Computational Statistics and Data Analysis 144 (2020) 106898

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

More powerful goodness-of-fit tests for uniform stochastic ordering

Dewei Wang [a,*], Chuan-Fa Tang [b], Joshua M. Tebbs [a]

[a] *Department of Statistics, University of South Carolina, Columbia, SC 29208, USA*
[b] *Department of Mathematical Sciences, University of Texas-Dallas, Richardson, TX 75080, USA*

ARTICLE INFO

Article history:
Received 13 September 2018
Received in revised form 2 December 2019
Accepted 4 December 2019
Available online 13 December 2019

Keywords:
Antitonic regression
Hazard rate ordering
Order restricted inference
Ordinal dominance curve
Star-shaped ordering

ABSTRACT

The ordinal dominance curve (ODC) is a useful graphical tool to compare two population distributions. These distributions are said to satisfy uniform stochastic ordering (USO) if the ODC for them is star-shaped. A goodness-of-fit test for USO was recently proposed when both distributions are unknown. This test involves calculating the L^p distance between an empirical estimator of the ODC and its least star-shaped majorant. The least favorable configuration of the two distributions was established so that proper critical values could be determined; i.e., to control the probability of type I error for all star-shaped ODCs. However, the use of these critical values can lead to a conservative test and minimal power to detect certain non-star-shaped alternatives. Two new methods for determining data-dependent critical values are proposed. Simulation is used to show both methods can provide substantial increases in power while still controlling the size of the distance-based test. The methods are also applied to a data set involving premature infants. An R package that implements all tests is provided.

1. Introduction

Two population distributions, whose distribution functions are denoted by F and G, are said to satisfy uniform stochastic ordering (USO) if and only if $\{1 - F(t)\}/\{1 - G(t)\}$ is nonincreasing over the support of G; we denote this ordering by $F \preceq_{\text{USO}} G$. When F and G are absolutely continuous, USO is also known as "hazard rate ordering", which is an important characterization in reliability, survival analysis, econometrics, and actuarial science (Boland et al., 1994; Shaked and Shanthikumar, 2007; El Barmi and McKeague, 2016; Balakrishnan et al., 2018; Whang, 2019). If X and Y are random variables whose distribution functions are F and G, respectively, then an equivalent and straightforward interpretation of $F \preceq_{\text{USO}} G$ is that $\text{pr}(X > t|X > t_0) \leq \text{pr}(Y > t|Y > t_0)$ for all $t \geq t_0$. Therefore, if one regards X and Y as survival times, USO implies that all residual life distributions are stochastically ordered.

Because of its practical utility, inferential methods for distributions satisfying USO have received substantial attention in the statistics literature. Rojo and Samaniego (1993) and Mukerjee (1996) proposed nonparametric estimators of distribution functions satisfying a USO constraint. Arcones and Samaniego (2000) further examined the asymptotic properties of these estimators and proposed a conservative goodness-of-fit testing procedure for USO when one distribution (e.g., G) is known. Dykstra et al. (1991) constructed a likelihood ratio test to test the equality of several distribution functions against a global USO alternative. El Barmi and McKeague (2016) and El Barmi (2016) later examined this same test by using empirical likelihood methods.

[*] Corresponding author.
E-mail address: deweiwang@stat.sc.edu (D. Wang).

https://doi.org/10.1016/j.csda.2019.106898
0167-9473/Published by Elsevier B.V.

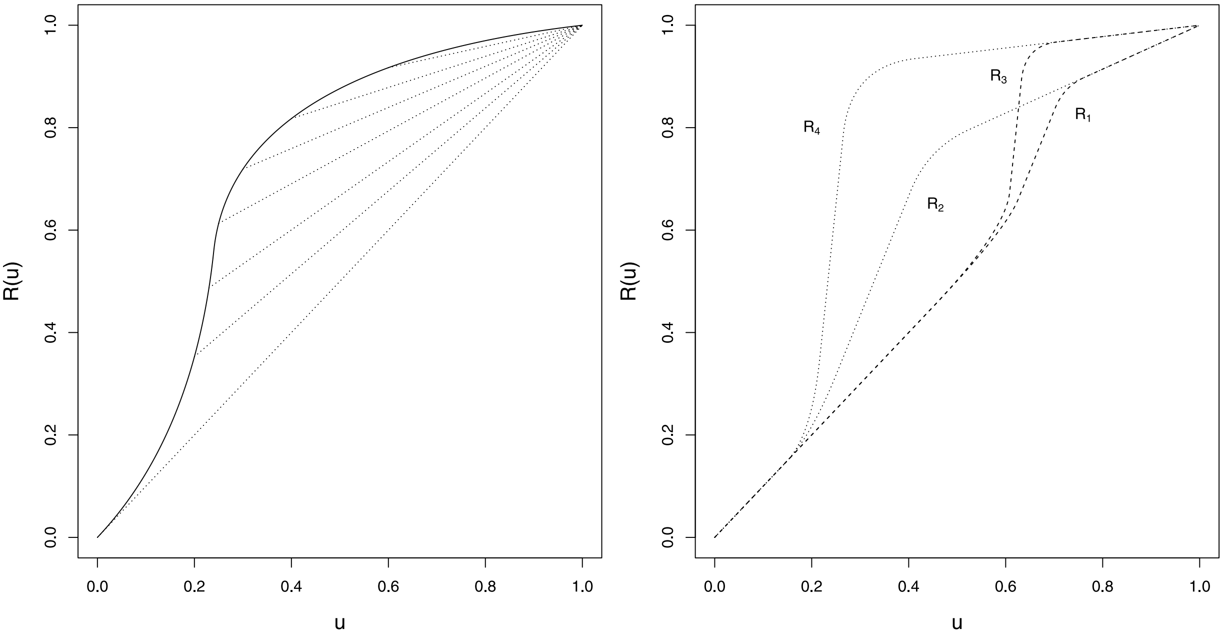

Fig. 1. Left: A star-shaped ODC $R(u)$ with dotted secant lines passing through $(1, 1)$. The slope of the secant lines, $r(u) = \{1 - R(u)\}/(1 - u)$, is nonincreasing in u. Right: Four star-shaped ODCs.

In this article, we consider testing $H_0 : F \preceq_{\text{USO}} G$ versus $H_1 : F \npreceq_{\text{USO}} G$; i.e., a goodness-of-fit test for USO. Tang et al. (2017) recently proposed a nonparametric test for H_0 versus H_1 when F and G are unknown, generalizing the earlier work of Dardanoni and Forcina (1998), Park et al. (1998), and Arcones and Samaniego (2000) for two populations. The approach in Tang et al. (2017) makes use of the ordinal dominance curve (ODC), which is a curve in $[0, 1]^2$ given by $R = FG^{-1}$, where $G^{-1}(u) = \inf\{t : G(t) \geq u\}$ is the quantile function of G. When F and G satisfy USO, the corresponding ODC R is "star-shaped" (Lehmann and Rojo, 1992), meaning that the slope of the secant line connecting $(1, 1)$ and $(u, R(u))$; i.e., $r(u) = \{1 - R(u)\}/(1 - u)$, is a nonincreasing function of u. Fig. 1 provides an illustration and examples of star-shaped ODCs.

The large-sample test statistic in Tang et al. (2017) is based on the L^p distance between an empirical estimator of the ODC and its least star-shaped majorant, with large distances indicating evidence against $H_0 : F \preceq_{\text{USO}} G$; see Section 2. When independent random samples are taken from F and G, Tang et al. (2017) showed the least favorable configuration among all star-shaped ODCs occurs when $F = G$, for all $p \geq 1$. The ODC that arises when $F = G$ is denoted by R_0 and satisfies $R_0(u) = u$, for $0 \leq u \leq 1$, the so-called equal distribution line. Therefore, critical values calculated by assuming $R = R_0$ provide an upper bound on the probability of type I error (asymptotically) for all star-shaped ODCs; i.e., for all ODCs that satisfy H_0.

