Queen's Economics Department Working Paper No. 1054

Inference via Kernel Smoothing of Bootstrap P Values

Jeff Racine
McMaster University

James G. MacKinnon
Queen's University

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

3-2006 (minor correction 9-2019)

INFERENCE VIA KERNEL SMOOTHING OF BOOTSTRAP P VALUES

JEFF RACINE AND JAMES G. MACKINNON

Abstract. Resampling methods such as the bootstrap are routinely used to estimate the finite-sample null distributions of a range of test statistics. We present a simple and tractable way to perform classical hypothesis tests based upon a kernel estimate of the CDF of the bootstrap statistics. This approach has a number of appealing features: i) it can perform well when the number of bootstraps is extremely small, ii) it is approximately exact, and iii) it can yield substantial power gains relative to the conventional approach. The proposed approach is likely to be useful when the statistic being bootstrapped is computationally expensive.

1. Introduction

The use of bootstrap methods for inference in finite samples is by now well established. In many cases, these methods yield more accurate inferences than methods based on asymptotic theory; see Hall [10] and Mammen [14], among many others. Moreover, they may be more robust to departures from distributional assumptions; for an example, see Bickel and Freedman [3].

The simplest bootstrap methods often do not work any better than asymptotic ones, in the sense that the mistakes made by both methods are of the same order in the sample size n. However, more sophisticated bootstrap methods often do work better than asymptotic methods. In particular, methods based on quantities that are asymptotically pivotal, such as *percentile-t* methods, work well in theory and often (but not always) work well in practice; see Hall [9] and Beran [2]. A statistic is (asymptotically) pivotal if its (limiting) distribution does not depend on unknown quantities (Hall, [10] page 83). For example, suppose we are interested in a parameter θ. The

Date: September 20, 2019.
Key words and phrases. Resampling, Monte Carlo test, percentiles.
JEL Classification: C12, C14, C15.
We are grateful to Aman Ullah and to seminar participants at the University of Montreal, the Canadian Economics Association, and the Canadian Econometric Study Group for comments on earlier versions.

estimator $\hat{\theta}$ is almost never asymptotically pivotal. However, the studentized statistic, $\hat{\tau} = (\hat{\theta} - \theta)/s(\hat{\theta})$, where $s(\hat{\theta})$ is a consistent estimate of the standard error of $\hat{\theta}$, will be asymptotically pivotal in most cases.

In this paper, we will be concerned with hypothesis tests based on asymptotically pivotal test statistics. The latter may be likelihood ratio statistics, Lagrange Multiplier statistics, Wald statistics, or indeed any sort of test statistic that is asymptotically pivotal. In most cases, these statistics will have known asymptotic distributions, although this is not essential. We will let $\hat{\tau} \equiv \hat{\tau}_n$ denote a test statistic computed from a sample of size n, and we will let $\tau_j^* \equiv \tau_{n,j}^*$ denote a bootstrap statistic computed from the j^{th} bootstrap sample, which is generated under the null hypothesis, where $j = 1, \ldots, B$. Since all samples are of size n, the "n" subscripts will often be omitted for simplicity.

We assume that the limiting distribution of the $\tau_{n,j}^*$ is the same as the limiting distribution of the $\hat{\tau}_n$ under the null hypothesis. Showing that this is so for any particular method of generating bootstrap samples in any specific case may require considerable effort, and the details of how the bootstrap samples are constructed can be extremely important. However, we abstract from these issues in this paper.

If computing the τ_j^* is not expensive, so that we can afford to make B a reasonably large number (for example, 999), then there is no need for the methods introduced in this paper. However, despite the astonishing advances in computing power over the past two decades, there are many cases in which calculating $\hat{\tau}$ and the τ_j^* is computationally demanding. For example, simulation-based methods may be required to estimate the model; see van Dijk and Monfort [20] and Gouriéroux and Monfort [8] for introductions to simulation-based methods in econometrics.

Even when simulation is not needed to compute parameter estimates, it may be required to compute standard errors, or, more generally, covariance matrices. As Efron and Tibshirani ([7], page 162) note, "standard error formulae exist for very few statistics." Thus, in order to calculate the τ_j^*, it may first be necessary to compute a bootstrap estimate of the standard error for each bootstrap sample. This can be computationally demanding. Even when standard errors can be calculated without simulation, they may not be reliable. In such cases, the double bootstrap proposed by Beran [2] can be used to obtain more accurate inferences. As its name implies, this method involves two levels of bootstrap. Formally, our analysis applies to the

double bootstrap, since we can think of $\hat{\tau}$ as being a first-level bootstrap P value and the τ_j^* as being second-level bootstrap P values.

We propose a simple and tractable approach that computes a P value based on a kernel estimate of the CDF of the bootstrap statistics. This approach requires fewer bootstrap replications than the conventional approach to attain any level of accuracy. It can yield reasonably accurate inferences even with fewer than 20 bootstrap replications. Existing bandwidth selection rules do not yield an exact test, though for many users the size distortions may be perfectly acceptable and be more than offset by the power gains. Data-driven bandwidth selection rules that deliver an exact test remain a topic for future research.

