ON A BIDIRECTIONAL TEST FOR STOCHASTIC DOMINANCE

CHRISTOPHER J. BENNETT †

Abstract. This paper introduces a sequential bidirectional test for stochastic dominance. In the first stage, we posit a null of equality between the distributions under study. Failure to reject the null of equality suggests that no statistically significant difference exists between the distributions and further testing is not pursued. If equality of the distributions is rejected, then a second test is performed based on the minimum of one-sided tests for dominance. Rejection of the initial test combined with acceptance/rejection of the second test and the relationship between the individual one-sided statistics allow us to distinguish between the various alternatives, namely dominance or crossing of the distributions. When the underlying CDFs under test are continuous, the asymptotic null distributions associated with our sequential procedure are distribution-free and we are thus able to provide tabulated critical values. For use more generally we also introduce a novel bootstrap procedure that requires only a single set of bootstrap samples. We show that the bootstrap test delivers greater power than the asymptotic test while also maintaining tighter control over the asymptotic misclassification rates.

Keywords: stochastic dominance, conditional test, bootstrap

JEL Classifications: C12, C14, C52

Revised: September 9, 2009

---

†Department of Economics, Vanderbilt University, Nasvhille TN. 37240

Email: chris.bennett@vanderbilt.edu

Web Address: http://people.vanderbilt.edu/ chris.bennett/.

1. Introduction

Let F and G denote cumulative distribution functions. We say that F dominates G stochastically at first-order and write $F \succeq G$ if $F(t) \leq G(t)$ for every t. The partial ordering on the space of distributions induced by the stochastic dominance criterion, and its strong implications in various subfields of economics are well studied. In the context of comparing income distributions, for example, dominance of F over G implies that there is greater poverty in G according to any poverty index that is symmetric and monotonically decreasing in income (Foster, Greer and Thorbecke (1984), Foster and Shorrocks (1988)). On the other hand, when F and G correspond to financial return distributions, dominance of F over G is consistent with F being strictly preferred to G by any expected-utility maximizer whose utility is strictly increasing in wealth; see, e.g., Levy (1992).

Despite their potential for delivering powerful unambiguous statements, orderings based on stochastic dominance suffer from the practical limitation that population CDFs are generally unobserved and thus inferences must be drawn from comparisons based solely on their empirical counterparts. Not surprisingly, an extensive literature on statistical inference for the SD partial ordering has therefore emerged. Anderson (1996) and Davidson and Duclos (2000), for instance, propose to test the null of weak dominance, i.e. $F \succeq G$, against the alternative of nondominance. Their statistics are based on a comparison of the CDFs (or functionals of the CDFs in the case second- and third-order dominance) at a finite number of points in the support. Consequently, their tests may lack the desirable property of consistency. Building on the earlier work of McFadden (1989), Barrett and Donald (2003) introduce a Kolmogorov-Smirnov type test for stochastic dominance, where again dominance is posited under the null with nondominance under the alternative. These procedures, and the extension in Bennett (2008) to tests based on weighted Cramér-von Mises type statistics, test the complete set of restrictions implied by stochastic dominance and consequently have the advantage of being consistent against the full range of alternatives. As noted by Davidson and Duclos (2006), however, rejection of a null of dominance is inconclusive in the sense that it fails to rank the two populations; moreover, non-rejection of dominance does not enable one to accept dominance as this result may simply be the product of having insufficient data.

In contrast to the above procedures, Kaur, Rao and Singh (1994) develop a test of whether $F \succeq G$ based on the minimum t statistic for the hypothesis $F = G$, whereby a positive statistically significant value of the minimum is taken as sufficient evidence against the null in the direction of the alternative of dominance. Davidson and Duclos (2006) also devise a test with nondominance under the null, albeit through a novel use of the empirical likelihood ratio. While these tests lead to an inference of dominance upon rejection of the null, an unfortunate drawback is that they apply only to the interior of the support (i.e., they test for restricted dominance) and will infer dominance only if the distribution F lies everywhere below G over the interior of the support. Even when F and G touch only at a single point in the interior and F is otherwise strictly below G, these tests will not identify dominance.

In related work, though in the context of testing for Lorenz dominance, Dardanoni and Forcina (1999) and Bhattacharya (2007) propose the use of sequential testing procedures. Their proposals consist of a first stage in which they test the null of $F \succeq_L G$ against an unrestricted alternative. Thus, only in the event of non-rejection do the authors pursue a second test; namely, testing the null of equality against that of strict dominance, i.e. $F \succ_L G$. Note that rejection of the null in the first stage is again inconclusive regarding the nature of rejection—do the distributions cross or is F dominated?

By construction, users of the above tests are presumed to have an *a priori* belief about the direction of dominance. In practice, however, this more likely to be the exception rather than the rule. In many situations, investigators may wish to rely solely on the data to distinguish between multiple alternatives, namely crossing or dominance in either direction, without advancing any directional hypothesis.

In this paper we depart from this literature and instead propose a two-stage bidirectional testing procedure for distinguishing between these multiple alternatives. In the first stage we posit a null of equality between the two distributions under consideration and construct a test based on the maximum of two one-sided Kolmogorov-Smirnov-type statistics. Failure to reject the null hypothesis of equality suggests that the distributions are statistically indistinguishable and further testing is not pursued. If, on the other hand, we reject the null of equality and we wish to infer the nature of this difference, then a second test is performed in which we consider the minimum of the same one-sided statistics. Rejection occurs if we observe a "large" value for the minimum of the statistics, and an inference of crossing is reached. Otherwise, the signs of the individual statistics can be used to infer the direction of dominance. Thus, our procedure, like that of Bishop, Formby and Thistle (1992), begins with equality and seeks to infer the nature of the difference between the distributions only when equality is rejected the data. Unlike their procedure, however, we construct a consistent and non-conservative test that properly accounts for the possibility of a Type I error being made in the first-stage. Note also that our second stage test is akin to the procedures of Linton, Maasoumi and Whang (2005) and Linton, Song and Whang (2008) who build on the work of Klecan, McFadden and McFadden (1991) to propose a test with dominance under the null, but with crossing of the statistical functionals as the alternative. On its own, their test does lead to a conclusive statement about the distributions under consideration, namely crossing, when the null is rejected; however the inferences which one may draw from non-rejection are still subject to the same limitations as those tests described above. By embedding their test in a two stage procedure, we are better able to discern the nature of non-rejection by coupling this with the information afforded to us in the first-stage of our sequential test.

An important and delicate consideration in the design of our test is appropriate error rate control. Since we are considering a sequential testing procedure, control of the error rates is complicated by the conditional nature of the second-stage test. In order to solve this problem we introduce both asymptotic and bootstrap procedures which, in the case of equal distributions, maintain asymptotic control of the Type-I error rate of the second-stage test conditional upon making a Type-I error in the first-stage. When the two distributions under test are unequal, the null of equality is rejected at the first stage with probability one in large samples and asymptotic control of the Type-I error rate of the second-stage test is still maintained. Overall, our procedures possess a remarkable ability to deliver explicit asymptotic control of the various misclassification error rates. Additionally, since our proposed procedures take into account the full set of restrictions implied by stochastic dominance we will correctly identify crossing with probability one in large samples whenever crossing exists in the population.

While our asymptotic test does have the advantage of delivering tabulated critical values, its asymptotic validity hinges critically on the assumption of continuity of the underlying CDFs. On the contrary, our bootstrap procedure is valid more generally—e.g., without the assumption of continuity—and is also shown to lead to tighter control over the misclassification error rates. Further, the power of our test to detect dominance (or crossing) depends on the power of the first-stage test to correctly reject the null of homogeneity or equality of the two distributions under consideration. For this classical problem in statistics we prove that our new bootstrap test weakly dominates existing procedures. This gain is power is achieved by deviating from the standard convention of imposing the null on the bootstrap generating process, and instead we allow for the bootstrap data generating process (DGP) to better reflect the true DGP. Asymptotically, our critical values will converge to those obtained by imposing the null on the bootstrap DGP *when the null is true*, otherwise our critical values will be strictly smaller and our test more powerful.

The remainder of the paper is structured as follows. In the next section we discuss the hypotheses of interest and outline our proposed testing procedure. Section 3 introduces the bootstrap procedure. Section 4 contains results on the finite sample performance. Section 5 concludes. Note that our focus throughout the paper is exclusively on testing for first-order stochastic dominance. Our limited focus here is merely for expositional simplicity and in no way should reflect upon the ability to generalize our procedure to tests of other partial orders including, for example, higher orders of stochastic dominance and Lorenz dominance. Indeed, given existing results on the processes involved in such tests, these types of extensions should prove to be rather straightforward.

2. Hypotheses

Let F and G denote cumulative distribution functions (cdfs) with common support on the non-negative reals. Our objective in this paper is to develop a test of the hypothesis

$$H_0 : F = G$$

against the multiple alternatives

$$H_{A1} : F \succ G$$
$$H_{A2} : G \succ F$$
$$H_{A3} : F \not\succ G \text{ and } G \not\succ F$$

where $\succ$ denotes the strict dominance relation and $\not\succ$ its negation. To aid in the development of our testing procedure we introduce the following functionals on the space $D[0, \infty)$ of cadlag functions on $[0, \infty)$

$$\Gamma_1(\nu) = \sup_t |\nu(t)|,$$
$$\Gamma_2(\nu) = \min\{\sup_t \nu(t), -\inf_t \nu(t)\},$$

and

$$\Gamma_+(\nu) = \sup_t \nu(t),$$

as well as the corresponding parameters

$$\gamma_1 = \Gamma_1(\Delta), \quad (1)$$

and

$$\gamma_2 = \Gamma_2(\Delta), \quad (2)$$

where $\Delta = F - G$. Note that $\gamma_1 \geq \gamma_2$, with equality holding only if $F = G$. Additionally, we define the parameters $\theta_1 = \Gamma_+(\Delta)$ and $\theta_2 = \Gamma_+(-\Delta)$. Given these parameter definitions, the hypothesis of interest may now be rewritten as

$$H_0 : \gamma_1 = 0$$

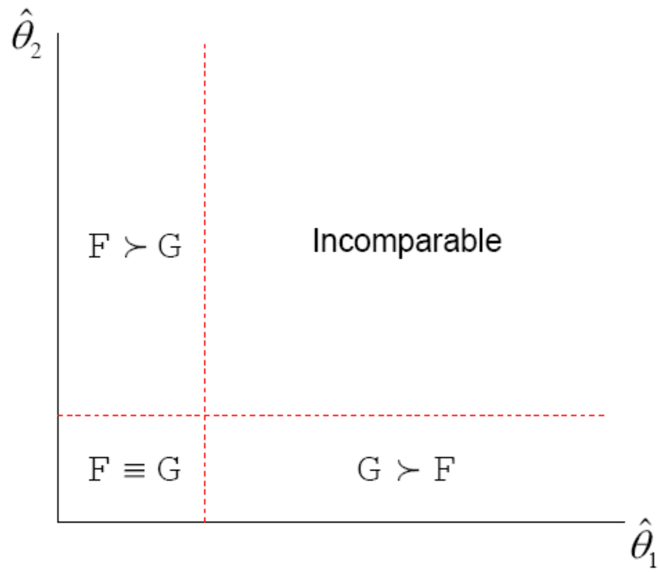


FIGURE 1. Decision Framework

against

$$
\begin{aligned}
H_{A1} :& \gamma_1 > 0,\ \gamma_2 \leq 0, \text{ and } \theta_2 > \theta_1 \\
H_{A2} :& \gamma_1 > 0,\ \gamma_2 \leq 0, \text{ and } \theta_1 > \theta_2 \\
H_{A3} :& \min\{\gamma_1, \gamma_2\} > 0
\end{aligned}
\tag{3}
$$