Establishing the least favorable configuration provides a structured way to determine critical values for implementation; however, in finite samples, these critical values can lead to a conservative test and one with low power to detect certain non-star-shaped alternatives. Therefore, we investigate approaches to calculate data-dependent critical values instead. Our motivation is simple. If we can determine critical values by using a star-shaped ODC (possibly different than R_0) that also controls the probability of type I error asymptotically, the resulting test should be more powerful than the one in Tang et al. (2017). In this article, we propose two data-dependent approaches to identify this "possibly more generous" configuration from which critical values can be obtained. Our simulation evidence suggests both methods can be highly successful at increasing power while not unduly inflating the size of the test.

Subsequent sections of this article are organized as follows. In Section 2, we review the salient features of the ODC-based test in Tang et al. (2017) and the calculation of critical values using the least favorable configuration. In Section 3, we describe our data-dependent methods to determine new critical values. In Section 4, we provide simulation evidence to investigate the finite-sample performance of our proposals. In Section 5, we implement our tests using data from a study examining the effects of administering caffeine to premature infants. In Section 6, we conclude with a discussion. Additional details and simulation evidence are provided in the Supplementary Material.

The methods in this article can be implemented using an R package we have created and placed on GitHub at https://github.com/harrindy/TestUSO.

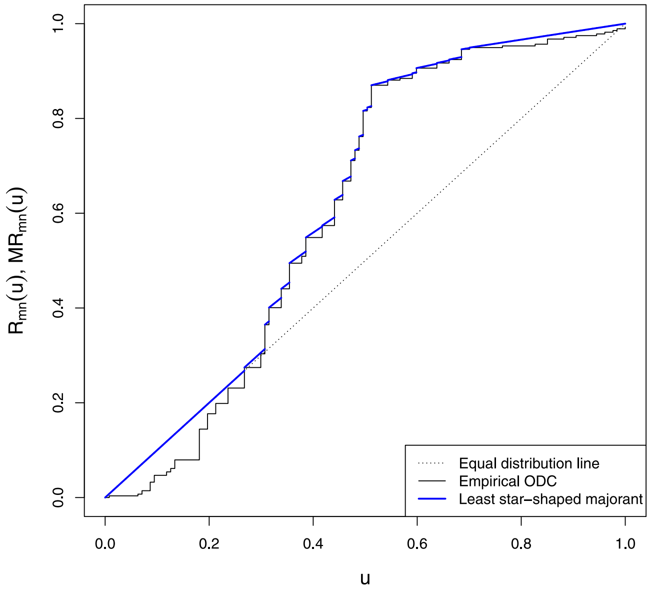

Fig. 2. Premature infant data. The empirical ODC R_{mn} is shown in black. The least star-shaped majorant $\mathcal{M}R_{mn}$ is shown in blue. The equal distribution line $R_0(u) = u$ is shown dotted. This data set is examined in Section 5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. ODC-based test and the least favorable configuration

We review the most important aspects of the large-sample test in Tang et al. (2017). Suppose we have two samples $\mathcal{X}_m = \{X_1, \ldots, X_m\}$ and $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ from unknown distributions F and G, respectively. All random variables are assumed to be mutually independent, and we assume F and G have continuous densities.

Because USO between F and G holds if and only if $R = FG^{-1}$ is star-shaped, the test for $H_0 : F \preceq_{\text{USO}} G$ versus $H_1 : F \npreceq_{\text{USO}} G$ can be written equivalently as

$$H_0 : R \text{ is star-shaped} \quad \text{versus} \quad H_1 : R \text{ is not star-shaped.}$$

The test statistic in Tang et al. (2017) is given by

$$M^p_{mn} = \{mn/(m+n)\}^{1/2} \|\mathcal{M}R_{mn} - R_{mn}\|_p,$$

where $\|\cdot\|_p$ denotes the L^p norm, $R_{mn} = F_m G_n^{-1}$ is the nonparametric maximum likelihood estimator of R under no restriction (Hsieh and Turnbull, 1996), and $\mathcal{M}R_{mn}$ is the least star-shaped majorant of R_{mn}; i.e., the smallest star-shaped function in $[0, 1]^2$ that is at least as large as the empirical ODC R_{mn}. A closed-form expression for calculating $\mathcal{M}R_{mn}$ is provided in Tang et al. (2017). Fig. 2 displays the estimates R_{mn} and $\mathcal{M}R_{mn}$ for the premature infant data we examine in Section 5.

Because M^p_{mn} measures a scaled distance between the two estimates R_{mn} and $\mathcal{M}R_{mn}$, it is easy to see that large values of M^p_{mn} will lead to the rejection of H_0. The challenging part is to determine what is meant by "large". Under certain assumptions which govern how sample sizes m and n increase without bound, Tang et al. (2017) proved that the limit of $\text{pr}_{R\in H_0}(M^p_{mn} \geq t)$ is at its maximum value when $F = G$, for all $p \geq 1$; i.e., the equal distribution line $R_0(u) = u$ is the least favorable configuration for the large-sample test of H_0 versus H_1. Therefore, critical values $c_{\alpha,p}$ calculated using R_0 maximize the (asymptotic) probability of type I error over all ODCs R which satisfy H_0, that is, rejecting H_0 when $M^p_{mn} \geq c_{\alpha,p}$ provides an asymptotic size α decision rule. For different values of α and p, Tang et al. (2017) approximated these critical values by simulating the large-sample distribution of M^p_{mn} under R_0 and tabled selected values in their supplementary article.

3. More powerful ODC-based tests

We propose two new approaches for selecting critical values when testing $H_0 : F \preceq_{\text{USO}} G$ versus $H_1 : F \npreceq_{\text{USO}} G$. Both are "data-dependent", meaning that critical values are determined by using the observed data $\mathcal{X}_m = \{X_1, \ldots, X_m\}$ and $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$. Ultimately, the goal of both approaches is to let the data identify a star-shaped configuration R^*, possibly different than R_0, that enhances our power to detect non-star-shaped alternatives while still maintaining the proper size. Critical values are then determined by using this new configuration R^*.

3.1. The antitonized slope (AS) method

The defining feature of a star-shaped ODC R is that its slope function $r(u) = \{1 - R(u)\}/(1 - u)$ is nonincreasing for $u \in [0, 1)$. This serves as the foundation for our first approach, which we call the *antitonized slope (AS) method*. Specifically, we use antitonic regression to estimate $r(u)$ from the observed data and then create a star-shaped configuration from this estimate. An outline of the AS method is below.

Step 1: Calculate the empirical ODC R_{mn} from the observed data $\mathcal{X}_m$ and $\mathcal{Y}_n$.

Step 2: For $i = 1, \dots, n-1$, calculate

$$r_{mn,i} = \frac{1 - R_{mn}\left(\frac{i}{n}\right)}{1 - \frac{i}{n}},$$

the $n-1$ secant line slopes for the empirical ODC R_{mn}.

Step 3: Calculate the antitonic regression of the slopes in Step 2; i.e., minimize

$$\sum_{i=1}^{n-1}(r_{mn,i} - \omega_i)^2 \text{ subject to } 1 \geq \omega_1 \geq \cdots \geq \omega_{n-1} \geq 0.$$

Denote the (constrained) minimizers by $\widehat{\omega}_1, \dots, \widehat{\omega}_{n-1}$. These solutions estimate $r(i/n)$, $i = 1, \dots, n-1$, subject to the constraint that R is star-shaped.