2. Conventional Approaches

The usual method for computing bootstrap P values is simply to find the proportion of the bootstrap statistics τ_j^* that are more extreme than the actual statistic $\hat{\tau}$. Call this proportion P_B^*. Then one rejects the null hypothesis if P_B^* is less than α, the level of the test. When B is chosen so that $\alpha(B+1)$ is an integer, and the test statistic is pivotal, this procedure yields an exact test. Such a test is sometimes called a "Monte Carlo test." The reason why Monte Carlo tests are exact is easy to see. If $\hat{\tau}$ and the τ_j^* all follow the same distribution, then the probability that $\hat{\tau}$ will be among the $\alpha(B+1)$ most extreme values by chance is precisely α. But this is precisely the situation in which P_B^* is less than α.

Formally, for a test that rejects in the upper tail,

$$P_B^* = 1 - \hat{F}(\hat{\tau}) = 1 - \frac{1}{B}\sum_{j=1}^{B} I(\tau_j^* \leq \hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} I(\tau_j^* > \hat{\tau}), \tag{1}$$

where $I(\cdot)$ is the indicator function, and $\hat{F}(\hat{\tau})$ is the empirical distribution function (EDF) of the bootstrap statistics. Thus this procedure effectively uses the EDF $\hat{F}(\hat{\tau})$ to estimate a critical value. For a test at level α that rejects when the test statistic is in the right-hand tail, the appropriate critical value is the $1-\alpha$ quantile. When $\alpha(B+1)$ is an integer, rejecting whenever $\hat{\tau}$ exceeds observation $(1-\alpha)(B+1)$ in the list of the $\hat{\tau}_j^*$ sorted from smallest to largest yields exactly the same test as rejecting when P_B^* is less than α.

There are at least two problems with this approach. First, it works very badly if B is small and $\alpha(B+1)$ is not an integer. To appreciate the size distortions in such instances, Figure 1 plots the empirical rejection frequencies for the EDF approach. These are very large for some (small) values of B. It is possible to avoid this problem by using a method proposed in Racine and MacKinnon [17], but it involves drawing an additional random number on which the outcome of the test may depend.

Second, the randomness introduced by simulation causes the test to lose power. This loss of power is proportional to $1/B$. For discussions of power loss in Monte Carlo and bootstrap tests, see Jöckel [13], Hall and Titterington [11], and Davidson and MacKinnon [6]. Because the EDF-based test ignores a great deal of information about the shape of the distribution that the τ_j^* potentially convey, it loses more power than it needs to lose. Consider a change in the value of one of the τ_j^*. Such a change will have no effect on P_B^* unless the sign of $\tau_j^* - \hat{\tau}$ changes. When the sign does change, P_B^* rises or falls by precisely $1/B$, regardless of how much τ_j^* changed.

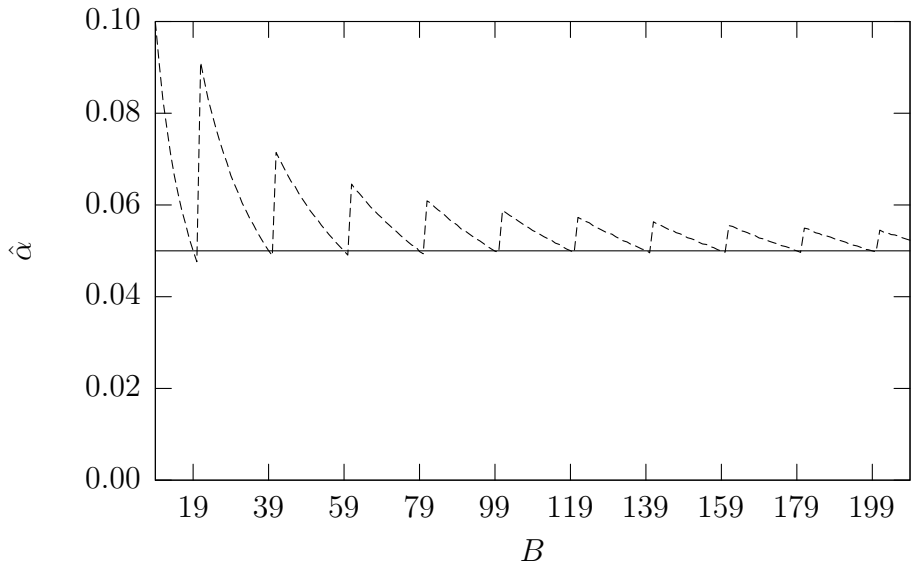

FIGURE 1. Empirical rejection frequencies for the conventional EDF test at the .05 level. For small B where $\alpha(B+1)$ is not an integer, the size distortions are upwards of 100%.

As an example, consider the situation in which $B = 19$ and $\hat{\tau}$ is more extreme than all 19 of the τ_j^*. This situation might often arise if the null hypothesis were false.

Formally, even though $P_{19}^* = 0$ in this case, we can only reject the null hypothesis at the .05 level. If several of the τ_j^* are very close to $\hat{\tau}$, even that conclusion may seem dubious. On the other hand, if $\hat{\tau}$ is much more extreme than all of the τ_j^*, it would seem reasonable to reject the null hypothesis quite decisively. The EDF-based estimator makes no distinction between these two cases. In contrast, the kernel-based procedures that we are about to describe would yield very different, and much more informative, P values.

Some authors use a modified version of the EDF approach in which

$$P_B^* = \frac{1}{B+1} + \frac{1}{B+1} \sum_{j=1}^{B} I(\tau_j^* > \hat{\tau}), \tag{2}$$

and reject whenever $P_{B*} \leq \alpha$. This *biased EDF* approach leads to precisely the same inferences as the usual EDF approach whenever $\alpha(B+1)$ is an integer, but it prevents overrejection caused by a poor choice of B when that condition is not satisfied. However, it can cause severe loss of power when B is small and badly chosen. Notice that a test based on (2) can never reject when B is less than the smallest value for which $\alpha(B+1)$ is an integer. Figure 2 plots the empirical rejection frequencies for the biased EDF approach.