Note that the distinction between the three alternatives depends only on γ_2 and the sign of $\theta_1 - \theta_2$; moreover, the value of γ_2 is of interest only after rejection of the null $\gamma_1 = 0$. This suggests a natural progression in the testing procedure; namely, we first test the null hypothesis $\gamma_1 = 0$ and then, only upon rejection do we proceed to test the second hypothesis. To be concrete, let $\hat{\gamma}_1$ and $\hat{\gamma}_2$ denote suitably scaled consistent estimators of γ_1 and γ_2, respectively. Further, let c_1 and c_2 denote "critical values" associated with the first and second test statistics (our choice of statistics as well as appropriate methods for determining c_1 and c_2 are discussed in the next section). The decision rule for the two-stage procedure may then be described as follows:

(1) If $\hat{\gamma}_1 \leq c_1$ we fail to reject H_0
(2) If $\hat{\gamma}_1 > c_1$ then we reject H_0 and
   (a) Infer $F \succ G$ if $\hat{\gamma}_2 \leq c_2$ and $\hat{\theta}_1 - \hat{\theta}_2 < 0$
   (b) Infer $G \succ F$ if $\hat{\gamma}_2 \leq c_2$ and $\hat{\theta}_1 - \hat{\theta}_2 > 0$
   (c) Infer that the distributions cross if $\hat{\gamma}_2 > c_2$

The basic intuition for the testing procedure is illustrated in Figure 1, where c denotes a "critical value" which determines the position of the dashed lines and, for simplicity, is assumed to be common to both tests; i.e. $c_1 = c_2 = c$. The set $\mathcal{A} = \{(x, y) : 0 \leq x \leq c, 0 \leq y \leq c\}$ forms the acceptance region of the null of equality and for which clearly $(x, y) \in \mathcal{A}$ if and only if $\max\{x, y\} \leq c$. According to the structure of the test, observing $\max\{\hat{\theta}_1, \hat{\theta}_2\} > c$ would lead us to infer only that $(\theta_1, \theta_2) \in \mathbb{R}^2_+ \backslash \mathcal{A}$, but otherwise would not generally be informative as to the nature of our rejection of the null of equality. If, however, the values of the statistics are such that their minimum is also greater than the critical value, then we may infer that the distributions cross. Dominance may be inferred in the intermediate cases where the maximum is significant but the minimum is not. For example, observing $\hat{\gamma}_1 > c_1$, $\hat{\gamma}_2 \leq c_2$, and $\hat{\theta}_2 > \hat{\theta}_1$ would lead us to infer that F stochastically dominates G.

In the design of our testing procedure we naturally seek to control the "size" of the overall test when equality holds between the two distributions. That is, we would like the ability to select (or estimate) the values of c_1 and c_2 for which

$$\lim_{n \to \infty} P\{\hat{\gamma}_1 > c_1 | H_0 \text{ True }\} = \alpha_1 \tag{4}$$

and

$$\lim_{n \to \infty} P\{\hat{\gamma}_2 > c_2, \hat{\gamma}_1 > c_1 | H_0 \text{ True}\} = \alpha_2 \alpha_1 \tag{5}$$

for given nominal levels α_1 and α_2. Any test for which both (4) and (5) hold will be said to be (asymptotically) correctly sized. In the next section, we discuss our choice of test statistics and introduce both asymptotic and bootstrap strategies for obtaining asymptotically valid critical values in the sense of (4) and (5).

3. Test Statistics and Asymptotics

Let $\{X_i, 1 \leq i \leq n\}$ and $\{Y_i, 1 \leq i \leq m\}$ be independent random samples generated according to the (univariate) distributions F and G, respectively; and define the marginal empirical distributions F_n and G_m in the usual way. Further, define $\Delta_{m,n} = F_n - G_m$. We will require the following assumption concerning the sampling process:

Assumption 3.1. *The sampling process is such that* $\lambda_N = \frac{m}{m+n} \to \lambda \in (0, 1)$.

Remark 3.1. *Assumption 3.1 essentially requires that the sample sizes m and n grow at the same rate.*

In place of $\hat{\gamma}_1$ and $\hat{\gamma}_2$, we take as our first and second-stage test statistics

$$T_{m,n,1} = \sqrt{\frac{mn}{m+n}} \Gamma_1(\Delta_{m,n}), \tag{6}$$

and

$$T_{m,n,2} = \sqrt{\frac{mn}{m+n}} \Gamma_2(\Delta_{m,n}). \tag{7}$$

Both $T_{m,n,1}$ and $T_{m,n,2}$ are familiar statistical objects: $T_{m,n,1}$ is nothing other than the two-sample Kolmogorov-Smirnov statistic—see, e.g., Stephens (1986)—whereas (a generalized version of) $T_{m,n,2}$ is employed by Klecan et al. (1991), Maasoumi and Heshmati (2000), Linton et al. (2005), and Linton et al. (2008) to test for maximality of a given collection of prospects. A similar statistic to $T_{m,n,2}$ also appears in Schechtman, Shelef, Yitzhaki and Zitikis (2008) where it is used to develop tests concerning absolute concentration curves.

It is well known, under H_0 and Assumption 3.1, that $T_{m,n,1}$ and $T_{m,n,2}$ converge weakly to $\Gamma_1(\mathcal{B}(F(t)))$ and $\Gamma_2(\mathcal{B}(F(t)))$, respectively, where $\mathcal{B}(\cdot)$ denotes the standard Brownian bridge process. Furthermore, under the additional assumption of continuity of the CDFs, the limiting distribution of $T_{m,n,1}$ may be shown to be distribution free; i.e., independent of F. A straightforward procedure for obtaining a valid α_1-level critical value for the first stage test in this case would therefore be to use the $(1-\alpha_1)$th quantile of the well-known Kolmogorov-Smirnov distribution (see p.143 of Shorack and Wellner (1986)).

What is perhaps surprising is that the structure of our test also enables us to obtain a closed-form expression for the asymptotic conditional distribution of $T_{m,n,2}$ under H_0 and continuity of the CDFs which—like the asymptotic distribution of $T_{m,n,1}$—is also independent of the underlying distribution F. This result is the content of Proposition 3.1 below.

Proposition 3.1. *Suppose H_0 is true and that F is continuous. Then, for $a \leq b$,*

$$P[\Gamma_2(\nu_{m,n}) \leq a | \Gamma_1(\nu_{m,n}) > b] \to \frac{2[G_1(a) - G(a,b)]}{1 - G_2(b)} \tag{8}$$

where

$$G_1(a) = 1 - e^{-2a^2},$$

$$G_2(b) = 1 + 2\sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 b^2},$$

and

$$G(a,b) = \sum_{k=-\infty}^{\infty} \exp(-2k^2(a+b)^2) - \sum_{k=-\infty}^{\infty} \exp(-2[b + k(a+b)]^2).$$

Proposition 3.1 together with the well-known asymptotic distribution of the Kolmogorov-Smirnov statistic may be used to derive asymptotically valid critical values in the sense of equations (4) and (5). Indeed, the critical value $c_1(\alpha_1)$ may be obtained directly from tabled probabilities associated with the distribution of $\Gamma_1(\mathcal{B})$ (Shorack and Wellner 1986, p. 143) whereas the value of $c_2(\alpha_2)$ may then be obtained from Proposition 3.1 for a fixed value of $c_1(\alpha_1)$. Approximate $c_2(\alpha_2)$ for a select combination of (α_1, α_2) pairs are recorded in Table 3. Hereafter, we will refer to an implementation of our test based on these critical value as an asymptotic test.

In contrast to the asymptotic test, we may also consider a resampling-based implementation, and indeed there are several reasons for doing so. First, the validity

Critical Values for the Asymptotic Test

		α_2				
		0.01	0.02	0.03	0.04	0.05
	0.01	(1.64, 0.66)	(1.64, 0.59)	(1.64, 0.54)	(1.64, 0.51)	(1.64, 0.48)
α_1	0.05	(1.36, 0.75)	(1.36, 0.67)	(1.36, 0.62)	(1.36, 0.58)	(1.36, 0.55)
	0.10	(1.22, 0.80)	(1.22, 0.71)	(1.22, 0.66)	(1.22, 0.63)	(1.22, 0.60)

TABLE 1. Approximate critical values (c_1, c_2) for the asymptotic test with nominal sizes α_1 and α_2. For fixed values of $c_1(\alpha_1)$, the values of $c_2(\alpha_2)$ are approximated via Monte Carlo simulation .

of the asymptotic test hinges on the additional assumption of continuity of the distribution functions which is necessary for the asymptotic distribution free property. On the other hand, the bootstrap can be used to deliver asymptotically valid critical values irrespective of whether the asymptotic distribution depends on the underlying CDFs under test. Thus, a bidirectional bootstrap test for first-order dominance is not only valid under more general conditions but it may also be adapted for use in testing dominance beyond first-order where the asymptotic distribution free property fails to hold. Second, as remarked earlier, the ability to correctly infer dominance or crossing hinges critically on the power of the first-stage test. We show that through the use of the bootstrap we may improve upon the power of the asymptotic test—or standard (e.g. fully recentered or pooled) bootstrap procedures that essentially mimic the asymptotic test—by employing a simple modification of the standard bootstrap recentering scheme.

For the purpose of introducing our bootstrap procedure, let F_n^* and G_m^* denote the bootstrap empirical distributions based on random samples from F_n and G_m, respectively; and let $\delta_{m,n}(t)$ denote a positive (possibly stochastic) array of the form $\{\delta_{m(n),n}(t) : n \geq 1, 1 \leq t \leq m(n)+n\}$ which is chosen in accordance with Conditions 3.1 i. and ii. below.

Condition 3.1. *$\delta_{m,n}(t)$ satisfies*

i. $\forall \epsilon > 0 \ P\Big[\sup_t \big[\big(\frac{mn}{m+n}\big)^{-1/2}\delta_{m,n}(t)\big] \geq \epsilon\Big] \to 0$ *as* $n \to \infty$

ii. $P\Big[\sup_t \big|\frac{\nu_{m,n}(t)}{\delta_{m,n}(t)}\big| \leq 1\Big] \to 1$ *as* $n \to \infty$

where $\nu_{m,n}(t) = \sqrt{\frac{mn}{m+n}}(\Delta_{m,n} - \Delta)$.