Step 4: Linearly interpolate the $n+1$ points

$$\left\{(0,0),\ \left(\frac{1}{n}, \widehat{R}_{\text{AS}}\left(\frac{1}{n}\right)\right), \dots, \left(\frac{n-1}{n}, \widehat{R}_{\text{AS}}\left(\frac{n-1}{n}\right)\right),\ (1,1)\right\},$$

where

$$\widehat{R}_{\text{AS}}\left(\frac{i}{n}\right) = 1 - \left(1 - \frac{i}{n}\right)\widehat{\omega}_i,$$

for $i = 1, \dots, n-1$. This builds a piecewise linear star-shaped curve, which we denote by $\widehat{R}_{\text{AS}}$.

We use the star-shaped configuration $R^* = \widehat{R}_{\text{AS}}$ to produce critical values similarly to how Tang et al. (2017) did so by using R_0. Specifically, we generate random samples $\mathcal{X}_m^\dagger = \{X_1^\dagger, \dots, X_m^\dagger\}$ from $\widehat{R}_{\text{AS}}$ via inverse transform sampling and $\mathcal{Y}_n^\dagger = \{Y_1^\dagger, \dots, Y_n^\dagger\}$ from a $\mathcal{U}(0,1)$ distribution and calculate the test statistic $(M_{mn,1}^{p\dagger})$ from these simulated data. We then repeat this procedure a large number of times (say, L) and select the $1-\alpha$ quantile from $\{M_{mn,l}^{p\dagger}\}_{l=1}^{L}$. Denote this quantile by $c_{\alpha,p}^{\text{AS}}$. The AS method's decision rule is to reject H_0 when the test statistic $M_{mn}^p \geq c_{\alpha,p}^{\text{AS}}$.

Our simulation results in Section 4 suggest the AS method performs quite well. In terms of power, using this method to determine critical values is guaranteed to provide a consistent test. This is true because the same ODC-based test is consistent when using the critical value $c_{\alpha,p}$ from Tang et al. (2017) and $c_{\alpha,p}^{\text{AS}}$ can never exceed $c_{\alpha,p}$. Furthermore, if a non-star-shaped $R \in H_1$ is "close" to H_0, yet differs greatly from R_0, the new critical value $c_{\alpha,p}^{\text{AS}}$ can be *much* smaller than $c_{\alpha,p}$ which will lead to large gains in power.

The AS method also does a good job at controlling the size. Recall that Tang et al. (2017) showed the non-degenerate part of the large-sample distribution of M_{mn}^p under H_0 depends on

(a) regions where $R \in H_0$ is non-strictly star-shaped; i.e., those regions where $r(u) = \{1 - R(u)\}/(1 - u)$ does not change, and
(b) the specific value of $r(u)$ over those regions.

When the samples $\mathcal{X}_m$ and $\mathcal{Y}_n$ are from R_0, the antitonic regression of the slopes in Step 3 produces an excellent estimator of $r_0(u) = \{1 - R_0(u)\}/(1 - u)$. Furthermore, our use of linear interpolation in Step 4 produces a configuration whose non-strictly star-shaped region is very often $[0, 1]$; i.e., $\widehat{R}_{\text{AS}} = R_0$, or a collection of non-strictly star-shaped regions whose union is nearly $[0, 1]$. Therefore, the limiting distribution of M_{mn}^p calculated at $\widehat{R}_{\text{AS}}$ and the one at R_0 will be identical or at least very similar. On the other hand, when the samples arise from $R \in H_0 - \{R_0\}$, antitonizing the slopes still provides an excellent estimator of $r(u)$. However, linear interpolation will generally produce a much wider non-strictly star-shaped region than the true R (which may have no such regions). Therefore, the limiting distribution of M_{mn}^p calculated at $\widehat{R}_{\text{AS}}$ will be stochastically larger than the one calculated at R, and, hence, the probability of type I error is automatically controlled.

Perhaps the nicest feature of the AS method is that it is straightforward to implement. The most challenging part is performing the antitonic regression in Step 3; however, this can be easily done by using the well known pool-the-adjacent-violators algorithm (Robertson et al., 1988) which we implement using the `OrdMonReg` package in R. The upshot is that, even when the sample sizes are large, determining $\widehat{R}_{\text{AS}}$ is nearly instantaneous. With this in mind, we now move to our second data-dependent method which utilizes the bootstrap to select a new configuration R^*.

3.2. The resample and tune (RT) method

Our second method bootstraps the observed data $\mathcal{X}_m = \{X_1, \ldots, X_m\}$ and $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ to construct resampled versions of the least star-shaped majorant $\mathcal{M}R_{mn}$. Note that our use of the bootstrap in this article is *not* to approximate the distribution of the test statistic M_{mn}^p. In fact, this is not possible because $\mathcal{M}$, when viewed as an operator on functions, is not always Hadamard directionally differentiable (Tang et al., 2017); for more details on this issue, see Dümbgen (1993) and Fang and Santos (2019). Instead, we use bootstrapping only to identify an initial candidate set of star-shaped ODCs. We then introduce a tuning parameter to select a new configuration R^* which controls the size of the ODC-based test. We call our second approach the *resample and tune (RT) method*. An outline of this method is below.

Step 1: Generate B bootstrap samples of $\mathcal{X}_m$ and $\mathcal{Y}_n$ from the empirical distributions F_m and G_n, respectively. Denote the bth bootstrap version of the least star-shaped majorant by $\mathcal{M}R^*_{mn,b}$, for $b = 1, \ldots, B$.

Step 2: For each $b = 1, \ldots, B$, calculate the slope

$$r_b^*\left(\frac{i}{n}\right) = \frac{1 - \mathcal{M}R^*_{mn,b}\left(\frac{i}{n}\right)}{1 - \frac{i}{n}},$$

for $i = 0, 1, \ldots, n-1$. For each i, calculate the $1-\gamma$ quantile of the B slopes $\{r_b^*(i/n)\}_{b=1}^B$. Denote this quantile by $\tilde{r}^\gamma(i/n)$.

Step 3: Sort the quantiles $\{\tilde{r}^\gamma(i/n)\}_{i=0}^{n-1}$ in Step 2 in descending order to obtain $\{\widehat{r}^\gamma_{\mathrm{RT}}(i/n)\}_{i=0}^{n-1}$.

Step 4: Linearly interpolate the $n+1$ points

$$\left\{(0,0),\ \left(\frac{1}{n},\ \widehat{R}^\gamma_{\mathrm{RT}}\left(\frac{1}{n}\right)\right), \ldots, \left(\frac{n-1}{n},\ \widehat{R}^\gamma_{\mathrm{RT}}\left(\frac{n-1}{n}\right)\right),\ (1,1)\right\},$$

where

$$\widehat{R}^\gamma_{\mathrm{RT}}\left(\frac{i}{n}\right) = 1 - \left(1 - \frac{i}{n}\right)\widehat{r}^\gamma_{\mathrm{RT}}\left(\frac{i}{n}\right),$$

for $i = 1, \ldots, n-1$. This builds a piecewise linear star-shaped curve, which we denote by $\widehat{R}^\gamma_{\mathrm{RT}}$.