We now propose a new approach that does not suffer from the size distortions associated with the conventional EDF and biased EDF approaches, one that is also capable of recapturing the loss in power associated with Monte Carlo and bootstrap tests. We do so by constructing smooth estimates of the distribution function, a subject going back to the pioneering work of Nadaraya [15].

3. Kernel estimation of P values

The kernel estimator of the cumulative distribution function (CDF) of $X \in \mathbb{R}$ at the point $x \in \mathbb{R}$, using a sample of B observations x_j, is given by

$$\hat{F}_h(x) = \frac{1}{B} \sum_{j=1}^{B} K(x_j, x, h), \tag{3}$$

where $K(x_j, x, h)$ is a cumulative kernel, for example, the cumulative Epanechnikov or Gaussian kernel, and h is the bandwidth. Properties of this estimator were originally examined by Nadaraya [15], who demonstrated that this kernel estimator has the same asymptotic mean and variance as the EDF $\hat{F}$. Azzalini [1], Reiss [18], and

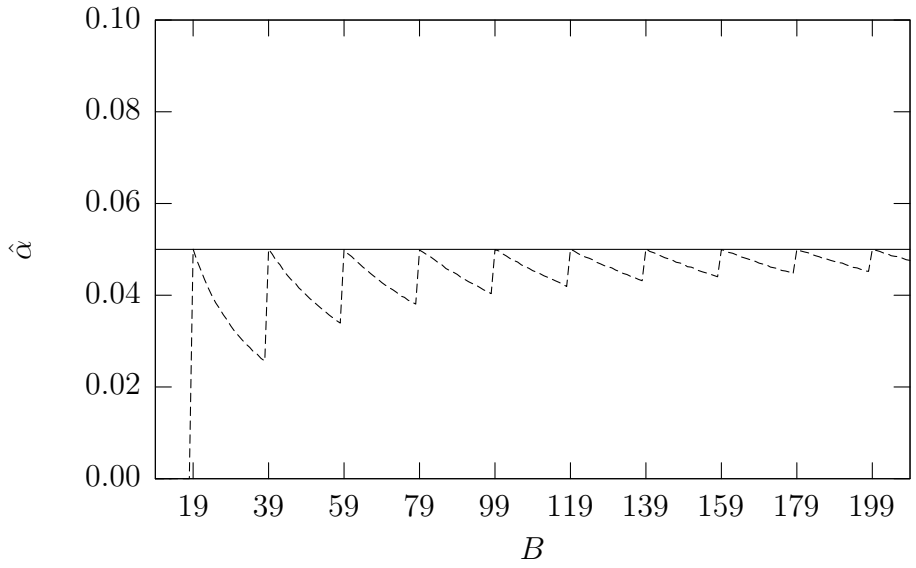

FIGURE 2. Empirical rejection frequencies for the conventional biased EDF test at the .05 level. For small B where $\alpha(B+1)$ is not an integer, the size distortions are downwards of 100%.

others have demonstrated how kernel smoothing reduces the variance but introduces bias.

By way of comparison, Figure 3 presents empirical and kernel estimates of a distribution function for a random sample of size $B = 25$ drawn from a $N(0, 1)$ distribution. For the empirical approach, the distribution is set equal to 1 for all values $x > x_{\max}$, where $x_{\max} = 2.71$ in this case. In contrast, the kernel estimator assigns values between $x_{\max}$ and roughly 3.7 a CDF value less than 1. For instance, for $X = 2.72$, the empirical probability of obtaining a draw this large is $1 - \hat{F}(2.72) = 0.000$, while that based upon the kernel estimate is $1 - \hat{F}_h(2.72) = 0.021$. Clearly, the kernel estimate is much better than the estimate based on the EDF.

Our proposal is to replace P_B^* defined in equation (1) by the smoothed bootstrap P value

$$\hat{P}_B^h = 1 - \hat{F}_h(\hat{\tau}) = 1 - \frac{1}{B}\sum_{j=1}^{B} K(\hat{\tau}_j^*, \hat{\tau}, h), \tag{4}$$

