REVSTAT – Statistical Journal
Volume 14, Number 4, October 2016, 349–394

DENSITY OF A RANDOM INTERVAL CATCH DIGRAPH FAMILY AND ITS USE FOR TESTING UNIFORMITY

Author: Elvan Ceyhan
– Department of Mathematics, Koç University,
Istanbul, Turkey
elceyhan@ku.edu.tr

Received: June 2014 Revised: February 2015 Accepted: February 2015

Abstract:

- We consider (arc) density of a parameterized interval catch digraph (ICD) family with random vertices residing on the real line. The ICDs are random digraphs where randomness lies in the vertices and are defined with two parameters, a centrality parameter and an expansion parameter, hence they will be referred as central similarity ICDs (CS-ICDs). We show that arc density of CS-ICDs is a U-statistic for vertices being from a wide family of distributions with support on the real line, and provide the asymptotic (normal) distribution for the (interiors of) entire ranges of centrality and expansion parameters for one dimensional uniform data. We also determine the optimal parameter values at which the rate of convergence (to normality) is fastest. We use arc density of CS-ICDs for testing uniformity of one dimensional data, and compare its performance with arc density of another ICD family and two other tests in literature (namely, Kolmogorov–Smirnov test and Neyman's smooth test of uniformity) in terms of empirical size and power. We show that tests based on ICDs have better power performance for certain alternatives (that are symmetric around the middle of the support of the data).

Key-Words:

- *asymptotic normality; class cover catch digraph; intersection digraph; Kolmogorov–Smirnov test; Neyman's smooth test; proximity catch digraph; random geometric graph; U-statistics.*

AMS Subject Classification:

- 05C80, 05C20, 60D05, 60C05, 62E20.

1. INTRODUCTION

Intersection graphs have received considerable attention in literature since their introduction. The main reasons for this attention are their applications in real life and their "tame" behavior, in the sense that many problems that are NP-hard for graphs in general are solvable in polynomial time for intersection graphs ([Prisner, 1994]). Intersection digraphs are introduced by [Beineke and Zamfirescu, 1982] who called them "connection digraphs". Let V be an index set and (S_v, T_v) be ordered pairs of sets associated with the elements v of V, where S_v is called the *source* and T_v is called the target or *sink* set ([Douglas, 1996]). The intersection digraph associated with this collection of ordered pairs is $D = (V, A)$ which has vertex set V and arc (i.e., directed edge) set A with $(u, v) \in A$ iff $S_u \cap T_v \neq \emptyset$. When the source and sink sets are intervals, we obtain interval digraphs ([Douglas, 1996]). If the set T_v resides in S_v for each $v \in V$, then the ordered pair set is a *nest representation* for the interval digraph, and if T_v is just a point residing in S_v, it is called a *catch representation*. A digraph is called an *interval catch digraph* (ICD), if it is an intersection digraph with a catch representation ([Prisner, 1994]). The set of ordered pairs, $\{S_v, T_v = \{p_v\}\}$, in the catch representation for the ICD is also called a "pointed set" where S_v is a set with base point p_v ([Prisner, 1989]). Equivalently, an ICD is the catch digraph of a family of pointed intervals of T if $(T, \leq)$ is a totally ordered set. Indeed, [Maehara, 1984] provides a simple characterization of ICDs for finite n, for which one can always take $T = \mathbb{R}$.

The ICDs we consider in this article are defined in a randomized setting. Our ICDs are vertex random digraphs in which each vertex corresponds to a random data point from a distribution, and arcs are defined by a bivariate relation using the regions based on these data points. Our ICDs are a special type of proximity graphs which were introduced by [Toussaint, 1980], and are closely related to the class cover problem of [Cannon and Cowen, 2000] and proximity catch digraphs (PCDs) which were introduced recently and have applications in spatial data analysis and statistical pattern classification ([Ceyhan and Priebe, 2005]).

