Statistica Sinica Preprint No: SS-2023-0034	
Title	Nonparametric Comparisons of Multiple Distributions under Uniform Stochastic Ordering
Manuscript ID	SS-2023-0034
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0034
Complete List of Authors	Chuan-Fa Tang and Dewei Wang
Corresponding Authors	Chuan-Fa Tang
E-mails	chuan-fa.tang@utdallas.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

Nonparametric Comparisons of Multiple Distributions under Uniform Stochastic Ordering

Chuan-Fa Tang and Dewei Wang

University of Texas at Dallas and University of South Carolina

Abstract:

Uniform stochastic ordering (USO), also known as hazard rate or failure rate ordering, has garnered significant interest across various applications. In this study, we present nonparametric approaches for comparing distributions within the framework of USO using the ordinal dominance curve. Our study consists of three main components. The first one offers new tests for equality among multiple distributions under USO assumptions. Secondly, we provide goodness-of-fit tests to investigate whether multiple distributions adhere to USO. Lastly, we identify distributions that exhibit significant statistical differences within the context of USO. We provide asymptotic properties and supporting numerical evidence for our proposed methods. To exemplify the application of our inferential techniques, we focus on a biomarker, microfibrillar-associated protein 4, and assess its potential for diagnosing fibrosis stages in hepatitis C patients.

Key words and phrases: Brownian bridge, Bonferroni correction, Hazard rate ordering, Ordinal dominance curve, Order-restricted inference.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

1. Introduction

Uniform Stochastic Ordering (USO) has held significant importance across various applications since it was introduced by Lehmann (1955). Define two independent random variables X_1 and X_2 with distributions F_1 and F_2, respectively. Denoted by $X_1 \preceq X_2$ or $F_1 \preceq F_2$, we say X_1 is smaller than X_2 in the sense of USO when the ratio $\{1-F_1(t)\}/\{1-F_2(t)\}$ is non-increasing in t whenever $F_2(t) < 1$. When both F_1 and F_2 are absolutely continuous, USO is also known as hazard rate ordering. Applications in actuarial science, biology, economics, reliability, and survival analysis further emphasize the versatility of USO, e.g., see Dykstra et al. (1991), Navarro and Shaked (2006), Da and Ding (2016), and Balakrishnan et al. (2018). USO is stronger than ordinary stochastic ordering but weaker than likelihood ratio ordering (Keilson and Sumita, 1982; Shaked and Shanthikumar, 2007).

When data are collected from $k > 2$ distributions, a well-ordered USO, say $F_1 \preceq F_2 \preceq \cdots \preceq F_k$, has garnered considerable attention in the literature. In this work, we also focus on $F_1 \preceq F_2 \preceq \cdots \preceq F_k$ and examine three fundamental questions: 1) testing if all distributions are identical under USO, 2) assessing the occurrence of USO, and 3) identifying which distributions differ in the context of USO.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

Statistical inference often initially involves testing the equality of distributions, denoted as $F_1 = F_2 = \cdots = F_k$. Under USO, Kochar (1979) provided a U-statistic-based test for equal distributions with $k = 2$. When $k \geq 2$, Dykstra et al. (1991) developed a likelihood ratio test by discretizing the support of the data. El Barmi and McKeague (2016) and El Barmi (2017) proposed empirical likelihood-based tests which localized hypotheses and accumulated local test statistics. Daradanoni and Forcina (1998) unified likelihood ratio tests for discrete distributions satisfying stochastic orderings, including USO. However, the discretization in likelihood-based approaches may be subjective and lose efficiency. On the other hand, in the case of equal sample sizes n, the empirical likelihood-based test may require computing $kn(kn+1)/2$ local test statistics and U-statistics of order two include $\{n(n+1)/2\}^2$ terms. Consequently, both approaches pose computational challenges for large sample sizes. We propose ordinal dominance curve (ODC) approaches to circumvent the issues mentioned above.

The ODC, also known as the probability-probability plot, is a graphical tool visualizing the relationship between two distributions. Given two distributions F_1 and F_2, the ODC is defined by $R_1(u) = F_1\{F_2^{-1}(u)\}$ for $0 \leq u \leq 1$ where F_2^{-1} is the quantile function associated with F_2. When $F_1 \preceq F_2$, the ODC R_1 exhibits star-shaped, that is, $\{1 - R_1(u)\}/(1-u)$ is

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

nonincreasing in u (Tang et al., 2017). In the case of $F_1 = F_2$, the ODC R_1 is the diagonal line in the unit square, denoted by R_0. Hence, ODC-based equality tests for multiple distributions utilize L^p differences between shape-restricted estimators of ODCs and equal distribution line R_0 as evidence of unequal distributions (Carolan and Tebbs, 2005; Davidov and Herman, 2012). Here we apply star-shaped estimators for ODCs $R_i(\cdot) = F_i\{F_{i+1}^{-1}(\cdot)\}$ proposed by Tang et al. (2017) for each $1 \leq i < k$ and accumulate the $k-1$ differences.

To check for the presence of USO, conducting goodness-of-fit (GOF) tests for $F_1 \preceq F_2 \preceq \cdots \preceq F_k$ is essential. Park et al. (1998) proposed a likelihood ratio test among multiple distributions by discretizing the data's support; however, how to properly discretize remains unsolved. In contrast, two-sample ODC-based GOF tests have been studied by Tang et al. (2017), Wang et al. (2020), and Wang et al. (2021). The L^p differences between star-shaped and empirical estimators of the ODC R_1 were used to test if $F_1 \preceq F_2$ holds. For k-sample GOF tests, it is a natural extension to consider ODCs $R_1, \ldots, R_{k-1}$ and accumulate the L^p differences. However, this approach requires extra care because ODC estimators become correlated ($R_i = F_i\{F_{i+1}^{-1}(u)\}$ and $R_{i+1} = F_{i+1}\{F_{i+2}^{-1}(u)\}$ both depend on F_{i+1}). We propose data-adaptive critical values addressing the dependency and

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

provide Bonferroni-corrected GOF tests for comparison.

Lastly, if distributions are known to differ in advance, it is crucial to determine which distributions change. For example, determining distinguishable distributions of a noninvasive biomarker over stages of a disease is a critical assessment when searching for a replacement of an invasive "gold standard" diagnosis; see Section 7. However, to the best of our knowledge, none of the equality tests can distinguish unequal distributions under USO. We propose ODC-based methods capable of detecting distributional changes to address this gap. We further offer a BIC-type method even if distributions are unknown to differ beforehand.

The rest of this paper is organized as follows. We propose ODC-based equality tests under USO in Section 2. Section 3 constructs GOF tests for USO with data-adaptive and Bonferroni-corrected critical values. Methods to distinguish distributions under USO are proposed in Section 4. We provide numerical justification in Section 5 to support theoretical results. Section 6 illustrates our methods when assessing a new serum biomarker for distinguishing hepatic fibrosis stages. Finally, we conclude with a discussion in Section 7. All proofs and additional numerical results are provided in the Supplementary Materials. The `R` codes on GitHub (https://github.com/cftang9/MSUSO) reproduce all numerical results.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

2. Equality Tests

With $k \geq 2$ continuous distributions $F_1, F_2, \ldots, F_k$, the relevant hypotheses for testing equality under USO are

$$H_0 : F_1 = F_2 = \cdots = F_k \quad \text{and} \quad H_1 : F_1 \preceq F_2 \preceq \cdots \preceq F_k \ \text{ but not } H_0.$$

Given the transitivity property of "$\preceq$" in USO, the alternative hypothesis H_1 can be expressed as $F_i \preceq F_{i+1}$ for all $1 \leq i < k$, excluding H_0. Thus, the analysis can focus on pairs of distributions F_i and F_{i+1}. We denote the corresponding ODC as $R_i(u) = F_i\{F_{i+1}^{-1}(u)\}$ for $0 \leq u \leq 1$, where $F_{i+1}^{-1}(u) = \inf\{t : F_{i+1}(t) \geq u\}$ is the quantile function associated with F_{i+1}. The ordering $F_i \preceq F_{i+1}$ is equivalent to R_i being star-shaped (i.e., the ratio $r_i(u) = \{1 - R_i(u)\}/(1-u)$ is nonincreasing in u). Therefore, H_1 can be stated equivalently as all R_i being star-shaped, with the condition that at least one R_i deviates from the equal distribution line R_0. Similarly, H_0 is equivalent to $R_i = R_0$ for all $1 \leq i < k$.

Our test statistics rely on the disparity between a star-shaped estimator of R_i and R_0. Given independent random samples $X_{j1}, \ldots, X_{jn_j}$ from distributions F_j for $1 \leq j \leq k$, we denote the empirical estimator of R_i as $\hat{R}_i(u) = \mathbb{F}_i\{\mathbb{F}_{i+1}^{-1}(u)\}$ for $0 \leq u \leq 1$. Here, $\mathbb{F}_i(t)$ represents the empirical

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

distribution of the ith random sample and $\mathbb{F}_{i+1}^{-1}(u)$ is the empirical quantile function of the $(i+1)$th random sample. The star-shaped estimator, following the approach of Tang et al. (2017), is defined as the *least star-shaped majorant* of $\hat{R}_i$; i.e., the smallest star-shaped function that is greater than $\hat{R}_i$. This estimator is denoted as $\mathcal{M}\hat{R}_i$ and has an explicit expression:

$$\mathcal{M}\hat{R}_i(u) = \begin{cases} 1-(1-u)\inf\limits_{0\le v\le u}\{1-\hat{R}_i(v)\}/(1-v), & \text{for } 0\le u<1, \\ 1, & \text{when } u=1. \end{cases}$$

From this expression, it follows that $\mathcal{M}\hat{R}_i(u) \ge R_0(u)$ holds for all $u\in[0,1]$ because $\hat{R}_i(0)=0$ so that $\inf\limits_{0\le v\le u}\{1-\hat{R}_i(v)\}/(1-v) \le \{1-\hat{R}_i(0)\}/(1-0)=1$.