Remark 3.2. *Examples of arrays satisfying Conditions i. and ii. above include* $\delta_{m,n}(t) = \sqrt{2\log\log\big(\frac{mn}{m+n}\big)}W_\epsilon(t)$, *where* $W_\epsilon(t) = \big[H(t)(1-H(t))\big]^\epsilon$ *for* $0 \leq \epsilon < 1/2$ *with* $H = \lambda F + (1-\lambda)G$*; see, e.g., James (1975).*

Given an array satisfying Condition 3.1, our proposal in this paper is to estimate $c_1(\alpha_1)$ based on the $1-\alpha_1$ quantile of

$$T^*_{m,n,1} = \sqrt{\frac{mn}{m+n}}\Gamma_1(\tilde{\Delta}^*_{m,n}) \tag{9}$$

where

$$\tilde{\Delta}^*_{m,n} = (\Delta^*_{m,n} - \Delta_{m,n})C_{m,n},$$
$$C_{m,n}(t) = \mathbb{1}\Big\{\sqrt{\frac{mn}{m+n}}|\Delta_{m,n}(t)| \le \delta_{m,n}(t)\Big\},$$

and $\mathbb{1}\{\}$ denotes the indicator function which takes on the value one if the argument in the parentheses is true, and zero otherwise. We denote the critical value obtained via (9) by $\hat{c}_1(\alpha_1)$. Notice here that unlike the standard two-sample bootstrap procedure which is based on

$$T^{FC}_{m,n,1} = \sqrt{\frac{mn}{m+n}}\Gamma_1(\Delta^*_{m,n} - \Delta_{m,n}), \tag{10}$$

we do not impose the null of equality between F and G. This technically subtle—yet methodologically significant—departure from conventional tests of equality has a dramatic impact on the power of the test. Upon comparing the two bootstrap procedures, for example, it is immediate that $\hat{c}_1(\alpha_1)$ is no bigger (and generally strictly smaller) than the $1-\alpha_1$ quantile (10). Consequently, the critical values based on (9) generally result in a more powerful omnibus test for equality of two distributions without sacrificing asymptotic control of the size of the test.

As for the second stage of our test, in which we wish to determine the nature of a rejection that occurs in the first stage, the situation is more delicate. In particular, we could arrive at the second stage by falsely identifying the distributions under test as being unequal, and hence we need to condition the second stage test on making a Type I error in the first stage. In order to maintain proper error control at level α_2 in the second stage of our testing procedure we propose using the critical value

$$\hat{c}_2(\alpha_2) = \inf_c\Big\{c : \frac{P\left[T^*_{m,n,2} \le c, T^{**}_{m,n,1} > \hat{c}_1(\alpha_1)\right]}{P\Big[T^{**}_{m,n,1} > \hat{c}_1(\alpha_1)\Big]]} \ge (1-\alpha_2)\Big\} \tag{11}$$

where $T^{**}_{m,n,1}$ is the uncentered bootstrap statistic

$$T^{**}_{m,n,1} = \sqrt{\frac{mn}{m+n}}\Gamma_1(\Delta^*_{m,n});$$

and where the bootstrap statistic $T^*_{m,n,2}$ is defined by

$$\begin{aligned} T^*_{m,n,2} = \sqrt{\frac{mn}{m+n}}\Big\{&\Gamma_2(\tilde{\Delta}^*_{m,n})\big[\mathbb{1}\{a_{m,n} > 0, b_{m,n} > 0\} + \mathbb{1}\{a_{m,n} < 0, b_{m,n} < 0\}\big] \\ &+ \Gamma_+(\tilde{\Delta}^*_{m,n})\mathbb{1}\{a_{m,n} < 0, b_{m,n} > 0\} + \Gamma_+(-\tilde{\Delta}^*_{m,n})\mathbb{1}\{a_{m,n} > 0, b_{m,n} < 0\}\Big\}, \end{aligned} \tag{12}$$

with $\Gamma_+(\nu) = \sup_{0\le t\le 1} \nu(t)$, and the stochastic sequences $a_{m,n}$ and $b_{m,n}$ being defined by

$$a_{m,n} = \sup_t \Big(\sqrt{\frac{mn}{m+n}} \Delta_{m,n} - \delta_{m,n}(t) \Big), \tag{13}$$

and

$$b_{m,n} = \sup_t \Big(-\sqrt{\frac{mn}{m+n}} \Delta_{m,n} - \delta_{m,n}(t) \Big). \tag{14}$$

Our proposed bootstrap as outlined through equations (12)– (14) combines several innovative features. First, notice that we have conditioned the bootstrap in (12) on the observed behavior of F and G: if $\sqrt{\frac{mn}{m+n}}\Delta_{m,n}$ is contained in the band $(-\delta_{m,n}(t), -\delta_{m,n}(t))$ for all t or crosses above *and* below the band, then $T^*_{m,n,2}$ is based on $\sqrt{\frac{mn}{m+n}}\Gamma_2(\tilde{\Delta}^*_{m,n})$. Otherwise, the bootstrap is akin to that which used in a one-sided test for dominance. An examination of the limiting distribution of $T_{m,n,2}$ (see Lemma A.5) reveals that our adaptive bootstrap design correctly mimics the limiting distribution under a wider set of configurations of F and G. As we show in the next section, the adaptive design endows our two-stage procedure with asymptotic power at least that of the nominal level of the second-stage test against Pitman alternatives converging to any configuration where $F \neq G$. This would not be true if we instead elected to employ $T^*_{m,n,2} = \sqrt{\frac{mn}{m+n}}\Gamma_2(\tilde{\Delta}^*_{m,n})$. In this case, the bootstrap test may have asymptotic power less than the nominal level of the second-stage test even for some configurations where $F \le G$ or $G \le F$ and $F \neq G$.

It is also noteworthy that we have replaced $T^*_{m,n,1}$ with the uncentered bootstrap statistic $T^{**}_{m,n1}$ in equation (11). The intuition for this choice is rather simple: when $F = G$, the difference (for large n) between $T^{**}_{m,n,1}$ and $T^{FC}_{m,n,1}$ or $T^*_{m,n,1}$ is negligible, thus enabling preservation of the correct size, at least asymptotically. When $F \neq G$ though, $T^{**}_{m,n1}$ is stochastically larger than $T^*_{m,n,1}$ from which it follows that $P(T^{**}_{m,n1} > \hat{c}(\alpha_1)) > \alpha_1$ for large n. In fact, when $F \neq G$, $P(T^{**}_{m,n1} > \hat{c}(\alpha_1)) \to 1$. Upon inspecting the form of (11), we see that this change has the effect of lowering the critical value of—or, equivalently, delivering greater power to— the second stage test without disrupting asymptotic control of the misclassification error rates.

Finally, we point out that the estimation of $c_1(\alpha_1)$ and $c_2(\alpha_2)$ via our bootstrap procedure is remarkably simple since we require only a single set of bootstrap samples, and for each bootstrap sample we can simultaneously compute the bootstrap observations $(T^*_{m,n,1}, T^{**}_{m,n,1}, T^*_{m,n,2})$. To see this, consider the following basic algorithm which may be used to implement our bootstrap test:[1]

Algorithm 3.1 (Bootstrap Algorithm)**.**

[1]An implementation of the test written in the R statistical programming language is available at http://people.vanderbilt.edu/~chris.bennett/research.htm

Misclassification Rates for the Asymptotic Test

	Inference			
Population	F=G	F≻G	G≻F	Cross
F=G	$1-\alpha_1$	$\frac{\alpha_1(1-\alpha 2)}{2}$	$\frac{\alpha_1(1-\alpha 2)}{2}$	$\alpha_1\alpha_2$
F≻G	0	$F_{\Gamma_+(\tilde{\nu})}(c_2(\alpha_2))$	0	$1-F_{\Gamma_+(\tilde{\nu})}(c_2(\alpha_2))$
G≻F	0	0	$F_{\Gamma_+(\tilde{\nu})}(c_2(\alpha_2))$	$1-F_{\Gamma_+(\tilde{\nu})}(c_2(\alpha_2))$
Cross	0	0	0	1

Table 2. Off-diagonal entries represent misclassification probabilities for the asymptotic test

1. *Draw random samples of size* n *and* m *from* F_n *and* G_m, *i.e.* $(X_1^*, X_2^*, \ldots, X_n^*) \sim F_n$ *and* $(Y_1^*, Y_2^*, \ldots, Y_m^*) \sim G_m$, *and compute the bootstrap statistics*

$$T^*_{m,n,1},\ T^*_{m,n,2},\ \text{and } T^{**}_{m,n,1} \tag{15}$$

2. *Repeat Step 1 B times.*
3. *Estimate the* $(1-\alpha_1)$ *quantile* $\hat{c}_1(\alpha_1)$ *from the bootstrap empirical distribution*

$$J_1(x) = B^{-1}\sum 1\{T^*_{m,n,1} \le x\}.$$

4. *If* $T_{m,n,1} \le \hat{c}_1(\alpha_1)$ *stop. Otherwise, estimate the* $(1-\alpha_2)$ *quantile* $\hat{c}_2(\alpha_2)$ *from the bootstrap empirical distribution*

$$J_{1,2}(x, \hat{c}_1(\alpha_1)) = \frac{B^{-1}\sum 1\{T^*_{m,n,2} \le x, T^{**}_{m,n,1} > \hat{c}_1(\alpha_1)\}}{B^{-1}\sum 1\{T^{**}_{m,n,1} > \hat{c}_1(\alpha_1)\}}.$$

5. *If* $T_{m,n,2} > \hat{c}_2(\alpha_2)$ *infer crossing. Else, infer the direction of dominance using the decision rule outlined in Section 2.*

3.1. **Asymptotic Properties.** A particularly attractive feature of our proposed testing procedures is the explicit characterization of the asymptotic misclassification error rates which they afford. For example, if α_1 and α_2 denote the pre-specified nominal levels for the first and second stage tests, then the large sample behavior of our proposed asymptotic test may be characterized according to Table 2. From the table we observe that the test is consistent in the sense that asymptotically we will detect crossing with probability one when the distributions F and G indeed cross one another. Also, the Type I error rate under H_0 is equal to the nominal level α_1 with the misclassifications of dominance in either direction being equally likely, and with an inference of crossing being the least likely misclassification. For instance, when $\alpha_1 = \alpha_2 = 0.05$ and H_0 is true, the asymptotic probability of inferring $F \succeq G$ is 0.02375, whereas the probability of inferring crossing is only 0.0025. These probabilities are in stark contrast to, say, the repeated application of the McFadden (1989) or Barrett and Donald (2003) tests at the %5 nominal level with the role of F

and G interchanged. In this case, we find that the likelihood of inferring dominance when $F = G$ is in fact closer to 10%.[2]

Misclassification Rates for the Bootstrap Test

	Inference			
Population	F=G	F≻G	G≻F	Cross
F=G	$1-\alpha_1$	$\frac{\alpha_1(1-\alpha 2)}{2}$	$\frac{\alpha_1(1-\alpha 2)}{2}$	$\alpha_1\alpha_2$
F≻G†	0	$1-\alpha_2$	0	α_2
G≻F†	0	0	$1-\alpha_2$	α_2
Cross	0	0	0	1

Table 3. Off-diagonal entries represent misclassification probabilities for the bootstrap test for a fixed pair of nominal levels α_1 and α_2. Results are stated under the implicit assumption that the quantile functions associated with the respective limiting distributions are continuous at the point $(1-\alpha_1, 1-\alpha_2)$. When the contact set is empty, the asymptotic misclassification rates in the cases marked with a dagger are zero.

