Consistent Tests For Stochastic Dominance

Garry Barrett
Department of Economics
University of New South Wales
Sydney, NSW, 2052
and
Stephen Donald*
Department of Economics
University of Texas at Austin
Austin TX, 78712

First Version: April, 1999
This Version: May, 2001

Abstract

Methods are proposed for testing stochastic dominance of any pre-specified order, with primary interest in the distributions of income. We consider consistent tests, that are similar to Kolmogorov Smirnov tests, of the complete set of restrictions that relate to the various forms of stochastic dominance. For such tests, in the case of tests for stochastic dominance beyond first order, we propose and justify a variety of approaches to inference based on simulation and the bootstrap. We compare these approaches to one another and to alternative approaches based on multiple comparisons in the context of a Monte Carlo experiment and an empirical example.

Keywords: Stochastic dominance, test consistency, simulation, bootstrap.

*We thank Ken Zhu for excellent research assistance, and a co-editor and three anonymous referees for constructive comments and suggestions. The research of Stephen Donald has been supported by a Alfred P. Sloan Foundation Research Fellowship and an NSF grant SES-0196372.

1 Introduction

Recent work on inequality and poverty analysis has emphasized the importance of various forms of stochastic dominance relationships between income distributions. In particular Anderson (1996) and Davidson and Duclos (2000) have discussed the importance of the concepts of first, second and third order stochastic dominance (SD1, SD2 and SD3 respectively) relationships between income distributions for social welfare and poverty rankings of distributions[1]. These papers have considered the problem of making inferences regarding various forms of stochastic dominance by comparing objects (usually the income distribution itself or partial integrals thereof) at an arbitrarily chosen and fixed number of income values. Anderson (1996) proposed tests for the various forms of stochastic dominance based on t-statistics comparing the objects calculated in two independent samples, while Davidson and Duclos (2000) suggested an approach based on tests of inequality constraints[2]. A key merit of these approaches is their practicality, since they are based on a small number of comparisons. However, as noted by Davidson and Duclos (2000, p.1446), the fact that the comparisons are made at a fixed number of arbitrarily chosen points is not a desirable feature, and introduces the possibility of test inconsistency. A more desirable approach would be based on comparison of the objects at all points in the support of income.

The aim of this paper is to consider tests of stochastic dominance of any pre-specified order that are based on Kolmogorov-Smirnov type tests which compare the objects at all points. The objects being compared are multiple partial integrals of some underlying income distribution, and since the objects are being compared at all points in the income range the tests have the potential to be consistent tests of the full set of restrictions implied by stochastic dominance. In contrast, tests based on a fixed number of comparisons are potentially inconsistent since only a subset of the restrictions implied by stochastic dominance are considered. Although the Kolmogorov-Smirnov type tests are based on

[1]Also note that Shorrocks (1983) has shown that second order dominance is equivalent to generalized Lorenz dominance. See Lambert (1993) for a nice exposition of this result.

[2]In Davidson and Duclos (2000) the tests for various forms of stochastic dominance are a prelude to estimating the smallest income level at which the distributions (or integrals thereof) cross.

a comparison at all income values, the exact value of the statistics can be calculated exactly with a finite number of calculations and so the tests should be useful in practice. In the case of first and second order stochastic dominance McFadden (1989) has considered the use of such tests with independent samples with equal numbers of observations. Unlike McFadden (1989) we allow for different sample sizes and we also consider tests for stochastic dominance of any pre-specified order, say SDj. The main difficulty with the tests, as noted by McFadden (1989), is in constructing appropriate rejection regions for conducting the tests, since the test statistics for testing SDj for j larger than 1 (i.e. SD2, SD3 and so on) have limiting distributions that depend on the underlying distributions. McFadden (1989) proposed a Monte Carlo based method to estimate an approximate asymptotic significance level (or p-value). In our case we also use a variety of simulation and bootstrap methods to estimate the exact asymptotic p-value, and are able to show that these methods can be theoretically justified. In addition we show that the methods give rise to tests with desirable size and power properties in finite samples.[3]

Other papers have considered the problem of testing for SD2 and have attempted to deal with the difficulty of conducting inference in a variety of ways. Schmid and Trede (1998) have proposed a test for SD2, for which critical values can be obtained, but require that one of the distributions be known and have a density that is monotonically decreasing. This would seem to be unsatisfactory in general and in particular would seem to rule out applications to income distributions. Kaur, Prakasa Rao and Singh (1994) proposed a test of SD2 that has the advantage of giving rise to a test statistic with a standard limiting distribution. While much of the literature formulates the null hypothesis as corresponding to SD2, they do not, and have the alternative hypothesis as being one of strong SD2 and the null being the converse. Therefore it is possible to have a distribution dominate another distribution (in a second order sense) almost everywhere and to fail to reject the null hypothesis. Another alternative, proposed by Eubank, Schechtman and Yitzhaki (1993), tests a necessary condition for SD2 and also

[3]The authors have written Gauss procedures that allow one to compute the test statistics and to obtain p-values. These are available on request.

does not test a null hypothesis that corresponds directly to SD2.

The remainder of the paper is structured as follows. In Section 2 we give a statement of the testing problems and provide a characterization of the limiting distributions of the test statistics under the null hypothesis in terms of well known stochastic processes. Additionally in Section 2, we provide critical values for SD1 tests. In Section 3 we present a variety of simulation and bootstrap methods for computing p-values for testing SDj with j larger than 1, and give a theoretical justifications for the methods. Section 4 considers a variety of approaches that are based on using a fixed number of comparisons. In Section 5 we conduct a small scale Monte Carlo experiment to examine the usefulness of the approach in small samples and compare the approach based on KS type tests with methods based on a fixed number of comparisons. In section 6 we illustrate the methods by comparing the Canadian income distributions for 1978 and 1986. Section 7 offers concluding remarks.

2 Hypotheses, Test Statistics and Limiting Distributions

2.1 Stochastic Dominance and Hypothesis Formulation

We focus on a situation in which we have independent samples of income (or some other measure of individual welfare) with possibly different sample sizes, from two populations that have associated cumulative distribution functions (CDFs) given by G and F. Stochastic dominance is closely related to social welfare comparisons as shown, for example, in Deaton (1997). In particular (weak) first order stochastic dominance (hereafter SD1) of G over F corresponds to $G(z) \leq F(z)$ for all z. As noted by Deaton (1997), when this occurs social welfare in the population summarized by G is at least as large as that in the F population for any social welfare function of the form

$$W(H) = \int U(z)dH(z)$$

where H is the distribution of income and U is *any* increasing monotonic function of z – i.e. $U'(z) \geq 0$. On the other hand (weak) second order stochastic dominance (SD2) of

G over F corresponds to $\int_0^z G(t)dt \leq \int_0^z F(t)dt$ for all z and has the implication that the social welfare in the population summarized by G is at least as large as that in the F population for any social welfare function of the form $W(H)$ where U is monotonically increasing and concave – that is $U'(z) \geq 0$ and $U''(z) \leq 0$. Finally, (weak) third order stochastic dominance (SD3) of G over F corresponds to $\int_0^z \int_0^s G(t)dtds \leq \int_0^z \int_0^s F(t)dtds$ for all z and has the implication that the social welfare in the population summarized by G is at least as large as that in the F population for any social welfare function of the form $W(H)$ where U satisfies $U'(z) \geq 0$, $U''(z) \leq 0$ and $U'''(z) \geq 0$.[4]

It is convenient notationally to represent the various orders of stochastic dominance using the integral operator, $\mathcal{I}_j(.;G)$ to be the function that integrates the function G to order $j-1$ so that for example,

$$\begin{aligned}
\mathcal{I}_1(z;G) &= G(z) \\
\mathcal{I}_2(z;G) &= \int_0^z G(t)dt = \int_0^z \mathcal{I}_1(t;G)dt \\
\mathcal{I}_3(z;G) &= \int_0^z \int_0^t G(s)dsdt = \int_0^z \mathcal{I}_2(t;G)dt
\end{aligned}$$

and so on. It is well known that there is a one way relationship between the different forms of stochastic dominance as suggested not only by the functions that are being compared but also by their implications for social welfare. In particular SDj implies SD$(j+1)$ – there is not necessarily a converse relationship.

The preceding discussion is suggestive of hypotheses that could be tested for the various forms of stochastic dominance. Before doing so, and in order to be precise we make the following assumption regarding the two distributions F and G.

Assumption 1: *Assume that*

(i) F and G have common support $[0, \bar{z}]$ where $\bar{z} < \infty$,

(ii) F and G are continuous functions on $[0, \bar{z}]$.

[4]The correspondence between SD and the properties of the social welfare function $W(H)$ extend to any order of SD. That is, SD of order j is equivalent to the quasi-ordering induced by $W(H)$ where U satisfies the set of conditions $(-1)^k.U^{(k)} \leq 0$ for $k = 1, .., j$.

In the context of income distributions it seems natural to have the lower bound on the support of the distribution be equal to zero. The results of the paper do extend to situations where the lower bound is any finite number. Additionally we do not require that any (measurable) set with positive Lebesgue measure have strictly positive probability under either F or G. Thus for instance we allow for the possibility that either $F(z) = 0$ (or $G(z) = 0$) for $z > 0$. It does appear to be crucial for testing SDj with $j \geq 2$ that $\bar{z}$ be finite. This is required because without this the multiple integrals of the CDFs in this case will be infinitely large. If this assumption appears unreasonable then the tests can be thought of as testing the full implications of stochastic dominance on a compact set.

Given our assumptions on the underlying distributions we now state the hypotheses that relate to the various forms of stochastic dominance that we consider. The general hypotheses for testing stochastic dominance of order j can be written compactly as,

$$
\begin{aligned}
H_0^j &: \mathcal{I}_j(z;G) \leq \mathcal{I}_j(z;F) \text{ for all } z \in [0,\bar{z}] \\
H_1^j &: \mathcal{I}_j(z;G) > \mathcal{I}_j(z;F) \text{ for some } z \in [0,\bar{z}]
\end{aligned}
$$

The way that we have formulated the hypotheses is the same as in McFadden (1989) and much of the literature that has considered stochastic dominance. An exception in the case of SD2 is Kaur, Prakasa Rao and Singh (1994) who (in a sense) reverse the roles of the hypotheses and have the alternative hypothesis as corresponding to strong second order dominance with the null being the converse. In such tests the situation where G dominates F (in a second order sense) at all but one point cannot be distinguished from the case where F and G are identical. On the other hand Eubank, Schechtman and Yitzhaki (1993) test a necessary (but not sufficient) condition for SD2.

We should note that weak stochastic dominance (of whatever order) of G over F implies that G is no larger than F (or for second and third order the integrals of these objects) for any value of income – this includes the case where the distributions are equal everywhere. Therefore the null hypotheses are composite in the sense that they are true for many different G functions (with F fixed). The alternative hypothesis in each case is simply the converse of the null and implies that there is at least some income value at which G (or its integral) is strictly larger than F (or its integral). In other words stochastic dominance fails at some point. As formulated, one can in principle distinguish between the case where F and G coincide and the case where G dominates F (in whatever sense) by reversing the roles they play in the hypotheses and redoing the tests. Also note that as stated we consider all values of $z \in [0, \bar{z}]$ the common support of incomes. All of our results can be extended to the case where we compare the objects for any (common) compact subinterval of income values. Such an approach may be useful in the context of poverty comparisons where one focuses on the welfare of the "poor" – see Davidson and Duclos (2000).

2.2 Test Statistics and Asymptotic Properties

In this paper we consider the case where we have independent samples from the two distributions discussed in the previous section. Since we will be allowing for different sample sizes we need to make assumptions about the way in which sample sizes grow. The following gives our assumption on the sampling process.

Assumption 2:

(i) $\{X_i\}_{i=1}^N$ *and* $\{Y_i\}_{i=1}^M$ *are independent random samples from distributions with c.d.f.'s* F *and* G *(respectively),*

(ii) the sampling scheme is such that as $N, M \to \infty$,

$$\frac{N}{N+M} \to \lambda$$

where $0 < \lambda < 1$.

Assumption 2(i) concerns the sampling scheme and would be satisfied if one had samples of incomes (or some other measure of well being) from different segments of a population or separate samples across time.[5] Note Assumption 2(ii) implies that the ratio of the sample sizes is finite and bounded away from zero. Throughout the paper all

[5] Also, for technical reasons the samples are from distributions on a measurable space $(\mathcal{X}, \mathcal{A})$.

limits are taken as N and M grow in such a way that Assumption 2(ii) holds. We have assumed that the supports are finite. Such an assumption is not really needed in the case of tests of first and second order stochastic dominance, but appears to be required in the case of third order stochastic dominance.

The empirical distributions used to construct the tests are respectively,

$$\hat{F}_N(z) = \frac{1}{N}\sum_{i=1}^{N} 1(X_i \leq z),\ \hat{G}_M(z) = \frac{1}{M}\sum_{i=1}^{M} 1(Y_i \leq z)$$

The test statistics for testing these hypotheses can be written compactly using the integration operator as follows:

$$\hat{S}_j = \left(\frac{NM}{N+M}\right)^{1/2} \sup_z (\mathcal{I}_j(z;\hat{G}_M) - \mathcal{I}_j(z;\hat{F}_N))$$

The operator $\mathcal{I}_j$ is a linear operator so one can show that,

$$\begin{aligned} \mathcal{I}_j(z;\hat{F}_N) &= \frac{1}{N}\sum_{i=1}^{N} \mathcal{I}_j(z;1_{X_i}) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(j-1)!} 1(X_i \leq z)(z-X_i)^{j-1} \end{aligned} \quad (1)$$

where the second line follows from Davidson and Duclos (2000) using the notation 1_{X_i} to denote the function $1(X_i \leq x)$. Thus the above statistics can be computed quite simply.