We consider scaled L^p differences between $\mathcal{M}\hat{R}_i$ and R_0:

$$\Delta_{ip} = C_i\|\mathcal{M}\hat{R}_i - R_0\|_p,$$

where $C_i = \{n_in_{i+1}/(n_i+n_{i+1})\}^{1/2}$ serves as a normalizing constant and $\|\cdot\|_p$ stands for the L^p functional norm with $p\in[1,\infty]$. For example, when $p=1$, $\Delta_{i1} = C_i(\int_0^1 \mathcal{M}\hat{R}_i(u)\,du - \frac{1}{2})$; when $p=\infty$, $\Delta_{i\infty} = C_i\sup\limits_{0\le u\le 1}\{\mathcal{M}\hat{R}_i(u)-u\}$. Clearly, large values of Δ_{ip} are evidence of $F_i \preceq F_{i+1}$ excluding $F_i = F_{i+1}$. Therefore, to test $H_0: F_1 = F_2 = \cdots = F_k$, one can aggregate all Δ_{ip} to

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

construct relevant test statistics. Here we choose

$$T_{kp} = \sum_{1 \le i < k} \Delta_{ip} \quad \text{and} \quad U_{kp} = \max_{1 \le i < k} \Delta_{ip}$$

and denote critical values by $t_{kp,\alpha}$ and $u_{kp,\alpha}$ for T_{kp} and U_{kp}, respectively. The critical values satisfy $\alpha = \text{pr}(T_{kp} > t_{kp,\alpha}) = \text{pr}(U_{kp} > u_{kp,\alpha})$ under H_0. Theorem 1 shows our proposed tests are consistent.

Theorem 1. *When* $F_1, \ldots, F_k$ *satisfy* H_1, $\lim_{n\to\infty} \text{pr}(T_{kp} > t_{kp,\alpha}) = 1$ *and* $\lim_{n\to\infty} \text{pr}(U_{kp} > u_{kp,\alpha}) = 1$, *where* $n = \min\{n_1, \ldots, n_k\}$.

In practice, we recommend approximating $t_{kp,\alpha}$ and $u_{kp,\alpha}$ by generating Monte Carlo samples under H_0. According to Remark S1.1 in the Supplementary Materials, the sampling distributions of both $\mathcal{M}\hat{R}_i$ and $\hat{R}_i$ are solely dependent on R_i and sample sizes, irrespective of the underlying distributions. Consequently, it suffices to generate independent random samples $X_{j1}, \ldots, X_{jn_j}$ from the uniform distribution $\mathcal{U}(0,1)$ for $j = 1, \ldots, k$. In our analysis, we implemented this procedure $10,000$ times and estimated $t_{kp,\alpha}$ and $u_{kp,\alpha}$ by using the α-th upper quantile of the simulated values of T_{kp} and U_{kp}, respectively. We provide selected values of $t_{kp,\alpha}$ and $u_{kp,\alpha}$ in Table S1.1 in the Supplementary Materials.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

3. Goodness-of-fit tests

It is important to highlight that rejecting H_0 (equality) in favor of H_1 (USO) does not necessarily imply the ordering holds. For further assessment, we explore the goodness-of-fit (GOF) test for

$$H_0^* : F_1 \preceq F_2 \preceq \cdots \preceq F_k \quad \text{versus} \quad H_1^* : \text{ not } H_0^*.$$

3.1 Bonferroni-corrected critical values

Using a Bonferroni correction is a straightforward way to extend the two-sample GOF tests proposed by Tang et al. (2017) to $k > 2$ samples. For the ith and the $(i+1)$th samples, the two-sample GOF test examines the hypotheses $H_{0i}^* : F_i \preceq F_{i+1}$ versus the opposite. The test statistic, denoted by M_{ip}, is defined as

$$M_{ip} = C_i \|\mathcal{M}\hat{R}_i - \hat{R}_i\|_p,$$

which is a scaled distance between $\mathcal{M}\hat{R}_i$ and $\hat{R}_i$. Because the hypothesis H_{0i}^* is composite, determining a critical value that controls the type I error probability is challenging. Tang et al. (2017) proved that equal distributions $F_i = F_{i+1}$ serve as the asymptotic *least favorable configuration*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

maximizing the type I error probability, that is, $\sup_{F_i \preceq F_{i+1}} \lim_{n\to\infty} \mathrm{pr}(M_{ip} > t) \leq \sup_{F_i = F_{i+1}} \lim_{n\to\infty} \mathrm{pr}(M_{ip} > t)$ for $t \geq 0$. Thus, for large sample sizes, considering the limiting distribution of M_{ip} under $F_i = F_{i+1}$ suffices. When $F_i = F_{i+1}$, M_{ip} converges in distribution to $\|\mathcal{D}\|_p$ as $n \to \infty$, where $\mathcal{D}(u) = (1-u) \sup_{0\leq v\leq u} \{\mathcal{B}(v)/(1-v)\} - \mathcal{B}(u)$ for $0 \leq u < 1$, $\mathcal{D}(1) = 0$, and $\mathcal{B}$ is a standard Brownian bridge. Let $b_{p,\alpha}$ denote the α-th upper quantile of $\|\mathcal{D}\|_p$. The type I error probability is asymptotically controlled under H_{0i}^*, that is, $\lim_{n\to\infty} \mathrm{pr}(M_{ip} > b_{p,\alpha}) \leq \mathrm{pr}(\|\mathcal{D}\|_p > b_{p,\alpha}) = \alpha$.

Applying a Bonferroni correction, we define the test statistic $W_{kp} = \max_{1\leq i<k} M_{ip}$ and reject H_0^* when $W_{kp} > b_{p,\alpha/(k-1)}$. Utilizing the definition of $\|\mathcal{D}\|_p$, and assuming $F_1, \ldots, F_k$ satisfy H_0^*, the following inequality holds

$$\begin{aligned}\lim_{n\to\infty} \mathrm{pr}\left(W_{kp} > b_{p,\alpha/(k-1)}\right) &\leq \lim_{n\to\infty} \sum_{i=1}^{k-1} \mathrm{pr}(M_{ip} > b_{p,\alpha/(k-1)}) \\ &\leq \sum_{i=1}^{k-1} \mathrm{pr}(\|\mathcal{D}\|_p > b_{p,\alpha/(k-1)}) = (k-1) \times \frac{\alpha}{k-1} = \alpha. \quad (3.1)\end{aligned}$$

Therefore, the Bonferroni-corrected tests have size α asymptotically.

3.2 Data-adaptive critical values

While theoretically sound, Bonferroni-corrected tests can be conservative in the finite-sample cases, especially when k is large. Therefore, we propose

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

a more sophisticated method to determine the critical values. Our method avoids using the summation of marginal upper bounds, as in (3.1), for controlling the type I error probability. Instead, we seek to employ data-adaptive critical values, steering clear of the need to search for a fixed and conservative value from the least favorable configuration.

We consider the test statistics $W_{kp} = \max_{1\le i<k} M_{ip}$ defined in Section 3.1 and $S_{kp} = \sum_{1\le i<k} M_{ip}$. The asymptotic behavior of M_{ip} inspires our construction of critical values. Lemma S1.3 in the Supplementary Materials shows that when R_i is star-shaped, the limiting distribution of M_{ip} relies on the segment of the process

$$\mathcal{L}_i(u) = \begin{cases} \lambda_i^{1/2}\mathcal{B}_i\{\mathcal{M}R_i(u)\} - (1-\lambda_i)^{1/2}\left\{\frac{1-\mathcal{M}R_i(u)}{1-u}\right\}\mathcal{B}_{i+1}(u), & 0 \le u < 1, \\ 0, & u = 1, \end{cases}$$

over certain regions of $[0,1]$, where λ_i is the limit of $n_{i+1}/(n_i + n_{i+1})$ as $n = \min_{1\le i<k}\{n_i\} \to \infty$ and the $\mathcal{B}_i$'s are independent standard Brownian bridges. These regions are subsets of $[0,1]$, known as the *non-strictly-star-shaped* regions of R_i (see Tang et al., 2017). Achieving precise control over the type I error probability necessitates consistent estimation of these regions, a task that proves to be exceedingly challenging. To bypass this challenge, we opt to expand these regions to $[0,1]$ as a pragmatic means of

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

bounding the type I error probability. We utilize $\mathcal{L}_i(u)$ to define

$$\tilde{M}_{ip} = \begin{cases} \left[\int_0^1 \left\{\sup_{0\le v\le u}\left(\frac{\mathcal{L}_i(v)}{1-v}\right)(1-u) - \mathcal{L}_i(u)\right\}^p du\right]^{1/p}, & 1 \le p < \infty, \\ \sup_{0\le u\le 1}\left\{\sup_{0\le v\le u}\left(\frac{\mathcal{L}_i(v)}{1-v}\right)(1-u) - \mathcal{L}_i(u)\right\}, & p = \infty, \end{cases}$$

in which the integral and supremum (of u) are both over $[0,1]$. In the Supplementary Materials (Remark S1.2), we show that, for $p \in [1,\infty]$,

$$\lim_{n\to\infty} \text{pr}(M_{ip} > t) \le \text{pr}(\tilde{M}_{ip} > t) \text{ holds at any } t \ge 0.$$

Hence, we can employ $\tilde{S}_{kp} = \sum_{1\le i<k} \tilde{M}_{ip}$ and $\tilde{W}_{kp} = \max_{1\le i<k} \tilde{M}_{ip}$ to find critical values, denoted by $\tilde{s}_{kp,\alpha}$ and $\tilde{w}_{kp,\alpha}$, respectively.