In this article, we define *central similarity (CS) ICDs* for one dimensional data which may also be viewed as one dimensional version of the PCDs considered in [Ceyhan *et al.*, 2007]. We derive the asymptotic distribution of the *arc density* of CS-ICDs for random data points. For undirected simple graphs, the *edge density* (also called *graph density*) is defined as the ratio of number of edges in the graph to the total number of edges possible with the same number of vertices. So the edge density is $2|E|/\big(n(n-1)\big)$ for a graph $G = (V, E)$ with $|V| = n$. The minimal density is 0, which is attained for empty graphs (i.e., for $E = \emptyset$) and the maximal density is 1, which is attained for complete graphs

([Coleman and Moré, 1983]). Based on the graph density concept, 'dense' and 'sparse' graphs are defined. For a dense graph, graph density is close to 1 and for sparse graphs it is close to 0. There are other quantities related to graph density, such as average degree which is defined as $2|E|/n$ ([Goldberg, 1984]); edge density of a graph is also defined as $|E|/n$ in literature (see, e.g., [Grünbaum, 1988]). Notice that both of these quantities are scaled versions of the edge or graph density, $2|E|/\big(n(n-1)\big)$. On the other hand, density of a digraph is the ratio of number of arcs in a given digraph with n vertices to the total number of arcs possible (i.e., to the number of arcs in a complete symmetric digraph of order n). Hence for a simple digraph $D = (V, A)$ with vertex set $|V| = n$ and arc set A, *digraph density* (or *arc density*) is $|A|/\big(n(n-1)\big)$, which is the quantity of interest in this article. Arc density is also referred to as *relative density* in literature. Properly scaled, the arc density of the ICDs is a U-statistic, which yields the asymptotic normality by the general central limit theory of U-statistics ([Lehmann, 2004]). Our ICDs can also be viewed as a generalization of class cover catch digraphs (CCCDs) which was introduced by [Priebe *et al.*, 2001]. CS-ICDs have two defining parameters, a centrality and an expansion parameter. Here, we derive the explicit form of the asymptotic normal distribution of the arc density of the CS-ICDs for the (interiors of) entire ranges of these parameters for uniform one dimensional data from a class whose support being partitioned by points from another class. We investigate the arc density of CS-ICDs for uniform data in one interval (in $\mathbb{R}$) and the analysis is generalized to uniform data in multiple intervals (see Remark 4.1). We determine the optimal parameters for the rate of convergence to normality and show that arc density of CS-ICDs has a faster rate than that of the respective optimal parameter values of another ICD family called proportional-edge (PE) ICDs which were introduced in [Ceyhan, 2012] (and therein referred to as proportional-edge proximity catch digraphs). We employ the arc density of CS-ICDs for testing uniformity of one dimensional data and compare its performance with two prevalent tests in literature (namely, Kolmogorov–Smirnov test and Neyman's smooth test) in terms of size and power as well as arc density of the PE-ICDs. Testing uniformity of one-dimensional data is of substantial importance in various fields, e.g., for assessing the goodness-of-fit problems ([Marhuenda *et al.*, 2005]). For this purpose, some graph theoretical tools are used in literature although not so commonly; e.g., minimum spanning trees are employed for testing uniformity of two-dimensional data ([Jain *et al.*, 2002]). However, to the best of author's knowledge, arc density is not previously employed for testing uniformity of one-dimensional data. The tests based on the arc density of the ICD families have been shown to have better power performance for certain types of alternatives (which are symmetric around the midpoint of the support of the distribution) against uniformity. CS-ICDs can also be used for testing spatial patterns between (two or more) classes of data points.

We define the ICDs and describe the random ICDs and CS-ICDs in Section 2, define their arc density and provide preliminary results in Section 3,

provide the distribution of the arc density for uniform data in one interval in Section 4, present the size and power analysis and comparison with other tests as well as some consistency results in Section 5, and discussion and conclusions in Section 6. Shorter proofs are given in the main body of the article; while longer proofs are deferred to the Appendix.