A somewhat disappointing feature of the asymptotic test is that the asymptotic misclassification rates of the second-stage test depends on the underlying CDFs. In particular, the asymptotic probability of incorrectly rejecting strict dominance, which is given $F_{\Gamma_+(\tilde{\nu})}(c_2(\alpha_2))$, depends on the contact set through the stochastic process $\tilde{\nu}(t) := \nu(t)\mathbb{1}\{t : F(t) = G(t)\}$, where ν denotes the limiting process of $\nu_{m,n}(t) = \sqrt{\frac{mn}{m+n}}[\Delta_{m,n}(t) - \Delta(t)]$. It is therefore possible for the asymptotic rejection probability to exceed the nominal level α_2 of the second-stage test. This dependence is also one of the main features differentiating the asymptotic and bootstrap tests. Specifically, our bootstrap implementation corrects this fault by effectively untangling the first- and second-stage tests. This is achieved by using the uncentered bootstrap version of $T_{m,n,1}$ when conditioning the bootstrap distribution of $T_{m,n,2}$ on a Type I error being made in the first-stage. Thus when $F \neq G$, and as the sample size increases, the bootstrap further discounts the likelihood of a Type I error being committed. Asymptotically, the bootstrap test treats the second-stage test as a stand alone test independent of the first-stage, thus enabling α_2-level control at the second stage; see Table 3 for complete details concerning the misclassification error rates of our bootstrap test.

From either of the two tables we also see that with probability one in the limit, an inference of strict dominance is either correct or has been misclassified as *strict* dominance instead of *weak* dominance. Thus, for instance, the probability of inferring dominance when only weak dominance holds is $\frac{\alpha_1(1-\alpha 2)}{2} < \alpha_1/2$. Of course without explicit knowledge concerning the finite sample power of either test, non-rejection at

[2]This point is illustrated via Monte Carlo simulation; see Section 4 for details.

the second stage can never provide conclusive evidence that strict dominance holds in the population. Nevertheless, our testing procedures provide knowledge and even some control over the misclassification probabilities, thus allowing us greater flexibility in setting our tolerance for making the various types of misclassification errors.

3.2. **Power Against Drifting Alternatives.** We have thus far considered the behaviour of our proposed tests under the null and against fixed alternatives. We now turn to an investigation of the power of our bootstrap test against $\sqrt{n}$ drifting alternatives. Since the two-sided Kolmogorov-Smirnov test—and by extension our first stage bootstrap test— is known to asymptotically unbiased,[3] we focus our analysis here on the power of the two-stage procedure to detect root-n Pitman alternatives of the form

$$\Delta_{N(m,n)} = \Delta + \left(\frac{mn}{m+n}\right)^{-1/2}\xi \tag{16}$$

where, for some non-empty set $A \subset \mathbb{R}_+$, $\xi(t) \in \{0, \alpha\}$ and $\Delta(t) = 0$ for all $t \in A$ with $\xi(t) = \alpha$ for at least one $t \in A$; otherwise $\xi(t) = 0$ and $\Delta(t) < 0$ for all $t \in \mathbb{R}_+ \backslash A$. In other words, we consider sequences of alternatives consisting of $F_{N(m,n)}$ and $G_{N(m,n)}$ which converge at the rate $\sqrt{\frac{nm}{m+n}}$ to G and F, where F strictly dominates G, and for which, given any finite N, the distributions $G_{N(m,n)}$ and $F_{N(m,n)}$ cross. The following theorem characterizes the power of our bootstrap test against such drifting alternatives:

Theorem 3.1. *Consider a sequence of alternatives satisfying (40) above, then*

$$\lim_{n\to\infty} P_{N(m,n)}\big(T_{m,n,1} > \hat{c}_1(\alpha_1), T_{m,n,2} > \hat{c}_2(\alpha_2)\big) \geq \alpha_2$$

Remark 3.3. *Theorem 3.1 shows that our two-stage bootstrap test has asymptotic power at least that of the nominal level α_2 against a sequence of alternatives as defined in (40). This remarkable property suggests that, at least asymptotically, the second test operates as if it were a stand alone test independent of the first stage whenever the distributions under consideration are unequal.*

4. Finite Sample Performance

In this section we conduct a small scale Monte Carlo experiment to assess the finite sample performance of our proposed tests. In our analysis, we employ the experimental design of Barrett and Donald (2003), and use as a benchmark their two-sample bootstrap test for first-order stochastic dominance.

[3]Asymptotic unbiasedness can be established by applying Anderson's lemma; see, e.g., Abadie (2002)

Design	Parameter Values (μ_1,σ_1)	(μ_2,σ_2)	(μ_3,σ_3)	c
(a)	(0.85,0.6)	(0.85,0.6)	·	0.0
(b)	(0.85,0.6)	(0.6,0.8)	·	0.0
(c)	(0.85,0.6)	(1.2,0.2)	·	0.0
(d)	(0.85,0.6)	(0.8,0.5)	(0.9,0.9)	0.1
(e)	(0.85,0.6)	(0.85,0.4)	(0.4,0.9)	0.1
(f)	(0.85,0.6)	(0.75,0.65)	·	0.0

TABLE 4. Monte Carlo Experimental Designs

In each of the experiments, independent random samples $\{X_i, 1 \leq i \leq n\}$ and $\{Y_i, 1 \leq i \leq m\}$ are generated according to

$$X_i = \exp(\sigma_1 Z_{1i} + \mu_1)$$

and

$$Y_i = 1(U_i \geq c)\exp(\sigma_2 Z_{2i} + \mu_2) + 1(U_i < c)\exp(\sigma_3 Z_{3i} + \mu_3)$$

where the Z_{ki} are independent $N(0,1)$ random variables, and U_i is a uniform $[0,1]$ random variable. The parameter c controls the mixture proportion of the distributions whereas the (μ_i, σ_i) pairs completely characterize the respective lognormal distributions. Considering samples which are generated as the mixture of two lognormal distribution enables the design of a broader class of alternatives in which the population curves cross at multiple points.

For our tests, the null hypothesis in each experiment is that the distributions generating the samples are equal. Our interest centers on the ability of our two-stage procedures to identify the correct classification, i.e. equality, crossing, or dominance. We compare these empirical classification probabilities to those obtained from repeated application of the two-sample bootstrap test for stochastic dominance of Barrett and Donald (2003) at the nominal level $\alpha = 0.05$. In each case, we use 2,500 Monte Carlo simulations and 9,999 bootstrap samples. The specific choice of parameter values to be used in each experimental design is summarized in Table 4.

In the case of design (a), the distributions are equivalent and so this design provides evidence as to actual size of the tests. According to the theoretical results presented in the previous section, we should expect that our tests should have classification rates with relative frequencies approximately equal to those predicted in Tables 2 and 3. For designs (b)-(e) the distributions cross one another and thus are clearly all in the alternative of both the first- and second-stage tests. We should therefore expect the relative frequency of the inference of crossing to approach one in large samples. Finally, under design (f), the distribution F strictly dominates G.

Tables 5 and 6 contain the empirical classification rates (ECRs) from the Monte Carlo simulations when $n = m = 500$ and $n = m = 1000$, with $\alpha_1 = \alpha_2 =$

TABLE 5. Empirical Classification Probabilities: $n = m = 500$. $\alpha_1 = \alpha_2 = \alpha = 0.05$.

		Inference			
Design	Test	$F = G$	$F \succ G$	$G \succ F$	*Crossing*
	2 Stage (asymptotic)	**0.958**	0.021	0.019	0.002
(a)	2 Stage (bootstrap)	**0.945**	0.026	0.028	0.001
	Repeated one-sided	**0.900**	0.051	0.049	0.000
	2 Stage (asymptotic)	0.000	0.834	0.000	**0.167**
(b)	2 Stage (bootstrap)	0.000	0.550	0.000	**0.450**
	Repeated one-sided	0.000	1.000	0.000	**0.000**
	2 Stage (asymptotic)	0.000	0.000	0.000	**1.000**
(c)	2 Stage (bootstrap)	0.000	0.000	0.000	**1.000**
	Repeated one-sided	0.000	0.000	0.112	**0.888**
	2 Stage (asymptotic)	0.705	0.210	0.002	**0.083**
(d)	2 Stage (bootstrap)	0.617	0.294	0.004	**0.085**
	Repeated one-sided	0.519	0.456	0.023	**0.002**
	2 Stage (asymptotic)	0.079	0.031	0.000	**0.890**
(e)	2 Stage (bootstrap)	0.050	0.064	0.000	**0.886**
	Repeated one-sided	0.020	0.494	0.052	**0.434**
	2 Stage (asymptotic)	0.384	**0.601**	0.000	0.015
(f)	2 Stage (bootstrap)	0.312	**0.667**	0.000	0.020
	Repeated one-sided	0.223	**0.777**	0.000	0.000

0.05. In design (a), the ECRs of equality for our tests closely approximates the theoretical prediction of 95% at both sample sizes. Thus, for instance, dominance is incorrectly inferred in only 5% of the simulations when $n = m = 1000$. In contrast, repeated application of the Barrett and Donald (2003)test leads to inference of equality in roughly only 90% of the simulations and incorrectly infers dominance in the remaining 10%.

In designs (b) through (e), where the distributions under consideration cross, both our tests correctly identify crossing more often in every case, and in some instances by a wide margin. For all tests, it is designs (b) and (d) for which the detection of crossing is most difficult. However, our tests generally outperform the test of Barrett and Donald (2003); the most notable difference occurring in design (b) for a sample of $n = m = 1000$ where our bootstrap test correctly identifies crossing in almost 70% of the trials, whereas repeated application of the BD test leads to an incorrect inference of dominance in *every* trial. Overall, in designs (b) through (e) we find that repeated application of the test of Barrett and Donald (2003) leads to an incorrect inference of dominance substantially more often than our sequential tests; and that our bootstrap test closely mimics our asymptotic test in every design except (b), where the bootstrap test clearly dominates.

TABLE 6. Empirical Classification Probabilities: $n = m = 1000$. $\alpha_1 = \alpha_2 = \alpha = 0.05$.

		Inference			
Design	Test	$F = G$	$F \succ G$	$G \succ F$	*Crossing*
(a)	2 Stage (asymptotic)	**0.953**	0.023	0.021	0.002
	2 Stage (bootstrap)	**0.951**	0.022	0.025	0.001
	Repeated one-sided	**0.904**	0.048	0.048	0.000
(b)	2 Stage (asymptotic)	0.000	0.741	0.000	**0.259**
	2 Stage (bootstrap)	0.000	0.310	0.000	**0.690**
	Repeated one-sided	0.000	1.000	0.000	**0.000**
(c)	2 Stage (asymptotic)	0.000	0.000	0.000	**1.000**
	2 Stage (bootstrap)	0.000	0.000	0.000	**1.000**
	Repeated one-sided	0.000	0.000	0.000	**1.000**
(d)	2 Stage (asymptotic)	0.346	0.341	0.008	**0.312**
	2 Stage (bootstrap)	0.334	0.393	0.002	**0.272**
	Repeated one-sided	0.214	0.759	0.018	**0.010**
(e)	2 Stage (asymptotic)	0.000	0.000	0.000	**1.000**
	2 Stage (bootstrap)	0.000	0.005	0.000	**0.995**
	Repeated one-sided	0.000	0.170	0.000	**0.830**
(f)	2 Stage (asymptotic)	0.076	**0.900**	0.000	0.024
	2 Stage (bootstrap)	0.065	**0.904**	0.000	0.030
	Repeated one-sided	0.040	**0.960**	0.000	0.000

Design (e) is the only case in which dominance holds in the population. Here, the test of Barrett and Donald (2003) outperforms our tests by correctly identifying dominance in 10% (resp., 6%) more trials in smaller (resp., larger) samples. This finding should not be surprising given that their test *always* leads to an inference of dominance more often than ours irrespective of whether dominance holds in the population.