If one wanted to select γ *a priori* (e.g., $\gamma = 0.05$, etc.), the star-shaped configuration $\widehat{R}^\gamma_{\mathrm{RT}}$ could then be used to produce critical values in the same way as $\widehat{R}_{\mathrm{AS}}$ was used in Section 3.1. That is, one could sample $\mathcal{X}_m^\dagger = \{X_1^\dagger, \ldots, X_m^\dagger\}$ from $\widehat{R}^\gamma_{\mathrm{RT}}$ and $\mathcal{Y}_n^\dagger = \{Y_1^\dagger, \ldots, Y_n^\dagger\}$ from a $\mathcal{U}(0,1)$ distribution a large number of times (say, L) and select $c^{\mathrm{RT}}_{\alpha,p}(\gamma)$, the $1-\alpha$ quantile of the collection of test statistics $\{M^{p\dagger}_{mn,l}\}_{l=1}^L$ calculated from $\mathcal{X}_m^\dagger$ and $\mathcal{Y}_n^\dagger$. The decision rule would be to reject H_0 when $M^p_{mn} \geq c^{\mathrm{RT}}_{\alpha,p}(\gamma)$.

It is important to understand how selecting γ can influence the star-shaped configuration identified by the RT method. To gain insight, refer to Fig. 3 which displays the empirical ODC R_{mn} for the premature infant data we examine in Section 5. Using $B = 1000$ bootstrap samples, we plot 999 star-shaped configurations $\widehat{R}^\gamma_{\mathrm{RT}}$ (in blue), one for each choice of $\gamma \in \{0.001, \ldots, 0.999\}$. Note that choosing γ to be large (i.e., $1-\gamma$ small) produces a star-shaped configuration far away from R_0, which would produce a much smaller critical value than the one in Tang et al. (2017). This would greatly improve the power of the test, but the probability of type I error would almost surely be inflated when $R = R_0$ (or even near R_0). On the other hand, selecting γ to be small (i.e., $1-\gamma$ large) would likely do an acceptable job at controlling the probability of type I error; however, it may offer only a small improvement in power.

Instead of selecting γ beforehand and determining $R^* = \widehat{R}^\gamma_{\mathrm{RT}}$ from it, an anonymous referee has suggested that one could actually view γ as a "tuning parameter" which can be selected perspicaciously to provide a potentially better configuration. Following this recommendation, our goal becomes to select γ in such a way that controls the size of the test yet also offers as much power gain as possible. We now describe an approach on how to make this selection.

Tuning parameter selection

Let $U_i = F_m(X_i)$ for $i = 1, \ldots, m$ and $V_j = G_n(Y_j)$ for $j = 1, \ldots, n$, where F_m and G_n are empirical distribution functions of $\mathcal{X}_m$ and $\mathcal{Y}_n$, respectively. Let $0 < \gamma_1 < \cdots < \gamma_T < 1$ be a grid on $[0,1]$.

Step TP1: Generate B bootstrap samples using $\mathcal{U}_m = \{U_1, \ldots, U_m\}$ and $\mathcal{V}_n = \{V_1, \ldots, V_n\}$. Denote the bth bootstrap version of the least star-shaped majorant by $\mathcal{M}R^{**}_{mn,b}$, for $b = 1, \ldots, B$.

Step TP2: For each $b = 1, \ldots, B$, calculate the slope

$$r_b^{**}\left(\frac{i}{n}\right) = \frac{1 - \mathcal{M}R^{**}_{mn,b}\left(\frac{i}{n}\right)}{1 - \frac{i}{n}},$$

for $i = 0, 1, \ldots, n-1$. For each i and $t = 1, \ldots, T$, calculate the $1-\gamma_t$ quantile of the B slopes $\{r_b^{**}(i/n)\}_{b=1}^B$. Denote these quantiles by $\tilde{r}^{*\gamma_t}(i/n)$.

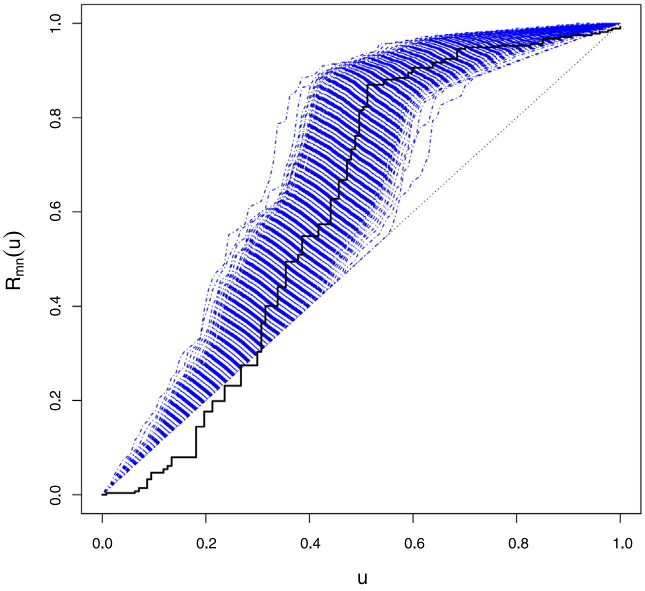

Fig. 3. Premature infant data. The empirical ODC R_{mn} is shown in black. The 999 dot-dashed curves (shown in blue) denote $\widehat{R}^{\gamma}_{\mathrm{RT}}$, as γ varies from 0.001 to 0.999 by 0.001. The equal distribution line $R_0(u) = u$ is shown dotted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Step TP3: For each $t = 1, \ldots, T$, sort the quantiles $\{\tilde{r}^{*\gamma_t}(i/n)\}_{i=0}^{n-1}$ in descending order to obtain $\{\widehat{r}^{*\gamma_t}_{\mathrm{RT}}(i/n)\}_{i=0}^{n-1}$. Let $\widehat{\gamma}$ denote the largest γ_t that solves

$$\widehat{r}^{*\gamma_t}_{\mathrm{RT}}\left(\frac{n-1}{n}\right) = 1.$$

We select $\gamma = \widehat{\gamma}$ and use this value in Steps 2, 3, and 4 of the RT method.

Several comments are in order. First, regardless of the true R that generates the observed data $\mathcal{X}_m$ and $\mathcal{Y}_n$, the transformed samples $\mathcal{U}_m$ and $\mathcal{V}_n$ are both approximately $\mathcal{U}(0, 1)$ and hence arise from a configuration close to R_0. Therefore, when $F = G$, the collection of ordered slopes $\{\widehat{r}^{\gamma_t}_{\mathrm{RT}}(i/n)\}_{i=0}^{n-1}$ based on the observed data and the collection $\{\widehat{r}^{*\gamma_t}_{\mathrm{RT}}(i/n)\}_{i=0}^{n-1}$ based on the transformed data should be similar—especially for large sample sizes. Second, the selection of $\widehat{\gamma}$ in Step TP3 is made strategically. For each value of γ_t, $t = 1, \ldots, T$, the slopes $\{\widehat{r}^{*\gamma_t}_{\mathrm{RT}}(i/n)\}_{i=0}^{n-1}$ form a nonincreasing sequence in [0, 1]. Therefore, if $\widehat{r}^{*\gamma_t}_{\mathrm{RT}}((n-1)/n) = 1$, then it must be true that $\widehat{r}^{*\gamma_t}_{\mathrm{RT}}(i/n) = 1$, for $i = 1, \ldots, n-1$. Applying linear interpolation in Step 4 of the RT method with any such value of γ_t will produce the equal distribution line R_0 (or a configuration extremely close to R_0), and the probability of type I error will be controlled when $R = R_0$. Finally, by choosing $\widehat{\gamma}$ to be the *largest* value of γ_t where $\widehat{r}^{*\gamma_t}_{\mathrm{RT}}((n-1)/n) = 1$, we are attempting to combine the "best of both worlds". This choice not only controls the size of the test but, as we illustrated in Fig. 3, a large value of γ also boosts the power when $R \in H_1$.