2. RANDOM INTERVAL CATCH DIGRAPHS

Let $(\Omega, \mathcal{F}, P_x)$ be a probability space equipped with a metric $d\colon \Omega\times\Omega \to [0,\infty)$. Our random catch digraphs will be defined in a randomized setting where vertices are randomly generated in Ω and the associated metric distance will be taken to be the Euclidean distance. Let $\mathcal{X}_n = \{X_1, X_2, ..., X_n\}$ and $\mathcal{Y}_m = \{Y_1, Y_2, ..., Y_m\}$ be two sets of Ω-valued random variables from classes $\mathcal{X}$ and $\mathcal{Y}$, respectively, whose joint probability distribution is $F_{X,Y}$ with marginals F_X and F_Y, respectively. Our random catch digraph will be based on $\mathcal{X}_n$ and $\mathcal{Y}_m$. More specifically, we choose $\mathcal{X}$ points to be the vertices and put an arc from X_i to X_j, based on a binary relation which measures the relative allocation of X_i and X_j with respect to $\mathcal{Y}$ points. In particular, in our setting, the $\mathcal{Y}$ points will be used to partition the support set Ω, and the relative position of X_i and X_j with respect to $\mathcal{Y}$ points will be determined by the Euclidean distances between X_i, X_j, and the $\mathcal{Y}$ points. Notice that the randomness is only on the vertices, hence our catch digraphs are *vertex random*. Given $\mathcal{Y}_m \subseteq \Omega$, let $\mathcal{P}(\Omega)$ represent the power set of Ω, then *proximity map* $N_{\mathcal{Y}}\colon \Omega \to \mathcal{P}(\Omega)$ maps each point $x \in \Omega$ to a *proximity region* $N_{\mathcal{Y}}(x) \subseteq \Omega$. A *vertex random catch digraph* has the vertex set $\mathcal{V} = \mathcal{X}_n$ and arc set $\mathcal{A}$ defined by $(X_i, X_j) \in \mathcal{A}$ if $X_j \in N_{\mathcal{Y}}(X_i)$ for $i \neq j$. Hence the binary relation defining the digraph is based on the proximity region, $N_{\mathcal{Y}}$, which indicates the relative allocation of $\mathcal{X}$ points with respect to $\mathcal{Y}$ points. Notice also that arcs of the form (X_i, X_i) (i.e., loops) are not allowed in our catch digraph definition. If loops were allowed, the corresponding digraph would have been called a *pseudodigraph* according to some authors (see, e.g., [Chartrand *et al.*, 2010]). We also define arc probability, denoted $p_a(i,j)$, between two vertices X_i and X_j as $p_a(i,j) := P((X_i, X_j) \in \mathcal{A})$ for all $i \neq j$, $i, j = 1, 2, ..., n$. If $\mathcal{X}_n$ is a random sample from F_X, then $p_a(i,j) = p_a$ for all $i \neq j$, $i, j = 1, 2, ..., n$. For calculations leading to the distribution of arc density of ICDs, we also need a concept which is dual to proximity regions. For a set $B \subseteq \Omega$, the Γ_1-*region* is the image of the map $\Gamma_1(\cdot, N_{\mathcal{Y}})\colon \mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ that assigns the region $\Gamma_1(B, N_{\mathcal{Y}}) := \{z \in \Omega\colon B \subseteq N_{\mathcal{Y}}(z)\}$ to the set B. For a point $x \in \Omega$, we denote $\Gamma_1(\{x\}, N_{\mathcal{Y}})$ as $\Gamma_1(x, N_{\mathcal{Y}})$. The concept of Γ_1-region is introduced in [Ceyhan and Priebe, 2005] and is associated with another graph invariant called domination number (which is denoted as γ). In a proximity graph, if a vertex falls in the Γ_1-region, then the domination number would equal to 1. For brevity, we drop the subscript $\mathcal{Y}$ in the notation, $N_{\mathcal{Y}}$, henceforth.