Theorem 2. *When $F_1,\ldots,F_k$ satisfy H_0^*, $\lim_{n\to\infty}\text{pr}(S_{kp} > \tilde{s}_{kp,\alpha}) \le \alpha$ and $\lim_{n\to\infty}\text{pr}(W_{kp} > \tilde{w}_{kp,\alpha}) \le \alpha$ hold for every $p \in [1,\infty]$. When $F_1,\ldots,F_k$ satisfy H_1^*, $\lim_{n\to\infty}\text{pr}(S_{kp} > \tilde{s}_{kp,\alpha}) = 1$ and $\lim_{n\to\infty}\text{pr}(W_{kp} > \tilde{w}_{kp,\alpha}) = 1$.*

Provided that $\tilde{s}_{kp,\alpha}$ or $\tilde{w}_{kp,\alpha}$ are known in advance, Theorem 2 shows they are desirable as new critical values in testing H_0^*, ensuring type I error probabilities are controlled asymptotically. Moreover, under H_1^*, both $\tilde{s}_{kp,\alpha}$ and $\tilde{w}_{kp,\alpha}$ still guarantee consistency.

Now we provide a numerical approximation for $\tilde{s}_{kp,\alpha}$ and $\tilde{w}_{kp,\alpha}$. It is important to note that both $\tilde{S}_{kp}$ and $\tilde{W}_{kp}$ depend on $\mathcal{L}_i$ with unknown $\mathcal{M}R_i$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

and λ_i. Therefore, we estimate $\mathcal{M}R_i$ by its consistent estimator $\mathcal{M}\hat{R}_i$ and approximate λ_i by $n_{i+1}/(n_i+n_{i+1})$. Lastly, the Brownian bridges $\mathcal{B}_1, \ldots, \mathcal{B}_k$ in $\mathcal{L}_1, \ldots, \mathcal{L}_{k-1}$ are replaced by a sequence of independent Brownian bridges $\mathcal{B}_1^*, \ldots, \mathcal{B}_k^*$, which are approximated by independently generating k random samples of size K from $\mathcal{U}(0,1)$. Thus, we approximate $\mathcal{L}_i$ by

$$\widehat{\mathcal{L}}_i^*(u) = \left(\frac{n_{i+1}}{n_i+n_{i+1}}\right)^{1/2} \mathcal{B}_i^*\{\mathcal{M}\hat{R}_i(u)\} - \left(\frac{n_i}{n_i+n_{i+1}}\right)^{1/2} \left\{\frac{1-\mathcal{M}\hat{R}_i(u)}{1-u}\right\} \mathcal{B}_{i+1}^*(u),$$

for $u \in [0,1)$ and $\widehat{\mathcal{L}}_i^*(1) = 0$ and plug $\mathcal{L}_i = \widehat{\mathcal{L}}_i^*$ into $\tilde{M}_{ip}$ to obtain both $\hat{S}_{kp}^*$ and $\hat{W}_{kp}^*$ accordingly. By generating the approximated Brownian bridges B times, we obtain B pairs of $\hat{S}_{kp}^*$ and $\hat{W}_{kp}^*$. It is advisable to select much larger values of K and B than the maximum of the k sample sizes (e.g., we set $K = B = 1000$ when sample sizes are less than or equal to 200). Lastly, upper α-th sample quantiles, denoted by $\hat{s}_{kp,\alpha}^*$ and $\hat{w}_{kp,\alpha}^*$, serve as the estimated critical values for S_{kp} and W_{kp}, respectively. Theorem 3 shows that both tests have size α asymptotically and are consistent under H_1^*.

Theorem 3. *When $F_1, F_2, \ldots, F_k$ satisfy H_0^*, $\lim_{n\to\infty} \mathrm{pr}(S_{kp} > \hat{s}_{kp,\alpha}^*) \le \alpha$ and $\lim_{n\to\infty} \mathrm{pr}(W_{kp} > \hat{w}_{kp,\alpha}^*) \le \alpha$. When $F_1, F_2, \ldots, F_k$ satisfy to the alternative H_1^*, $\lim_{n\to\infty} \mathrm{pr}(S_{kp} > \hat{s}_{kp,\alpha}^*) = \lim_{n\to\infty} \mathrm{pr}(W_{kp} > \hat{w}_{kp,\alpha}^*) = 1$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

4. Distinguishing Distributions

When $H_0^* : F_1 \preceq F_2 \preceq \cdots \preceq F_k$ is true, it is still possible that some F_i's are equal and not strictly ordered. The final component of our work focuses on identifying all potential indices i where $F_i \preceq F_{i+1}$ while $F_i \neq F_{i+1}$. To simplify, we refer to this relationship as $F_i \prec F_{i+1}$ and label such an index i as a "jump point." We collect the set of true jump points as $J = \{i : F_i \prec F_{i+1}\}$. In essence, our objective is to pinpoint all jump points under H_0^*.

To identify J, we can effectively leverage the statistic Δ_{ip} as defined in Section 2. Recall that large values of Δ_{ip} indicate that R_i differs from R_0 in the context of USO, suggesting $F_i \prec F_{i+1}$. Consequently, it is logical to designate i as a jump point when Δ_{ip} exceeds the cutoff value $u_{kp,\alpha}$ defined in Section 2. We then gather the identified jump points as $J_p^0 = \{i : \Delta_{ip} > u_{kp,\alpha}\}$. Theorem 4 provides the probability of J_p^0 containing J and accurately determining J under both H_0 and H_1.

Theorem 4. *Under* H_1, $\mathrm{pr}(J_p^0 \supseteq J)$ *converges to* 1 *and* $\mathrm{pr}(J_p^0 = J) \geq 1-\alpha$ *as* $n \to \infty$. *Under* H_0, $\mathrm{pr}(J_p^0 \supseteq J) = 1$ *and* $\mathrm{pr}(J_p^0 = J) = 1 - \alpha$ *for all finite sample sizes.*

Theorem 4 indicates that, under $H_0^* = H_0 \cup H_1$, J_p^0 guarantees containing all the true jump points J with probability approaching 1. However, J_p^0

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

may falsely include some non-jumping points with probability smaller than or equal to α, preventing the correct rate from approaching 1, especially when H_0 is true. To improve the correct rate, we propose a distinguishing method with the correct rate converging to 1 under H_0. We introduce threshold values denoted by $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{k-1})$, where δ_i increases with sample sizes such that all $\delta_i \to \infty$ as $n \to \infty$. We then gather the identified jump points using $J_p^*(\boldsymbol{\delta}) = \{i : \Delta_{ip} > \delta_i\}$. Theorem 5 suggests that as long as δ_i grows slower than C_i (i.e., $\delta_i/C_i \to 0$ as $n \to \infty$), $J_p^*(\boldsymbol{\delta})$ will correctly determine all jump points in the limit.

Theorem 5. *Under H_0^*, for any $\emptyset \subseteq J \subseteq \{1, \ldots, k-1\}$, $\lim_{n\to\infty} \mathrm{pr}(J_p^*(\boldsymbol{\delta}) = J) = 1$, provided that $\delta_i \to \infty$ and $\delta_i/C_i \to 0$ as $n \to \infty$, for $1 \le i < k$.*

In practice, one can simplify the threshold values by setting all $\delta_i = \eta$ and defining $J_p^*(\eta) = \{i : \Delta_{ip} > \eta\}$. We utilize BIC-type criteria to determine η by minimizing the loss function

$$Q_p(\eta) = \sum_{i \notin J_p^*(\eta)} \|\mathcal{M}\hat{R}_i - R_0\|_p + \sum_{i \in J_p^*(\eta)} \left(\|\mathcal{M}\hat{R}_i - \hat{R}_i\|_p + d_{ip} \frac{\log C_i}{C_i} \right).$$

Here, R_0 is viewed as a "parsimonious model," and a BIC-type penalty term $d_{ip}(\log C_i)/C_i$ is added to account for the departure from a star-shaped configuration. We suggest using $d_{ip} = \log\log C_i\{(1+p)/(2+p)\}$, as the term

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

$\log\log C_i$ is recommended by Wang et al. (2009) and the term $(1+p)/(2+p)$ is introduced to appropriately adjust the magnitude for the L^p errors. Denote the minimizer of $Q_p(\eta)$ by η_p^*, Theorem 6 shows that $J_p^* := J_p^*(\eta_p^*)$ effectively identifies all jump points with probability converging to 1.