In our second set of simulations we manipulate our choices of α_1 and α_2 to bring our theoretical misclassification error rates under H_0 in line with the empirical rates reported by the test of Barrett and Donald (2003). To do this we set $\alpha_1 = 0.1$ and $\alpha_2 = 0.04$. The corresponding results are outlined in Tables 7 and 8. Our tests now closely mimic the performance of the Barrett and Donald (2003) test under H_0. However, our tests continue to outperform the BD test in designs (b)-(e) where the distributions cross in the population; and, perhaps more importantly, our test now reports empirical classification rates very close to those of the BD test in the cases where dominance holds in the population.

TABLE 7. Empirical Classification Probabilities: $n = m = 500$. $\alpha_1 = 0.1$, $\alpha_2 = 0.04$, and $\alpha = 0.05$.

Design	Test	Inference			
		$F = G$	$F \succ G$	$G \succ F$	*Crossing*
(a)	2 Stage (asymptotic)	**0.910**	0.045	0.043	0.002
	2 Stage (bootstrap)	**0.903**	0.042	0.052	0.003
	Repeated one-sided	**0.900**	0.051	0.049	0.000
(b)	2 Stage (asymptotic)	0.000	0.886	0.000	**0.114**
	2 Stage (bootstrap)	0.000	0.590	0.000	**0.410**
	Repeated one-sided	0.000	1.000	0.000	**0.000**
(c)	2 Stage (asymptotic)	0.000	0.000	0.000	**1.000**
	2 Stage (bootstrap)	0.000	0.000	0.000	**1.000**
	Repeated one-sided	0.000	0.000	0.119	**0.881**
(d)	2 Stage (asymptotic)	0.518	0.354	0.006	**0.122**
	2 Stage (bootstrap)	0.500	0.370	0.006	**0.124**
	Repeated one-sided	0.516	0.467	0.015	**0.002**
(e)	2 Stage (asymptotic)	0.021	0.049	0.000	**0.930**
	2 Stage (bootstrap)	0.017	0.070	0.000	**0.913**
	Repeated one-sided	0.014	0.474	0.054	**0.458**
(f)	2 Stage (asymptotic)	0.223	**0.760**	0.000	0.015
	2 Stage (bootstrap)	0.216	**0.764**	0.000	0.019
	Repeated one-sided	0.230	**0.770**	0.000	0.000

TABLE 8. Empirical Classification Probabilities: $n = m = 1000$. $\alpha_1 = 0.1$, $\alpha_2 = 0.04$, and $\alpha = 0.05$.

Design	Test	Inference			
		$F = G$	$F \succ G$	$G \succ F$	*Crossing*
(a)	2 Stage (asymptotic)	**0.910**	0.047	0.039	0.004
	2 Stage (bootstrap)	**0.898**	0.048	0.051	0.004
	Repeated one-sided	**0.902**	0.048	0.051	0.000
(b)	2 Stage (asymptotic)	0.000	0.869	0.000	0.131
	2 Stage (bootstrap)	0.000	0.340	0.000	**0.660**
	Repeated one-sided	0.000	1.000	0.000	**0.000**
(c)	2 Stage (asymptotic)	0.000	0.000	0.000	**1.000**
	2 Stage (bootstrap)	0.000	0.000	0.000	**1.000**
	Repeated one-sided	0.000	0.000	0.000	**1.000**
(d)	2 Stage (asymptotic)	0.236	0.476	0.001	**0.277**
	2 Stage (bootstrap)	0.224	0.451	0.002	**0.323**
	Repeated one-sided	0.225	0.735	0.019	**0.021**
(e)	2 Stage (asymptotic)	0.000	0.004	0.00	**0.996**
	2 Stage (bootstrap)	0.000	0.005	0.000	**0.995**
	Repeated one-sided	0.000	0.173	0.001	**0.826**
(f)	2 Stage (asymptotic)	0.040	**0.948**	0.000	0.012
	2 Stage (bootstrap)	0.038	**0.942**	0.000	0.020
	Repeated one-sided	0.033	**0.966**	0.000	0.000

5. Conclusion

In this paper we have combined the maximum and minimum of one-sided Kolmogorov-Smirnov type statistics to construct bidirectional tests for stochastic dominance. Our first test, which we refer to as an asymptotic test, is based on an explicit representation of the asymptotic distributions of the first- and second-stage statistics under the null of equality of the distributions under test. Our asymptotic approximations to the finite sample distributions, which are valid whenever the distributions under test are continuous, allow us to provide tabulated critical values for our sequential testing procedure.

For use more generally, we have also developed a computationally feasible bootstrap procedure. The resulting bootstrap test is valid even when the distributions are not continuous, and is shown analytically to provide tighter control of the asymptotic misclassification rates than that provided by our asymptotic test. These results are confirmed by our simulation results, where we also demonstrate the failure of repeated application of the Barrett and Donald (2003) test to maintain appropriate control over the misclassification rates. In particular, our results suggest that the latter testing strategy is far to liberal in its inclination towards dominance.

Finally, we note that while we have concentrated in this paper on testing for first order stochastic dominance, our bootstrap procedure can be generalized in a natural way to develop tests of higher orders of dominance. Furthermore, our bootstrap procedure can be generalized to other situations involving tests of partial orderings, an obvious application being the Lorenz ordering.

Appendix A. Mathematical Appendix

A.1. **Preliminary Results.** We begin by introducing some notation and collecting some useful results. Let $\alpha_n := \sqrt{n}(F_n - F)$ and $\beta_m := \sqrt{m}(G_m - G)$. It is well known (Billingsley 1968, p. 141) that the empirical processes α_n and β_m converge weakly to independent Guassian processes $\mathcal{B}_1(F(t))$ and $\mathcal{B}_2(G(t))$, where, for example, $E[\mathcal{B}_1(F(t))] = 0$ and $E[\mathcal{B}(F(s))\mathcal{B}(F(t))] = F(s)(1-F(t))$ for $s \le t$. From these results, together with Assumption 3.1, it is straightforward to show that

$$\begin{aligned} \nu_{m,n}(t) &= \lambda_{m,n}^{1/2}\alpha_n(t) + (1-\lambda_{m,n})^{1/2}\beta_m(t) \\ &\Rightarrow \lambda^{1/2}\mathcal{B}_1(F(t)) + (1-\lambda)^{1/2}\mathcal{B}_2(G(t)) \qquad (17) \\ &= \nu(t) \end{aligned}$$

From (17) and an application of the continuous mapping theorem we obtain

$$\Gamma_1(\nu_{m,n}) \Rightarrow \Gamma_1(\nu), \qquad (18)$$

$$\Gamma_2(\nu_{m,n}) \Rightarrow \Gamma_2(\nu), \qquad (19)$$

and

$$\Gamma_+(\nu_{m,n}) \Rightarrow \Gamma_+(\nu). \qquad (20)$$

Further, we note that the limiting distributions in (18), (19) , and (20) are continuous. Finally, by Theorem 3.7.6 of van der Vaart and Wellner (1996), we also have that the bootstrap empirical processes $\nu_{m,n}^*$ converges weakly to ν given $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ in probability. Weak convergence for the bootstrap analogues of (18) and (19) also hold by Proposition 10.7 of Kosorok (2008).

Lemma A.1. *Suppose Assumption 3.1 holds and that F is continuous. Then, under H_0,*

$$\sqrt{\frac{mn}{N}}(F_n(x) - G_m(x)) \Rightarrow \mathcal{B}(t),$$

the standard Brownian Bridge on $[0,1]$.

Proof of Lemma A.1. The fact that

$$\begin{aligned} F_n(x) - G_m(x) &= \frac{N}{m}\left(\frac{(m+n)F_n}{N} - \frac{nF_n}{N} - \frac{mG_m}{N}\right) \\ &= \frac{N}{m}\left(F_n(x) - H_N(x)\right) \end{aligned}$$

permits us to write

$$\begin{aligned}\sqrt{\frac{mn}{N}}(F_n(Z_{(i)})-G_m(Z_{(i)})) =& \sqrt{\frac{m}{n}}\sqrt{N}\left(F_n(Z_{(i)})-H_N(Z_{(i)})\right)\\ =& \sqrt{\frac{\lambda_{m,n}}{1-\lambda_{m,n}}}\sqrt{N}\left(F_n(Z_{(i)})-H_N(Z_{(i)})\right) \qquad (21)\\ =& \sqrt{\frac{\lambda_{m,n}}{1-\lambda_{m,n}}}\mathbb{L}_{m,n}(i/N)\end{aligned}$$

where $Z_{(i)}$ denotes the ith order statistic from the pooled sample, and $\mathbb{L}_N$ denotes the empirical process defined by

$$\mathbb{L}_N = \sqrt{N}\left(F_nH_N^{-1} - H_NH_N^{-1}\right)$$

Writing

$$\mathbb{L}_N = \sqrt{N}\left(F_nH_N^{-1} - FH^{-1} + FH^{-1} - H_NH_N^{-1}\right)$$

and noting that $HH^{-1} - H_NH_N^{-1} \leq 1/N$, we have under the null hypothesis that

$$\mathbb{L}_N = \sqrt{N}\left(F_nH_N^{-1} - FH^{-1}\right) + o_p(1)$$

The process defined by the first term is shown in Pyke and Shorack (1968) to satisfy

$$\sqrt{N}\left(F_nH_N^{-1} - FH^{-1}\right) \Rightarrow \lambda\left[(1-\lambda)^{-1/2}b_0\mathcal{B}_1(FH^{-1}) - \lambda^{-1/2}a_0\mathcal{B}_2(GH^{-1})\right]$$

where a_0 and b_0 denote the derivatives of FH^{-1} and GH^{-1}. Since

$$HH^{-1}(t) = \lambda FH^{-1}(t) + (1-\lambda)GH^{-1}(t) = t,$$

we have $\lambda a_0 + (1-\lambda)b_0 = 1$. Additionally, since $F = G = H$ under the null, it follows that $a_0 = b_0 = 1$ and hence

$$\begin{aligned}\sqrt{N}\left(F_nH_N^{-1} - FH^{-1}\right) &\Rightarrow \lambda\left[(1-\lambda)^{-1/2}\mathcal{B}_1(t) - \lambda^{-1/2}\mathcal{B}_2(t)\right] &(22)\\ &= \left(\frac{\lambda}{1-\lambda}\right)^{1/2}\left[\lambda^{1/2}\mathcal{B}_1(t) - (1-\lambda)^{1/2}\mathcal{B}_2(t)\right] &(23)\\ &=_d \left(\frac{\lambda}{1-\lambda}\right)^{1/2}\mathcal{B}(t) &(24)\end{aligned}$$

The desired result is then obtained upon combining the weak convergence result in (22) with (21). □

Lemma A.2. *Suppose Assumption 3.1 holds and that F is continuous. Then, under H_0,*

$$\begin{pmatrix} T_{m,n,1} \\ T_{m,n,2} \end{pmatrix} \Rightarrow \begin{pmatrix} \Gamma_1(\mathcal{B}) \\ \Gamma_2(\mathcal{B}) \end{pmatrix}. \qquad (25)$$

Proof. For an arbitrary $(h_1, h_2) \in \mathbb{R}^2$, the map Φ_h from $D[0,\infty)$ into $\mathbb{R}$ defined by

$$z \mapsto h_1 \max\{\sup_t z(t), -\inf_t z(t)\} + h_2 \min\{\sup_t z(t), -\inf_t z(t)\}$$

is continuous with respect to the Skorokhod metric. Consequently, by Lemma A.1 and the continuous mapping theorem,

$$\Phi_h(\mathcal{B}_{m,n}) \Rightarrow \Phi_h(\mathcal{B}),$$

where $\mathcal{B}_{m,n} := \sqrt{\frac{mn}{N}}(F_n - G_m)$. Since the value of h above is an arbitrary element of $\mathbb{R}^2$, we may further conclude via the Cramer-Wold device that (25) holds.