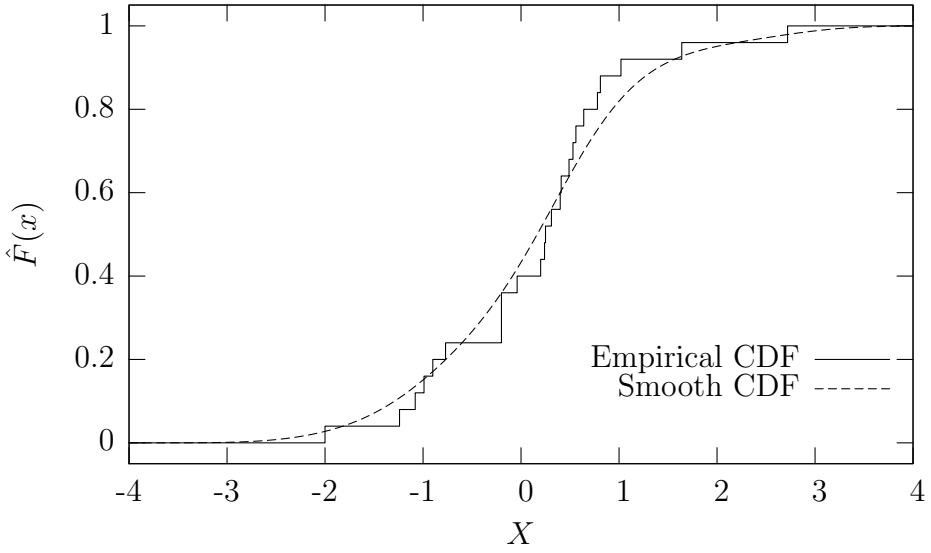

FIGURE 3. Empirical and smooth kernel estimates of a distribution function for $B = 25$ ($h = 1.587\hat{\sigma}B^{-1/3}$).

where $\hat{F}_h(\hat{\tau})$ is the smoothed CDF. We call this the "smoothed P value approach." Although both approaches yield identical results in the limit as $B \to \infty$, their performance will differ when B is finite. When computing the bootstrap test statistics is computationally burdensome, so that it is not feasible to make B a large number, there can be substantial gains in power from using the smoothed P value approach.

As is always the case with kernel estimation, the choice of h is extremely important, but the choice of kernel does not seem to matter very much. For what follows, we restrict attention to the cumulative Gaussian kernel, that is, the standard normal CDF.

3.1. **Bandwidth selection for CDFs.** Data-driven methods of bandwidth selection for kernel estimators of CDFs have been studied by a number of authors, notably Sarda [19] and Bowman, Hall, and Prvan [4], who studied cross-validation methods, and Polansky and Baker [16], who examined plug-in methods. These methods produce IMSE-optimal bandwidths, that is, bandwidths which *globally* minimize the MSE of the resulting estimator. It can readily be shown that the globally-optimal reference bandwidth assuming an underlying normal distribution when using a Gaussian kernel

is $h = 1.587\sigma(x)B^{-1/3}$. This often differs substantially from the reference bandwidth that is optimal for kernel estimation of a PDF, which is $h = 1.059\sigma(x)B^{-1/5}$.

Alternatively, if one is interested in obtaining a pointwise optimal estimate (that is, an estimate of $F(x)$ at a *particular* point x rather than over its entire support), one would instead select a bandwidth that is MSE-optimal. Azzalini [1] has demonstrated that the MSE-optimal bandwidth is again of the form $h = c\sigma(x)B^{-1/3}$, similar to the IMSE-optimal bandwidth, and that the value of c appropriate for estimating the tails of $F(x)$ is roughly 1.30 for a range of distributions.

Both the pointwise (MSE-optimal) and the globally optimal bandwidths (that is, $1.587\sigma(x)B^{-1/3}$ or $1.30\sigma(x)B^{-1/3}$) deliver kernel estimators of $\hat{P}_B^h = 1 - \hat{F}_h(\hat{\tau})$ that have desirable squared error properties.

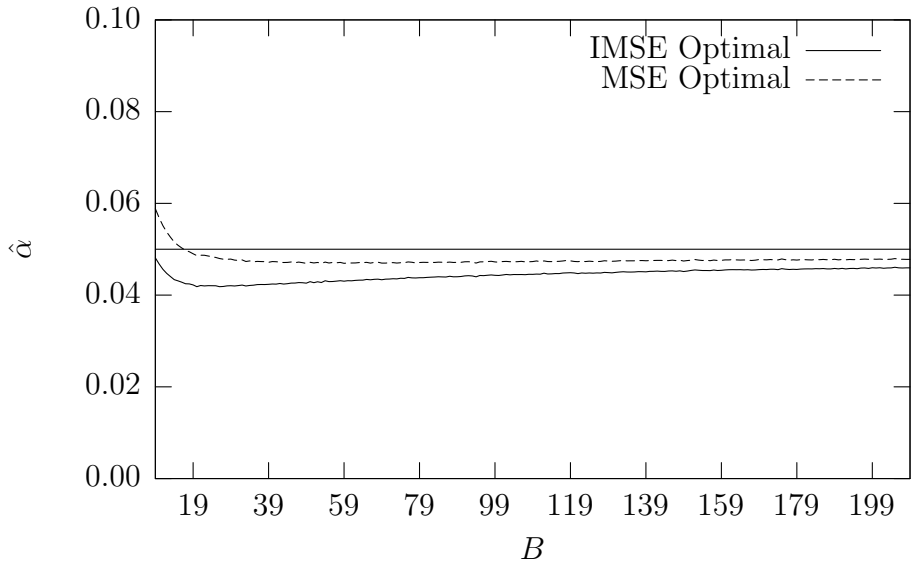

FIGURE 4. Empirical rejection frequencies for the kernel-based test using IMSE and MSE optimal bandwidths.

The performance of conventional bandwidth selection rules is shown graphically in Figure 4, which plots rejection frequencies for tests at the .05 level against B. It is based on two million replications. Both the actual test statistic and the "bootstrap" statistics are $N(0, 1)$; note that a different choice of distribution would have led to different results. There are two choices of bandwidth, one which is IMSE-optimal and one which is MSE-optimal. We use an estimate of the standard deviation of the

bootstrap statistics to compute the optimal bandwidths. In both cases, the smoothed bootstrap tests underreject for most values of B and overreject for very small values. For instance, the MSE optimal rule yields an empirical size of .059 for $B = 9$, .050 for $B = 15$, .048 for $B = 25$, and .047 for $B = 50$, while the IMSE optimal rule yields an empirical size of .048 for $B = 9$, .043 for $B = 15$, .042 for $B = 25$, and .043 for $B = 50$. For comparison, the conventional EDF approach yields empirical sizes of .100 for $B = 9$, .062 for $B = 15$, .077 for $B = 25$, and .059 for $B = 50$, but it yields exact tests when $B = 19$, $B = 39$, and so on.