Theorem 6. *Under H_0^*, for any $\emptyset \subseteq J \subseteq \{1, \ldots, k-1\}$, $\lim_{n\to\infty} \mathrm{pr}(J_p^* = J) = 1$.*

Because $Q_p(\eta)$ is a step function, η_p^* can be efficiently computed. By setting $\eta_i^\dagger = \Delta_{ip}$ and $\eta_0^\dagger = 0$, we compute $Q_p(\eta)$ over the values $\eta_i^\dagger$ for $i = 0, \ldots, k-1$ and select the one that yields the smallest $Q_p(\eta)$. It is important to highlight that while the probability of correct identification approaches 1, our finite-sample evaluation in Section 5.5 reveals that J_p^* tends to be a bit more conservative when compared to J_p^0.

5. Simulation

5.1 ODC Examples

Throughout the assessment, we consider a family of ODCs denoted as G_q for $-1 \le q \le 1$, depicted in Figure 1(a). The configuration of these ODCs is controlled by the parameter q. Specifically, G_q exhibits a star-shaped pattern when q is non-negative and a non-star-shaped pattern otherwise.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

To evaluate the sensitivity and robustness of our tests, we measure the degree of departure of the corresponding null hypothesis. The departure of an ODC R from the equal distribution line R_0 is measured by $D_0(R,p) = \|\mathcal{M}R - R_0\|_p$. Therefore, $D_0(G_0,p) = 0$ because $G_0 = R_0$. For $q < 0$, $D_0(G_q,p) = 0$ as $\mathcal{M}G_q = R_0$. As q increases from 0 to 1, $\mathcal{M}G_q$ progressively distances itself from R_0, resulting in an increase in $D_0(G_q,p)$.

The departure of an ODC R from being star-shaped is measured by $D^*(R,p) = \|\mathcal{M}R - R\|_p$. When $q \geq 0$, G_q is star-shaped (i.e., $\mathcal{M}G_q = G_q$), leading to $D^*(G_q,p) = 0$. As q increases from -1 to 0, G_q approaches $\mathcal{M}G_q$, causing a decrease in $D^*(G_q,p)$. We refer to Figure 1(b) for plots of $D_0(G_q,p)$ and $D^*(G_q,p)$ over $q \in [-1,1]$ for $p = 1, 2, \infty$.

5.2 Data generation under specified ODCs

To generate independent samples that adhere to specified ODCs, we assume all distributions have the same support $[0,1]$ and that the inverse of F_j exists and coincides with the quantile function F_j^{-1} for each j. We write $F_j^{-1}(u) = F_1^{-1} \circ R_1 \circ R_2 \circ \cdots \circ R_{j-1}(u)$ where "$\circ$" denotes function composition. Indeed, it suffices to rewrite $F_j^{-1} = R_1 \circ R_2 \circ \cdots \circ R_{j-1}$ by choosing F_1 as $\mathcal{U}(0,1)$. Hence, the inverse cumulative distribution method can generate the jth random sample. For example, considering F_1, F_2, and F_3 with $(R_1, R_2) =$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

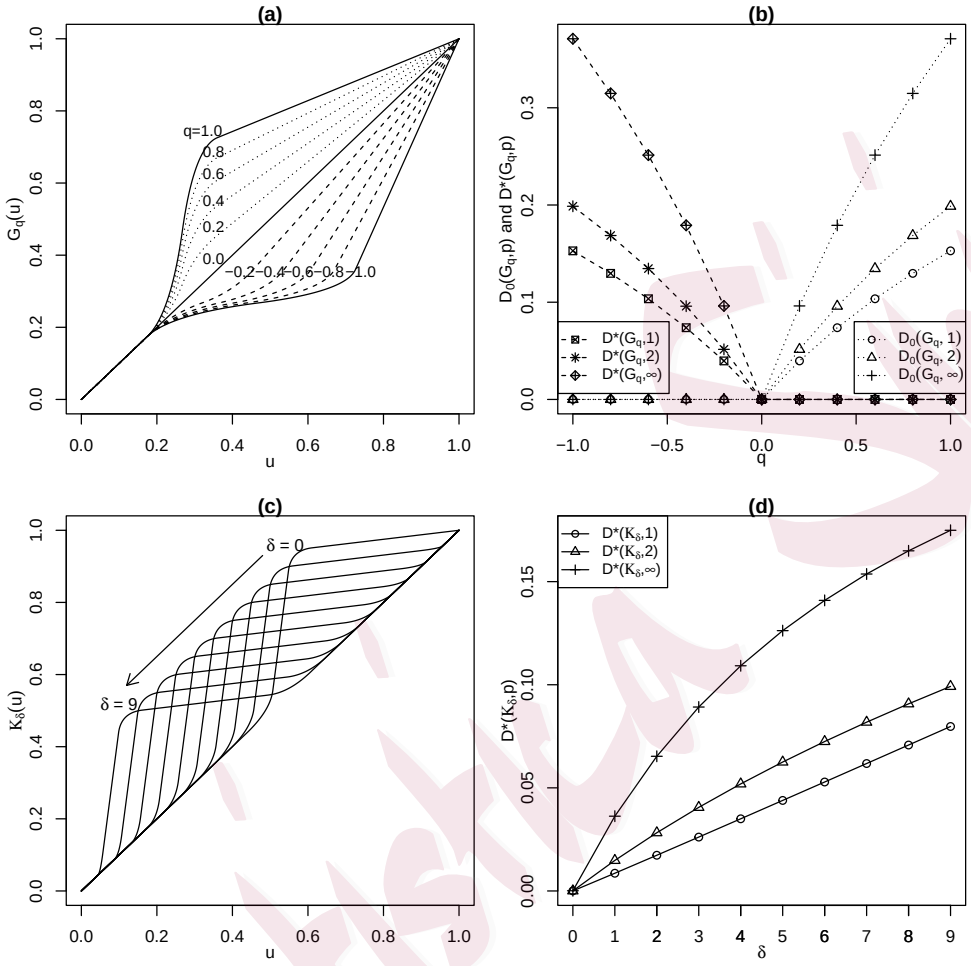

Figure 1: Departures $D_0(\cdot, p)$ and $D^*(\cdot, p)$ for ODCs G_q and K_δ.

(G_{q_1}, G_{q_2}), we generate independent random samples from each F_i by $X_{1l_1} = U_{1l_1}$, $X_{2l_2} = R_1(U_{2l_2}) = G_{q_1}(U_{2l_2})$, and $X_{3l_3} = R_1 \circ R_2(U_{3l_3}) = G_{q_1} \circ G_{q_2}(U_{3l_3})$, for $1 \le l_j \le n_j$, where $U_{j1}, \ldots, U_{jn_j}$ are random samples from $\mathcal{U}(0, 1)$.

Finally, we are prepared to introduce numerical evaluations of the proposed methods. In each configuration, we fixed $\alpha = 0.05$, considered equal

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

sample sizes ($n_j = n$ for all $1 \le j \le k$), and used 1000 Monte Carlo replications. In the manuscript, we focused on $k = 3$ and used different settings of $R_1 = G_{q_1}$ and $R_2 = G_{q_2}$, denoted by $(R_1, R_2) = (G_{q_1}, G_{q_2})$ for simplicity. Results for $k > 3$ are in Section S2 of the Supplementary Materials.

5.3 Equality tests

To assess the size, we set $R_1 = R_2 = R_0$ as specified by H_0. For power assessment, we choose R_1 and R_2 from the star-shaped G_q's (i.e., $0 \le q \le 1$), where at least one of them differs from R_0. With the chosen (R_1, R_2), we generated data of equal sample size $n \in \{60, 100, 200\}$.

The results are summarized in Table 1. Under H_0 with $(q_1, q_2) = (0.0, 0.0)$, both T_{3p} and U_{3p} maintain type I error probability close to $\alpha = 0.05$ with errors less than 0.018 for $p \in \{1, 2, \infty\}$ and $n \in \{60, 100, 200\}$. The value 0.018 is the margin of error for estimating the nominal rejection rate 0.05 with 1000 Monte Carlo replications at the 99% confidence level. Under $H_1 : F_1 \preceq F_2 \preceq F_3$ with $(q_1, q_2) \neq (0.0, 0.0)$ and $q_1, q_2 \ge 0$, the probability of rejecting H_0 approaches 1 with growing sample sizes increase, aligning with Theorem 1. When the sample size is fixed, the power increases as either q_1 or q_2 increases, which is expected as the departure $D_0(G_q, p)$ increases in q. When comparing T_{kp} and U_{kp}, we see that T_{kp} ex-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

hibits superior power across almost all settings. The bold numbers in Table 1 mark the maximum value in each row. It is clear that $p = \infty$ performed better than $p = 1$ and $p = 2$, and $T_{k\infty}$ is consistently the best.