□

Proof of Proposition 3.1. Let $X_1 = -\inf_t \mathcal{B}(t)$, $X_2 = \sup_t \mathcal{B}(t)$, $Y_1 = \min\{X_1, X_2\}$, and $Y_2 = \max\{X_1, X_2\}$, where $\mathcal{B}$ denotes the standard Brownian Bridge process. That, for $a \leq b$,

$$P[[T_{m,n,2} \leq a | T_{m,n,1} > b] \rightarrow P[Y_1 \leq a | Y_2 > b]$$

is a direct consequence of Lemma A.2. Thus, it suffices to derive the distribution of the right-hand side. Accordingly, we have

$$\begin{aligned} P[Y_1 \leq a | Y_2 > b] &= \frac{P[Y_1 \leq a, Y_2 > b]}{1 - P[Y_2 \leq b]} \\ &= \frac{P[Y_1 \leq a] - P[Y_1 \leq a, Y_2 \leq b]}{1 - P[Y_2 \leq b]} \end{aligned} \tag{26}$$

Treating each term in the numerator separately we have

$$\begin{aligned} P[Y_1 \leq a] &= P[\min\{X_1, X_2\} \leq a] \\ &= P[X_1 \leq a] + P[X_2 \leq a] - P[X_1 \leq a, X_2 \leq a] \\ &= 2P[X_1 \leq a] - P[Y_2 \leq a] \end{aligned} \tag{27}$$

and

$$\begin{aligned} P[Y_1 \leq a, Y_2 \leq b] &= P[X_1 \leq a, X_2 \leq b] + P[X_1 \leq b, X_2 \leq a] - P[X_1 \leq a, X_2 \leq a] \\ &= P[X_1 \leq a, X_2 \leq b] + P[X_1 \leq b, X_2 \leq a] - P[Y_2 \leq a] \end{aligned} \tag{28}$$

The desired result is then obtained from

$$P[Y_1 \leq a | Y_2 > b] = \frac{2P[X_1 \leq a] - P[X_1 \leq a, X_2 \leq b] - P[X_1 \leq b, X_2 \leq a]}{1 - P[Y_2 \leq b]} \tag{29}$$

which is comprised entirely of well known distributions; see, e.g., p. 79 of Billingsley (1968). □

Lemma A.3. *Suppose Assumption 3.1 holds. Then,*

$$T_{m,n,1} \Rightarrow \begin{cases} \Gamma_1(\nu) & \text{if } F = G \\ \infty & \text{if } F(t) \neq G(t) \text{ for some } t \in [0,\infty) \end{cases} \tag{30}$$

Proof of Lemma A.3. It is immediate upon writing

$$T_{m,n,1} = \sup_t \Big| \nu_{m,n}(t) + \sqrt{\frac{nm}{n+m}} \Delta(t) \Big|$$

that $T_{m,n,1}$ diverges to ∞ w.p.1 whenever $|\Delta(t)| > 0$ for some t. That $T_{m,n,1} \Rightarrow \Gamma_1(\nu)$ under H_0 follows from our earlier results, namely from (17) and (18).

□

Lemma A.4. *Suppose Assumption 3.1 holds and that $\delta_{m,n}$ is chosen in accordance with Conditions 3.1(i-ii). Then,*

$$T^{**}_{m,n,1} \Rightarrow \Gamma_1(\tilde{\nu}) \tag{31}$$

in probability conditional on $X_1, X_2, \ldots, Y_1, Y_2, \ldots,$ where $\tilde{\nu}(t) = \nu(t)\mathbb{1}\{t : F(t) = G(t)\}$.

Proof of A.4. Writing

$$T^*_{m,n,1} = \Gamma_1\Big(\sqrt{\frac{nm}{n+m}} (\Delta^*_{m,n} - \Delta_{m,n}) C_{m,n} \Big)$$

the desired result will be obtained via an application of the continuous mapping theorem if we can show that

$$\sqrt{\frac{nm}{n+m}} (\Delta^*_{m,n} - \Delta_{m,n}) C_{m,n} \Rightarrow \tilde{\nu}. \tag{32}$$

Now,

$$\begin{aligned} \sup_t \Big| \sqrt{\frac{nm}{n+m}} (\Delta^*_{m,n} - \Delta_{m,n}) C_{m,n} - \sqrt{\frac{nm}{n+m}} (\Delta^*_{m,n} - \Delta_{m,n}) 1_{\{t:F(t)=G(t)\}} \Big| \\ = \sup_t \Big| \sqrt{\frac{nm}{n+m}} (\Delta^*_{m,n} - \Delta_{m,n}) [C_{m,n} - 1_{\{t:F(t)=G(t)\}}] \Big| \\ \leq \sup_t \Big| \sqrt{\frac{nm}{n+m}} (\Delta^*_{m,n} - \Delta_{m,n}) \Big| \sup_t \Big| C_{m,n} - 1_{\{t:F(t)=G(t)\}} \Big| \end{aligned} \tag{33}$$

The first term converges weakly to $\Gamma_1(\nu)$ and is therefore $O_p(1)$, whereas the second term is $o_P(1)$ as a direct consequence of Condition 3.1. It follows that (32) converges weakly to the same limiting process as

$$\sqrt{\frac{nm}{n+m}} (\Delta^*_{m,n} - \Delta_{m,n}) 1_{\{t:F(t)=G(t)\}},$$

which due to the deterministic nature of the function $1_{\{t:F(t)=G(t)\}}$ is easily seen to satisfy

$$\sqrt{\frac{nm}{n+m}}(\Delta^*_{m,n} - \Delta_{m,n})1_{\{t:F(t)=G(t)\}} \Rightarrow \tilde{\nu}$$

□

Lemma A.5. *Suppose Assumption 3.1 holds. Then,*

$$T_{m,n,2} \Rightarrow \begin{cases} \Gamma_2(\nu) & \text{if } F = G \\ \infty & \text{if } F(t_1) > G(t_1) \text{ and } F(t_2) < G(t_2) \text{ for some } t_1, t_2 \in [0,\infty) \\ \Gamma_+(\tilde{\nu}) & \text{if } F \leq G \text{ or } F \geq G, F \neq G, \text{ and } \mathcal{C} \neq \emptyset \\ -\infty & \text{if } F \leq G \text{ or } F \geq G, \text{ and } \mathcal{C} = \emptyset \end{cases} \tag{34}$$

where $\mathcal{C} = \{t \in [0,\infty) : F(t) = G(t)\}$

Proof of A.5 . First, write

$$T_{m,n,2} = \min\{\sup_t[\nu_{m,n}(t) + \sqrt{\frac{mn}{m+n}}\Delta(t)], -\inf_t[\nu_{m,n}(t) + \sqrt{\frac{mn}{m+n}}\Delta(t)]\}.$$

Note that $\Delta(t) = 0$ for all t under H_0 and so the first line follows from (17) and (19). In the second case we have $\Delta(t_1) > 0$ and $\Delta(t_2) < 0$. It follows that the minimum diverges to ∞ w.p.1 in this case.

Now suppose $F(t) \leq G(t)$ for all t, $F \neq G$, and $\mathcal{C} \neq \emptyset$. In this case we may write

$$\begin{aligned} &\sup_t\big[\nu_{m,n}(t) + \sqrt{\frac{mn}{m+n}}\Delta(t)\big] \\ &\quad = \sup\Big\{\sup_t\Big[\nu_{m,n}(t)1_{\{t\in\mathcal{C}\}}\Big], \sup_t\Big[\nu_{m,n}(t)1_{\{t\in\mathcal{C}^c\}} + \sqrt{\frac{mn}{m+n}}\Delta(t)\Big]\Big\} \\ &\quad = \sup_t[\nu_{m,n}(t)1_{\{t\in\mathcal{C}\}}] + o_P(1) \end{aligned} \tag{35}$$

And since $-\inf_t[\nu_{m,n}(t) + \sqrt{\frac{mn}{m+n}}\Delta(t)]$ diverges to ∞, it follows also that

$$T_{m,n,2} = \sup_t[\nu_{m,n}(t)1_{\{t\in\mathcal{C}\}}] + o_P(1),$$

whence

$$T_{m,n,2} \Rightarrow \sup_t \tilde{\nu}(t).$$

The proof in the case $F(t) \geq G(t)$ for all t, $F \neq G$, and $\mathcal{C} \neq \emptyset$ follows by symmetry. Lastly, $\mathcal{C} = \emptyset$ implies that either $\Delta(t) > 0$ or $\Delta(t) < 0$ for all t. Suppose the former to be true. Then $-\inf_t[\nu_{m,n}(t) + \sqrt{\frac{mn}{m+n}}\Delta(t)]$ diverges to $-\infty$ and hence so to does $T_{m,n,2}$. Clearly, the same is also true of $T_{m,n,2}$ when $\Delta(t) < 0$ and the result follows. □

Lemma A.6. *Suppose Assumption 3.1 holds and that $\delta_{m,n}$ is chosen in accordance with Conditions 3.1(i-ii). Then,*

$$T^*_{m,n,2} \Rightarrow \begin{cases} \Gamma_2(\nu) & \text{if } F = G \\ \Gamma_2(\tilde{\nu}) & \text{if } F(t_1) > G(t_1) \text{ and } F(t_2) < G(t_2) \text{ for some } t_1, t_2 \in \mathbb{R} \\ \Gamma_+(\tilde{\nu}) & \text{if } F \le G \text{ or } F \ge G, \text{and} F \neq G \end{cases} \tag{36}$$

in probability conditional on $X_1, X_2, \ldots, Y_1, Y_2, \ldots,$ where $\tilde{\nu}(t) = \nu(t)\mathbb{1}\{t : F(t) = G(t)\}$.