2.1. Central Similarity ICDs

For one dimensional data, we have $\Omega = \mathbb{R}$, then there is a natural partitioning of the real line based on $\mathcal{Y}$ points. Let $Y_{(i)}$ be the i^{th} order statistic of $\mathcal{Y}_m$ for $i = 1, 2, ..., m$, with the extension that $-\infty =: Y_{(0)}$ and $Y_{(m+1)} := \infty$ and assume $Y_{(i)}$ values are distinct (which happens with probability one for continuous distributions). The $Y_{(i)}$ values partition $\mathbb{R}$ into $(m+1)$ intervals, with $\left(-\infty, Y_{(1)}\right)$ and $\left(Y_{(m)}, \infty\right)$ being the *end intervals*, and $\left(Y_{(i-1)}, Y_{(i)}\right)$ for $i = 2, ..., m$ being the *middle intervals*. For one dimensional data sets, $\mathcal{X}_n$ and $\mathcal{Y}_m$, we define the CS-ICD with expansion parameter $\tau > 0$ and centrality parameter $c \in (0, 1)$ as follows. For $x \in \left(Y_{(i-1)}, Y_{(i)}\right)$ (i.e., for x in a middle interval) with $i \in \{2, ..., m\}$ and $M_{c,i} = Y_{(i-1)} + c\left(Y_{(i)} - Y_{(i-1)}\right) \in \left(Y_{(i-1)}, Y_{(i)}\right)$, that is $c \times 100\,\%$ of $\left(Y_{(i)} - Y_{(i-1)}\right)$ is to the left of $M_{c,i}$, we define the CS proximity region as follows:

$$N(x, \tau, c) = \begin{cases} \left(x - \tau\left(x - Y_{(i-1)}\right),\ x + \frac{\tau(1-c)}{c}\left(x - Y_{(i-1)}\right)\right) \cap \left(Y_{(i-1)}, Y_{(i)}\right) & \text{if } x \in \left(Y_{(i-1)}, M_{c,i}\right), \\ \left(x - \frac{c\,\tau}{1-c}\left(Y_{(i)} - x\right),\ x + \tau\left(Y_{(i)} - x\right)\right) \cap \left(Y_{(i-1)}, Y_{(i)}\right) & \text{if } x \in \left(M_{c,i}, Y_{(i)}\right). \end{cases} \tag{2.1}$$

Notice that dependence on $\mathcal{Y}$ points is explicit in the definition of the CS proximity region. Furthermore, the Euclidean distance is implicit in the terms $\left(x - Y_{(i-1)}\right)$ and $\left(Y_{(i)} - x\right)$, where the former is $d\left(x, Y_{(i-1)}\right)$ and the latter is $d\left(x, Y_{(i)}\right)$. This definition yields two types of regions for $N(x, \tau, c)$, one with $\tau \in (0, 1]$ and the other with $\tau > 1$. For $\tau \in (0, 1]$, we have

$$N(x, \tau, c) = \begin{cases} \left(x - \tau\left(x - Y_{(i-1)}\right),\ x + \frac{\tau(1-c)}{c}\left(x - Y_{(i-1)}\right)\right) & \text{if } x \in \left(Y_{(i-1)}, M_{c,i}\right), \\ \left(x - \frac{c\,\tau}{1-c}\left(Y_{(i)} - x\right),\ x + \tau\left(Y_{(i)} - x\right)\right) & \text{if } x \in \left(M_{c,i}, Y_{(i)}\right), \end{cases} \tag{2.2}$$

and with $\tau > 1$, we have

$$N(x, \tau, c) = \begin{cases} \left(Y_{(i-1)}, x + \frac{\tau(1-c)}{c}\left(x - Y_{(i-1)}\right)\right) & \text{if } x \in \left(Y_{(i-1)}, \frac{c\,Y_{(i)} + \tau(1-c)\,Y_{(i-1)}}{c + \tau(1-c)}\right), \\ \left(Y_{(i-1)}, Y_{(i)}\right) & \text{if } x \in \left(\frac{c Y_{(i)} + \tau(1-c)\,Y_{(i-1)}}{c + \tau(1-c)}, \frac{(1-c)\,Y_{(i-1)} + c\,\tau Y_{(i)}}{1 - c + c\,\tau}\right), \\ \left(x - \frac{c\,\tau}{1-c}\left(Y_{(i)} - x\right), Y_{(i)}\right) & \text{if } x \in \left(\frac{(1-c)\,Y_{(i-1)} + c\,\tau Y_{(i)}}{1 - c + c\,\tau}, Y_{(i)}\right). \end{cases} \tag{2.3}$$

For an illustration of $N(x, \tau, c)$ in the middle interval case, see Figure 1 (left) where $\mathcal{Y}_2 = \{y_1, y_2\}$ with $y_1 = 0$ and $y_2 = 1$ (hence $M_{c,2} = c$).