Table 1: Sizes and powers for testing H_0 versus H_1.

n	(q_1, q_2)	T_{31}	U_{31}	T_{32}	U_{32}	$T_{3\infty}$	$U_{3\infty}$
60	(0.0,0.0)	0.051	0.048	0.043	0.049	0.044	0.051
	(0.4,0.0)	0.446	0.341	0.482	0.392	**0.506**	0.469
	(0.6,0.0)	0.678	0.573	0.746	0.673	**0.781**	0.750
	(1.0,0.0)	0.948	0.887	0.973	0.936	**0.990**	0.973
	(0.2,0.2)	0.428	0.230	0.462	0.269	**0.466**	0.312
	(0.4,0.2)	0.700	0.412	0.734	0.482	**0.747**	0.567
	(0.6,0.4)	0.954	0.758	0.974	0.856	**0.974**	0.907
100	(0.0,0.0)	0.067	0.051	0.067	0.059	0.065	0.047
	(0.4,0.0)	0.621	0.463	0.674	0.569	**0.709**	0.623
	(0.6,0.0)	0.901	0.781	0.944	0.863	**0.957**	0.911
	(1.0,0.0)	1.000	1.000	1.000	1.000	**1.000**	1.000
	(0.2,0.2)	0.599	0.337	0.648	0.403	**0.657**	0.399
	(0.4,0.2)	0.893	0.600	0.929	0.711	**0.934**	0.741
	(0.6,0.4)	0.998	0.940	1.000	0.975	**1.000**	0.990
200	(0.0,0.0)	0.048	0.053	0.046	0.056	0.050	0.054
	(0.4,0.0)	0.914	0.824	0.958	0.896	**0.970**	0.942
	(0.6,0.0)	0.998	0.984	1.000	0.995	**1.000**	0.999
	(1.0,0.0)	1.000	1.000	1.000	1.000	**1.000**	1.000
	(0.2,0.2)	0.889	0.574	0.921	0.678	**0.927**	0.720
	(0.4,0.2)	0.995	0.924	0.997	0.966	**0.997**	0.989
	(0.6,0.4)	1.000	1.000	1.000	1.000	**1.000**	1.000

5.4 GOF tests

The first part of our assessment of the GOF tests proposed in Section 3 considers three scenarios:

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

(A) $H_0^* : F_1 \preceq F_2 \preceq F_3$ with $(R_1, R_2) = (G_{q_1}, G_{q_2})$, where $q_1, q_2 \geq 0$;

(B) $H_1^* : F_1 \not\preceq F_2 \preceq F_3$ with $(R_1, R_2) = (G_{q_1}, G_{q_2})$, where $q_1 < 0, q_2 \geq 0$;

(C) $H_1^* : F_1 \not\preceq F_2 \not\preceq F_3$ with $(R_1, R_2) = (G_{q_1}, G_{q_2})$, where $q_1, q_2 < 0$.

Scenario (A) evaluates the size of the test, while Scenarios (B) and (C) assess the power. Under the hypothesis specified in each scenario, we generated datasets of equal sizes $n \in \{60, 100, 200\}$ and conducted our tests with S_{kp} and W_{kp} and also the Bonferroni-corrected test, denoted by B_{kp}, for $p \in \{1, 2, \infty\}$. Table 2 summarizes the results.

Under H_0^*, the setting $(q_1, q_2) = (0.0, 0.0)$ aligns with the null configuration that $F_1 = F_2 = F_3$. It is evident that all tests maintain type I errors at the nominal level (i.e., within a 0.018 margin of error for $\alpha = 0.05$). As either q_1 or q_2 increases from 0 under Scenario (A), the ODCs progressively take on a more star-shaped form. Consequently, the type I error probability decreases and consistently remains below 0.05. This pattern coincides with the theoretical findings in Theorem 3.

Under H_1^* in both Scenarios (B) and (C), all proposed tests have powers approaching 1 as the sample size n increases, which illustrates the consistency of our tests established in Theorem 3. When q_i decreases, the power increases as the departure $D^*(R_i, p)$ increases. Among S_{3p}, W_{3p}, and B_{3p},

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

we observe that S_{32} or $W_{3\infty}$ often achieves the highest power (marked in bold) while $B_{3\infty}$ performs in between.

To gain a clear picture of the comparison among S_{3p}, W_{3p}, and B_{3p}, we proceed to the second part of our assessment. We now study the power of these tests as the sample size remains constant while the deviation from the null hypothesis increases. We consider a sequence of 10 ODCs, denoted by K_δ for $\delta \in \{0, 1, \ldots, 9\}$, as previously used in Wang et al. (2020). When $\delta = 0$, K_0 is star-shaped. As δ deviates form 0, K_δ progressively becomes more non-star-shaped, leading to an increase in $D^*(K_\delta, p)$. We refer to Figures 1(c)-(d) for plots of the K_δ's and their corresponding $D^*(K_\delta, p)$.

We fixed $n = 200$ and considered two sequences of ODC pairs for (R_1, R_2), denoted by $\{(K_\delta, R_0)\}_{\delta=0}^9$ and $\{(K_\delta, K_\delta)\}_{\delta=0}^9$. Figure 2 illustrates the estimated power of each test. When $\delta = 0$, all three tests' type I error probabilities are well controlled, as anticipated. As δ increases, the powers of the GOF tests rise and approach 1. Comparing the three testing procedures reveals clear results. For the sequence $\{(K_\delta, R_0)\}_{\delta=0}^9$, where $F_2 = F_3$ is fixed, W_{kp} exhibits superior power to S_{kp} since only (F_1, F_2) departs from USO as δ increases. However, for the sequence $\{(K_\delta, K_\delta)\}_{\delta=0}^9$, S_{kp} accumulates the departures from USO of both (F_1, F_2) and (F_2, F_3) better, and thus achieves better power. Lastly, both S_{kp} and W_{kp} dominate the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

Bonferroni-corrected tests.

5.5 Distinguishing distributions

We now assess our methods in Section 4. For data generation, we utilize configurations $(R_1, R_2) = (G_{q_1}, G_{q_2})$ with $q_1, q_2 \geq 0$ and take $n \in \{100, 200\}$. We introduce the following measures. Let $\hat{J}$ be a general jump point collector (in our case, either J_p^0 or J_p^*). We evaluate the performance of $\hat{J}$ by the correct rate $\text{pr}(\hat{J} = J)$, the true positive average $E(\#\{\hat{J} \cap J\})$, and the false positive average $E(\#\{\hat{J} \cap J^c\})$, where $\#\{A\}$ means the size of a set A and $J^c = \{i : F_i = F_{i+1}\}$ is the collection of non-jump points. A well-performing $\hat{J}$ is expected to have high correct rates and low false positive averages, and the positive averages are about the size of underlying J.

Our investigations utilize G_q's to configure various combinations of $(R_1, R_2) = (G_{q_1}, G_{q_2})$. Start by examining the scenario with $q_1, q_2 > 0$, we observe the order $F_1 \prec F_2 \prec F_3$ with $J = \{1, 2\}$ and $J^c = \emptyset$. In this case, we anticipate that $E(\#\{\hat{J} \cap J\}) \approx 2$ and $\#\{\hat{J} \cap J^c\}$ must be 0. Moving on to the cases with $q_1 > 0$ and $q_2 = 0$, we find that $F_1 \prec F_2 = F_3$ with $J = \{1\}$ and $J^c = \{2\}$. Here, we expect $E(\#\{\hat{J} \cap J\}) \approx 1$. Lastly, $q_1 = q_2 = 0$ corresponds to $F_1 = F_2 = F_3$ with $J = \emptyset$ and $J^c = \{1, 2\}$, we have $\#\{\hat{J} \cap J\} = 0$. For the second and third cases where J^c is nonempty,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

we expect that $E(\#\{\hat{J} \cap J^c\}) \approx 0$ as $\hat{J}$ identifies J.

From Table 3, when J remains the same, the cases with larger departures $D_0(R_i, p)$ have better performance for both proposed methods J_p^0 and J_p^*. The sup-norm version performed the best for both J_p^0 and J_p^* in terms of the correct rate and true positive rate. In addition, both methods have false positive averages $E(\#\{\hat{J} \cap J^c\})$ lower than 0.025. Comparing J_p^0 and J_p^*, when all distributions are the same, J_p^0 has correct rates around 0.95 while J_p^* has slightly better rates from 0.95 to 1. Under H_1, when at least one jump point exists, J_p^0 outperforms J_p^* in terms of the correct rates and true positive average. Therefore, when H_1 is known in advance, we suggest using J_p^0. On the other hand, although J_p^* is more conservative than J_p^* under H_1, J_p^* has its own practical merits under H_0. Notably, J_p^* has very low false positive rates when H_0 is true. Consequently, J_p^* is recommended when the false positive rates must be strictly controlled and/or where H_1 is unknown in advance.

6. Data analysis

Liver fibrosis is a change in the microscopic structure of the liver, reflecting the body's response to injury, often associated with conditions like hepatitis C virus infection. The progression of this disease and the assessment of

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

therapeutic effectiveness are crucial considerations. The METAVIR scoring system categorizes fibrosis into five stages ranging from mild to severe: F_1 (no fibrosis), F_2 (portal fibrosis without septa), F_3 (few septa), F_4 (numerous septa without cirrhosis), and F_5 (cirrhosis). While liver biopsy has long been considered the "gold standard" for determining fibrosis stages, its invasive nature, potential complications, and lack of easy repeatability pose challenges (Adams, 2011; Berger et al., 2019). The pursuit of safe and efficient diagnostic methods remains a crucial focus of ongoing research.

Serum markers serve as non-invasive methods for diagnosing liver fibrosis, offering advantages such as reproducibility, applicability, and cost-effectiveness (Manning and Afdhal, 2008; Pinzani et al., 2008; Castera, 2009; Smith and Sterling, 2009; Nallagangula et al., 2017; Li et al., 2018). Given its relevance in disease-related tissue remodeling, human microfibrillar-associated protein 4 (MFAP4) emerges as a potential candidate for a serum marker (Mölleken et al., 2009). The viability of using MFAP4 as a biomarker for diagnosing liver fibrosis has been assessed in Bracht et al. (2016). However, concerns arise regarding the subjective transformation of MFAP4 measurements and the assumption of normality with equal variances, prompting consideration for a nonparametric approach.