We will be characterizing the limiting distributions of the test statistics under the null hypothesis using the fact that,

$$\sqrt{N}(\hat{F}_N - F) \Rightarrow \mathcal{B}_F \circ F,\ \sqrt{M}(\hat{G}_M - G) \Rightarrow \mathcal{B}_G \circ G$$

where $\mathcal{B}_F \circ F$ and $\mathcal{B}_G \circ G$ are independent Brownian Bridge processes.[6] The following result proves useful in characterizing the behavior of the test statistics and concerns the asymptotic properties of the process that involves integrals of the Brownian Bridges.

[6] Technically we have joint convergence. This type of result can be shown either using Billingsley (1968) or the recent approach outlined in Van der Vaart and Wellner (1996) to show marginal convergence for each process and then using Theorem 1.4.8 of Van der Vaart and Wellner (1996) which shows that since the processes $\mathcal{B}_F \circ F$ and $\mathcal{B}_G \circ G$ are separable joint convergence is equivalent to marginal convergence of each process. It is also straighforward to show that each empirical c.d.f. converges jointly and uniformly to the corresponding population c.d.f.

Lemma 1: *Under Assumption 1 we can show that for $j \geq 2$,*

$$\sqrt{N}(\mathcal{I}_j(.;\hat{F}_N) - \mathcal{I}_j(.;F)) \Rightarrow \mathcal{I}_j(.;\mathcal{B}_F \circ F)$$

where the limit process is mean zero Gaussian with covariance kernel given by (for $z_2 > z_1$),

$$\begin{aligned}\Omega_j(z_1, z_2; F) &= E(\mathcal{I}_j(z_1;\mathcal{B}_F \circ F)\mathcal{I}_j(z_2;\mathcal{B}_F \circ F)) \\ &= \sum_{l=0}^{j-1} \theta_l^j \frac{1}{l!}(z_2 - z_1)^l \mathcal{I}_{2j-1-l}(z_1;F) - \mathcal{I}_j(z_1;F)\mathcal{I}_j(z_2;F)\end{aligned}$$

where $\theta_0^1 = 1$ and

$$\begin{aligned}\theta_0^j &= 2\sum_{k=0}^{j-2} \theta_k^{j-1} \\ \theta_l^j &= \sum_{k=l-1}^{j-2} \theta_k^{j-1} \text{ for } l = 1, .., j-1\end{aligned}$$

The result is presented in terms of the process corresponding to F with the analogous result holding for G. The characterization in Lemma 1 is in a recursive form that is useful for results that follow. An alternative characterization could be obtained using Binomial coefficients and the representation in (1). In particular one can show that

$$\theta_l^j = \binom{2j-2-l}{j-1}$$

and the result in Lemma 1 follows from the recurrent formulae for Binomial coefficients.

We consider tests based on a decision rule of the form,

$$\text{“reject } H_0^j \text{ if } \hat{S}_j > c_j\text{”}$$

where c_j is some critical value that will be discussed later. It is convenient to define the following random variables[7],

$$\begin{aligned}\bar{S}_j^F &= \sup_z \mathcal{I}_j(z;\mathcal{B}_F \circ F) \\ \bar{S}_j^{G,F} &= \sup_z(\sqrt{1-\lambda}\mathcal{I}_j(z;\mathcal{B}_G \circ G) - \sqrt{\lambda}\mathcal{I}_j(z;\mathcal{B}_F \circ F))\end{aligned}$$

The following result characterize the properties of these tests.

[7] For instance $\bar{S}_2^F = \sup_z \int_{-\infty}^z \mathcal{B}(F(t))dt$

Proposition 1: *Given Assumptions 1, 2 and that c_j is a positive finite constant, then*

(A)*(i) if H_0^j is true,*

$$\lim_{N,M\to\infty} P(\text{reject } H_0^j) \leq P(\bar{S}_j^F > c_j) = \alpha_F(c_j),$$

with equality when $F(z) = G(z)$ for all $z \in [0, \bar{z}]$

(ii) if H_0^j is true,

$$\lim_{N,M\to\infty} P(\text{reject } H_0^j) \leq P(\bar{S}_j^{G,F} > c_j) = \alpha_{G,F}(c_j),$$

with equality when $F(z) = G(z)$ for all $z \in [0, \bar{z}]$,

(B) *if H_0^j is false,*

$$\lim_{N,M\to\infty} P(\text{reject } H_0^j) = 1.$$

The result provides two random variables that dominate the limiting random variables corresponding to the test statistic under the null hypothesis. The first is of a simpler form but is harder to prove than the second which is of a more complicated form. The proof of A(i) involves characterizing the distribution of the test statistic and then using the covariance structure shown in Lemma 1 to prove an inequality involving suprema of random Gaussian random variables with a certain covariance structure. One can show the result in A(i) holds for SD1 using the finite sample monotonicity of the power function under transformations from a random variable with a distribution G to another random variable with distribution G^* where G first order stochastically dominates G^*.[8] The finite sample power function for testing SDj for $j \geq 2$ is also monotonic under such transformations, but is not monotonic for transformations from one distribution to another that it dominates to an order that is higher than one. Therefore the asymptotic approach is required.[9]

[8] By this we mean that if $Y_i \sim G$ and $G(z) \leq G^*(z)$ for all z then $Y_i^* = G^{*-1}(G(Y_i)) \sim G^*$ and $Y_i^* \leq Y_i$ and the test statistic based on the Y_i^* will be at least as large as that based on Y_i. See Randles and Wolfe (1979, Theorem 4.3.3) for instance.

[9] To the best of our knowledge a result such as A(i) for SDj with $j \geq 2$ has always been stated without proof or else ignored – see McFadden (1989, p.121) and Schmid and Trede (1998, p.185-186) for instance.

The two random variables will coincide under the null when the distributions coincide. The inequalities in A(i) and A(ii) imply that the tests will never reject more often than $\alpha_F(c_j)$ (respectively $\alpha_{G,F}(c_j)$) for any G satisfying the null hypothesis. As noted in the result, when $F = G$ the probability of rejection will asymptotically be exactly $\alpha_F(c_j)$ (respectively $\alpha_{G,F}(c_j)$) and, moreover, $\alpha_F(c_j) = \alpha_{G,F}(c_j)$ because of the fact that $\bar{S}_j^{G,F} \stackrel{d}{=} \bar{S}_j^F$ (see Shorack and Wellner (1986) for instance). It also interesting to note that if $\mathcal{I}_j(z;G) < \mathcal{I}_j(z;F)$ for all z above $\inf_z\{z : F(z) > 0\}$ (which under the null hypothesis must be no larger than $\inf_z\{z : G(z) > 0\}$ for any order) then asymptotically the probability of rejection will be zero. The results in A(i) and A(ii) imply that if one could find a c_j to set the $\alpha_F(c_j)$ (respectively $\alpha_{G,F}(c_j)$) to some desired level (say 0.05 or 0.01) then this would be the significance level for composite null hypotheses in the sense described in Lehmann (1986). Moreover, the result in B implies that the tests are capable of detecting any violation of the full set of implications of the null hypothesis.

In order to make the result operational we need to find, in each case, an appropriate critical value, say c_j, to satisfy $P(\bar{S}_j^F > c_j) = \alpha$ or $P(\bar{S}_j^{G,F} > c_j) = \alpha$. As has been noted elsewhere (see McFadden (1989) for instance) however, this is only easily done in the case of SD1 tests for the limiting random variable in A(i). Since first order stochastic dominance is invariant to monotone transformations one can show that,[10]

$$P(\bar{S}_1^F > c) = P(\sup_{p\in[0,1]} \mathcal{B}(p) > c) = \exp(-2c^2) \tag{2}$$

Thus one can either compute a p-value by $\exp(-2(\hat{S}_1)^2)$ or else critical values can be obtained by inversion using $c_1(\alpha) = (-\frac{1}{2}\log\alpha)^{1/2}$. Some important critical values are 1.073, 1.2239 and 1.5174 for the 10%, 5% and 1% levels of significance respectively. The characterization in A(ii) is less useful in the case of SD1 because in general the distribution of $\bar{S}_1^{G,F}$ will depend on F and G and although the simulation methods proposed in the next section can be used to obtain approximate p-values, there seems little point when one already has an alternative with an analytic asymptotic distribution.

[10]See Billingsley (1968, p.85) for details. Note that the asymptotic distribution implied by equation 6 of McFadden (1989) differs from that presented here and appears to be due to a typographical error.

For testing orders of dominance beyond the first the distributions will depend on the underlying distributions. In particular $\bar{S}_j^F$ will depend on F while $\bar{S}_j^{G,F}$ will depend on both G and F. The approach taken in this paper is to use simulation methods as well as bootstrap methods to simulate p-values. Because of the fact that in general one cannot compare the random variables $\bar{S}_j^F$ and $\bar{S}_j^{G,F}$ (except in the case $G = F$) one cannot tell *a priori* which bound will result in a better test in terms of power as well as size. This will be addressed when we examine the performance of the various tests in the context of a Monte Carlo experiment. Finally, before proceeding, we note that the bound based on $\bar{S}_j^F$ is of a simpler form and that performing inference based on this bound will be less demanding computationally and that one can potentially test H_0^j with F fixed for a number of other distributions using one set of simulations.

3 Simulating P-Values

3.1 Multiplier Methods

In this section we consider the use of a simulation or Monte Carlo method for conducting inference for the tests that is similar to that used in Hansen (1996). It involves the use of artificial random numbers and exploits the multiplier central limit theory discussed in Van der Vaart and Wellner (1996) to simulate a process that is identical to but (asymptotically) independent of $\mathcal{B}(F(z))$. To do this let $\{U_i\}_{i=1}^N$ denote a sequence of i.i.d. $N(0,1)$ random variables that are independent of the samples. For each value of $z \in [0, \bar{z}]$ let, $\mathcal{B}^* \circ \hat{F}_N$ evaluated at z be given by,

$$\mathcal{B}^*(z; \hat{F}_N) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (1(X_i \leq z) - \hat{F}_N(z)) U_i \tag{3}$$

We can similarly define a simulated version of the Brownian Bridge corresponding to G using an independent set of draws (say $\{V_i\}_{i=1}^M$ with $V_i \sim i.i.d.\ N(0,1)$) and we denote the process by $\mathcal{B}^* \circ \hat{G}_M$. The method for doing inference consists of obtaining p-values from appropriate functionals of the simulated processes. These p-values can be obtained using either of the following two calculations that correspond respectively to Proposition

1A(i) and Proposition 1A(ii),

$$\hat{p}_j^F = P_U(\sup_z \mathcal{I}_j(z; \mathcal{B}^* \circ \hat{F}_N) > \hat{S}_j) \quad (4)$$

$$\hat{p}_j^{F,G} = P_U(\sup_z(\sqrt{1-\hat{\lambda}}\mathcal{I}_j(z; \mathcal{B}^* \circ \hat{G}_M) - \sqrt{\hat{\lambda}}\mathcal{I}_j(z; \mathcal{B}^* \circ \hat{F}_N) > \hat{S}_j) \quad (5)$$

where $P_U(.)$ is the probability function associated with the normal random variables U_i and is conditional on the realized sample(s). Note that these p-values depend on the sample sizes N and M although we have suppressed the dependence for notational convenience. The following result provides a justification for this approach.

Proposition 2: *Given Assumptions 1, 2 and assuming that* $\alpha < 1/2$, *a test for SDj based on either the rule,*

"reject H_0^j if $\hat{p}_j^F < \alpha$"

or,

"reject H_0^j if $\hat{p}_j^{F,G} < \alpha$"

satisfies the following,

$$\lim P(\text{reject } H_0^j) \leq \alpha \text{ for } F, G \text{ satisfying } H_0^j$$

$$\lim P(\text{reject } H_0^j) = 1 \text{ for } F, G \text{ satisfying } H_1^j$$

The p-value method can be justified by showing that these simulated processes converge weakly (almost surely)[11] to identical independent copies of the respective Brownian Bridge and an application of the continuous mapping theorem which shows that we have simulated copies of the bounding random variables that appear in Proposition 1. The result is obtained in a manner that is similar to a part of the proof of Theorem 2 of Hansen (1996). The main difference is that in our case we must deal with the fact that

[11] We show weak convergence conditional on the original samples of observations on X and Y and show that the convergence is for almost all samples. We call this weak convergence (almost surely). In connection with the bootstrap we consider the concept of weak convergence (in probability) as used in Hansen (1986). Formal definitions of these convergence notions are contained in Van der Vaart and Wellner (1996).

we have a one sided composite null and the fact that the test statistic may have a degenerate distribution at zero for some cases satisfying the null hypothesis. The result implies that a test based on the decision rule "reject SD2 if $\hat{p}_j < \alpha$" will reject a true null hypothesis with probability that is (asymptotically) no larger than α. The probability will be (asymptotically) equal to α when in fact $F = G$ (in which case the inequalities in the statement of Proposition 1 hold with equality).