When selecting $\gamma = \widehat{\gamma}$ in the manner described above, Steps 2, 3, and 4 of the RT method identify the star-shaped configuration $R^* = \widehat{R}^{\widehat{\gamma}}_{\mathrm{RT}}$, and the decision rule is to reject H_0 when $M^p_{mn} \geq c^{\mathrm{RT}}_{\alpha,p}(\widehat{\gamma})$. The $1-\alpha$ quantile $c^{\mathrm{RT}}_{\alpha,p}(\widehat{\gamma})$ is determined by simulating the distribution of M^p_{mn} under $\widehat{R}^{\widehat{\gamma}}_{\mathrm{RT}}$.

4. Simulation evidence

We use simulation to evaluate our proposed data-dependent methods. To assess size properties, we use the equal distribution line R_0 and the four star-shaped ODCs R_1, R_2, R_3 and R_4 shown in Fig. 1 (right). These four ODCs are members of a larger family of star-shaped ODCs described in the supplement to Tang et al. (2017). All simulation results are based on 1000 Monte Carlo data sets. To generate these data sets, we sample $\mathcal{X}_m$ from R_i, $i = 0, 1, \ldots, 4$, using inverse transform sampling and $\mathcal{Y}_n$ from a $\mathcal{U}(0, 1)$ distribution. AS and RT critical values are determined using the configurations $\widehat{R}_{\mathrm{AS}}$ and $\widehat{R}^{\widehat{\gamma}}_{\mathrm{RT}}$, respectively, which are identified for each data set separately. We use $B = 1000$ bootstrap samples for the RT method and $L = 1000$ for both methods to simulate $c^{\mathrm{AS}}_{\alpha,p}$ and $c^{\mathrm{RT}}_{\alpha,p}(\widehat{\gamma})$, respectively. The critical value from Tang et al. (2017), $c_{\alpha,p}$, is based on the asymptotic distribution of M^p_{mn} under R_0.

Table 1 displays the results for sample sizes $m = n \in \{50, 100, 150, 200\}$, distances $p \in \{1, 2, \infty\}$, and significance level $\alpha = 0.05$. Of primary interest is the performance of AS and RT when $F = G$; i.e., when $R = R_0$. This is true because if either method cannot control the probability of type I error at the least favorable configuration, then any power gains

Table 1
Simulation results. Estimated probability of rejecting $H_0 : F \preceq_{\text{USO}} G$ when $\alpha = 0.05$ and $p \in \{1, 2, \infty\}$ for different sample sizes (m, n). The equal distribution line configuration is R_0. The remaining ODCs $R_1, \dots, R_4$ are shown in Fig. 1 (right). The results are presented in this order: Tang et al. (2017), AS (Section 3.1), and RT (Section 3.2). Size estimates (under R_0) outside the margin of error are shown bolded.

ODC	p	$m = n = 50$			$m = n = 100$			$m = n = 150$			$m = n = 200$		
		Tang	AS	RT	Tang	AS	RT	Tang	AS	RT	Tang	AS	RT
R_0	1	0.063	0.051	0.053	**0.071**	0.057	0.059	0.056	0.050	0.051	0.061	0.058	0.061
	2	0.063	0.055	0.056	0.063	0.058	0.058	0.055	0.051	0.051	0.059	0.057	0.059
	∞	0.041	0.048	0.048	0.033	0.048	0.052	0.035	0.046	0.047	0.041	0.055	0.058
R_1	1	0.022	0.020	0.015	0.011	0.019	0.010	0.011	0.025	0.013	0.009	0.024	0.010
	2	0.027	0.024	0.024	0.023	0.032	0.023	0.019	0.031	0.019	0.018	0.037	0.027
	∞	0.031	0.036	0.037	0.023	0.036	0.039	0.026	0.038	0.036	0.031	0.042	0.041
R_2	1	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.002	0.001	0.000	0.000	0.000
	2	0.000	0.002	0.000	0.000	0.002	0.000	0.000	0.004	0.001	0.000	0.006	0.000
	∞	0.000	0.007	0.003	0.001	0.007	0.001	0.002	0.014	0.004	0.000	0.014	0.006
R_3	1	0.012	0.019	0.009	0.005	0.027	0.010	0.009	0.027	0.013	0.003	0.030	0.010
	2	0.022	0.028	0.020	0.016	0.033	0.027	0.014	0.034	0.022	0.013	0.038	0.029
	∞	0.025	0.033	0.033	0.021	0.038	0.038	0.025	0.040	0.033	0.028	0.041	0.039
R_4	1	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.006	0.000	0.000	0.006	0.000
	2	0.000	0.003	0.000	0.000	0.005	0.000	0.000	0.013	0.001	0.000	0.011	0.001
	∞	0.000	0.008	0.001	0.001	0.016	0.003	0.001	0.023	0.011	0.000	0.024	0.007

would be at best specious. Note that with 1000 data sets, the margin of error in the probability of type I error estimates when $F = G$, assuming a 99 percent confidence level, is approximately 0.018. Therefore, any estimate (under R_0) outside 0.050 ± 0.018 suggests the method is not conferring the nominal size. The results in Table 1 demonstrate that, regardless of which distance is used, both data-dependent methods for determining critical values are successful at controlling the size of the ODC-based test. Furthermore, for the star-shaped ODCs R_1, R_2, R_3 and R_4, the rejection rates for AS and RT are often less conservative than when using the fixed critical values in Tang et al. (2017).

We now provide a detailed power comparison which illustrates the potential benefit of using our sample-based methods. Fig. 4 displays four sequences of ODCs. Each sequence is indexed by a parameter δ whose ODC when $\delta = 0$ is star-shaped, and increasing the value of δ produces non-star-shaped ODCs that "move farther away" from H_0. The mathematical expressions for each sequence are complex, so we relegate these to the Supplementary Material. A summary of each sequence is provided below.

- The first sequence (Fig. 4, upper left) is created by rotating the line segment AB on the equal distribution configuration R_0 (which corresponds to $\delta_a = 0$) counter-clockwise towards the segment A_1B_1, where $\delta_a = 9$. The second sequence (Fig. 4, upper right) is created analogously except one rotates AB ($\delta_b = 0$) clockwise to A_2B_2 ($\delta_b = 9$). All ODCs in both sequences, excluding the $\delta_a = 0$ and $\delta_b = 0$ members, are non-star-shaped.
- The third sequence (Fig. 4, lower left) starts with the ODC labeled with segment CD ($\delta_c = 0$), which is star-shaped. As δ_c increases, the sequence moves towards the ODC labeled with C_1D_1. ODCs corresponding to $\delta_c \in \{1, \dots, 9\}$ are not star-shaped. Each ODC in this sequence is at least as large as the equal distribution configuration R_0; i.e., F and G satisfy usual stochastic ordering but not USO when $\delta_c > 0$.
- The fourth sequence (Fig. 4, lower right) starts with R_0 ($\delta_d = 0$). The $\delta_d = 1$ ODC connects the points labeled $(1, 1)$, K, H, E, and $(0, 0)$, which admits a very small non-star-shaped region. As δ_d increases, this region increases in size; e.g., the $\delta_d = 9$ ODC connects $(1, 1)$, K, H_1, E_1, and $(0, 0)$. Similar to the third sequence, F and G satisfy usual stochastic ordering but not USO when $\delta_d > 0$.

Fig. 5 displays the power results for distances $p \in \{1, 2, \infty\}$, significance level $\alpha = 0.05$, and sample sizes $m = n = 100$. The same figures for $m = n = 50$ and $m = n = 200$ are given in the Supplementary Material. As in the previous investigation examining size, all powers are estimated by using 1000 simulated data sets. Our results in Fig. 5 are depicted by using "power curves", which are formed by letting the corresponding δ parameter in each ODC sequence increase from 0 to 9 (see Fig. 4). Each subfigure contains three curves—one for the AS method (Section 3.1), one for the RT method (Section 3.2), and one using the fixed critical values in Tang et al. (2017).