Additionally, for x in an end interval, i.e., $x \in \left(Y_{(i-1)}, Y_{(i)}\right)$ with $i \in \{1, m+1\}$, the central similarity proximity region depends on the expansion parameter only.

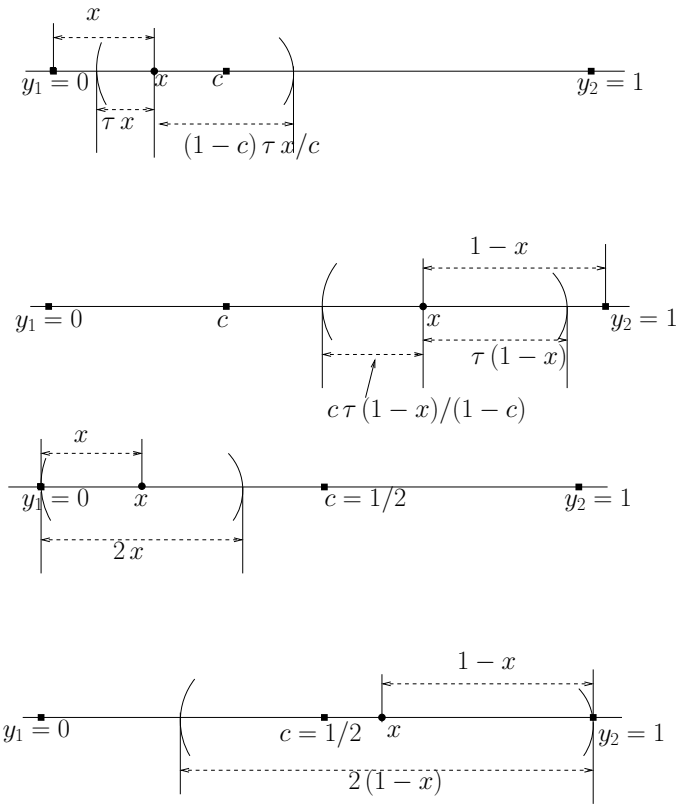

Figure 1: Plotted in the top two rows are illustrations of the construction of central similarity proximity regions, $N(x, \tau, c)$ with $\tau \in (0, 1]$, $\mathcal{Y}_2 = \{y_1, y_2\}$ with $y_1 = 0$ and $y_2 = 1$ (hence $M_{c,2} = c$) and $x \in (0, c)$ (first row) and $x \in (c, 1)$ (second row); and in the bottom two rows are the proximity regions associated with CCCD, i.e., $N(x, \tau = 1, c = 1/2)$ for an $x \in (0, 1/2)$ (third row) and $x \in (1/2, 1)$ (fourth row).

So we denote the central similarity proximity region for an x in an end interval as $N_e(x, \tau)$. Then with $\tau \in (0, 1]$, we have

$$N_e(x,\tau) = \begin{cases} \left(x - \tau\left(Y_{(1)} - x\right),\, x + \tau\left(Y_{(1)} - x\right)\right) & \text{if } x < Y_{(1)}\,, \\ \left(x - \tau\left(x - Y_{(m)}\right),\, x + \tau\left(x - Y_{(m)}\right)\right) & \text{if } x > Y_{(m)}\,, \end{cases} \tag{2.4}$$

and with $\tau > 1$, we have

$$N_e(x,\tau) = \begin{cases} \left(x - \tau\left(Y_{(1)} - x\right),\, Y_{(1)}\right) & \text{if } x < Y_{(1)}\,, \\ \left(Y_{(m)},\, x + \tau\left(x - Y_{(m)}\right)\right) & \text{if } x > Y_{(m)}\,. \end{cases} \tag{2.5}$$