In order to compute the p-values in practice we must deal with the fact that the probabilities in (4) and (5) must be calculated and that the suprema that define the relevant random variables must be calculated. As suggested by Hansen (1996) we use Monte-Carlo methods to approximate the probability and use a grid to approximate the suprema. Since these are under the control of the econometrician one can make the approximations as accurate as one wants given time and computer constraints.

More specifically let $\{U_i^r\}_{i=1}^N$ $\{V_i^r\}_{i=1}^M$denote the rth samples of U_i and V_i where we will let $r = 1, ..., R$ where R will denote the number of replications that will be used in the Monte Carlo simulation. Select a grid of values on $[0, \bar{z}]$ such as $0 = t_0 < t_1 < ... < t_K = \bar{z}$, where K will denote the number of subintervals. Using (1) we can approximate the rth realization of the statistic by,

$$
\begin{aligned}
\bar{S}_{j,r}^F &= \max_{t_k} \frac{1}{\sqrt{N}} \sum_{i=1}^N (\mathcal{I}_j(t_k; 1_{X_i}) - \mathcal{I}_j(t_k; \hat{F}_N))U_i^r \\
\bar{S}_{j,r}^{F,G} &= \max_{t_k} \sqrt{\frac{NM}{N+M}} \sum_{i=1}^N \Big((\mathcal{I}_j(t_k; 1_{Y_i}) - \mathcal{I}_j(t_k; \hat{G}_M))V_i^r - (\mathcal{I}_j(t_k; 1_{X_i}) - \mathcal{I}_j(t_k; \hat{F}_N))U_i^r \Big)
\end{aligned}
$$

Then the p-values can be approximated by,

$$\hat{p}_j^F \simeq \frac{1}{R} \sum_{r=1}^R 1(\bar{S}_{j,r}^F > \hat{S}_j). \tag{6}$$

$$\hat{p}_j^{F,G} \simeq \frac{1}{R} \sum_{r=1}^R 1(\bar{S}_{j,r}^{G,F} > \hat{S}_j) \tag{7}$$

As indicated by Hansen (1996), an appeal to the Central Limit Theorem suggests that the error in approximating $\hat{p}_j$ should have a standard error that is approximately no larger than $(4R)^{-1/2}$ so that if $R = 1000$ (or say 10,000) for instance the standard error in this approximation is roughly 0.015 (or 0.005 when $R = 10,000$) and much smaller in cases where $\hat{p}_j$ is close to zero.

3.2 Bootstrap

A natural alternative to the p-value simulation method is to conduct inferences using a form of the bootstrap. A possible advantage of this is that, although existence of a limiting distribution (for the test statistic) is generally needed, one does not necessarily need to be able to characterize it in the way that we were able to in the previous section. Therefore the bootstrap may be applicable in more complicated situations. As in the previous section we provide methods for bootstrapping based on the results in Proposition 1A(i) and 1A(ii). The first method, based on Proposition 1A(i), is to simulate the random variable corresponding to $\bar{S}_F^j$. In this case we define the sample as $\mathcal{X} = \{X_1, ..., X_N\}$ and compute the distribution of the random quantity,

$$\bar{S}_j^F = \sqrt{N} \sup_z (\mathcal{I}_j(z; \hat{F}_N^*) - \mathcal{I}_j(z; \hat{F}_N)) \tag{8}$$

where,

$$\hat{F}_N^*(z) = \frac{1}{N} \sum_{i=1}^{N} 1(X_i^* \leq z)$$

for a random sample drawn from $\mathcal{X}$. To simulate the random variable corresponding to $\bar{S}_{F,G}^j$ from Proposition 1A(ii) we follow Van der Vaart and Wellner (1996) and resample from the combined samples. Define the combined samples as $\mathcal{Z} = \{X_1, ..., X_N, Y_1, ..., Y_M\}$. Let $\hat{G}_M^*$ denote the empirical CDF of a random sample of size M from $\mathcal{Z}$ and let $\hat{F}_N^*$ denote the empirical CDF of an independently drawn random sample of size N from $\mathcal{Z}$. Then compute the distribution of the random quantity[12],

$$\bar{S}_{j,1}^{F,G} = \sqrt{\frac{NM}{N+M}} \sup_z (\mathcal{I}_j(z; \hat{G}_M^*) - \mathcal{I}_j(z; \hat{F}_N^*)) \tag{9}$$

One can justify a third method of bootstrapping by drawing samples of size N from $\mathcal{X}$ (with replacement) to construct an estimate $\hat{F}_N^*$, and independently drawing samples of

[12] Abadie (2000) has considered the use of this method of bootstrapping for the case of SD1 and SD2. Maasoumi and Heshmati (2000) have also considered bootstrapping for similar tests of stochastic dominance.

size M from $\mathcal{Y}$ to construct an estimate $\hat{G}^*_M$ and computing the statistic,

$$\bar{S}^{F,G}_{j,2} = \sqrt{\frac{NM}{N+M}} \sup_z((\mathcal{I}_j(z;\hat{G}^*_M) - \mathcal{I}_j(z;\hat{G}_M)) - (\mathcal{I}_j(z;\hat{F}^*_N) - \mathcal{I}_j(z;\hat{F}_N))) \tag{10}$$

In each case we are interested in computing the probability that the random variables exceed the value of the statistic given the respective samples. These can be approximated by Monte Carlo simulation in a manner that is exactly analogous to (6) and (7). Denote the respective p-values by the notation $\tilde{p}^F_j$, $\tilde{p}^{F,G}_{j,1}$ $\tilde{p}^{F,G}_{j,2}$ The following result provides a justification for this approach.

Proposition 3: *Let Assumptions 1, 2 hold and assume that $\alpha < 1/2$ then a test for SDj based on any of the rules,*

"reject H^j_0 if $\tilde{p}^F_j < \alpha$"

"reject H^j_0 if $\tilde{p}^{F,G}_{j,1} < \alpha$"

"reject H^j_0 if $\tilde{p}^{F,G}_{j,2} < \alpha$"

satisfies the following,

$$\begin{aligned} \lim P(\text{reject } H^j_0) &\leq \alpha \text{ if } H^j_0 \text{ is true} \\ \lim P(\text{reject } H^j_0) &= 1 \text{ if } H^j_0 \text{ is false} \end{aligned}$$

4 Tests Based on Multiple Comparisons

We can now contrast our approach with a variety of approaches based on Anderson (1996) and Davidson and Duclos (2000). Defining $\Delta_j(z_l) = \mathcal{I}_j(z_l;G) - \mathcal{I}_j(z_l;F)$, the methods considered in Davidson and Duclos (2000) are designed to test

$$\begin{aligned} H^j_0 &: \Delta_j(z_l) \leq 0 \text{ for all } l \in \{1,...,k\} \\ H^j_1 &: \Delta_j(z_l) > 0 \text{ for some } l \in \{1,...,k\} \end{aligned}$$

where the subscript j indicates the order of stochastic dominance being tested. It is clear that the hypothesis being tested only relates to dominance at a fixed number of points and hence is different from the hypothesis tested in the previous section. Nevertheless, these tests have been used to draw conclusions as to the truth or falsehood of the hypotheses described in Section 2.1. However, tests based on multiple comparisons will lack power in some situations since they fail to examine all of the implications of stochastic dominance. Specifically, the multiple comparisons paper will have low power where there is a violation of the inequality in the null hypothesis on some subinterval lying between income evaluation values – i.e. on some subinterval in (z_l, z_{l+1}).[13]

Davidson and Duclos (2000) consider two types of tests. The first is essentially a Wald test[14]. Define $\hat{\Delta}_j$ as the k vector of estimates of $\Delta_j(z_l)$ and $\hat{\Omega}_j$ as the estimate of the variance covariance matrix[15] of $\hat{\Delta}_j$, then the Wald test can be obtained by

$$\hat{W}_j = \min_{\Delta \in R_+^k} \{(\hat{\Delta}_j - \Delta)' \hat{\Omega}_j^{-1} (\hat{\Delta}_j - \Delta) : \Delta \leq 0\}$$

As has been shown in Wolak (1989) the Wald statistic has an asymptotic distribution that is a mixture of chi-squared random variables. As with the consistent tests proposed above, simulation is required in order for inference to be possible unless k is small. In particular, one must compute a large number of quadratic programming problems in order to estimate the weights that appear in the chi-squared mixture limiting distribution (see Wolak (1989, p. 213)).

A simpler approach to testing the hypotheses is to simply use the t-statistics that have been calculated for testing whether each $\Delta_j(z_l)$ is zero against the alternative that it is larger than zero. Let the individual t-statistics be given by $\hat{t}_j(z_l) = \hat{\Delta}_l(z_j)/\sqrt{\hat{\Omega}_{jj}}$. A simple test can then be performed by rejecting the null hypothesis if the largest t-statistic is large. Thus one could use the test statistic

$$\hat{S}_j^{MT} = \max_l \{\hat{t}_j(z_l)\}.$$

[13]Indeed, it is also possible (although perhaps a somewhat perverse case) that F stochastically dominates G (to what ever order) almost everywhere and to still have this implicit null hypothesis (in terms of multiple comparisons) being true. This would occur if $\Delta_l(z_j) = 0$ for all j and $\Delta_l(z) > 0$ for all $z \neq z_j$.

[14]See also Kudo (1963), Perlman (1969) and Gourieroux, Holly and Monfort (1982).

[15]Davidson and Duclos (2000) propose using an estimate of the variance that is the sum of the variances of the components that make up the difference $\hat{\Delta}_j(z_l)$. This can be justified if one has independent samples.

where the superscript MT indicates that this is a maximal t-statistic. This is analogous to the KS type tests in the context of the simpler hypotheses considered in this section. As noted by Davidson and Duclos (2000), this statistic has a non-standard distribution.[16] Given that the Wald test considered by Davidson and Duclos (2000) requires simulation for inference it only seems fair to consider the possibility of simulating p-values for the maximal t statistic test. A simple procedure for simulating the p-value for the maximal t-statistic is to use,

$$\hat{p}_j^{MT} = \frac{1}{R}\sum_{s=1}^{R} 1(\max\{\hat{\Gamma}_j^{1/2} Z_s\} > \hat{S}_j^{MT})$$

where $\hat{\Gamma}_j^{1/2}$ is the Cholesky decomposition of a consistent estimate of Γ_j (the correlation matrix corresponding to Ω), Z_s are multivariate standard normal pseudo-random numbers that can be generated on a computer, R is the number of random draws used to estimate the p-value and the max operator takes the largest value in the vector $\hat{\Gamma}_j^{1/2} Z_s$. Such an approach can be justified using the arguments presented in Section 3.

The approach of Anderson (1996) is similar in spirit to the approach based on the maximal t-statistic. One minor difference is that Anderson proposes estimating the variance under the assumption that the $\Delta_1(z_l)$ are all zero. A more important difference lies in the way that Anderson (1996) computes the $\hat{\Delta}_j(z_l)$ for $j = 2, 3$. In particular Anderson (1996) approximates the integrals that define $\Delta_j(z_l)$ (for $j = 2, 3$) by using an approximation (specifically a trapezoidal rule, as in Goodman (1967)) of the integral of an approximation to the (differences in the) empirical CDF. The method of Davidson and Duclos (2000) and the method we have proposed in Section 2 are both based on integrating the empirical CDF directly – which provide unbiased estimates. In contrast, the approximations used in Anderson produce estimates at the evaluation points that are potentially biased and inconsistent. Although these potential biases will generally not lead one to reject a true null they do introduce a further possibility that one could fail to reject a false null hypothesis for orders of dominance beyond SD1. It should be

[16]Davidson and Duclos suggested that a conservative critical value can be obtained from the studentized maximum modulus (SMM) distribution. The SMM distribution is the distribution of the maximum of a certain number of independent t statistics which provides the distribution of the quantity $\max_j |Z_j|$ when there are infinite degrees of freedom. See Stoline and Ury (1979).

noted that conducting inferences for the Anderson tests is exactly analogous to the case of the maximal t statistic approach based on the DD calculations – one can obtain widely applicable conservative critical values or simulate p-values.