The first and second rows in Fig. 5 correspond to the δ_a and δ_b ODC sequences, respectively, whose $\delta = 0$ members coincide with the equal distribution line R_0. When comparing the data-dependent approaches, the AS method does best for the δ_a sequence while the RT method is preferred for the δ_b sequence. Because of the shape of the δ_b sequence, the AS method will usually identify $R^* = \widehat{R}_{\text{AS}} = R_0$ as its configuration; hence, its power gains are at best minimal when compared to Tang et al. (2017). For the δ_a sequence, the largest gains in power are seen for the L_1 distance statistic (i.e., $p = 1$; Fig. 5, left); for the δ_b sequence, all three distances provide similar results.

Perhaps the most promising results are observed with the δ_c and δ_d ODC sequences whose $\delta > 0$ members represent stochastic ordering between F and G; i.e., $1 - F(t) \leq 1 - G(t)$, but not USO. For these cases in Fig. 5 (bottom two rows), the

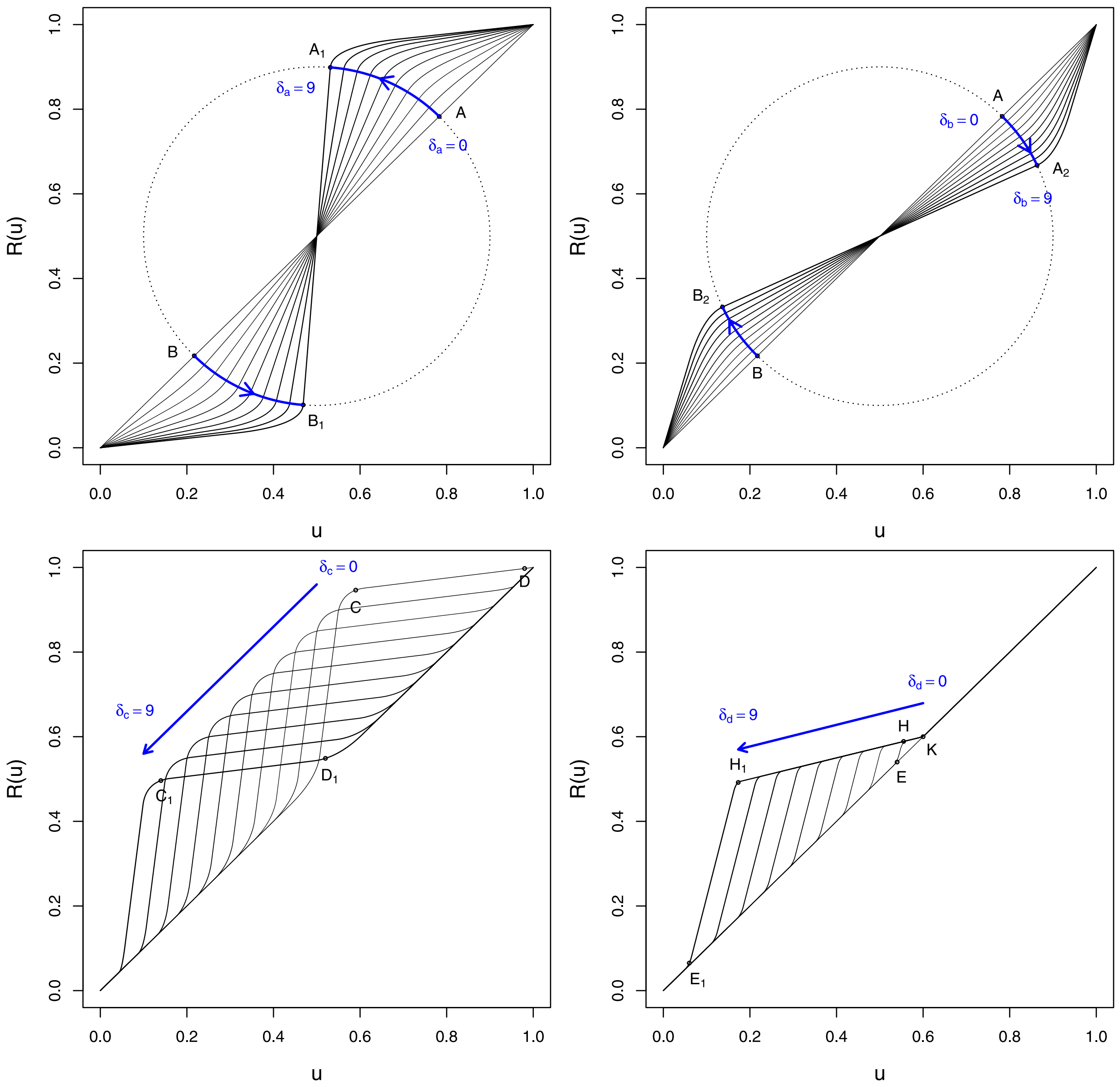

Fig. 4. ODC sequences. Each sequence is indexed by a parameter δ_a (upper left), δ_b (upper right), δ_c (lower left), and δ_d (lower right). In each subfigure, 10 ODCs are plotted corresponding to values of $\delta_a, \delta_b, \delta_c, \delta_d \in \{0, 1, \ldots, 9\}$. Explicit formulas for each sequence are given in the Supplementary Material.

RT approach is preferred among the data-dependent methods, and both AS and RT can offer enormous gains in power. This finding is especially encouraging on practical grounds. For example, suppose a researcher is confident that F and G already satisfy stochastic ordering but wants to assess whether F and G satisfy USO—a much stronger condition. Our results suggest that both AS and RT can be highly successful at discriminating between these two orderings, especially when compared to the conservative testing procedure in Tang et al. (2017).

5. Premature infant data

Cox et al. (2015) describe a retrospective study that examined the impact of administering caffeine to premature infants in Columbia, South Carolina. The primary goal of the investigation was to assess whether treating infants with caffeine was associated with an increased risk of developing necrotizing enterocolitis, a harmful medical condition characterized by bacterial infection and inflammation of the intestines. A secondary goal was to determine if exposing infants to caffeine