We demonstrate our equality tests and GOF tests using the MFAP4

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

data presented in Bracht et al. (2016). Additionally, we assess the potential of MFAP4 for distinguishing among the various fibrosis stages. We denote ODCs for consecutive distributions by $R_i = F_i \circ F_{i+1}^{-1}$ for $i = 1, \ldots, 4$. The number of patients collected for each fibrosis stage are $n_1 = 97$, $n_2 = 176$, $n_3 = 135$, $n_4 = 67$, and $n_5 = 67$. Empirical ODCs $\hat{R}_1, \ldots, \hat{R}_4$ and star-shaped ODC estimators $\mathcal{M}\hat{R}_1, \ldots, \mathcal{M}\hat{R}_4$ are plotted in Figure 3. The L^p differences Δ_{ip} and M_{ip} with $p = 1, 2$, and ∞, are reported in Table 4.

Set $\alpha = 0.05$. For testing H_0 versus H_1, both the test statistics T_{kp} and U_{kp} (as shown in Table 4) exceed their corresponding critical values $t_{kp,\alpha}$ and $u_{kp,\alpha}$ for all p, respectively. Consequently, the proposed methods consistently indicate that the MFAP4 distributions are not equal, aligning with the findings in Mölleken et al. (2009). Turning to the goodness-of-fit (GOF) test for H_0^* versus H_1^*, the test statistics S_{kp} and W_{kp} fall below their corresponding critical values $\hat{s}^*_{kp,\alpha}$ and $\hat{w}^*_{kp,\alpha}$ for all p. This suggests the presence of USO is strongly supported.

Under USO, we apply our distinguishing methods J_p^0 and J_p^*. For J_p^*, the optimized η_p^* are $0.913, 0.976$, and 1.555 for $p = 1, 2$, and ∞, respectively, and all result in the same set of jump points, $J_p^* = \{2\}$. Therefore, we conclude the MFAP4 distributions adhere to the order $F_1 = F_2 = F_3 \prec F_4 = F_5$, aligning with the findings in Bracht et al. (2016). In addition,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

if we view H_1 as established from our GOF tests, we can apply J_p^0 to detect the jump points. In this case, the analysis provides $J_p^0 = \{2, 3\}$ for all $p = 1, 2, \infty$, suggesting that MFAP4 distributions F_2, F_3, and F_4 are distinguishable, and the order of distributions is $F_1 = F_2 \prec F_3 \prec F_4 = F_5$. This is a stronger conclusion when compared to that in Bracht et al. (2016).

7. Discussion

In this article, we have provided a comprehensive nonparametric toolkit for comparing multiple distributions under USO. Extensive numerical studies via both synthetic and real data have been conducted to showcase the efficacy of our methods and their counterparts. In the following, we would like to discuss three additional topics.

In practical scenarios where the overarching goal is to distinguish among distributions, one can combine the three components of our toolkit using modified significance levels. We list four potential cases:

- If both the equality of distributions and presence of USO are unknown, one can perform the equality tests for H_0 at a significance level $\alpha/3$ and the GOF tests for H_0^* at $\alpha/3$. If H_0 is rejected but not H_0^*, one can perform J_p^0 with $\alpha/3$. Another option is to test for H_0 at $\alpha/2$ and for H_0^* at $\alpha/2$. If H_0 is rejected but not H_0^*, one can calculate

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

J^*_∞ since J^*_∞ has the correct rate approaching 1.

- If USO is known in advance, but it is unknown whether the distributions are identical, one can perform equality tests for H_0 first at $\alpha/2$ and then perform J^0_p with $\alpha/2$ if H_0 is rejected. Or one can calculate J^*_∞ directly without performing equality tests for H_0.

- If USO is unknown, but distributions are known to be non-identical, then one can perform the GOF tests for H^*_0 at $\alpha/2$ and then J^0_p with $\alpha/2$ if H^*_0 is not rejected. Or one can perform the GOF test with α and then J^*_∞ if H^*_0 is not rejected.

- If one knows the distributions adhere to USO and not all of them are equal, we suggest performing J^0_p with α directly to distinguish them.

Next, we want to highlight our contribution in comparison to the existing methods. For the equality test, El Barmi and McKeague (2016) used empirical likelihood (EL) to propose a nonparametric test for H_0 versus H_1. We have included their test in our R codes and conducted comparisons with our approach (see Section S2.1 of the Supplementary Materials). The results reveal that EL-based tests outperformed ours when the departure from equality is substantial. Our tests have better power when the departure is mild and are more computationally feasible in most cases.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

For GOF tests involving multiple distributions, to the best of our knowledge, our method stands as the sole sophisticated solution. We believe developing EL-based GOF tests has its own merit and could be an interesting future topic. We have numerically compared our method with the naive Bonferroni approach. We have also considered other types of tests that combine and adjust p-values, such as the Cauchy combination test (Liu and Xie, 2020) and Benjamini and Yekutieli methods (Benjamini and Yekutieli, 2001). Our comparative analysis in Section S2.2 of the Supplementary Materials reveals that Bonferroni's corrected test and Cauchy's combination test stand out from the p-value-adjusted methods across various scenarios. Nevertheless, none of the p-value-adjusted methods can surpass the effectiveness of our proposed methods.

In conclusion, we present some intriguing avenues for future exploration. First, our proposed ODC-based methods lend themselves to adaptation for partially ordered distributions such as tree-shape orderings ($F_0 \preceq F_i$ for $i = 1, \ldots, k$) and umbrella orderings ($F_1 \preceq F_2 \preceq \cdots \preceq F_i$ and $F_k \preceq \cdots \preceq F_{i+1} \preceq F_i$). In general, partially ordered distributions can be represented by ordered pairs $F_i \preceq F_j$ for $(i, j) \in \mathcal{O}$, where $\mathcal{O}$ collects all pairs of USO ordered distributions. One could consider corresponding L^p differences which accumulate these differences among $(i, j) \in \mathcal{O}$. With similar

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

arguments, these modified procedures are expected to exhibit good performance, such as well-controlled size, consistency, and favorable correct identification rates, at least asymptotically. Second, an interesting future topic involves investigating the ordering of observations taken in vector or matrix form. Third, as USO is also known as hazard ratio ordering, extending our methods to handle censored data would be valuable. Undoubtedly, these topics pose significant theoretical challenges.

Supplementary Material

The supplementary materials contain proofs of the theorems, required lemmas, and additional numerical results for all proposed methods.

Acknowledgments

This work is partially supported by NIH (R03 AI135614) and NSF (DMS 2311292). The authors thank Dr. Joshua M. Tebbs for his helpful suggestions, which significantly improved the presentation of this work.

References

Adams, L. (2011). Biomarkers of liver fibrosis. *J. Gastroen. Hepatol.* **66**, 802–809.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

Balakrishnan, N., Zhang, Y., and Zhao, P. (2018). Ordering the largest claim amounts and ranges from two sets of heterogeneous portfolios. *Scand. Actuar. J.* **1** ,23–41.

Beare, B. and Moon, J. (2015). Nonparametric tests of density ratio ordering. *Economet. Theor.* **31**, 471–492.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **19**, 1165–1188.

Berger, D., Desai, V. and Janardhan, S. (2019). Con: liver biopsy remains the gold standard to evaluate fibrosis in patients with nonalcoholic fatty liver disease. *Clin. Liver. Dis. (Hoboken)* **13**, 114-116.

Bracht, T., Mölleken, C., Ahrens, M., Poschmann, G., Schlosser, A., Eisenacher, M., Stühler, K., Meyer, H., Schmiegel, W., Holmskov, U., Sorensen G., and Sitek, B. (2016). Evaluation of the biomarker candidate for non-invasive assessment of hepatic fibrosis in hepatitis C patients. *J. Transl. Med.* **14**, 1-9.

Carolan, C. and Tebbs, J. (2005). Nonparametric tests for and against likelihood ratio ordering in the two-sample problem. *Biometrika* **92**, 159–171.

Castera, L. (2009). Transient elastography and other noninvasive tests to assess hepatic fibrosis in patients with viral hepatitis. *J. Viral. Hepatitis.* **16**, 300–314.

Da, G. and Ding, W. (2016). Component level versus system level k-out-of-n assembly systems. *IEEE T. Reliab.* **65**, 425–33.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

Dardanoni, V. and Forcina, A. (1998). A unified approach to likelihood inference on stochastic orderings in a nonparametric context. *J. Am. Stat. Assoc.* **93**, 1112–1123.

Davidov, O. and Herman, A. (2012). Ordinal dominance curve based inference for stochastically ordered distributions. *J. R. Stat. Soc. B.* **74**, 825–847.

Dykstra, R., Kochar, S., and Robertson, T. (1991). Statistical inference for uniform stochastic ordering in several populations. *Ann. Stat.* **19**, 870–888.

Dykstra, R., Kochar, S., and Rojo, J. (1997). Stochastic comparison of parallel systems of heterogeneous exponential components. *J. Stat. Plan. Infer.* **65**, 203–221.

El Barmi, H. (2017). Testing for uniform stochastic ordering via empirical likelihood under right censoring. *Stat. Sinica.* **27**, 645–664.

El Barmi, H. and McKeague, I. (2016). Testing for uniform stochastic ordering via empirical likelihood. *Ann. I. Stat. Math.* **68**, 955–976.

Hsieh, F., and Turnbull, B. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann. Stat.* **24**, 25–40.

Keilson, J. and Sumita, U. (1982). Uniform stochastic ordering and related inequalities. *Can. J. Stat.* **10**, 181–198.