5 Monte Carlo Results

In this section we consider a small scale Monte Carlo experiment in which we gauge the extent to which the preceding asymptotic arguments in Sections 2 and 3 hold in small samples. In addition we compare the tests proposed in Sections 2 and 3 (referred to as KS or Kolmogorov-Smirnov tests) with the other tests considered in Section 4. In a sense such a comparison is unfair because, as noted in the previous section, the KS tests and the tests based on multiple comparisons are testing different null hypotheses. In terms of the worth of the multiple comparison tests we would expect them to do well when the income values (at which comparisons are made) are chosen so that the differences in the objects being compared are representative of the overall ranking. One would expect the multiple comparison test to have good power (and indeed dominate the KS tests) when the differences at the income values used are close to the largest overall difference.[17]

We designed our experiment in an effort to mimic reality by using a class of distributions with shapes similar to those that have been found to work well in income distribution studies. In particular we used the log-normal distribution and in each experiment generate the two samples using the following,

$$\begin{aligned} X_i &= \exp(\sigma_1 Z_{1i} + \mu_1) \\ Y_i &= \exp(\sigma_2 Z_{2i} + \mu_2) \end{aligned}$$

where the Z_{1i} and Z_{2i} are independent $N(0,1)$ random variables, $(\mu_1, \mu_2, \sigma_1, \sigma_2)$ are parameters that will be varied across the different experiments. Five different cases were

[17] This is similar to the situation of examining tests of parametric restrictions with nonparametric alternatives. One would expect a test based on a parametric alternative to do better when the chosen alternative is close to the truth. Nevertheless, given that the alternative is unknown in practice, the lack of consistency of such a testing is a cause for concern. In our case this would correspond to having prior information as to the location of the largest difference between the objects being compared.

considered. Case 1 uses the values $\mu_1 = \mu_2 = 0.85$ and $\sigma_1 = \sigma_2 = 0.6$.[18] Our results suggest that if this is the case then the tests (designed for a particular nominal significance level) should reject the various null hypotheses with a relative frequency that is close to the nominal significance level. The extent to which this is satisfied gives us an idea of the extent to which the asymptotic theory holds, and the extent to which the p-value simulation and bootstrap methods work, in small samples. The second case, Case 2, involves the specification $\mu_2 = 0.6$ and $\sigma_2 = 0.8$ with μ_1 and σ_1 being the same as in Case 1. With this specification all three null hypotheses are false since there are regions of the support of the distributions over which the inequality in the null hypothesis is invalid. Case 3 has $\mu_2 = 1.2$ and $\sigma_2 = 0.2$. In this case H_0^1 fails by a very small amount but both H_0^2 and H_0^3 are true so we should expect to reject SD1 but not reject SD2 or SD3. For Case 4 Y is generated as a mixture of log-normal variables such that

$$Y_i = 1(U_i \geq 0.1)\exp(\sigma_2 Z_{2i} + \mu_2) + 1(U_i < 0.1)\exp(\sigma_3 Z_{3i} + \mu_3)$$

where U_i is a uniform $[0,1]$ random variable, Z_{2i} and Z_{3i} are independent standard normal random variables with $\mu_2 = 0.8$, $\sigma_2 = 0.5$, $\mu_3 = 0.9$ and $\sigma_3 = 0.9$. In this case the relevant curves for SD1-SD3 exhibit a single crossing and hence all of the null hypotheses are false. Finally, for Case 5 Y is also generated as a mixture of log-normal variables with $\mu_2 = 0.85$, $\sigma_2 = 0.4$, $\mu_3 = 0.4$ and $\sigma_3 = 0.9$. In this case the distribution functions exhibit multiple crossings and all of the null hypotheses are false.

For the alternative tests outlined in Section 4 we need to decide on the income values at which the various objects will be calculated. It is clear that such a choice will determine the extent to which these tests will agree or disagree with the KS tests. Following Anderson (1996) we use income values that were equal to the income deciles in the combined samples.[19] Since the last decile is the largest income value (at which both

[18]With this choice of parameters the variables X_i and Y_i have a mean of 2.8 and standard deviation equal to 1.8 which implies a mean standard deviation ratio that is similar to what is found in empirical studies (see Table I of Anderson (1996) for instance).

[19]The income values selected for the multiple comparison tests are sample determined (based on deciles or quintiles) and hence are stochastic. The multiple comparison tests do not take this source of randomness into account. An alternative approach, supported by the distributional theory underlying

empirical CDFs are one) the SD1 test is then based on the comparison of empirical CDFs at nine income deciles while the other tests are based on all 10. The tests based on these income deciles and using p-value simulations are referred to as MT10, W10 and MTA10 for the maximal t-test, the Wald test and the Anderson computation of the maximal t-test, respectively. We also considered these tests based on quintiles, which we refer to as MT5, W5 and MTA5, to gauge the effect of altering the comparison values on the tests properties.

The KS test for SD1 is performed using critical values obtained from (2). In performing the tests using the p-values for SD2 and SD3 we use the decision rule,

$$\text{“reject } H_0^j \text{ if } \hat{p}_j < \alpha\text{”}$$

where $\hat{p}_j$ is the p-value for the test statistic. In computing the p-values for the simulation based KS tests of SD2 and SD3 the grid was chosen as $0 < t_1 < t_2 < ... < t_K$, with the values being evenly spaced and where t_K is the largest value in the X_i sample. The number of gridpoints was fixed at $K = 100$. The simulation methods are referred to as KS1 and KS2 for the methods that are based on (4) and (5) respectively. The bootstrap methods are referred to as KSB1, KSB2 and KSB3 for the p-value calculations based on (8), (9) and (10) respectively.

A total of 1000 Monte Carlo replications were performed and the rejection rates were computed for each test and for the two conventional significance levels of 0.05 and 0.01. We considered two sample sizes of $N = M = 50$ and $N = M = 500$.[20] The six tables I(A)-III(B) report the results with the label I, II or III referring to the type of test (SD1, SD2 or SD3 respectively) and the label A and B referring to the nominal significance levels 0.05 and 0.01 respectively.

---

these tests, is to use a fixed set of values covering the range of income. We also implemented this approach in the Monte Carlo experiments and found the power of the MT, W and MTA tests to be considerably less than that of the KS tests.

[20] A total number of $R = 1000$ replications was used to simulate the the p-values for the tests.

Table I-A:SD1 Tests, $\alpha = 0.05$

	N=M=50					N=M=500				
	Case1	Case2	Case3	Case4	Case5	Case1	Case2	Case3	Case4	Case5
KS1	0.033	0.477	0.002	0.071	0.097	0.050	1.000	0.830	0.469	0.923
MT(10)	0.053	0.635	0.139	0.134	0.111	0.046	1.000	0.916	0.511	0.622
MT(5)	0.034	0.544	0.015	0.103	0.086	0.054	1.000	0.011	0.464	0.554
W(10)	0.065	0.648	0.139	0.141	0.229	0.052	1.000	0.872	0.495	0.984
W(5)	0.043	0.595	0.020	0.128	0.168	0.049	1.000	0.005	0.452	0.916
MTA(10)	0.046	0.622	0.160	0.139	0.103	0.044	1.000	0.786	0.566	0.651
MTA(5)	0.037	0.531	0.011	0.096	0.067	0.046	1.000	0.020	0.521	0.580

Table I-B:SD1 Tests, $\alpha = 0.01$

	N=M=50					N=M=500				
	Case1	Case2	Case3	Case4	Case5	Case1	Case2	Case3	Case4	Case5
KS1	0.008	0.216	0.000	0.018	0.021	0.012	1.000	0.379	0.224	0.729
MT(10)	0.014	0.361	0.041	0.032	0.030	0.008	1.000	0.788	0.290	0.391
MT(5)	0.012	0.333	0.005	0.025	0.026	0.013	1.000	0.003	0.234	0.298
W(10)	0.018	0.408	0.040	0.035	0.072	0.009	1.000	0.709	0.283	0.932
W(5)	0.013	0.369	0.005	0.031	0.051	0.014	1.000	0.001	0.229	0.786
MTA(10)	0.009	0.302	0.039	0.020	0.016	0.010	1.000	0.786	0.299	0.415
MTA(5)	0.010	0.296	0.003	0.031	0.015	0.009	1.000	0.005	0.276	0.330

Table II-A:SD2 Tests, $\alpha = 0.05$

	N=M=50					N=M=500				
	Case1	Case2	Case3	Case4	Case5	Case1	Case2	Case3	Case4	Case5
KS1	0.034	0.312	0.000	0.090	0.120	0.042	0.992	0.000	0.449	0.865
KS2	0.048	0.254	0.000	0.136	0.249	0.050	0.960	0.000	0.433	0.911
KSB1	0.042	0.334	0.000	0.086	0.172	0.043	0.996	0.000	0.479	0.875
KSB2	0.060	0.241	0.000	0.138	0.241	0.047	0.983	0.000	0.457	0.911
KSB3	0.067	0.269	0.000	0.163	0.269	0.045	0.983	0.000	0.475	0.927
MT(10)	0.053	0.729	0.000	0.084	0.143	0.055	1.000	0.000	0.309	0.553
MT(5)	0.048	0.721	0.000	0.097	0.145	0.053	1.000	0.000	0.297	0.564
W(10)	0.055	0.731	0.000	0.085	0.098	0.058	1.000	0.000	0.298	0.788
W(5)	0.047	0.733	0.000	0.092	0.112	0.054	1.000	0.000	0.288	0.811
MTA(10)	0.047	0.719	0.000	0.069	0.144	0.049	1.000	0.000	0.209	0.754
MTA(5)	0.052	0.684	0.000	0.074	0.152	0.049	1.000	0.000	0.226	0.680

Table II-B:SD2 Tests, $\alpha = 0.01$

	N=M=50					N=M=500				
	Case1	Case2	Case3	Case4	Case5	Case1	Case2	Case3	Case4	Case5
KS1	0.005	0.091	0.000	0.017	0.014	0.009	0.874	0.000	0.174	0.598
KS2	0.012	0.090	0.000	0.043	0.066	0.013	0.713	0.000	0.177	0.761
KSB1	0.007	0.123	0.000	0.017	0.039	0.007	0.892	0.000	0.194	0.593
KSB2	0.013	0.083	0.000	0.041	0.065	0.008	0.742	0.000	0.195	0.664
KSB3	0.016	0.099	0.000	0.043	0.079	0.009	0.755	0.000	0.197	0.710
MT(10)	0.014	0.451	0.000	0.020	0.037	0.012	1.000	0.000	0.134	0.306
MT(5)	0.011	0.483	0.000	0.022	0.038	0.014	1.000	0.000	0.126	0.304
W(10)	0.014	0.468	0.000	0.017	0.018	0.011	1.000	0.000	0.129	0.534
W(5)	0.010	0.483	0.000	0.019	0.023	0.012	1.000	0.000	0.126	0.582
MTA(10)	0.010	0.423	0.000	0.017	0.030	0.007	1.000	0.000	0.066	0.533
MTA(5)	0.010	0.372	0.000	0.017	0.026	0.011	1.000	0.000	0.067	0.399

Table III-A:SD3 Tests, $\alpha = 0.05$

	N=M=50					N=M=500				
	Case1	Case2	Case3	Case4	Case5	Case1	Case2	Case3	Case4	Case5
KS1	0.050	0.359	0.000	0.151	0.128	0.051	0.933	0.000	0.638	0.790
KS2	0.048	0.342	0.000	0.132	0.177	0.048	0.898	0.000	0.435	0.831
KSB1	0.039	0.377	0.000	0.130	0.138	0.046	0.982	0.000	0.620	0.823
KSB2	0.058	0.332	0.000	0.137	0.178	0.052	0.901	0.000	0.439	0.815
KSB3	0.056	0.331	0.000	0.136	0.183	0.045	0.904	0.000	0.436	0.825
MT(10)	0.044	0.729	0.000	0.078	0.131	0.060	1.000	0.000	0.282	0.431
MT(5)	0.046	0.745	0.000	0.085	0.137	0.059	1.000	0.000	0.298	0.442
W(10)	0.048	0.731	0.000	0.082	0.067	0.058	1.000	0.000	0.273	0.674
W(5)	0.049	0.746	0.000	0.085	0.074	0.060	1.000	0.000	0.293	0.695
MTA(10)	0.053	0.734	0.000	0.073	0.158	0.051	1.000	0.000	0.187	0.747
MTA(5)	0.053	0.682	0.000	0.067	0.157	0.050	1.000	0.000	0.200	0.686

Table III-B:SD3 Tests, $\alpha = 0.01$

	N=M=50					N=M=500				
	Case1	Case2	Case3	Case4	Case5	Case1	Case2	Case3	Case4	Case5
KS1	0.011	0.160	0.000	0.053	0.031	0.010	0.771	0.000	0.408	0.574
KS2	0.014	0.141	0.000	0.036	0.053	0.011	0.685	0.000	0.188	0.635
KSB1	0.011	0.168	0.000	0.046	0.043	0.008	0.945	0.000	0.409	0.501
KSB2	0.010	0.123	0.000	0.032	0.039	0.009	0.706	0.000	0.188	0.511
KSB3	0.013	0.136	0.000	0.045	0.066	0.011	0.700	0.000	0.190	0.521
MT(10)	0.010	0.468	0.000	0.015	0.028	0.011	1.000	0.000	0.120	0.206
MT(5)	0.010	0.484	0.000	0.019	0.036	0.009	1.000	0.000	0.130	0.221
W(10)	0.011	0.467	0.000	0.016	0.016	0.014	1.000	0.000	0.112	0.418
W(5)	0.010	0.488	0.000	0.018	0.017	0.010	1.000	0.000	0.117	0.449
MTA(10)	0.012	0.429	0.000	0.015	0.036	0.007	1.000	0.000	0.056	0.534
MTA(5)	0.010	0.383	0.000	0.022	0.027	0.010	1.000	0.000	0.065	0.402

Some basic observations can be made regarding the properties of the different test procedures. First, the p-value simulation method works quite well for all tests. The Case 1 columns in all the Tables contain rejection rates that are close to the nominal significance levels for all the tests for almost all the hypotheses. If anything there seems to be slight under-rejection but even this is small if and when it exists.[21]

All tests perform very well in Case 2. Recall that in this case the null hypothesis is false and one should expect to reject the null hypothesis often. This appears to happen in even small samples for all the tests. Indeed, when the sample size is set to 500 all the multiple comparison tests reject the null 100 per cent of the time. It is interesting to note that in this case using fewer income values does lead to a deterioration in the power of the MT, W or MTA tests – this is because the maximal values of the $\Delta_l(z_j)$ occur near the overall quintiles so that going from evaluation at deciles to quintiles has little effect on the power of these procedures. Although the KS tests have rejection rates which are lower than that of the multiple comparison tests in this case, all the KS tests exhibit very good power. Among the KS tests, the KS1 and KSB1 tests have greater power than the KS2, KSB2 and KSB3 tests in detecting the violation of the null of both SD2 and SD3 in this case.