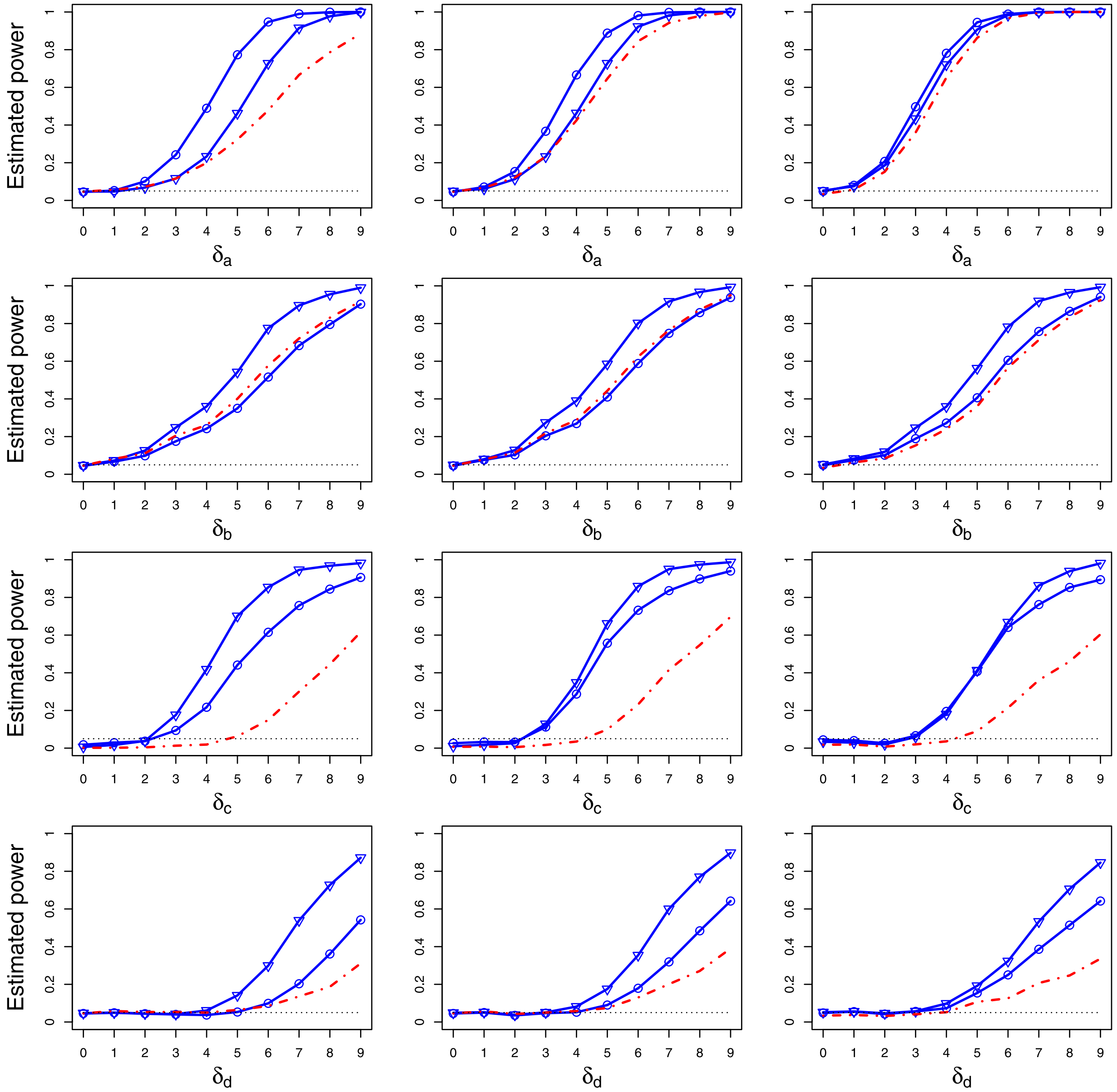

Fig. 5. Simulation results. Estimated power curves for sample sizes $m = n = 100$ and $\alpha = 0.05$. The top row corresponds to the δ_a ODC sequence; the second row corresponds to the δ_b sequence; the third row corresponds to the δ_c sequence; the bottom row corresponds to the δ_d sequence. Results are shown for distances $p = 1$ (left), $p = 2$ (middle), and $p = \infty$ (right). AS curves (Section 3.1) are shown in blue using circles; RT curves (Section 3.2) are shown in blue using triangles. The dot-dashed lines in red correspond to using the fixed critical values in Tang et al. (2017). The same figures for $m = n = 50$ and $m = n = 200$ are given in the Supplementary Material.

simultaneously conferred beneficial outcomes, such as reducing the time spent in the neonatal intensive care unit (NICU) and reducing the risk of apnea (due to underdeveloped lungs) and extubation failure.

To illustrate our data-dependent methods for selecting critical values when testing against USO, we use the same set of infants analyzed by Tang et al. (2017). Let X and Y denote the times from admission to discharge from the NICU for those infants treated with caffeine and for those not treated with caffeine, respectively. Our data set consists of $m = 127$ infants from the caffeine group and $n = 277$ infants from the no-caffeine group. As in Tang et al. (2017), we treat these groups as independent random samples from F and G, respectively. All infants were alive at the time of discharge from the NICU, that is, no times were censored.

Using the fixed critical values $c_{\alpha,p}$ from the least favorable configuration R_0, Tang et al. (2017) showed $H_0 : F \preceq_{\text{USO}} G$ would not be rejected at any reasonable significance level. Therefore, it is of interest to determine if our data-dependent

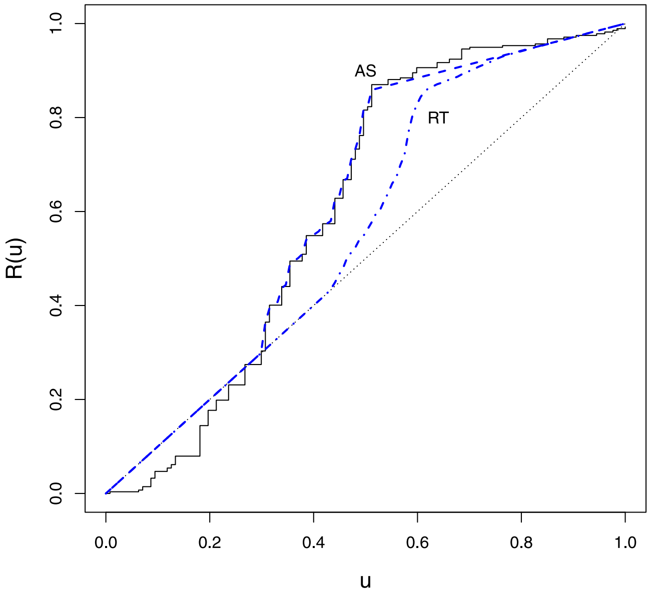

Fig. 6. Premature infant data. The empirical ODC R_{mn} is shown in black. The data-dependent configurations $\widehat{R}_{\text{AS}}$ and $\widehat{R}_{\text{RT}}^{\widehat{\gamma}}$ (labeled AS and RT) are shown dashed and dot-dashed, respectively, in blue. The equal distribution line $R_0(u) = u$ is shown dotted.

Table 2
Premature infant data analysis. Test statistics for testing $H_0 : F \preceq_{\text{USO}} G$ versus $H_1 : F \npreceq_{\text{USO}} G$ and critical values for $p \in \{1, 2, \infty\}$ when $\alpha = 0.05$. Critical values are presented in this order: Tang et al. (2017), AS (Section 3.1), and RT (Section 3.2).

Distance	Test statistic	Critical values		
		Tang	AS	RT
$p = 1$	$M_{mn}^1 = 0.170$	0.580	0.251	0.382
$p = 2$	$M_{mn}^2 = 0.263$	0.676	0.380	0.517
$p = \infty$	$M_{mn}^\infty = 0.949$	1.353	1.071	1.190

methods might reject H_0, because the corresponding critical values are likely to be smaller. In Fig. 6, we display the star-shaped configurations identified by AS and RT, respectively. The configuration $\widehat{R}_{\text{RT}}^{\widehat{\gamma}}$ was determined by using $B = 1000$ bootstrap samples, and the tuning parameter was selected to be $\widehat{\gamma} = 0.012$. For $\alpha = 0.05$, critical values $c_{0.05,p}^{\text{AS}}$ and $c_{0.05,p}^{\text{RT}}(\widehat{\gamma})$ were found by simulating the distribution of the test statistic M_{mn}^p $L = 1000$ times under $\widehat{R}_{\text{AS}}$ and $\widehat{R}_{\text{RT}}^{\widehat{\gamma}}$, respectively, and then selecting the 95th percentiles of these empirical distributions. This was done for values of $p \in \{1, 2, \infty\}$.