Proof of A.6. First, we have the following four mutually exclusive cases to consider where

(1) $F = G$ implies

$$\lim_{n\to\infty} P[a_{m,n} < 0, b_{m,n} < 0] = 1;$$

(2) $F \le G$, $F \neq G$ implies

$$\lim_{n\to\infty} P[a_{m,n} < 0, b_{m,n} > 0] = 1;$$

(3) $F \ge G$, $F \neq G$ implies

$$\lim_{n\to\infty} P[a_{m,n} > 0, b_{m,n} < 0] = 1;$$

and

(4) $F(t_1) > G(t_1)$ and $F(t_2) > G(t_2)$ for some $t_1, t_2 \in [0, \infty)$ implies

$$\lim_{n\to\infty} P[a_{m,n} > 0, b_{m,n} > 0] = 1.$$

Note that the above are direct implications of Condition 3.1. From these we obtain

$$T^*_{m,n,2} = \Gamma_2\Big(\sqrt{\frac{mn}{m+n}}\tilde{\Delta}^*_{m,n}\Big) + o_P(1)$$

in case (1) or (4) from which it follows that $T^*_{m,n,2} \Rightarrow \Gamma_2(\tilde{\nu})$. Alternatively, when $F \le G$ with $F \neq G$ we have

$$T^*_{m,n,2} = \Gamma_+\Big(\sqrt{\frac{mn}{m+n}}\tilde{\Delta}^*_{m,n}\Big) + o_P(1)$$

whence

$$T^*_{m,n,2} \Rightarrow \Gamma_+(\tilde{\nu}).$$

Case (3) is analogous to case (2) and thus omitted. □

A.2. **Proofs of Main Results.** We are now in a position to prove the the first and fourth assertions made on line 1 of Table 2 (the second and third assertions follow directly from these by symmetry). Since only minor modifications are necessary for a treatment of the asymptotic test, for the sake of brevity we prove the result only for the bootstrap implementation. Prior to our statement of the theorem we first introduce the following notation: let

$$c_L(\alpha_1) = \inf\{x : P[\Gamma_1(\nu) \leq x] \geq 1 - \alpha_1\},$$

and

$$c_U(\alpha_1) = \sup\{x : P[\Gamma_1(\nu) \leq x] \leq 1 - \alpha_1\};$$

Also, let

$$c_{L,2}(\alpha_1, \alpha_2) = \inf\{x : P[\Gamma_2(\nu) \leq x, \Gamma_1(\nu) > c_U(\alpha_1)] \geq \alpha_1 - \alpha_1\alpha_2\},$$

$$c_{U,2}(\alpha_1, \alpha_2) = \sup\{x : P[\Gamma_2(\nu) \leq x, \Gamma_1(\nu) > c_L(\alpha_1)] \leq \alpha_1 - \alpha_1\alpha_2\},$$

and

$$c_{m,n,2} = \inf\{x : P[T^*_{m,n,2} \leq x, T^{**}_{m,n,1} > \hat{c}_1(\alpha_1)] \geq \alpha_1 - \alpha_1\alpha_2\}.$$

Theorem A.1. *Suppose Assumption 3.1 holds. Then, if H_0 is true,*

i.

$$\begin{aligned}\lim_{x \uparrow c_L(\alpha_1)} P[\Gamma_1(\nu) \leq x] &\leq \liminf_{n\to\infty} P[T_{m,n,1} \leq \hat{c}_1(\alpha_1)] \\ &\leq \limsup_{n\to\infty} P[T_{m,n,1} \leq \hat{c}_1(\alpha_1)] \\ &\leq P[\Gamma_1(\nu) \leq c_U(\alpha_1)]\end{aligned}$$

with

$$\lim_{n\to\infty} P[T_{m,n,1} \leq \hat{c}_1(\alpha_1)] = 1 - \alpha_1$$

whenever $c_L(\alpha_1) = c_U(\alpha_1)$; and

ii.

$$\begin{aligned}\lim_{x \uparrow c_{L,2}(\alpha_1,\alpha_2)} H(x, c_U(\alpha_1)) &\leq \liminf_{n\to\infty} P[T_{m,n,2} \leq c_{m,n,2}(\alpha_1,\alpha_2), T_{m,n,1} > \hat{c}_1(\alpha_1)] \\ &\leq \limsup_{n\to\infty} P[T_{m,n,2} \leq c_{m,n,2}(\alpha_1,\alpha_2), T_{m,n,1} > \hat{c}_1(\alpha_1)] \\ &\leq H(c_{U,2}(\alpha_1,\alpha_2), c_L(\alpha_1))\end{aligned}$$

with

$$\lim_{n\to\infty} P[T_{m,n,2} \leq \hat{c}_2(\alpha_2), T_{m,n,1} > \hat{c}_1(\alpha_1)] = \alpha_1 - \alpha_1\alpha_2$$

whenever $c_L(\alpha_1) = c_U(\alpha_1)$ and $c_{L,2}(\alpha_1,\alpha_2) = c_{U,2}(\alpha_1,\alpha_2)$.

Proof. (i) The proof follows directly from Theorem 1 of Beran (1984).

(ii) For an arbitrary $(h_1, h_2) \in \mathbb{R}^2$, the map Φ_h from $D[0,\infty)$ into $\mathbb{R}$ defined by

$$z \mapsto h_1 \max\{\sup_t z(t), -\inf_t z(t)\} + h_2 \min\{\sup_t z(t), -\inf_t z(t)\}$$

is continuous with respect to the Skorokhod metric. Consequently, by the continuous mapping theorem,

$$\Phi_h(\nu_{m,n}) \Rightarrow \Phi_h(\nu).$$

Since the value of h above is an arbitrary element of $\mathbb{R}^2$, we may further conclude via the Cramer-Wold device that

$$\begin{pmatrix} T_{m,n,1} \\ T_{m,n,2} \end{pmatrix} \Rightarrow \begin{pmatrix} \Gamma_1(\nu) \\ \Gamma_2(\nu) \end{pmatrix}.$$

Using weak convergence results for the bootstrap empirical processes, an analogous argument may be used to establish that

$$\begin{pmatrix} T^*_{m,n,1} \\ T^*_{m,n,2} \end{pmatrix} \Rightarrow \begin{pmatrix} \Gamma_1(\nu) \\ \Gamma_2(\nu) \end{pmatrix}$$

in probability. Noting that $|T^{**}_{m,n,1} - T^*_{m,n1}| = o_p(1)$ under H_0, and letting $X = \Gamma_2(\nu)$ and $Y = \Gamma_1(\nu)$, we then have

$$\begin{aligned} H_{m,n}(x,y) &:= P(T^*_{m,n,2} \le x, T^{**}_{m,n} > y) \\ &\to H(x,y) := P(X \le x, Y > y) \end{aligned} \tag{37}$$

in probability. Noting that

$$H(x, c_U(\alpha_1)) \le H(x, c_L(\alpha_1)), \tag{38}$$

where $c_U(\alpha_1)$ and $c_L(\alpha_1)$ are defined above, and letting $\epsilon > 0$ be such that $c_{L,2}(\alpha_1,\alpha_2) - \epsilon$ and $c_{U,2}(\alpha_1,\alpha_2) + \epsilon$ are continuity points of $H(x, c_U(\alpha_1))$ and $H(x, c_L(\alpha_1))$ respectively, we have, as a consequence of weak convergence of the bootstrap distribution, that

$$P[H_{m,n}(c_{L,2}(\alpha_1,\alpha_2) - \epsilon, c_U(\alpha_1)) < \alpha_1 - \alpha_1\alpha_2 < H_{m,n}(c_{U,2}(\alpha_1,\alpha_2) + \epsilon, c_L(\alpha_1))] \to 1,$$

and hence

$$P[c_{L,2}(\alpha_1,\alpha_2) - \epsilon \le c_{m,n,2}(\alpha_1,\alpha_2) \le c_{U,2}(\alpha_1,\alpha_2) + \epsilon] \to 1,$$

as $n \to \infty$. The proof is complete upon combining this result with the inequality in (38) and taking the limit as $n \to \infty$ and $\epsilon \to 0$.

□

Theorem A.2. *Suppose Assumption 3.1 holds. Then, if $F \le G$ and H_0 is false,*

i $\lim_{n\to\infty} P[T_{m,n,1} > c^a_1(\alpha_1)] = 1$

ii $\lim_{n\to\infty} P[T_{m,n,2} > c^a_2(\alpha_2), T_{m,n,1} > c^a_1(\alpha_1)] \in \{0, 1 - F_{\Gamma_+(\tilde{\nu})}(c^a_2(\alpha_2))\}$

with zero being attained only when $\mathcal{C} = \emptyset$, and where $c^a_1(\alpha_1)$ and $c^a_2(\alpha_2)$ denote the α_1 and α_2 level critical values of the asymptotic test.

Proof. (i) The result is a direct consequence of the finiteness of $c_1^a(\alpha_1)$ and the fact that $T_{m,n,1}$ diverges to infinity under the alternative.

(ii) Since

$$\lim_{n\to\infty} P[T_{m,n,2} > c_2^a(\alpha_2), T_{m,n,1} > c_1^a(\alpha_1)] = P[\Gamma_+(\tilde{\nu}) > c_2^a(\alpha_2)]$$

we may write the probability on the right-hand side as

$$1 - F_{\Gamma_+(\tilde{\nu})}(c_2^a(\alpha_2)).$$

Finally, when $\mathcal{C} = \emptyset$, then $T_{m,n,2}$ diverges to $-\infty$ under the maintained assumptions and so the rejection probability is zero. □

For the statement of and proof of our next theorem we introduce the following objects

$$c_L(\alpha_2) = \inf\{x : P[\Gamma_2(\tilde{\nu}) \leq x] \geq 1 - \alpha_1\},$$

and

$$c_U(\alpha_2) = \sup\{x : P[\Gamma_2(\tilde{\nu}) \leq x] \leq 1 - \alpha_1\};$$

Theorem A.3. *Suppose Assumption 3.1 holds. Then, if $F \leq G$ and H_0 is false,*

i.

$$\lim_{n\to\infty} P[T_{m,n,1} > \hat{c}_1(\alpha_1)] = 1,$$

and

ii.

$$\begin{aligned}\lim_{x\uparrow c_L(\alpha_2)} P[\Gamma_2(\tilde{\nu}) \leq x] &\leq \liminf_{n\to\infty} P[T_{m,n,2} \leq \hat{c}_2(\alpha_2), T_{m,n,1} > \hat{c}_1(\alpha_1)] \\ &\leq \limsup_{n\to\infty} P[T_{m,n,2} \leq \hat{c}_2(\alpha_2), T_{m,n,1} > \hat{c}_1(\alpha_1)] \\ &\leq P[\Gamma_2(\tilde{\nu}) \leq c_U(\alpha_2)]\end{aligned}$$

with

$$\lim_{n\to\infty} P[T_{m,n,2} \leq \hat{c}_2(\alpha_2), T_{m,n,1} > \hat{c}_1(\alpha_1)] = 1 - \alpha_2$$

whenever $c_L(\alpha_2) = c_U(\alpha_2)$.

Proof. (i) The proof trivially follows from the finiteness of $\hat{c}_1(\alpha_1)$ and the fact that $T_{m,n,1}$ diverges to infinity under the alternative.