One could reasonably argue that these size distortions are quite acceptable, particularly when compared with the distortions associated with many bootstrap tests even when $B = \infty$ and with the distortions for the conventional approach when B is chosen poorly. Furthermore, when coupled with the power gains associated with smoothing that we investigate later in this paper, a reasonable researcher may well be convinced of the overall advantages of using the proposed approach based upon existing bandwidth selection rules, especially in light of established data-driven methods that approximate the IMSE optimal rules used above. We now investigate whether or not a bandwidth selection rule exists that is capable of yielding an exact test.

3.2. **Existence of an optimal bandwidth.** A bandwidth selection rule for smooth bootstrap P values ought to yield an approximately exact test if the test statistic is a pivot. It is evident that the existing MSE and IMSE optimal bandwidth rules outlined above are not optimal for the task at hand.

The tests we are proposing will reject the null hypothesis whenever $\hat{P}_B^h < \alpha$. Let $\hat{\alpha}$ denote the empirical rejection frequency of such a test. The expectation of $\hat{\alpha}$ is

$$\begin{aligned} \mathrm{E}(\hat{\alpha}) &= \mathrm{E}\big(I(\hat{P}_B^h < \alpha)\big) \\ &= \Pr(\hat{P}_B^h < \alpha) = \Pr\big(\hat{F}_h(\hat{\tau}) > 1 - \alpha\big). \end{aligned} \tag{5}$$

In order for our test to be exact, therefore, the bandwidth h must be chosen to ensure that $\Pr\big(\hat{F}_h(\hat{\tau}) > 1 - \alpha\big) = \alpha$. That is, we must select an h such that, when we conduct a test at level α, the expected rejection frequency of the test is equal to the nominal level at which we conduct the test. This is a major departure from the squared error bandwidth selection rules that are commonly used, and this departure has two implications. First, h is a function of the nominal level of the test. Therefore, when conducting a test at, say, the .05 level, one would use a different h than when

conducting the same test at the .10 level. Second, it will be seen that h is no longer proportional to $B^{-1/3}$.

There must exist a value of h such that $\mathrm{E}(\hat{\alpha}) = \alpha$ if the test statistic is a pivot. To show this, we demonstrate that a smoothed P value test will overreject when h is sufficiently small and underreject when h is sufficiently large. The continuity in h of the kernel function then ensures that it will reject the correct proportion of the time for some value of h.

The cumulative kernel used for the estimation of a CDF is given by $K(z) = \int_{-\infty}^{z} k(Z)\, dZ$, where $k(Z)$ is a symmetric function that integrates to unity, and where $z = (\hat{\tau} - \tau_j^*)/h$. In practice, $k(Z)$ is often taken to be the standard Gaussian, and $K(z)$ is then the cumulative standard Gaussian. Such kernel functions therefore possess the following properties: i) $K(z) \to 0$ when $z \to -\infty$; ii) $K(z) \to 1$ when $z \to \infty$; and iii) $K(0) = 0.5$.

First, consider the behavior of a right-tailed test based upon $\hat{P}_B^h$ for large h. When $h \to \infty$, $z = (\hat{\tau} - \tau_j^*)/h \to 0$, and therefore $K(z) \to 0.5$. This implies that $\hat{F}_h(\hat{\tau}) \to 0.5$ and $\hat{P}_B^h \to 0.5$. Therefore, $\mathrm{E}\big(I(\hat{P}_B^h < \alpha)\big)$ tends to zero for conventional levels of α. In consequence, tests based on $\hat{P}_B^h$ must underreject when h is sufficiently large.

Next, consider the behavior of the test for small h. As $h \to 0$, $\hat{F}_h(\hat{\tau})$ tends to the EDF, $\hat{F}(\hat{\tau})$. There are three cases, depending on how B is related to α. When neither $\alpha(B+1)$ nor αB is an integer, a test based on the EDF will overreject. However, when $\alpha(B+1)$ is an integer, such a test is exact, and when αB is an integer, it underrejects. Nevertheless, it appears that, for h somewhat larger than 0, the test always overrejects.

Figure 5 shows rejection frequencies as a function of c for bandwidths of the form $h = cB^{-4/9}$ for three values of B, namely, 19, 20, and 21. Once again, the test statistic is distributed as $N(0,1)$. When $B = 19$, the EDF-based test is exact, when $B = 20$ it underrejects, and when $B = 21$ it overrejects. We used $B^{-4/9}$ rather than $B^{-1/3}$ based on extensive simulation evidence. However, we have no theoretical justification for this choice. For all three values of B, it is clear that there exists an optimal bandwidth such that the test rejects precisely 5% of the time.