Kochar, S. (1979). Distribution-free comparison of two probability distributions with reference to their hazard rates. *Biometrika* **66**, 437–441.

Lehmann, E. (1955). Ordered Families of Distributions. *Ann. Stat.* **26**, 399–419.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

Li, C., Li, R., and Zhang, W. (2018). Progress in non-invasive detection of liver fibrosis. *Cancer Biology & Medicine* **15**, 124–136.

Liu, Y. and Xie, J. (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* **115**, 393–402.

Manning, D. and Afdhal, N. (2008). Diagnosis and quantitation of fibrosis. *Gastroenterology* **134**, 1670–1681.

Mölleken, C., Sitek, B., Henkel, C., Poschmann, G. and Sipos, B., Wiese, S., Warscheid, B., Broelsch, C., Reiser, M., Friedman, S., Tørnoe, I., Schlosser, A., Klöppel, G., Schmiegel, W., Meyer, H. E., Holmskov, U., and Stühler, K. (2009). Detection of novel biomarkers of liver cirrhosis by proteomic analysis. *Hepatology* **49**, 1257–1266.

Nallagangula, K., Nagaraj, S., Venkataswamy, L., and Chandrappa, M. (2017). Liver fibrosis: a compilation on the biomarkers status and their significance during disease progression. *Future Science OA* **4**, FSO250.

Navarro, J. and Shaked, S. (2006). Hazard rate ordering of order statistics and systems. *J. Appl. Probab.* **43**, 391–408.

Park, C., Lee, C., and Robertson, T. (1998). Goodness-of-fit test for uniform stochastic ordering among several populations. *Can. J. Stat.* **26**, 69–81.

Pinzani, M., Vizzutti, F., Arena, U. and Marra, F. (2008). Technology Insight: noninvasive assessment of liver fibrosis by biochemical scores and elastography. *Nat. Clin. Pract. Gastr* **5**, 95–106.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

Shaked, M. and Shanthikumar, J. (2007). *Stochastic Orders*. Springer-Verlag, New York.

Smith, J. and Sterling, R. (2009). Systematic review: non-invasive methods of fibrosis analysis in chronic hepatitis C. *Aliment. Pharm. Therap.* **30**, 557–576.

Tang, C., Wang, D. and Tebbs, J. (2017). Nonparametric goodness-of-fit tests for uniform stochastic ordering. *Ann. Stat.* **48**, 2565–2589.

Wang, H. and Li, B. and Leng, C. (2020). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. B.* **71**, 671–683.

Wang, D., Tang, C. and Tebbs, J. (2020). More powerful goodness-of-fit tests for uniform stochastic ordering. *Comput. Stat. Data. An.* **144**, 106898.

Wang, D. and Tang, C. (2021). Testing against uniform stochastic ordering with paired observations. *Bernoulli* **27**, 2556–2563.

Department of Mathematical Sciences, University of Texas at Dallas

E-mail: chuan-fa.tang@utdallas.edu

Department of Statistics, University of South Carolina

E-mail: deweiwang@stat.sc.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

Table 2: Size and power comparisons for GOF tests with test statistics S_{kp}, W_{kp}, and Bonferroni-corrected method B_{kp} with $p = 1, 2$ and ∞.

n	(q_1, q_2)	S_{31}	W_{31}	B_{31}	S_{32}	W_{32}	B_{32}	$S_{3\infty}$	$W_{3\infty}$	$B_{3\infty}$
60	(0.0,0.0)	0.066	0.067	0.047	0.062	0.062	0.046	0.027	0.053	0.053
	(0.2,0.0)	0.033	0.045	0.028	0.028	0.037	0.026	0.012	0.032	0.030
	(0.4,0.2)	0.004	0.012	0.006	0.005	0.011	0.006	0.001	0.008	0.007
	(-0.4,0.2)	0.425	0.450	0.371	0.430	**0.505**	0.449	0.343	0.484	0.497
	(-0.6,0.2)	0.713	0.712	0.643	0.747	0.766	0.716	0.681	**0.790**	0.787
	(-0.4,0.0)	0.516	0.453	0.390	**0.537**	0.948	0.466	0.435	0.488	0.517
	(-0.6,0.0)	0.791	0.711	0.669	**0.819**	0.765	0.749	0.768	0.798	0.815
	(-0.2,-0.2)	0.449	0.309	0.254	**0.466**	0.326	0.291	0.357	0.284	0.310
	(-0.4,-0.2)	0.754	0.533	0.473	**0.772**	0.574	0.563	0.696	0.568	0.613
	(-0.6,-0.4)	0.980	0.871	0.835	**0.986**	0.909	0.910	0.979	0.920	0.933
100	(0.0,0.0)	0.048	0.060	0.048	0.041	0.057	0.048	0.023	0.048	0.043
	(0.2,0.0)	0.023	0.037	0.026	0.018	0.037	0.029	0.014	0.034	0.028
	(0.4,0.2)	0.001	0.004	0.003	0.002	0.006	0.003	0.005	0.006	0.004
	(-0.4,0.2)	0.619	0.645	0.574	0.662	0.720	0.672	0.609	**0.751**	0.735
	(-0.6,0.2)	0.908	0.912	0.866	0.940	0.947	0.924	0.929	**0.967**	0.961
	(-0.4,0.0)	0.722	0.650	0.593	**0.758**	0.714	0.691	0.703	0.745	0.749
	(-0.6,0.0)	0.953	0.915	0.890	**0.974**	0.950	0.936	0.968	0.970	0.973
	(-0.2,-0.2)	0.653	0.419	0.379	**0.669**	0.456	0.433	0.607	0.426	0.453
	(-0.4,-0.2)	0.927	0.744	0.701	**0.950**	0.799	0.787	0.924	0.815	0.831
	(-0.6,-0.4)	0.998	0.982	0.974	**0.999**	0.993	0.990	1.000	0.995	0.995
200	(0.0,0.0)	0.063	0.059	0.046	0.052	0.057	0.044	0.049	0.048	0.043
	(0.2,0.0)	0.018	0.036	0.023	0.022	0.037	0.023	0.021	0.038	0.027
	(0.4,0.2)	0.000	0.001	0.000	0.001	0.002	0.001	0.004	0.009	0.006
	(-0.4,0.2)	0.905	0.906	0.884	0.933	0.945	0.937	0.937	**0.976**	0.972
	(-0.6,0.2)	0.997	0.996	0.994	**1.000**	0.999	0.999	0.998	**1.000**	1.000
	(-0.4,0.0)	0.955	0.911	0.893	0.971	0.949	0.945	0.974	**0.976**	0.975
	(-0.6,0.0)	1.000	0.999	0.997	**1.000**	1.000	1.000	1.000	**1.000**	1.000
	(-0.2,-0.2)	0.905	0.649	0.633	**0.939**	0.740	0.715	0.928	0.786	0.788
	(-0.4,-0.2)	0.999	0.959	0.953	**0.999**	0.982	0.980	0.999	0.995	0.996
	(-0.6,-0.4)	1.000	1.000	1.000	**1.000**	1.000	1.000	1.000	1.000	1.000