Case 3 illustrates nicely the potential sensitivity of the MT and Wald tests to the points at which the statistics are evaluated. In particular it is noteworthy that when one goes from comparisons at deciles to comparisons at quintiles the test for SD1 loses power completely – recall that Case 3 is one where SD1 fails but where the other hypotheses are true. Indeed, when the sample size is 500 the test goes from rejecting at a rate of about 90% to a rejection rate that is less than the nominal size of the test. When deciles are used, however, the test does seem to have a higher rejection rate than the KS tests.

[21] The KS1 and KSB1 methods were implemented taking the supremum of the simulated process over $[0, \max_i\{\max\{X_i, Y_i\}\}]$ rather than the region $[0, \max_i\{X_i\}]$. This only has consequences for the tests of SD3 (more generally SDj for $j > 2$) where, unlike SD1 and SD2, the value of the test statistic will depend on the relative size of the objects beyond the largest value in the X_i sample. Taking the supremum over the range of values in the X sample tended to lead to over-rejection of the null hypothesis for SD3. However, taking the supremum over the range of values in the combined X and Y sample resulted in the KS1 and KSB1 tests of SD3 having nominal size very close to the actual size.

In Case 3 the population values of the $\Delta_l(z_j)$ are all negative and the theory in Section 2 suggests that one should reject the null at a rate that is less than the nominal size of the test. This appears to be supported by the results of the tests of the SD2 and SD3 hypotheses in this case. We never reject the null hypothesis using any test.

Some interesting features of the tests are evident in Case 4 . The violation of SD1 occurs at an income value close to the first quintile. The crossing of the functions for SD2 and SD3 occur at progressively higher income values, with the violation of SD3 near the fourth quintile. The extent of the violation of the null hypothesis is also progressively less for higher orders of SD. All the tests perform reasonably well in detecting the violation of SD1, with the multiple comparison tests evaluated at the sample deciles having the highest rejection rates. The KS tests perform better than the other tests at detecting the violation of SD2, particularly when the sample size is 500. The KS tests clearly have superior performance in detecting the violation of SD3, with higher rejection rates than the multiple comparison tests. Among the KS tests, the KS1 and KSB1 tests, which are easier to compute, perform relatively better than the other KS.

Among the multiple comparison tests of SD2 and SD3 for Case 4, the MT and W have similar performance while the MTA tests have the lowest rejection rates. The inferior performance of the MTA tests can be attributed to the fact that the approximations used in the calculation of the MTA tests introduced some bias and inconsistency into the calculation of the integrals that define the objects compared in the SD2 and SD3 tests. Thus with the MTA tests there is the potential inconsistency arising from comparing objects at a fixed number of income values plus there is the additional source of inconsistency introduced through the approximation method used in their calculation.

For Case 5, with multiple crossings of the functions defining SD1 and SD2, and a single crossing in the lower tail of the functions defining SD3, the KS tests appear to have very good power. The W tests perform marginally better than the KS tests for SD1, but the KS tests have greater power at detecting violations for SD2 and especially the violation for SD3. The W tests generally perform better than the MT and MTA tests at detecting the violations of SD1-SD3 in this case.

Overall the results suggest that the KS tests have some merit. While there may be some cost in terms of power and computational time, the tests do a fairly good job of detecting any departure from the properly specified null hypothesis. In addition they circumvent the need for one to arbitrarily choose a set of income values at which objects are to be compared as required by the other methods. Among the other methods the results suggest that there is little to distinguish between the maximal t-test and the Wald test either in terms of size or power across most of the cases considered, though the Wald test had more desirable properties in the case of multiple crossings of distribution (and related) functions. The MTA tests based on the calculations of Anderson (1996), while simpler still than the MT (and W) test, appear to have some undesirable properties when testing SD2 and SD3 hypotheses, principally because of the way they are calculated.[22]

6 Empirical Example

In this section we consider the use of the different methods in the context of an empirical example. The data we use comes from the Canadian Family Expenditure Survey for the years 1978 and 1986.[23] We consider a comparison of the income distributions in 1978 and 1986 using the methods that were compared in the previous section. In Table IV we have supplied some basic descriptive statistics for these data. In addition in Figures 1(A) and 2(A) we have plotted the empirical CDF for the before and after tax income data respectively with the 1978 distribution being the solid line. The Figures 1(B) and 2(B) contain the difference between the 1978 and 1986 empirical CDFs plotted against income values and give a much clearer picture. Note that in all of these figures income has been normalized to lie between 0 and 20 so that (for instance) a value of 2 on the z axis represents an income value of around $20,000. As indicated by the latter figures

[22]As expected, in all cases the use of conservative critical values for the MT and MTA test statistics resulted in less powerful tests compared to the simulation of their p-values.

[23]In fact we analyzed data from the years 1974, 1978, 1982, 1986 and 1990. In comparing the distributions across time all tests were generally in agreement that 1990 dominates 1986, 1982 dominates 1986 and 1978 and finally that 1978 dominates 1974. Therefore, as noted by Anderson (1996) (who referred to these data as the Family Income surveys with years that differ by one in each case), with the exception of 1986, the income distribution has unambiguously been improving over time.

the difference between these distributions is quite erratic even though the distributions themselves are quite regular looking. The plots of the differences also give one an idea of the importance of selecting income values for evaluating the MT and W tests – one may miss out on important differences between the distributions depending on where one computes and compares the empirical CDFs. Similar issues arise for tests of higher order stochastic dominance.

Table IV: Descriptive Statistics

	Before Tax		After Tax	
	1978	1986	1978	1986
Sample	8,526	9,470	8,526	9,470
Mean	35,535	36,975	29,840	30,378
Std Dev.	22,098	24,767	16,873	18,346
Median	32,423	32,658	27,813	27,337

In Tables V and VI we present p-values for all the tests considered in this paper for the 1978/1986 income distribution comparison. In Table V we have the results for before tax income while Table VI contains the after tax income results. The panel labelled "1986 versus 1978" contains p-values for testing whether the 1986 income distribution stochastically dominates the 1978 income distribution (to the specified order) while the other panel tests the opposite hypothesis.[24] In Table V there is agreement between the KS tests, the MT test and the W test (for either 5 or 10 evaluation points) that the 1986 (before tax) income distribution dominates the 1978 distribution in both a second order and third order sense. The first panel indicates that one cannot reject that 1986 dominates 1978 in both a second and third order sense while the second panel indicates that the converse can easily be rejected since the p-values are essentially zero. The MTA test, on the other hand leads to a different conclusion since the p-values in the second panel are larger than any conventionally used significance level. Based on the MTA tests one would seem to conclude the neither year dominates the other in a second or third order sense. This difference between results for SD2 and SD3 tests between the MTA test

[24] Therefore the null hypothesis for the SD1 column of the panel labelled "1986 versus 1978" is that the CDF in 1986 is less than or equal to that in 1978. Similar interpretations hold for the other tests.

and the others is similar to the finding in the Monte Carlo experiments in Case 4. With respect to first order dominance the KS test suggest that one can reject the nulls of SD1 in both cases while the other tests are less clear on this with p-values falling between 0.01 and 0.05. It is interesting to note that when only 5 values are used to compute the MT, MTA and W tests, the p-values are all larger than conventional significance levels when testing the null that the 1986 distribution stochastically dominates (in a first order sense) the distribution in 1978. This appears to occur because in this case the tests are based on values that excludes the largest difference between the CDFs that occurs around the value of 2 (see Figures 1(B) and 2(B) for instance).

In Table VI similar results hold, although there is not as much agreement between the KS tests and the MT and W tests in terms of their implications for second and third order stochastic dominance. In particular the KS tests suggests that one cannot reject that 1986 dominates 1978 in a second or third order sense but there is only weak evidence against the converse hypothesis with p-values falling near conventional levels of significance. The MT and W tests provide a clearer result with easy rejection of the null that 1978 dominates 1986 in a second and third order sense and less evidence against the reverse hypotheses. Note, however, that this result depends on the use of these tests with 10 evaluation points and that with only 5 one has a more cloudy picture similar to the KS tests. Again the MTA test gives strikingly different results for tests of SD2 and SD3 particularly when the null is that 1978 stochastically dominates 1986 at both second and third order. Again, as was the case with the before tax income distributions, the KS tests suggest that in a first order sense there is no stochastic dominance and that for some (low) income values 1978 stochastically dominates 1986 while for other (larger) income values the converse is true. On the other hand, the other tests all easily reject the null that 1978 (first order) stochastically dominates 1986 but offer only weak evidence against the converse hypothesis.

Table V: Stochastic Dominance In Canadian Before Tax Family Income

	1986 versus 1978			1978 versus 1986		
	SD1	SD2	SD3	SD1	SD2	SD3
KS1	0.010	0.380	0.550	0.000	0.000	0.000
KS2	0.010	0.350	0.503	0.000	0.000	0.000
KSB1	0.010	0.370	0.540	0.000	0.000	0.000
KSB2	0.010	0.370	0.570	0.000	0.000	0.000
KSB3	0.010	0.280	0.480	0.000	0.000	0.000
MT(10)	0.018	0.216	0.388	0.000	0.001	0.001
MT(5)	0.156	0.194	0.353	0.000	0.000	0.000
W(10)	0.038	0.228	0.412	0.000	0.000	0.000
W(5)	0.157	0.189	0.369	0.000	0.000	0.001
MTA(10)	0.022	0.130	0.182	0.000	0.148	0.243
MTA(5)	0.155	0.121	0.152	0.000	0.169	0.289

Table VI: Stochastic Dominance In Canadian After Tax Family Income

	1986 versus 1978			1978 versus 1986		
	SD1	SD2	SD3	SD1	SD2	SD3
KS1	0.005	0.220	0.520	0.001	0.010	0.060
KS2	0.005	0.224	0.471	0.001	0.022	0.073
KSB1	0.005	0.200	0.470	0.001	0.030	0.060
KSB2	0.005	0.200	0.480	0.001	0.030	0.070
KSB3	0.005	0.240	0.460	0.001	0.010	0.050
MT(10)	0.019	0.068	0.184	0.000	0.002	0.001
MT(5)	0.014	0.070	0.204	0.000	0.056	0.033
W(10)	0.025	0.077	0.198	0.000	0.000	0.003
W(5)	0.017	0.083	0.213	0.000	0.054	0.027
MTA(10)	0.023	0.048	0.062	0.000	0.743	0.745
MTA(5)	0.033	0.185	0.356	0.000	0.744	0.855

7 Conclusion

In this paper we have considered Kolmogorov Smirnov type tests for an arbitrary degree of stochastic dominance. We have proposed a variety of simulation and bootstrap methods for conducting inference for degrees of stochastic dominance beyond the first degree and shown that the approach behaves well asymptotically. In addition we have shown that the tests perform well in finite samples. The way that the p-value approach was implemented in both the Monte Carlo and empirical example suggest that one does not need to perform too many computations to obtain reasonable inferences. The main advantage of the approach is that the tests are consistent, being based on an examination of the complete set of restrictions that follow from stochastic dominance. The main disadvantage of the approach is that simulation or resampling is required for inference. However, this is not a major issue given modern computing capabilities. Moreover, the main competitor proposed in the literature appears to be the Wald test of (a fixed number of) inequality restrictions which also requires simulation for inference but has the potential for inconsistent test results. Finally, the methods developed in this paper can be extended to other situations where one is interested in comparing curves and testing for dominance relations. An obvious application is to the case of Lorenz curves and testing for Lorenz dominance relations in the analysis of economic inequality.