Ideally, bandwidth selection would be data-driven. As mentioned above, we require a bandwidth selection rule that depends on i) the sample size, ii) the underlying data generating process, and iii) the nominal rejection frequency of the test, α, when the

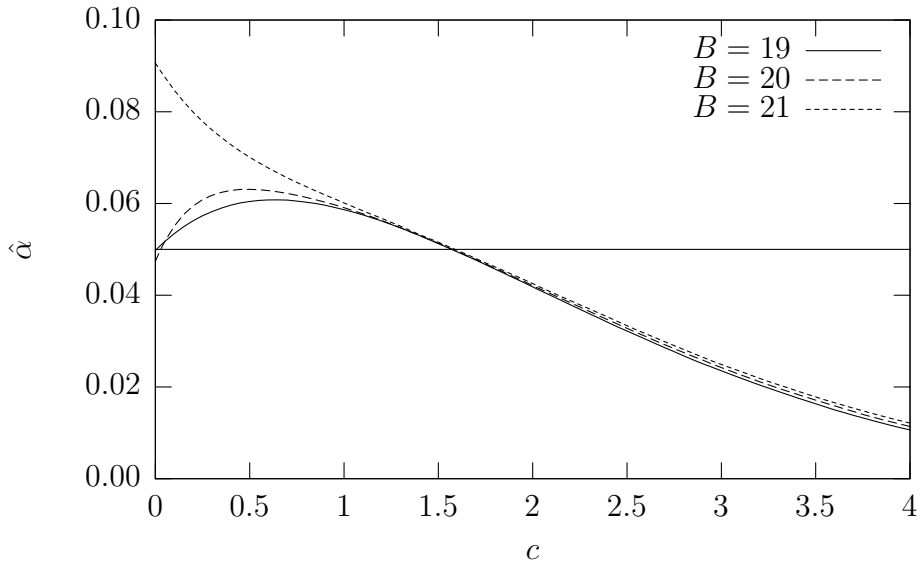

FIGURE 5. Empirical rejection frequencies for tests at the .05 level as a function of the bandwidth.

test statistic is a pivot. Existing CDF bandwidth selection rules such as those of Bowman, Hall, and Prvan [4] and Polansky and Baker [16] satisfy i) and ii) but not iii). The last of these requirements introduces an additional level of complexity. We need to find neither a globally optimal bandwidth nor a pointwise optimal one, but rather one that is tailored to the tail of the distribution, where the tail probability is α. At this point, it is unclear to us how to modify existing methods of data-driven bandwidth selection to incorporate this additional requirement. We remain optimistic that a solution to this problem can be obtained, and we hope that the simulations below may provide inspiration for the interested reader.

3.3. **Finding the optimal bandwidth.** In order to find the optimal bandwidths corresponding to $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$, we performed another set of experiments. Once again, there were two million replications. We also used control variates based on the known distribution of the test statistic to further reduce experimental error; see Davidson and MacKinnon [5]. For each replication and each choice of B, we generated B "bootstrap" statistics. We then chose the bandwidth to minimize the squared difference between the estimated rejection frequency and the nominal level

of the test. The bandwidths we obtained in this way were accurate to at least 5, and usually 6, decimal digits.

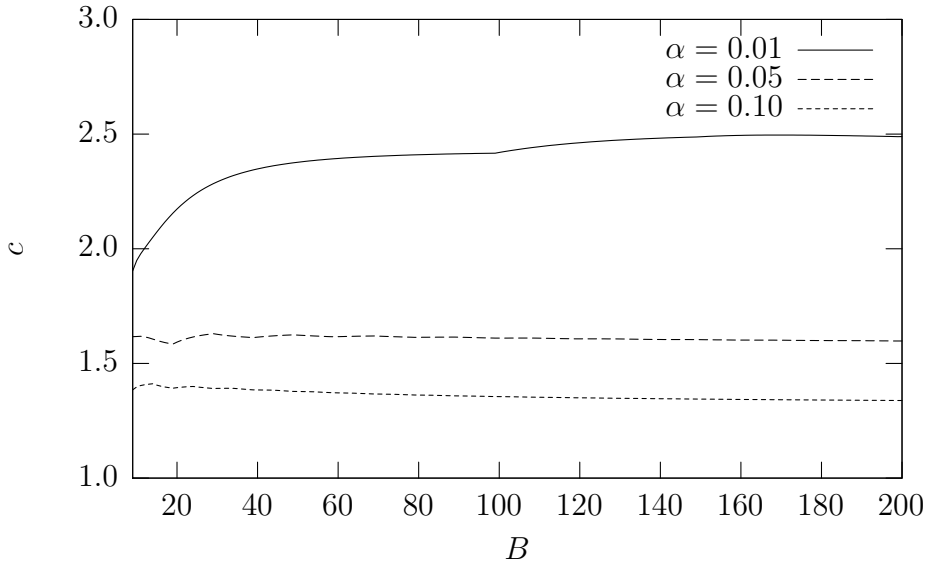

FIGURE 6. Optimal bandwidths for smoothed bootstrap P values ($h = c\hat{\sigma}B^{-4/9}$).

Despite the very large number of replications, the optimal bandwidths, when plotted against B, displayed a good deal of apparently random variation, especially for larger values of B. This reflects the fact that the smoothed bootstrap P values become less sensitive to h as B increases. Therefore, the function that is being minimized to find the optimal value of h becomes less curved near its minimum. We therefore estimated response surface regressions, in which the optimal bandwidths (expressed in terms of the factor c, where $h = cB^{-4/9}$) were regressed on a constant, several powers of $1/B$, and the variable

$$\frac{|B - B'|}{B^2}, \tag{6}$$

where B' is the closest value of B for which $\alpha(B' + 1)$ is an integer. The fitted values from these regressions are shown in Figure 6. It is evident that the optimal value of c is nearly constant for tests at the .05 and .10 levels, but it is not at all constant for tests at the .01 level. As $B \to \infty$, the optimal values of c tend to 2.418, 1.575, and 1.3167 for $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$, respectively.

