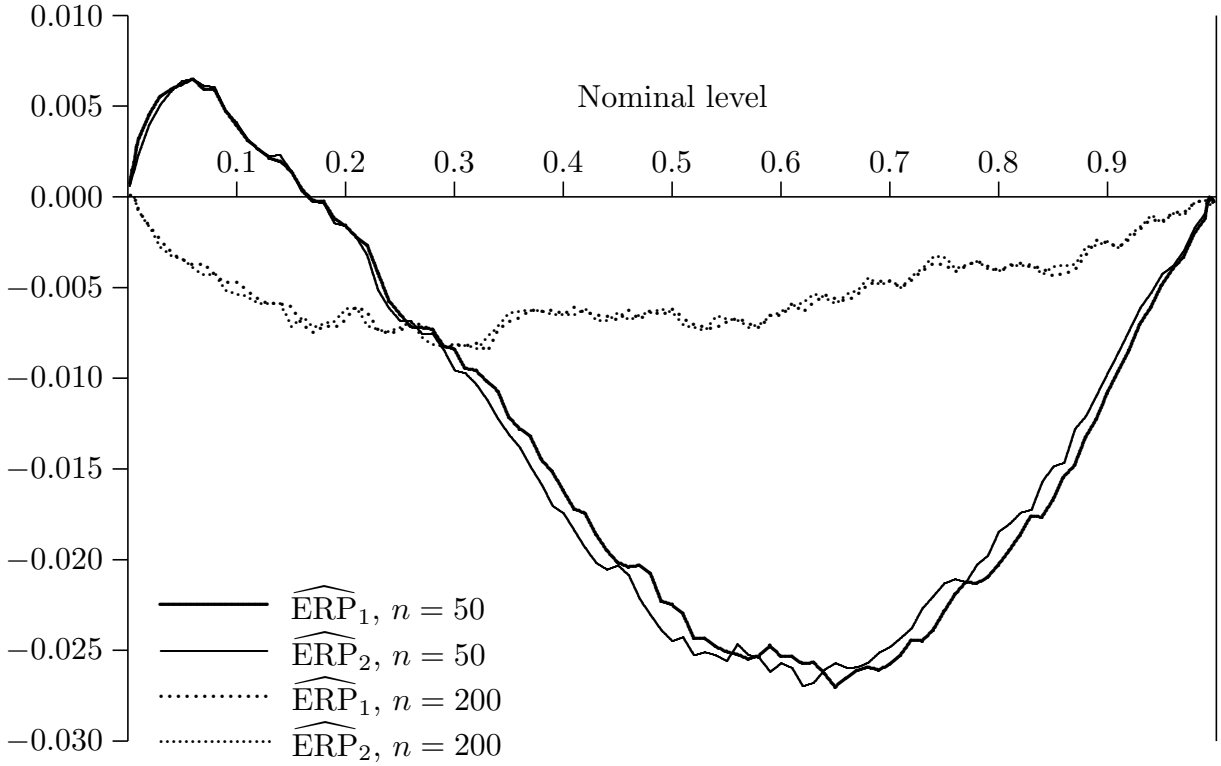

Figure 7. *P* value discrepancy plots, Case 4