Table 2 shows the results. Based on the selected configurations in Fig. 6, it is not surprising that the AS method provides the smallest critical values, followed by the RT method, and that the critical values calculated from R_0 are the largest. However, when $\alpha = 0.05$, we still fail to reject $H_0 : F \preceq_{\text{USO}} G$ using either data-dependent method. This reaffirms and, in fact, strengthens the conclusion in Tang et al. (2017). That is, these data suggest the time to discharge for infants treated with caffeine is uniformly stochastically smaller than the time to discharge for infants not treated with caffeine.

6. Discussion

We have presented two data-dependent approaches to calculate improved critical values for the nonparametric test of $H_0 : F \preceq_{\text{USO}} G$ versus $H_1 : F \npreceq_{\text{USO}} G$ presented in Tang et al. (2017). The AS method uses antitonic regression to estimate the ODC of F and G under H_0, whereas the RT method uses bootstrapping to select a potentially more generous star-shaped configuration. Both methods are shown to confer the nominal size in our simulation study, and both can provide large gains in power when compared to using the fixed critical values in Tang et al. (2017). For practical implementation, our R package `TestUSO` automates the entire process of performing the test, including determining the AS and RT configurations and critical values. This package is hosted at the GitHub address given in Section 1.

It may be possible to modify our AS and RT methods in an effort to improve the power even more, but any attempt at this may be computationally overwhelming and ultimately not helpful. For example, the antitonic regression step in the AS method (Step 3) could be altered to minimize $\sum_{i=1}^{n-1}(r_{mn,i} - \omega_i)^2 w_i$, where the w_i's are user-specified weights (our Step 3 uses $w_i = 1$ for each i). A different configuration of these weights might produce a more powerful star-shaped configuration from which to select critical values. Of course, any "optimal" selection would almost surely depend on the true R, and it is unclear if all possible weight selections would control the size. In addition, one could generalize the RT method by selecting a different value of γ at each point $1/n, \ldots, (n-1)/n$ in Step 2 and then order different quantiles in Step 3. This would certainly produce a different configuration than selecting $\widehat{R}^{\gamma}_{\mathrm{RT}}$ based on a common quantile. However, the complexity involved with selecting $n-1$ tuning parameters, say $\gamma^{(1)}, \ldots, \gamma^{(n-1)}$, may not be worth the effort when compared to the approach we have taken.

A more promising avenue for future research may be to borrow ideas from the econometrics literature examining tests for likelihood ratio ordering (LRO), a stronger stochastic order arising when $R = FG^{-1}$ is concave (Carolan and Tebbs, 2005; Beare and Moon, 2015). Similar to our problem, fixed critical values from the least favorable configuration can be used; however, the resulting goodness-of-fit test for LRO is conservative and lacks power to detect non-concave alternatives. Beare and Shi (2019) overcome this problem by estimating a "contact set" associated with R (R concave) and then formulating a bootstrap procedure to calculate more generous critical values. We believe this approach could be attempted to improve the power of the test in Tang et al. (2017), but overwhelming technical challenges could arise because the least star-shaped majorant operator $\mathcal{M}$ does not enjoy certain differential properties.

Acknowledgments

We are grateful to two anonymous referees who provided insightful comments on an earlier version of this article. This research was supported by Grant R03 AI135614 and Grant R01 AI121351 from the National Institutes of Health, USA.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2019.106898.

References

Arcones, M., Samaniego, F., 2000. On the asymptotic distribution theory of a class of consistent estimators of a distribution satisfying a uniform stochastic ordering constraint. Ann. Statist. 28, 116–150. http://dx.doi.org/10.1214/aos/1016120367.

Balakrishnan, N., Zhang, Y., Zhao, P., 2018. Ordering the largest claim amounts and ranges from two sets of heterogeneous portfolios. Scand. Actuar. J. 1, 23–41. http://dx.doi.org/10.1080/03461238.2017.1278717.

Beare, B., Moon, J., 2015. Nonparametric tests of density ratio ordering. Econometric Theory 31, 471–492. http://dx.doi.org/10.1017/S0266466614000401.

Beare, B., Shi, X., 2019. An improved bootstrap test of density ratio ordering. Econom. Stat. 10, 9–16. http://dx.doi.org/10.1016/j.ecosta.2018.08.002.

Boland, P., El-Neweihi, E., Proschan, F., 1994. Applications of the hazard rate ordering in reliability and order statistics. J. Appl. Probab. 31, 180–192. http://dx.doi.org/10.2307/3215245.

Carolan, C., Tebbs, J., 2005. Nonparametric tests for and against likelihood ratio ordering in the two-sample problem. Biometrika 92, 159–171. http://dx.doi.org/10.1093/biomet/92.1.159.

Cox, C., Hashem, N., Tebbs, J., Bookstaver, B., Iskersky, V., 2015. Evaluation of caffeine and the development of necrotizing enterocolitis. J. Neonatal-Perinat. Med. 8, 339–347. http://dx.doi.org/10.3233/NPM-15814059.

Dardanoni, V., Forcina, A., 1998. A unified approach to likelihood inference on stochastic orderings in a nonparametric context. J. Amer. Statist. Assoc. 93, 1112–1123. http://dx.doi.org/10.1080/01621459.1998.10473772.

Dümbgen, L., 1993. On nondifferentiable functions and the bootstrap. Probab. Theory Related Fields 95, 125–140. http://dx.doi.org/10.1007/BF01197342.

Dykstra, R., Kochar, S., Robertson, T., 1991. Statistical inference for uniform stochastic ordering in several populations. Ann. Statist. 19, 870–888. http://dx.doi.org/10.1214/aos/1176348125.

El Barmi, H., 2016. Testing for uniform stochastic ordering via empirical likelihood under right censoring. Statist. Sinica 27, 645–664. http://dx.doi.org/10.5705/ss.202015.0042.

El Barmi, H., McKeague, I., 2016. Testing for uniform stochastic ordering via empirical likelihood. Ann. Inst. Statist. Math. 68, 955–976. http://dx.doi.org/10.1007/s10463-015-0523-z.

Fang, Z., Santos, A., 2019. Inference on directionally differentiable functions. Rev. Econom. Stud. 86, 337–412. http://dx.doi.org/10.1093/restud/rdy049.

Hsieh, F., Turnbull, B., 1996. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. Ann. Statist. 24, 25–40. http://dx.doi.org/10.1214/aos/1033066197.

Lehmann, E., Rojo, J., 1992. Invariant directional orderings. Ann. Statist. 20, 2100–2110. http://dx.doi.org/10.1007/978-1-4614-1412-4_62.

Mukerjee, H., 1996. Estimation of survival functions under uniform stochastic ordering. J. Amer. Statist. Assoc. 91, 1684–1689. http://dx.doi.org/10.1080/01621459.1996.10476738.

Park, C., Lee, C., Robertson, T., 1998. Goodness-of-fit test for uniform stochastic ordering among several populations. Canad. J. Statist. 26, 69–81. http://dx.doi.org/10.2307/3315674.

Robertson, T., Wright, F., Dykstra, R., 1988. Order Restricted Statistical Inference. Wiley, New York.

Rojo, J., Samaniego, F., 1993. On estimating a survival curve subject to a uniform stochastic ordering constraint. J. Amer. Statist. Assoc. 88, 566–572. http://dx.doi.org/10.1080/01621459.1993.10476308.

Shaked, M., Shanthikumar, J., 2007. Stochastic Orders. Springer, New York.

Tang, C., Wang, D., Tebbs, J., 2017. Nonparametric goodness-of-fit tests for uniform stochastic ordering. Ann. Statist. 45, 2565–2589. http://dx.doi.org/10.1214/16-AOS1535.

Whang, Y., 2019. Econometric Analysis of Stochastic Dominance: Concepts, Methods, Tools, and Applications. Cambridge University Press, New York.