4. Performance of the Smoothed Bootstrap Approach

4.1. **Size.** Figure 7 presents empirical rejection frequency results for two forms of smoothed bootstrap test at the .05 level. One form uses the fully optimal values of h implied by the values of c that are plotted in Figure 6 for $\alpha = .05$. A second, simpler, form uses $h = 1.575B^{-4/9}$; recall that 1.575 is the value to which c tends as $B \to \infty$. It can be seen that the smoothed bootstrap approach works perfectly, except for experimental error, when optimal bandwidths are used, and it works almost perfectly when $h = 1.575B^{-4/9}$.

Of course, one cannot expect to obtain quite such superlative results in practice. The optimal bandwidths inevitably depend on the actual distribution of the test statistic. The ones that we have used are based on the standard normal distribution. When the actual distribution differs substantially from the latter, the optimal bandwidths might be quite different. In experiments that are not reported here, we obtained rejection frequencies as high as .0596 for $B = 10$ with $h = 1.575B^{-4/9}$ when the test statistic actually followed the $\chi^2(2)$ distribution. But these rejection frequencies declined towards .05 fairly rapidly as B increased. Moreover, if an investigator knew that a test statistic was asymptotically $\chi^2(2)$, he or she could choose an optimal bandwidth based on that distribution instead of one based on the standard normal distribution. Of course, if one wishes to avoid overrejection, then the safest course is to err in the direction of making h too large. This is evident from Figure 5.

4.2. **Power.** The main reason to use the smoothed bootstrap approach is that it should yield higher power when B is small. Our experimental results strongly support this conjecture. Figure 8 shows power loss for bootstrap tests at the .05 level as a function of δ, the parameter that is zero under the null. There is a substantial gain from using the smoothed bootstrap approach when $B = 19$ and a modest gain when $B = 99$. For tests at the .01 level (not shown in the figure), there is a substantial gain when $B = 99$. Smoothed bootstrap tests using $B = 50$, a value that is too small to use for EDF-based tests, are somewhat more powerful than the latter with $B = 99$.

5. Summary

We propose a new method for the nonparametric estimation of bootstrap P values which are then used to perform a classical test. The proposed approach has a number

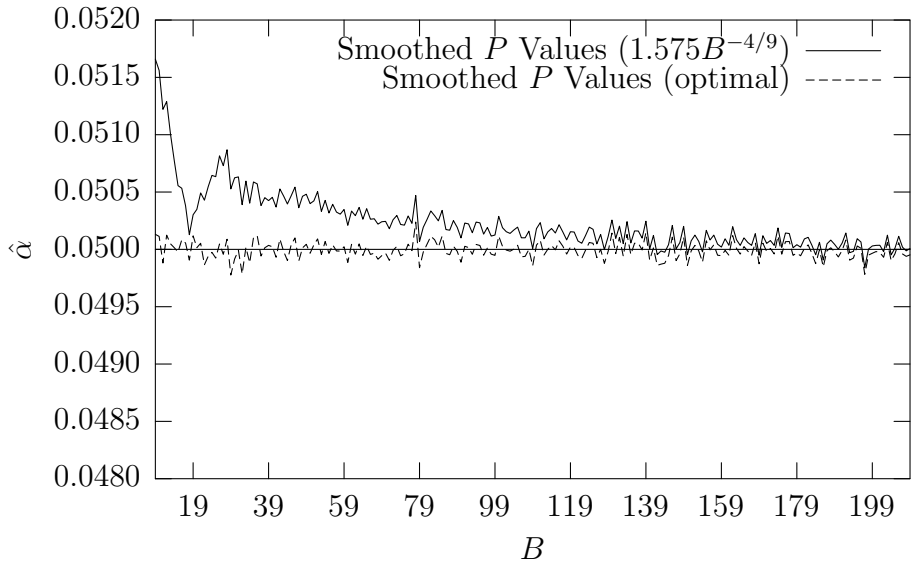

FIGURE 7. Empirical rejection frequencies for two forms of smoothed bootstrap test at the .05 level.

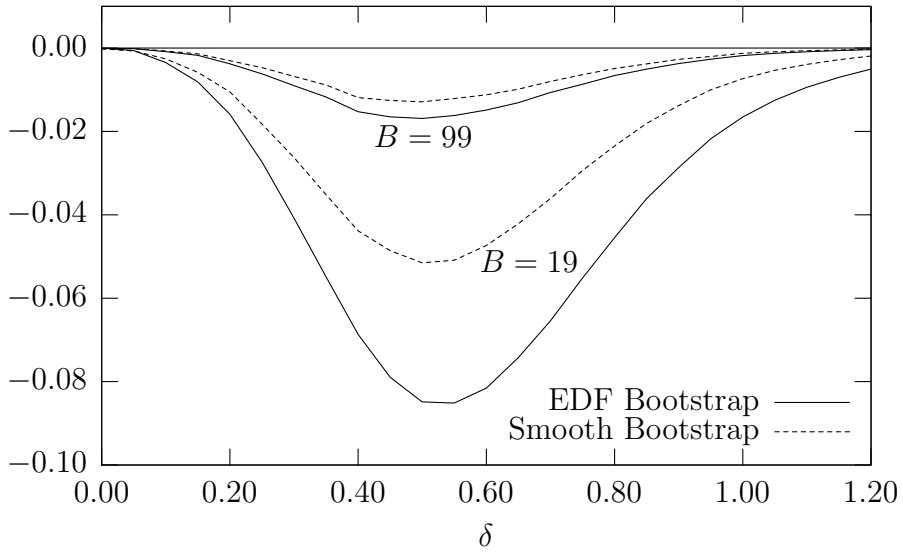

FIGURE 8. Power loss for the conventional EDF and proposed tests at the .05 level.