(ii) With $F \leq G$ and H_0 false, we have

$$\begin{aligned}H_{m,n}(x) &= \frac{P\left[T^*_{m,n,2} \leq c, T^{**}_{m,n,1} > \hat{c}_1(\alpha_1)\right]}{P\left[T^{**}_{m,n,1} > \hat{c}_1(\alpha_1)\right]]} \\ &\to H(x) = P[\Gamma_+(\tilde{\nu}) \leq c]\end{aligned} \tag{39}$$

since $\hat{c}_1(\alpha_1)$ is bounded in probability whereas $T^{**}_{m,n,1}$ diverges to infinity. Letting $\epsilon > 0$ be such that $c_L(\alpha_2) - \epsilon$ and $c_U(\alpha_2) + \epsilon$ be continuity points of $H(x)$ we have

$$P[H_{m,n}(c_L(\alpha_2) - \epsilon) < 1 - \alpha_2 < H_{m,n}(c_U(\alpha_2) + \epsilon)] \to 1$$

as $n \to \infty$, from which we obtain

$$P[c_L(\alpha_2) - \epsilon \leq \hat{c}_2(\alpha_2) < c_U(\alpha_2) + \epsilon] \to 1.$$

The desired result then follows by considering the limit as $\epsilon \to \infty$.

It follows that

$$\hat{c}_2(\alpha_2) \to^P c_2(\alpha) = \inf\{c : P[\Gamma_+(\tilde{\nu}) \leq c] \geq 1 - \alpha_2\},$$

the $(1-\alpha_2)$ quantile of the limiting distribution of $P[T_{m,n,2} \leq \hat{c}_2(\alpha_2), T_{m,n,1} > \hat{c}_1(\alpha_1)]$ when $F \leq G$, H_0 is false and $\mathcal{C} \neq \emptyset$. If $\mathcal{C} = \emptyset$, then $\hat{c}_2(\alpha_2) \to^P 0$, $T_{m,n,2}$ diverges to minus infinity, in which case

$$P[T_{m,n,2} > \hat{c}_2(\alpha_2), T_{m,n,1} > \hat{c}_1(\alpha_1)] \to 0$$

as $n \to \infty$. □

Theorem A.4. *Suppose Assumption 3.1 holds. Then, if F and G cross,*

$$\lim_{n\to\infty} P[T_{m,n,2} > \hat{c}_2(\alpha_2), T_{m,n,1} > \hat{c}_1(\alpha_1)] = 1$$

Proof. Under this alternative, the conclusion follows from the fact that both $T_{m,n,2}$ and $T_{m,n,1}$ diverge to infinity whereas the critical values $\hat{c}_2(\alpha_2)$ and $\hat{c}_1(\alpha_1)$ are both bounded in probability □

Proof of Theorem 3.1. Recall that the sequence of Pitman alternatives under consideration is given by

$$\Delta_{N(m,n)} = \Delta + \left(\frac{mn}{m+n}\right)^{-1/2} \xi \tag{40}$$

where, for some non-empty set $A \subset \mathbb{R}$, $\xi(t) \in \{0, \alpha\}$ and $\Delta(t) = 0$ for all $t \in A$ with $\xi(t) = \alpha$ for at least one $t \in A$; otherwise $\xi(t) = 0$ and $\Delta(t) < 0$ for all $t \in \mathbb{R}\backslash A$.

To prove the theorem we first write

$$\sqrt{\frac{mn}{m+n}}\Delta_{m,n} = \sqrt{\frac{mn}{m+n}}(\Delta_{n,m} - \Delta_{N(n,n)}) + \sqrt{\frac{mn}{m+n}}\left[\Delta + \left(\frac{mn}{m+n}\right)^{-1/2}\xi\right]$$

from which we obtain $T_{m,n1} \to_P \infty$, $T^{**}_{m,n1} \to_P \infty$, and

$$T_{m,n,2} \Rightarrow \Gamma_+(\nu 1_{t\in A} + \xi).$$

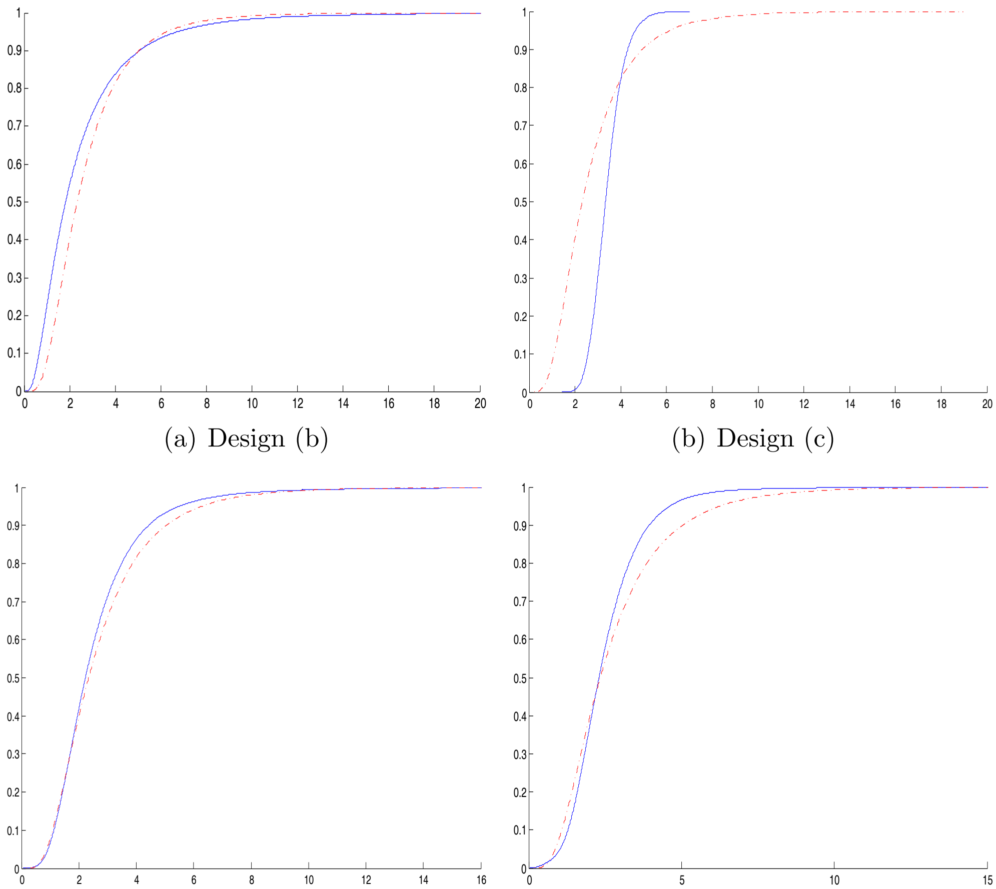
On the other hand, we have

$$T^*_{m,n,2} \Rightarrow \Gamma_+(\nu 1_{t\in A}).$$

Since

$$\hat{c}_2(\alpha_2) \rightarrow_P F^{-1}_{\Gamma_+(\nu 1_{t \in A})}(1 - \alpha_2),$$

the desired result follows from the fact that $\Gamma_+(\nu 1_{t\in A} + \xi)$ is stochastically larger than $\Gamma_+(\nu 1_{t\in A})$. $\square$



(a) Design (b) (b) Design (c)

(c) Design (d) (d) Design (e)

FIGURE 2. Plots of the distribution functions F (solid line) and G (dashed line) under each experimental design.

References

Abadie, A. (2002), 'Bootstrap tests for distributional treatment effects in instrumental variable models', *Journal of the American Statistical Association* **97**(457), 1893–1905.

Anderson, G. (1996), 'Nonparametric tests of stochastic dominance in income distributions', *Econometrica* **64**(5), 1183–1193.

Barrett, G. F. and Donald, S. G. (2003), 'Consistent tests for stochastic dominance', *Econometrica* **71**(1), 71–104.

Bennett, C. J. (2008), Consistent integral-type tests for stochastic dominance, Mimeo, Vanderbilt University.

Beran, R. (1984), 'Bootstrap methods in statistics', *Jahresbericht der Deutschen Mathematiker-Vereinigung* **86**(1), 14–30.

Bhattacharya, D. (2007), 'Inference on inequality from household survey data', *Journal of Econometrics* **137**(2), 674–707.

Billingsley, P. (1968), *Probability and Measure*, John Wiley & Sons, New York.

Bishop, J. A., Formby, J. P. and Thistle, P. D. (1992), 'Convergence of the south and non-south income distributions, 1969-1979', *American Economic Review* (82), 262–272.

Dardanoni, V. and Forcina, A. (1999), 'Inference for lorenz curve orderings', *Econometrics Journal* **2**, 4975.

Davidson, R. and Duclos, J.-Y. (2000), 'Statistical inference for stochastic dominance and for the measurement of poverty and inequality', *Econometrica* **68**(6), 1435–1464.

Davidson, R. and Duclos, J.-Y. (2006), 'Testing for restricted stochastic dominance', *IZA Discussion Paper No. 2047* .

Foster, J., Greer, J. and Thorbecke, E. (1984), 'A class of decomposable poverty measures', *Econometrica* **52**(3), 761–66.

Foster, J. and Shorrocks, A. (1988), 'Poverty orderings', *Econometrica* (56), 173–177.

James, B. R. (1975), 'A functional law of the iterated logarithm for weighted empirical distributions', *The Annals of Probability* **3**(5), 762–772.

Kaur, A., Rao, B. L. S. P. and Singh, H. (1994), 'Testing for second-order stochastic dominance of two distributions', *Econometric Theory* **10**, 849–866.

Klecan, L., McFadden, R. and McFadden, D. (1991), 'A robust test for stochastic dominance', *Working Paper, Department of Economics, MIT* .

Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, Springer, New York.

Levy, H. (1992), 'Stochastic dominance and expected utility: Survey and analysis', *Management Science* **38**(4), 555.

Linton, O., Maasoumi, E. and Whang, Y.-J. (2005), 'Consistent testing for stochastic dominance under general sampling schemes', *Review of Economic Studies* **72**(1), 735–765.

Linton, O., Song, K. and Whang, Y.-J. (2008), Bootstrap tests of stochastic dominance with asymptotic similarity on the boundary, CeMMAP working papers CWP08/08, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Maasoumi, E. and Heshmati, A. (2000), 'Stochastic dominance amongst Swedish income distributions', *Econometric Reviews* **19**(3), 287–320.

McFadden, D. (1989), Testing for Stochastic Dominance, *in* T. B. Fomby and T. K. Seo, eds, 'Studies in the Economics of Uncertainty. In Honour of Josef Hadar', Springer Verlag, New York, pp. 113–134.

Pyke, R. and Shorack, G. (1968), 'Weak convergence of a two-sample empirical process and a new approach to Chernoff-Savage theorems', *Annals of Mathematical Statistics* (39).

Schechtman, E., Shelef, A., Yitzhaki, S. and Zitikis, R. (2008), 'Testing hypotheses about absolute concentration curves and marginal conditional stochastic dominance', *Econometric Theory* **24**(04), 1044–1062.

Shorack, G. R. and Wellner, J. A. (1986), *Empirical Processes with Applications to Statistics*, John Wiley & Sons, New York ;; Toronto.

Stephens, M. (1986), *Tests based on EDF statistics*, Goodness-of-Fit Techniques, Marcel Dekker, pp. 97–193.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer, New York.