Appendix A: Proofs of Results

Proof of Lemma 1: To show the result we first note that we can write,

$$\mathcal{B}_F(F(z)) = \mathcal{W}_F(F(z)) - F(z)\mathcal{W}_F(1)$$

for some appropriately defined Weiner process $\mathcal{W}_F$. Then,

$$\mathcal{I}_j(z;\mathcal{B}_F \circ F) = \mathcal{I}_j(z;\mathcal{W}_F \circ F) - \mathcal{I}_j(z;F)\mathcal{W}_F(1)$$

Consequently,

$$\begin{aligned}
E(\mathcal{I}_j(z_1;\mathcal{B}_F \circ F)\mathcal{I}_j(z_2;\mathcal{B}_F \circ F)) &= E(\mathcal{I}_j(z_1;\mathcal{W}_F \circ F)\mathcal{I}_j(z_2;\mathcal{W}_F \circ F)) \\
&\quad -\mathcal{I}_j(z_1;F)E(\mathcal{I}_j(z_2;\mathcal{W}_F \circ F)\mathcal{W}_F(1)) \\
&\quad -\mathcal{I}_j(z_2;F)E(\mathcal{I}_j(z_1;\mathcal{W}_F \circ F)\mathcal{W}_F(1)) + \mathcal{I}_j(z_2;F)\mathcal{I}_j(z_1;F) \\
&= E(\mathcal{I}_j(z_1;\mathcal{W}_F \circ F)\mathcal{I}_j(z_2;\mathcal{W}_F \circ F)) - \mathcal{I}_j(z_2;F)\mathcal{I}_j(z_1;F)
\end{aligned}$$

using the definition of the integral operator and the fact that $E(\mathcal{W}_F(F(z))\mathcal{W}_F(1)) = F(z)$. Let,

$$\Omega_1(z_1,z_2;F) = \Omega_1(z_2,z_1;F) = E(\mathcal{W}_F(F(z_1))\mathcal{W}_F(z_2)) = F(z_1) \text{ for } z_1 \leq z_2.$$

Thus it remains to compute $E(\mathcal{I}_j(z_1;\mathcal{W}_F \circ F)\mathcal{I}_j(z_2;\mathcal{W}_F \circ F))$. Let $\theta_0^1 = 1$. For $j = 2$ we have, using Lemma A1 and Lemma A2 that,

$$\begin{aligned}
E(\int_0^{z_1} \mathcal{W}_F(F(t))dt \int_0^{z_2} \mathcal{W}_F(F(t))dt) &= \int_0^{z_1}\int_0^t E(\mathcal{W}_F(F(t))\mathcal{W}_F(F(s))dsdt \\
&\quad + \int_0^{z_1}\int_0^{z_2} E(\mathcal{W}_F(F(t))\mathcal{W}_F(F(s))dsdt \\
&= \int_0^{z_1}\int_0^t \Omega_1(s,t;F)dsdt + \int_0^{z_1}\int_t^{z_2} \Omega_1(s,t;F)dsdt \\
&= \int_0^{z_1}\int_0^t F(s)dsdt + \int_0^{z_1}\int_t^{z_2} F(t)dsdt \\
&= 2\mathcal{I}_3(z_1;F) + (z_2 - z_1)\mathcal{I}_2(z_1;F) \\
&= 2\theta_0^1\mathcal{I}_{2+1}(z_1;F) + \theta_0^1(z_2 - z_1)\mathcal{I}_2(z_1;F) \\
&= \Omega_2(z_1,z_2;F) = \Omega_2(z_2,z_1;F)
\end{aligned}$$

where $\theta_0^2 = 2\theta_0^1 = 2\sum_{l=0}^{2-2}\theta_l^{j-1}$ and $\theta_1^2 = \theta_0^1 = \sum_{l=1-1}^{2-2}\theta_l^{j-1}$. Next consider the case where $j = 3$. Here,

$$\begin{aligned}
E(\mathcal{I}_3(z_1;\mathcal{W}_F\circ F)\mathcal{I}_3(z_2;\mathcal{W}_F\circ F)) &= E(\int_0^{z_1}\mathcal{I}_2(t;\mathcal{W}_F\circ F)\int_0^{z_2}\mathcal{I}_2(z_2;\mathcal{W}_F\circ F)) \\
&= \int_0^{z_1}\int_0^t \Omega_2(s,t;F)dsdt + \int_0^{z_1}\int_t^{z_2}\Omega_2(s,t;F)dsdt \\
&= \int_0^{z_1}\int_0^t \{\theta_0^2\mathcal{I}_3(s;F) + \theta_1^2(t-s)\mathcal{I}_2(s;F)\}dsdt \\
&\quad + \int_0^{z_1}\int_t^{z_2}\{\theta_0^2\mathcal{I}_3(t;F) + \theta_1^2(s-t)\mathcal{I}_2(t;F)\}dsdt \\
&= 2(\theta_0^2+\theta_1^2)\mathcal{I}_5(z_1;F) + (\theta_0^2+\theta_1^2)(z_2-z_1)\mathcal{I}_4(z_1;F) \\
&\quad +\theta_1^2\frac{1}{2}(z_2-z_1)^2\mathcal{I}_3(z_1;F)
\end{aligned}$$

The result then follows by induction. **Q.E.D.**

Lemma A1: *Under Assumption 1,*

$$\int_0^z\int_0^t \frac{1}{l!}(t-s)^l\mathcal{I}_j(s;F)dsdt = \mathcal{I}_{j+l+2}(z;F)$$

Proof: Follows by integration by parts and the facts that,

$$\mathcal{I}_j(0;F) = 0 \text{ for all } j.$$

Q.E.D.

Lemma A2: *Under Assumption 1, for* $z_2 \geq z_1$,

$$\int_{-\infty}^{z_1}\int_t^{z_2}\frac{1}{l!}(s-t)^l\mathcal{I}_j(t;F)dsdt = \sum_{k=0}^{l+1}\frac{1}{k!}(z_2-z_1)^k\mathcal{I}_{j+l+2-k}(z_1;F)$$

Proof: Note that,

$$\int_0^{z_1}\int_t^{z_2}\frac{1}{l!}(s-t)^l\mathcal{I}_j(s;F)dsdt = \int_0^{z_1}\frac{1}{(l+1)!}(z_2-t)^{l+1}\mathcal{I}_j(t;F)dt$$

Then repeated integration by parts and induction give the result.**Q.E.D.**

Proof of Proposition 1: The proof is based on a characterization for the limiting distribution and the application of an inequality. From the Glivenko-Cantelli, Donsker and Continuous Mapping Theorems and the fact that $\bar{z} < \infty$ we have,

$$\sup_z |\mathcal{I}_j(z;\hat{G}_M) - \mathcal{I}_j(z;G)| \stackrel{a.s.}{\to} 0, \ \sqrt{M}(\mathcal{I}_j(.;\hat{G}_M) - \mathcal{I}_j(.;G)) \Rightarrow \mathcal{I}_j(.;\mathcal{B}_G \circ G) \quad (11)$$
$$\sup_z |\mathcal{I}_j(z;\hat{F}_N) - \mathcal{I}_j(z;F)| \stackrel{a.s.}{\to} 0, \ \sqrt{N}(\mathcal{I}_j(.;\hat{F}_N) - \mathcal{I}_j(.;F)) \Rightarrow \mathcal{I}_j(.;\mathcal{B}_F \circ F) \quad (12)$$

in the space $D([0,\bar{z}])$ for $j = 1$ and the space $C([0,\bar{z}])$ for $j \geq 2$. Use the notation $Z = [0,\bar{z}]$. Immediate implications are that,

$$\sup_z |(\mathcal{I}_j(z;\hat{G}_M) - \mathcal{I}_j(z;\hat{F}_N)) - (\mathcal{I}_j(z;G) - \mathcal{I}_j(z;F))| \stackrel{a.s.}{\to} 0 \quad (13)$$

and, using Assumption 2(ii),

$$\begin{aligned} \hat{T}_j(.) &= \sqrt{\frac{NM}{N+M}}(\mathcal{I}_j(.;\hat{G}_M) - \mathcal{I}_j(.;G)) - \sqrt{\frac{NM}{N+M}}(\mathcal{I}_j(.;\hat{F}_N) - \mathcal{I}_j(.;F)) \quad (14) \\ &\Rightarrow \lambda^{1/2}\mathcal{I}_j(.;\mathcal{B}_G \circ G) - (1-\lambda)^{1/2}\mathcal{I}_j(.;\mathcal{B}_F \circ F) \\ &\equiv \bar{T}_j(.) \end{aligned}$$

Use the notation $\hat{T}_j(z)$ for $\hat{T}_j$ evaluated at the specific point $z \in Z$. An implication of the weak convergence result is that for any $\gamma, \varepsilon > 0$ that there exists a $\delta > 0$ such that the following stochastic equicontinuity condition holds,

$$\limsup P(\sup_{|z_1 - z_2| < \delta} |\hat{T}_j(z_1) - \hat{T}_j(z_2)| > \varepsilon) < \gamma \quad (15)$$

(see for instance Pollard (1984, Chapter V, Theorem 3)).

To show the result in A(ii) we note that,

$$\begin{aligned} \hat{S}_j &\leq \sup_z \hat{T}_j(z) + \sup_z \left(\frac{NM}{N+M}\right)^{1/2} (\mathcal{I}_j(z;G) - \mathcal{I}_j(z;F)) \\ &\leq \sup_z \hat{T}_j(z) \end{aligned}$$

by the definitions of $\hat{S}_j$ and $\hat{T}_j(z)$ and the fact that under H_0^j, $\mathcal{I}_j(z;G) - \mathcal{I}_j(z;F) \leq 0$ for all z. Therefore the result in (i) follows using (14) and the fact that $\bar{S}_j^{G,F} = \sup_z \bar{T}_j(z)$.

To show A(i), noting that $\mathcal{I}_j(z;G) - \mathcal{I}_j(z;F) \leq 0$ for all z we denote by Z^* the set of z values for which $\mathcal{I}_j(z;G) = \mathcal{I}_j(z;F)$. Then for any $z \in Z^*$ we have that

$$\hat{T}_j(z) = \left(\frac{NM}{N+M}\right)^{1/2} (\mathcal{I}_j(z;\hat{G}_M) - \mathcal{I}_j(z;\hat{F}_N))$$

It is easily seen that Z^* is a compact set because of Assumption 1. We aim to show that for $c > 0$,

$$P(\hat{S}_j > c) \to P(\sup_{z \in Z^*} \bar{T}_j(z) > c) \tag{16}$$

To show this we first note that,

$$\begin{aligned} \hat{S}_j &= \left(\frac{NM}{N+M}\right)^{1/2} \sup_{z \in Z}(\mathcal{I}_j(z; \hat{G}_M) - \mathcal{I}_j(z; \hat{F}_N)) \\ &\geq \sup_{z \in Z^*} \hat{T}_j(z) \\ &\Rightarrow \sup_{z \in Z^*} \bar{T}_j(z) \end{aligned}$$

because of the fact that $Z^* \subset Z$ and using the Continuous Mapping Theorem (CMT). Consequently,

$$\limsup P(\hat{S}_j \leq c) \leq P(\sup_{z \in Z^*} \bar{T}_j(z) \leq c) \tag{17}$$

Let $\hat{z}$ denote any value of z that solves the problem,

$$\sup_{z \in Z}(\mathcal{I}_j(z; \hat{G}_M) - \mathcal{I}_j(z; \hat{F}_N))$$

and note that $\hat{z} \in Z$. We suppress the dependence of $\hat{z}$ on N and M for ease of notation. Then, for any non-empty $Z^+ \subset Z^*$ we have that,

$$\begin{aligned} \hat{S}_j &= \left(\frac{NM}{N+M}\right)^{1/2} (\mathcal{I}_j(\hat{z}; \hat{G}_M) - \mathcal{I}_j(\hat{z}; \hat{F}_N)) \\ &\leq \sup_{z \in Z^*} \hat{T}_j(z) + \left(\frac{NM}{N+M}\right)^{1/2} (\mathcal{I}_j(\hat{z}; G) - \mathcal{I}_j(\hat{z}; F)) + \hat{T}_j(\hat{z}) - \inf_{z \in Z^+} \hat{T}_j(z) \\ &\leq \sup_{z \in Z^*} \hat{T}_j(z) + \sup_{z \in Z^+} (\hat{T}_j(\hat{z}) - \hat{T}_j(z)) \\ &\leq \sup_{z \in Z^*} \hat{T}_j(z) + \sup_{z \in Z^+} |\hat{T}_j(\hat{z}) - \hat{T}_j(z)| \end{aligned} \tag{18}$$

where the second line follows from the fact that,

$$\inf_{z \in Z^+} \hat{T}_j(z) \leq \sup_{z \in Z^*} \hat{T}_j(z)$$

the third line follows from the fact that under the null hypothesis,

$$(\mathcal{I}_j(\hat{z}; G) - \mathcal{I}_j(\hat{z}; F)) \leq 0$$

Now pick any $\varepsilon^* > 0$. Let c' be such that $c' < c$,

$$P(\sup_{z \in Z^*} \bar{T}_j(z) \leq c) - P(\sup_{z \in Z^*} \bar{T}_j(z) \leq c') < \varepsilon^* \tag{19}$$

Let ε_1 be a positive number such that $0 < \varepsilon_1 < c - c'\}$ and then pick a $\delta > 0$ such that (15) holds with $\varepsilon = \varepsilon_1$ and $\gamma = \varepsilon^*$. Define the set $Z^+ = Z^* \cap B(\hat{z}, \delta)$ where $B(\hat{z}, \delta)$ is a ball of radius δ around $\hat{z}$, and let $A_{N,M}$ denote the event that Z^+ is nonempty. We first demonstrate that $P(A_{N,M}) \to 1$. Let $\bar{Z}^*_\delta = \{z \in Z : d(z, Z^*) \geq \delta\}$ where $d(z, Z^*) = \inf_{z' \in Z^*} |z - z'|$ is a measure of the distance of the point z from the compact set Z^*. It is only necessary to consider the case that $\bar{Z}^*_\delta$ is nonempty because otherwise $P(A_{N,M}) = 1$ for all N, M. It is easy to show that $\bar{Z}^*_\delta$ is a compact set by Assumption 1. Consequently for some $\eta > 0$,

$$\sup_{z \in \bar{Z}^*_\delta} (\mathcal{I}_j(z; G) - \mathcal{I}_j(z; F)) = -2\eta < 0. \tag{20}$$