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

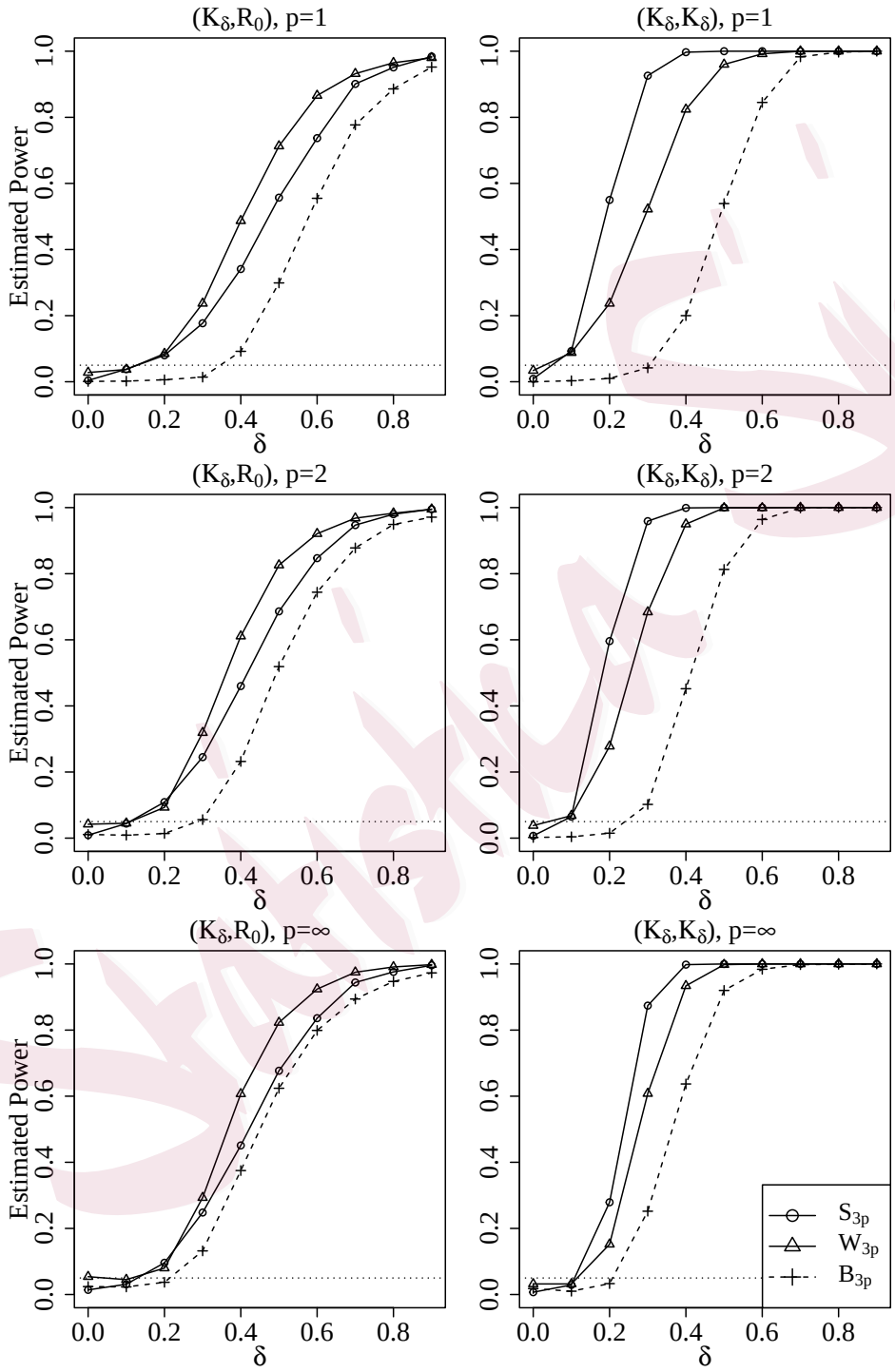

Figure 2: Power curves comparison for GOF tests with $\{(K_\delta, R_0)\}_{\delta=0}^{9}$ (left column) and $\{(K_\delta, K_\delta)\}_{\delta=0}^{9}$ (right column).

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

Table 3: Performance of $\hat{J} = J_p^0$ and $\hat{J} = J_p^*$ using the estimated correct rate $\text{pr}(\{\hat{J} = J\})$ denoted by C, true positive average $E(\#\{\hat{J} \cap J\})$ denoted by TA, and false positive average $E(\#\{\hat{J} \cap J^c\})$ denoted by FA.

			$p=1$			$p=2$			$p=\infty$		
	n	(q_1, q_2)	C	TA	FA	C	TA	FA	C	TA	FA
J_p^0	100	(0.0,0.0)	0.949	0.000	0.051	0.941	0.000	0.059	**0.953**	0.000	0.047
		(0.4,0.0)	0.456	0.456	0.027	0.556	0.558	0.029	**0.609**	0.614	0.024
		(0.8,0.0)	0.903	0.919	0.035	0.943	0.967	0.036	**0.972**	0.991	0.025
		(1.0,0.0)	0.967	0.983	0.017	0.981	0.998	0.017	**0.982**	0.999	0.017
		(0.6,0.4)	0.320	1.248	0.000	0.472	1.450	0.000	**0.554**	1.541	0.000
		(0.8,0.6)	0.691	1.690	0.000	0.836	1.836	0.000	**0.898**	1.898	0.000
		(1.0,1.0)	0.961	1.961	0.000	0.997	1.997	0.000	**0.999**	1.999	0.000
	200	(0.0,0.0)	**0.947**	0.000	0.053	0.944	0.000	0.056	0.946	0.000	0.054
		(0.4,0.0)	0.781	0.788	0.024	0.845	0.856	0.026	**0.898**	0.913	0.023
		(0.8,0.0)	0.975	1.000	0.025	0.978	1.000	0.022	**0.978**	1.000	0.022
		(1.0,0.0)	0.983	1.000	0.017	0.980	1.000	0.020	**0.981**	1.000	0.019
		(0.6,0.4)	0.757	1.757	0.000	0.852	1.852	0.000	**0.904**	1.904	0.000
		(0.8,0.6)	0.983	1.983	0.000	0.996	1.996	0.000	**0.999**	1.999	0.000
		(1.0,1.0)	1.000	2.000	0.000	1.000	2.000	0.000	**1.000**	2.000	0.000
J_p^*	100	(0.0,0.0)	0.996	0.000	0.004	**0.997**	0.000	0.003	0.995	0.000	0.005
		(0.4,0.0)	0.098	0.098	0.002	0.112	0.112	0.002	**0.216**	0.216	0.002
		(0.8,0.0)	0.550	0.550	0.002	0.642	0.642	0.001	**0.808**	0.809	0.003
		(1.0,0.0)	0.779	0.779	0.002	0.857	0.858	0.001	**0.948**	0.949	0.001
		(0.6,0.4)	0.006	0.414	0.000	0.012	0.474	0.000	**0.080**	0.774	0.000
		(0.8,0.6)	0.121	0.857	0.000	0.190	1.019	0.000	**0.426**	1.381	0.000
		(1.0,1.0)	0.559	1.542	0.000	0.740	1.735	0.000	**0.888**	1.888	0.000
	200	(0.0,0.0)	1.000	0.000	0.000	1.000	0.000	0.000	**1.000**	0.000	0.000
		(0.4,0.0)	0.030	0.030	0.000	0.047	0.047	0.000	**0.173**	0.173	0.000
		(0.8,0.0)	0.528	0.528	0.000	0.701	0.701	0.000	**0.924**	0.924	0.000
		(1.0,0.0)	0.800	0.800	0.000	0.918	0.918	0.000	**0.987**	0.987	0.000
		(0.6,0.4)	0.001	0.217	0.000	0.002	0.320	0.000	**0.076**	0.780	0.000
		(0.8,0.6)	0.054	0.722	0.000	0.149	0.986	0.000	**0.565**	1.553	0.000
		(1.0,1.0)	0.631	1.617	0.000	0.859	1.858	0.000	**0.987**	1.987	0.000

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0034

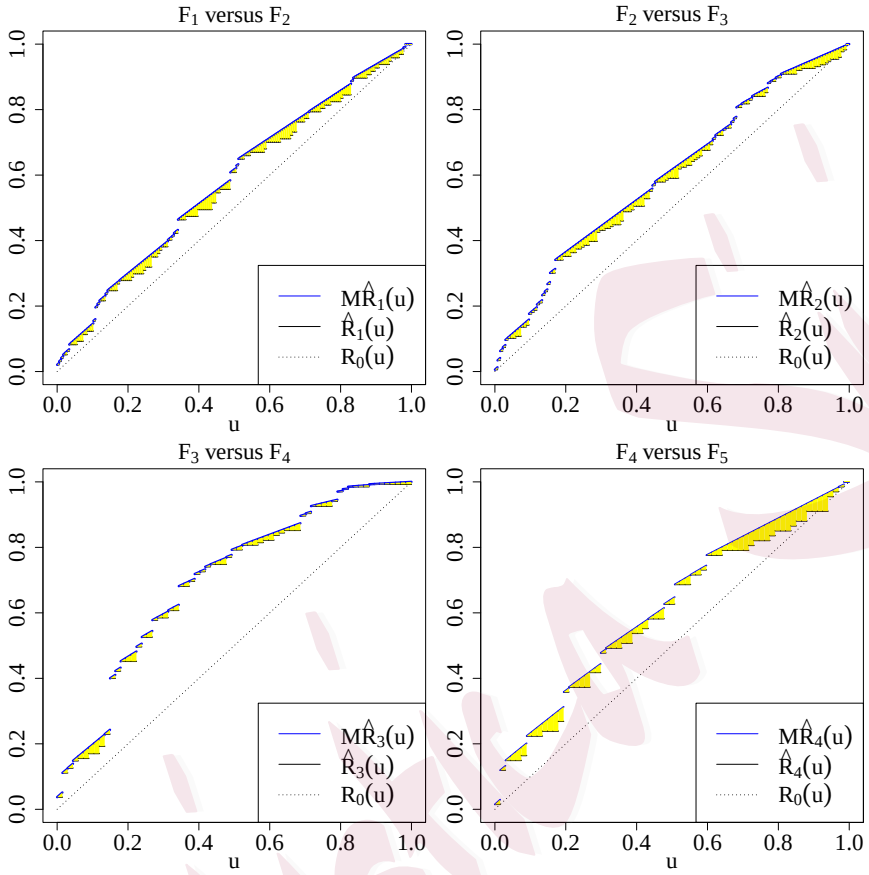

Figure 3: Empirical and star-shaped ODC estimates for R_1(top left), R_2(top right), R_3(bottom left), and R_4(bottom right) with the MFAP4 data.

Table 4: List of the scaled L^p differences, test statistics for equality and GOF, and corresponding cutoff values for the MFAP4 data.

	Δ_{1p}	Δ_{2p}	Δ_{3p}	Δ_{4p}	T_{kp}	$t_{kp,\alpha}$	U_{kp}	$u_{kp,\alpha}$
$p=1$	0.654	0.913	1.407	0.738	3.712	1.712	1.407	0.826
$p=2$	0.704	0.976	1.533	0.787	4.000	1.924	1.533	0.910
$p=\infty$	1.092	1.555	2.311	1.037	5.995	3.322	2.311	1.475
	M_{1p}	M_{2p}	M_{3p}	M_{4p}	S_{kp}	$\hat{s}^*_{kp,\alpha}$	W_{kp}	$\hat{w}^*_{kp,\alpha}$
$p=1$	0.130	0.134	0.059	0.154	0.477	1.291	0.154	0.592
$p=2$	0.154	0.159	0.080	0.182	0.575	1.581	0.182	0.690
$p=\infty$	0.382	0.354	0.304	0.369	1.297	3.447	0.382	1.299