of advantages relative to the usual procedure of constructing a bootstrap P value based on the EDF of the bootstrap statistics. First, it can validly be used for any

number of bootstrap replications. There is no restriction that $\alpha(B+1)$ must be an integer. When calculating the test statistic is very expensive, this restriction can be problematical, especially for tests at the .01 level. Second, because it uses the information in the bootstrap statistics more efficiently than the conventional approach based on the EDF, it yields noticeable gains in power relative to the latter, especially for tests at the .01 level. Finally, it can yield quite accurate tests even when B is very small. However, these tests will rarely be exact, even if the underlying test statistic is pivotal. We do not regard this as a major drawback, because test statistics that are extremely expensive to compute are almost never pivotal.

For the small numbers of bootstrap repetitions for which the proposed approach is intended, existing bandwidth selection rules may be perfectly acceptable. MSE optimal rules seem to work better the IMSE optimal ones, and we have proposed distribution-dependent rules that work even better in some cases. The smoothed bootstrap tests yield reasonable size and deliver solid improvements in power, particularly when compared with the size distortions and power loss associated with the conventional approach. Ideally, we would like to find optimal data-driven bandwidth selection rules capable of yielding exact tests when the underlying test statistic is pivotal. This remains a subject for future research.

The procedure we propose is easy to use. Once the bootstrap sampling has been completed, it can be applied in a matter of microseconds on a modern desktop computer. Code written in the R language (Ihaka and Gentleman [12]) based upon existing data-driven bandwidth selection rules is available from the authors upon request.

References

1. A. Azzalini, *A note on the estimation of a distribution function and quantiles by a kernel method*, Biometrika **68** (1981), no. 1, 326–8.
2. R. Beran, *Prepivoting test statistics: A bootstrap view of asymptotic refinements*, Journal of the American Statistical Association **83** (1988), 687–697.
3. P.J. Bickel and D.A Freedman, *Bootstrapping regression models with many parameters*, A Festschrift for Erich Lehmann (Belmont, California), Wadsworth, 1983, pp. 24–48.
4. Adrian Bowman, Peter Hall, and Tania Prvan, *Bandwidth selection for the smoothing of distribution functions*, Biometrika **85** (1998), 799–808.
5. R. Davidson and J. G. MacKinnon, *Regression-based methods for using control variates in Monte Carlo experiments*, Journal of Econometrics **54** (1992), 203–222.
6. _____, *Bootstrap tests: How many bootstraps?*, Econometric Reviews **19** (2000), 55–68.
7. B. Efron and R.J. Tibshirani, *An introduction to the bootstrap*, Chapman and Hall, New York, London, 1993.
8. C. Gouréroux and A. Monfort, *Simulation-based econometric methods*, Oxford University Press, Oxford, 1997.
9. P. Hall, *On the bootstrap and confidence intervals*, The Annals of Statistics **14** (1986), 1431–1452.
10. _____, *The bootstrap and edgeworth expansion*, Springer Series in Statistics, Springer-Verlag, New York, 1992.
11. P. Hall and D.M. Titterington, *The effect of simulation order on level accuracy and power of Monte Carlo tests*, Journal of the Royal Statistical Society **B 51** (1989), 459–467.
12. Ross Ihaka and Robert Gentleman, *R: A language for data analysis and graphics*, Journal of Computational and Graphical Statistics **5** (1996), no. 3, 299–314.
13. K.-H. Jöckel, *Finite sample properties and asymptotic efficiency of Monte Carlo tests*, Annals of Statistics **14** (1986), 336–347.
14. E. Mammen, *When does bootstrap work? asymptotic results and simulations*, Springer-Verlag, New York, 1992.
15. E. A. Nadaraya, *Some new estimates for distribution functions*, Theory of Probability and its Applications **9** (1964), 497–500.
16. A. M. Polansky and E. R. Baker, *Multistage plug-in bandwidth selection for kernel distribution function estimates*, Journal of Statistical Computation and Simulation **65** (2000), 63–80.
17. J. Racine and J. G. MacKinnon, *Simulation-based tests that can use any number of simulations*, Communications in Statistics: Simulation and Computation **36** (2007), forthcoming.
18. R. D. Reiss, *Nonparametric estimation of smooth distribution functions*, Scandinavian Journal of Statistics **9** (1981), 65–78.
19. P. Sarda, *Smoothing parameter selection for smooth distribution functions*, Journal of Statistical Planning and Inference **35** (1993), 65–75.
20. H. van Dijk and A. Monfort, *Econometric inference using simulation techniques*, John Wiley and Sons, Chichester, 1995.

Jeff Racine, Department of Economics, McMaster University, Hamilton, Ontario, Canada, L8S 4L8

James G. MacKinnon, Department of Economics, Queen's University, Kingston, Ontario, Canada, K7L 3N6