Pick an arbitrary $z^* \in Z^*$ and note that the event $A_{N,M}$ is implied by the event,

$$\begin{aligned} \sup_{z \in \bar{Z}^*_\delta} (\mathcal{I}_j(z; \hat{G}_M) - \mathcal{I}_j(z; \hat{F}_N)) &< -\eta \\ (\mathcal{I}_j(z^*; \hat{G}_M) - \mathcal{I}_j(z^*; \hat{F}_N)) &> -\eta \end{aligned}$$

so that,

$$\begin{aligned} P(A_{N,M}) &\geq P(\sup_{z \in \bar{Z}^*_\delta} (\mathcal{I}_j(z; \hat{G}_M) - \mathcal{I}_j(z; \hat{F}_N)) < (\mathcal{I}_j(z^*; \hat{G}_M) - \mathcal{I}_j(z^*; \hat{F}_N))) \\ &\geq P(\{\sup_{z \in \bar{Z}^*_\delta} (\mathcal{I}_j(z; \hat{G}_M) - \mathcal{I}_j(z; \hat{F}_N)) < -\eta\} \cap \{(\mathcal{I}_j(z^*; \hat{G}_M) - \mathcal{I}_j(z^*; \hat{F}_N)) > -\eta\}) \\ &\geq P((\mathcal{I}_j(z^*; \hat{G}_M) - \mathcal{I}_j(z^*; \hat{F}_N)) > -\eta) - P(\sup_{z \in \bar{Z}^*_\delta} (\mathcal{I}_j(z; \hat{G}_M) - \mathcal{I}_j(z; \hat{F}_N)) > -\eta) \\ &\to 1 \end{aligned}$$

using (11), (12), (13), (20) and CMT which implies that,

$$\begin{aligned} \sup_{z \in \bar{Z}^*_\delta} (\mathcal{I}_j(z; \hat{G}_M) - \mathcal{I}_j(z; \hat{F}_N)) &= \sup_{z \in \bar{Z}^*_\delta} \{(\mathcal{I}_j(z; \hat{G}_M) - \mathcal{I}_j(z; \hat{F}_N)) \\ &\quad -(\mathcal{I}_j(z; G) - \mathcal{I}_j(z; F)) + (\mathcal{I}_j(z; G) - \mathcal{I}_j(z; F))\} \\ &\leq \sup_{z \in \bar{Z}^*_\delta} |(\mathcal{I}_j(z; \hat{G}_M) - \mathcal{I}_j(z; \hat{F}_N)) - (\mathcal{I}_j(z; G) - \mathcal{I}_j(z; F))| \end{aligned}$$

$$+\sup_{z\in\bar{Z}_\delta^*}(\mathcal{I}_j(z;G)-\mathcal{I}_j(z;F))$$
$$\xrightarrow{a.s.} -2\eta$$

Then,

$$P(\hat{S}_j\le c)=P(\{\hat{S}_j\le c\}\cap A_{N,M})+P(\{\hat{S}_j\le c\}\cap \bar{A}_{N,M}) \tag{21}$$
$$\ge P(\{\sup_{z\in Z^*}\hat{T}_j(z)+\sup_{z\in Z^+}|\hat{T}_j(\hat{z})-\hat{T}_j(z)|\le c\}\cap A_{N,M})+P(\{\hat{S}_j\le c\}\cap \bar{A}_{N,M})$$
$$\ge P(\{\sup_{z\in Z^*}\hat{T}_j(z)+\sup_{|z_1-z_2|<\delta}|\hat{T}_j(z_1)-\hat{T}_j(z_2)|\le c\}\cap A_{N,M})+P(\{\hat{S}_j\le c\}\cap \bar{A}_{N,M})$$
$$\ge P(\sup_{z\in Z^*}\hat{T}_j(z)+\sup_{|z_1-z_2|<\delta}|\hat{T}_j(z_1)-\hat{T}_j(z_2)|\le c)-P(\bar{A}_{N,M})$$
$$+P(\{\hat{S}_j\le c\}\cap \bar{A}_{N,M})$$

where the second line follows from the fact that in the event $A_{N,M}$ the inequality in (18) holds and the third line follows from the fact that,

$$\sup_{z\in Z^+}|\hat{T}_j(\hat{z})-\hat{T}_j(z)|\le \sup_{|z_1-z_2|<\delta}|\hat{T}_j(z_1)-\hat{T}_j(z_2)|$$

For the first term we use the inequality $P(A\cap B)\ge P(A)-P(\bar{B})$ for events A and B (with $\bar{B}$ being the complement of B) to show that (given $c'+\varepsilon_1\le c$) the event,

$$\left\{\sup_{z\in Z^*}\hat{T}_j(z)\le c'\right\}\cap\left\{\sup_{|z_1-z_2|<\delta}|\hat{T}_j(z_1)-\hat{T}_j(z_2)|\le\varepsilon_1\right\}$$

implies the event,

$$\left\{\sup_{z\in Z^*}\hat{T}_j(z)+\sup_{|z_1-z_2|<\delta}|\hat{T}_j(z_1)-\hat{T}_j(z_2)|\le c\right\}$$

so that,

$$\begin{aligned} P(\sup_{z\in Z^*}\hat{T}_j(z) &\le c')-P(\sup_{|z_1-z_2|<\delta}|\hat{T}_j(z_1)-\hat{T}_j(z_2)|>\varepsilon_1) \\ &\le P(\sup_{z\in Z^*}\hat{T}_j(z)+\sup_{|z_1-z_2|<\delta}|\hat{T}_j(z_1)-\hat{T}_j(z_2)|\le c)\end{aligned} \tag{22}$$

Then we have that,

$$\liminf\left(P(\sup_{z\in Z^*}\hat{T}_j(z)\le c')-P(\sup_{|z_1-z_2|<\delta}|\hat{T}_j(z_1)-\hat{T}_j(z_2)|>\varepsilon_1)\right)>P(\sup_{z\in Z^*}\bar{T}_j(z)\le c)-2\varepsilon^*$$

using (15)and (19). Combine this result, (22) the last line of (21) and the fact that, $P(A_{N,M}) \to 1$ implies that both $P(\bar{A}_{N,M}) \to 0$ and,

$$P(\{\hat{S}_j \leq c\} \cap \bar{A}_{N,M}) \to 0$$

and we have,

$$\liminf P(\hat{S}_j \leq c) \geq P(\sup_{z \in Z^*} \bar{T}_j(z) \leq c) - 2\varepsilon^*$$

Since ε^* is arbitrary we have using (17) that,

$$\lim P(\hat{S}_j \leq c) = P(\sup_{z \in Z^*} \bar{T}_j(z) \leq c).$$

To show the result in A(i) of Proposition 1 fix F. For any G satisfying the null hypothesis we have that,

$$\mathcal{I}_{j+l}(z; G) \leq \mathcal{I}_{j+l}(z; F) \text{ for all } z \text{ for all } l \geq 0 \tag{23}$$

Compare the situation where $\mathcal{I}_j(z; F) \equiv \mathcal{I}_j(z; G)$ for all z (and hence $F(z) = G(z)$ and $\mathcal{I}_l(z; F) \equiv \mathcal{I}_l(z; G)$ for all z and all $l \geq 1$) to that where $\mathcal{I}_j(z; F) \equiv \mathcal{I}_j(z; G)$ for all $z \in Z^* \subset Z$. Denote the limiting random variable corresponding to $\bar{T}_j(z)$ in the case where $\mathcal{I}_j(z; F) \equiv \mathcal{I}_j(z; G)$ for all z by $\bar{T}^0_j(z)$. The result will follow from the inequalities,

$$P(\sup_{z \in Z^*} \bar{T}_j(z) > c) \leq P(\sup_{z \in Z^*} \bar{T}^0_j(z) > c) \leq P(\bar{S}^F_j > c) \tag{24}$$

The second inequality is obvious from the fact that $Z^* \subset Z$ and the fact that $\bar{S}^F_j = \sup_z \bar{T}^0_j(z)$. To show the first inequality let $\bar{T}^0_j(z)$ denote the process that is identical to $\bar{T}_j(z)$ in every respect except that $G = F$. Then consider (for $z_2 > z_1$ with $z_2, z_1 \in Z^*$),

$$\begin{aligned} E((\bar{T}_j(z_2) - \bar{T}_j(z_1))^2) &= \lambda(\Omega_j(z_2, z_2; G) + \Omega_j(z_1, z_1; G) - 2\Omega_j(z_2, z_1; G)) \\ &\quad +(1-\lambda)(\Omega_j(z_2, z_2; F) + \Omega_j(z_1, z_1; F) - 2\Omega_j(z_2, z_1; F)) \end{aligned}$$

Now by Lemma 1 and the fact that, $\mathcal{I}_j(z; F) \equiv \mathcal{I}_j(z; G)$ for $z = z_2$ and $z = z_1$ we can write,

$$\begin{aligned} \Omega_j(z_1, z_1; G) &= \Omega_j(z_1, z_1; F) - a_1 \\ \Omega_j(z_2, z_2; G) &= \Omega_j(z_2, z_2; F) - (a_1 + a_2) \end{aligned}$$

where $a_1 = a_2 = 0$ for $j = 1$ and,

$$\begin{aligned} a_1 &= \theta_0^j(\mathcal{I}_{2j-1}(z_1;F) - \mathcal{I}_{2j-1}(z_1;G)) \geq 0 \\ a_2 &= \theta_0^j(\mathcal{I}_{2j-1}(z_2;F) - \mathcal{I}_{2j-1}(z_1;F)) - \theta_0^j(\mathcal{I}_{2j-1}(z_2;G) - \mathcal{I}_{2j-1}(z_1;G)) \\ &= \theta_0^j \int_{z_1}^{z_2} (\mathcal{I}_{2j-1-1}(t;F) - \mathcal{I}_{2j-1-1}(t;G))dt \geq 0 \end{aligned}$$

when $j \geq 2$ by $2j - 2 \geq j$ and (23). Similarly we can write,

$$\Omega_j(z_2, z_1; G) = \Omega_j(z_2, z_1; F) - (a_1 + a_3)$$

where, $a_3 = 0$ for $j = 1$ and for $j \geq 2$,

$$\begin{aligned} a_3 &= \sum_{l=1}^{j-1} \theta_l^j \frac{1}{l!}(z_2 - z_1)^l(\mathcal{I}_{2j-1-l}(z_1;F) - \mathcal{I}_{2j-1-l}(z_1;G)) \\ &= \sum_{l=1}^{j-2} \theta_l^j \frac{1}{l!}(z_2 - z_1)^l(\mathcal{I}_{2j-1-l}(z_1;F) - \mathcal{I}_{2j-1-l}(z_1;G)) \geq 0 \end{aligned}$$

where the second line follows by $\mathcal{I}_j(z_1;F) = \mathcal{I}_j(z_1;G)$. Now by (23) and using a mean value expansion for $z^* \in (z_1, z_2]$ we have that,

$$\begin{aligned} & a_2 = \theta_0^j\left((\mathcal{I}_{2j-1}(z_2;F) - \mathcal{I}_{2j-1}(z_2;G)) - (\mathcal{I}_{2j-1}(z_1;F) - \mathcal{I}_{2j-1}(z_1;G))\right) \\ = \; & \theta_0^j \sum_{l=1}^{j-2} \frac{1}{l!}(z_2 - z_1)^l(\mathcal{I}_{2j-1-l}(z_1;F) - \mathcal{I}_{2j-1-l}(z_1;G)) \\ & + \frac{1}{(j-1)!}(z_2 - z_1)^{j-1}(\mathcal{I}_j(z^*;F) - \mathcal{I}_j(z^*;G)) \\ \geq \; & \sum_{l=1}^{j-2} 2\theta_l^j \frac{1}{l!}(z_2 - z_1)^l(\mathcal{I}_{2j-1-l}(z_1;F) - \mathcal{I}_{2j-1-l}(z_1;G)) \\ \geq \; & 2a_3 \end{aligned}$$

where the third line follows by Lemma 1 and (23). Consequently we have,

$$\begin{aligned} E((\bar{T}_j(z_2) - \bar{T}_j(z_1))^2) &= E((\bar{T}_j^0(z_2) - \bar{T}_j^0(z_1))^2) - \lambda(a_1 + a_1 + a_2 - 2(a_1 + a_3)) \\ &\leq E((\bar{T}_j^0(z_2) - \bar{T}_j^0(z_1))^2) \end{aligned}$$

Since the stochastic processes are separable, mean zero and Gaussian then Proposition A.2.6 of Van der Vaart and Wellner (1996) (the Slepian, Fernique, Marcus and Shepp inequality) implies that the first inequality in (24) holds and the result in (ii) follows since $P(\sup_{z\in Z} \bar{T}_j^0(z) > c)$ is the asymptotic probability of rejection in the case where $F(z) \equiv G(z)$ for all $z \in Z$.

To show the result in B we note that if the alternative is true then there is some z, say $\bar{z} \in Z$ for which,

$$\mathcal{I}_j(\bar{z}; G) - \mathcal{I}_j(\bar{z}; F) = \delta > 0$$

Then, the result follows using the inequality,

$$\hat{S}_j \geq \left(\frac{NM}{N+M}\right)^{1/2} (\mathcal{I}_j(\bar{z}; \hat{G}_M) - \mathcal{I}_j(\bar{z}; \hat{F}_N))$$

and the results in (11), (12) and (13). **Q.E.D.**

Proof of Proposition 2: Write,

$$\begin{aligned}
\mathcal{B}^*(z; \hat{F}_N) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (1(X_i \leq z) - \hat{F}_N(z))U_i \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (1(X_i \leq z) - F(z))U_i - (\hat{F}_N(z) - F(z)) \frac{1}{\sqrt{N}} \sum_{i=1}^{N} U_i
\end{aligned}$$

First consider the second term. Note that almost every sample (of X_i) has the property that $\sup_z |\hat{F}_N(z) - F(z)| \to 0$. Then using the facts that the U_i are mean zero independent Gaussian random variables we have that conditional on the sample,

$$\begin{aligned}
P_U(\sup_z |(\hat{F}_N(z) - F(z)) \frac{1}{\sqrt{N}} \sum_{i=1}^{N} U_i| &> \varepsilon) = P_U(\sup_z |(\hat{F}_N(z) - F(z))| | \frac{1}{\sqrt{N}} \sum_{i=1}^{N} U_i| > \varepsilon) \\
&\leq \frac{(\sup_z |(\hat{F}_N(z) - F(z)|)^2 E(\frac{1}{N} \sum_{i=1}^{N} U_i^2)}{\varepsilon} \\
&\to 0
\end{aligned}$$

Consequently for this sample we have that

$$(\hat{F}_N(z) - F(z)) \frac{1}{\sqrt{N}} \sum_{i=1}^{N} U_i \xrightarrow{p} 0$$

(where 0 is the zero function, a member of the space D) which implies that for the particular sample,

$$(\hat{F}_N(z) - F(z)) \frac{1}{\sqrt{N}} \sum_{i=1}^{N} U_i \Rightarrow 0$$

But this holds for almost all samples so that we have,

$$(\hat{F}_N(z) - F(z))\frac{1}{\sqrt{N}}\sum_{i=1}^{N} U_i \overset{a.s.}{\Rightarrow} 0$$

For the first term, Corollary 2.9.3 of Van der Vaart and Wellner (1996) implies that the process $\mathcal{B}^* \circ F$ which at z is given by,

$$\mathcal{B}^*(z; F) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(1(X_i \leq z) - F(z))U_i$$

satisfies $\mathcal{B}^* \circ F \Rightarrow \mathcal{B}' \circ F$ for almost all samples, where $\mathcal{B}' \circ F$ is an independent copy of $\mathcal{B}(F(z))$. Combining these results we have that, $\mathcal{B}^* \circ \hat{F}_N \overset{a.s.}{\Rightarrow} \mathcal{B}' \circ F$. Similar arguments can be used to show that $\mathcal{B}^* \circ \hat{G}_M \overset{a.s.}{\Rightarrow} \mathcal{B}' \circ G$.

To show the results concerning the asymptotic behavior of the p-values we prove the result for $\hat{p}_j^F$ with the result for $\hat{p}_j^{G,F}$ being analogous. Let $\hat{P}_{j,N}(t)$ be the CDF of the process (conditional on the original sample of X_i) generated by $\sup_z \mathcal{I}_j(z; \mathcal{B}^* \circ \hat{F}_N)$. By $\mathcal{B}^* \circ \hat{F}_N \overset{a.s.}{\Rightarrow} \mathcal{B}' \circ F$ and the CMT we have that,

$$\sup_z \mathcal{I}_j(z; \mathcal{B}^* \circ \hat{F}_N) \overset{a.s.}{\Rightarrow} \sup_z \mathcal{I}_j(z; \mathcal{B}' \circ F) \tag{25}$$

where the random $\sup_z \mathcal{I}_j(z; \mathcal{B}' \circ F)$ is an independent copy of $\bar{S}_j^F$. Note that the median of the distribution of $\sup_z \mathcal{I}_j(z; \mathcal{B}' \circ F)$ (denoted $\bar{P}_j^0(t)$) is strictly positive and finite. By Tsirel'son (1975) $\bar{P}_j^0(t)$ is absolutely continuous on $(0, \infty)$ and, moreover, $c_j(\alpha)$ (defined by $P(\bar{S}_j^F > c_j(\alpha)) = \alpha$) is finite and positive for any fixed $\alpha < 1/2$ using (for instance) Proposition A.2.7 of Van der Vaart and Wellner (1996). Note that event $\{\hat{p}_j^F < \alpha\}$ is equivalent to the event that $\{\hat{S}_j > \hat{c}_j(\alpha)\}$ where,

$$\inf\{t : \hat{P}_{j,N}(t) > 1 - \alpha\} = \hat{c}_j(\alpha) \overset{a.s.}{\to} c_j(\alpha) \tag{26}$$

by (25) and the properties of $\bar{P}_2^0(t)$ noted above. Then,

$$\begin{aligned}
\lim P(\text{reject } H_0^j | H_0^j) &= \lim P(\hat{S}_j > \hat{c}_j(\alpha)) \\
&= \lim P(\hat{S}_j > c_j(\alpha)) + \lim(P(\hat{S}_j > \hat{c}_j(\alpha)) - P(\hat{S}_j > c_j(\alpha))) \\
&\leq P(\bar{S}_j^F > c_j(\alpha)) \\
&= \alpha
\end{aligned}$$

where the last line follows from (26), A(i) of Proposition 1 and the fact that $c_j(\alpha)$ is a continuity point of the distribution $\bar{P}_j(t)$. On the other hand Proposition 1 B and finiteness of $c_j(\alpha)$ imply that $\lim P(\text{reject } H_0^j | H_1^j) = 1$. **Q.E.D.**

Proof of Proposition 3: By Theorem 3.6.3 of Van der Vaart and Wellner (1996) we have that for independent samples drawn from $\mathcal{X}$,

$$\sqrt{N}(\hat{F}_N^* - \hat{F}_N) \overset{p}{\Rightarrow} \mathcal{B}_{F,s}^* \circ F \overset{d}{=} \mathcal{B}_F \circ F \tag{27}$$

while from $\mathcal{Y}$,

$$\sqrt{M}(\hat{G}_M^* - \hat{G}_M) \overset{p}{\Rightarrow} \mathcal{B}_{G,s}^* \circ G \overset{d}{=} \mathcal{B}_G \circ G \tag{28}$$

where the processes are independent. Similarly Theorem 3.7.6 gives that with independent random samples from $\mathcal{Z}$

$$\sqrt{\frac{NM}{N+M}}(\hat{G}_M^* - \hat{F}_N^*) \overset{p}{\Rightarrow} \sqrt{\lambda}\mathcal{B}_{G,s}^* \circ G - \sqrt{1-\lambda}\mathcal{B}_{F,s}^* \circ F \tag{29}$$

$$\overset{d}{=} \sqrt{\lambda}\mathcal{B}_G \circ G - \sqrt{1-\lambda}\mathcal{B}_F \circ F \tag{30}$$

This convergence is in the sense that (for instance),

$$\sup_{h \in BL_1} |E_C(h(\sqrt{N}(\hat{F}_M^* - \hat{F}_N))) - E(h(\mathcal{B}_F \circ F))| \overset{p}{\to} 0$$

where BL_1 is the space of bounded Lipschitz functions mapping $C[0,1]$ into $[0,1]$, and where E_C is the expectation given the sample $\mathcal{X}$ and $\mathcal{Z}$ respectively. We can see that the functional, $\mathcal{I}_j(.;F)$ is Hadamard differentiable with derivative $\mathcal{I}_{j-1}(.;F)$ by induction. This starts by noting that $\mathcal{I}_1(.;F)$ is Hadamard differentiable being the identity mapping and therefore $\mathcal{I}_2(.;F)$ is Hadamard differentiable since it is linear. Consequently we have that $\mathcal{I}_j(.;F)$ is a linear functional of a Hadamard differentiable mapping $\mathcal{I}_{j-1}(.;F)$. Using this fact, the results in (27) and (29) and Theorem 3.9.11 of Van der Vaart and Wellner (1996) gives the result that,

$$\sqrt{N}(\mathcal{I}_j(.;\hat{F}_N^*) - \mathcal{I}_j(.;\hat{F}_N)) \overset{p}{\Rightarrow} \mathcal{I}_j(.;\mathcal{B}_{F,s}^* \circ F)$$

$$\sqrt{\frac{NM}{N+M}}(\mathcal{I}_j(.;\hat{G}_M^*) - \mathcal{I}_j(.;\hat{F}_N^*)) \overset{p}{\Rightarrow} \mathcal{I}_j(.;\sqrt{\lambda}\mathcal{B}_{G,s}^* \circ G - \sqrt{1-\lambda}\mathcal{B}_{F,s}^* \circ F)$$

$$\sqrt{\frac{NM}{N+M}}((\mathcal{I}_j(.;\hat{G}_M^*) - \mathcal{I}_j(.;\hat{G}_M)) - (\mathcal{I}_j(.;\hat{F}_N^*) - \mathcal{I}_j(.;\hat{F}_N)))$$

$$\overset{p}{\Rightarrow} \mathcal{I}_j(.;\sqrt{\lambda}\mathcal{B}_{G,s}^* \circ G - \sqrt{1-\lambda}\mathcal{B}_{F,s}^* \circ F)$$

The remainder of the proof follows the proof of Proposition 2 (using $\overset{p}{\Rightarrow}$ instead of $\overset{a.s}{\Rightarrow}$).

Q.E.D.

References

[1] Abadie, A. (2000) 'Bootstrap Tests for the Effect of a Treatment on the Distribution of an Outcome Variable' mimeo.

[2] Anderson, G. (1996) 'Nonparametric Tests for Stochastic Dominance', *Econometrica*, 64, 1183-1193.

[3] Billingsley, P. (1968) *Convergence of Probability Measures*, 1st Edition, New York: John Wiley and Sons.

[4] Davidson, R. and J-Y. Duclos (2000) 'Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality' *Econometrica*, 68, 1435-1464.

[5] Deaton, A. (1997) *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy,* Baltimore: John Hopkins University Press.

[6] Eubank R., Schechtman E. and S. Yitzhaki (1993) 'A Test for 2nd Order Stochastic Dominance' *Communications in Statistics: Theory and Methods,* 22, 1893-1905.

[7] Goodman, A. W. (1967) *Modern Calculus with Analytical Geometry*, Vol. 1, New York: MacMillan, .

[8] Gourieroux, C., Holly, A. and A. Monfort (1982) 'Likelihood Ratio Test, Wald Test, and Kuhn-Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters' *Econometrica* 50, 63-80.

[9] Hansen, B. E. (1996) 'Inference When a Nuisance Parameter Is Not Identified under the Null Hypothesis' *Econometrica*, 64, 413-430.

[10] Kaur A., Prakasa Rao B.L.S. and H. Singh (1994) 'Testing for 2nd-Order Stochastic Dominance of 2 Distributions' *Econometric Theory,* 10, 849-866.

[11] Kudo, A. (1963) 'A Multivariate Analogue of One-Sided Tests' *Biometrika,* 50, 403-418.

[12] Lambert, P.J. (1993) *The Distribution and Redistribution of Income: A Mathematical Analysis*, 2nd edition, Manchester: Manchester University Press.

[13] Lehmann, E. L. (1986) *Testing Statistical Hypotheses*, 2nd Edition, New York: John Wiley and Sons.

[14] Maasoumi, E. and A. Heshmati (2000) 'Stochastic Dominance Among Swedish Income Distributions' *Econometric Reviews* 19, 287-320.

[15] McFadden, D.(1989) 'Testing for Stochastic Dominance' in *Studies in the economics of uncertainty: In honor of Josef Hadar,* ed. by T.B. Fomby and T.K. Seo, New York; Berlin; London and Tokyo: Springer.

[16] Perlman, M. D. (1969) 'One Sided Problems in Multivariate Analysis' *Annals of Mathematical Statistics,* 40, 549-567.

[17] Pollard, D. (1984) *Convergence of Stochastic Processes,* New York: Springer-Verlag.

[18] Randles, R. H. and D. A. Wolfe (1979) *Introduction to the Theory of Nonparametric Statistics,* Wiley: New York.

[19] Schmid F. and Trede M. (1998) 'A Kolmogorov-type test for second-order stochastic dominance' *Statistics and Probability Letters,* 37, 183-193.

[20] Shorack, G. R. and J. A. Wellner (1986) *Empirical Processes with Applications in Statistics,* John Wiley and Sons: New York.

[21] Shorrocks, A.F. (1983) 'Ranking Income Distributions' *Economica*, 50, 1-17.

[22] Stoline, M. R. and H. A. Ury (1979) Tables of Studentised Maximum Modulus Distribution and an Application to Multiple Comparisons Among Means' *Technometrics,* 21, 87-93.

[23] Tsirel'son, (1975) "The Density of the Distribution of the Maximum of a Gaussian Process", *Theory of Probability and its Applications,* 16, 847-856.

[24] Tong, Y. L. (1990) *The Multivariate Normal Distribution,* New York: Springer-Verlag.

[25] Van der Vaart, A. W. and J. A. Wellner (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer-Verlag.

[26] Wolak, F. A. (1989) 'Testing Inequality Constraints in Linear Econometric Models' *Journal of Econometrics,* 41, 205-235.