If $x \in \mathcal{Y}_m$, then we define $N(x, \tau, c) = \{x\}$ and $N_e(x, \tau) = \{x\}$ for all $\tau > 0$, and if $x = M_{c,i}$, then in Equation (2.1), we arbitrarily assign $N(x, \tau, c)$ to be one of

the two defining intervals. For X from a continuous distribution, these special cases in the construction of central similarity proximity region — $X \in \mathcal{Y}_m$ and $X = M_{c,i}$ — occur with probability zero. Notice that $\tau > 0$ implies $x \in N(x, \tau, c)$ for all $x \in \left[Y_{(i-1)}, Y_{(i)}\right]$ with $i \in \{2, ..., m\}$ and $x \in N_e(x, \tau)$ for all $x \in \left[Y_{(i-1)}, Y_{(i)}\right]$ with $i \in \{1, m+1\}$. Furthermore, $\lim_{\tau \to \infty} N(x, \tau, c) = \left(Y_{(i-1)}, Y_{(i)}\right)$ for all $x \in \left(Y_{(i-1)}, Y_{(i)}\right)$ with $i \in \{2, ..., m\}$, so we define $N(x, \infty, c) = \left(Y_{(i-1)}, Y_{(i)}\right)$ for all such x. Similarly, $\lim_{\tau \to \infty} N_e(x, \tau) = \left(Y_{(i-1)}, Y_{(i)}\right)$ for all $x \in \left(Y_{(i-1)}, Y_{(i)}\right)$ with $i \in \{1, m+1\}$, so we define $N_e(x, \infty) = \left(Y_{(i-1)}, Y_{(i)}\right)$ for all such x. In the special case of $c = 1/2$ and $\tau = 1$, central similarity proximity region $N(x, \tau, c)$ is identical to the proportional edge proximity region with centrality parameter 1/2 and expansion parameter 2 (see [Ceyhan, 2012]).

In a vertex random CS-ICD, the vertex set is $\mathcal{X}_n$ and arc set $\mathcal{A}$ is defined by $(X_i, X_j) \in \mathcal{A} \iff X_j \in N(X_i, \tau, c)$ for X_i, X_j with $i \neq j$ in the middle intervals and $(X_i, X_j) \in \mathcal{A} \iff X_j \in N_e(X_i, \tau)$ for X_i, X_j with $i \neq j$ in the end intervals. We denote such digraphs as $\mathbf{D}_{n,m}(\tau, c)$. When $\tau = 1$ and $c = 1/2$ $\left(\text{i.e., } M_{c,i} = \left(Y_{(i-1)} + Y_{(i)}\right)/2\right)$ we have $N(x, 1, 1/2) = B(x, r(x))$ for an x in a middle interval and $N_e(x, 1) = B(x, r(x))$ for an x in an end interval where $r(x) = d(x, \mathcal{Y}_m) = \min_{y \in \mathcal{Y}_m} d(x, y)$ and the corresponding ICD is the CCCD of [Priebe *et al.*, 2001] or the proportional-edge PCD (PE-PCD) of [Ceyhan, 2012] with expansion parameter 2 and centrality parameter 1/2. See also Figure 1 (right).

3. ARC DENSITY OF CS-ICDS

For a digraph $D_n = (\mathcal{V}, \mathcal{A})$ with vertex set $\mathcal{V}$ and arc set $\mathcal{A}$, the arc density of D_n which is of order $|\mathcal{V}| = n \geq 2$, denoted $\rho(D_n)$, is defined as $\rho(D_n) = \frac{|\mathcal{A}|}{n(n-1)}$ ([Janson *et al.*, 2000]) where $|\cdot|$ stands for the set cardinality function. So $\rho(D_n)$ is the ratio of the number of arcs in the digraph D_n to the number of arcs in the complete symmetric digraph of order n, which is $n(n-1)$. For $n \leq 1$, we set $\rho(D_n) = 0$.

Let $\mathbb{I}_{ij} = \mathbb{I}\big((X_i, X_j) \in \mathcal{A}\big) = \mathbb{I}\big(X_j \in N(X_i)\big)$. Then for an ICD (hence for a CS-ICD), we can write the arc density as

$$\rho(D_n) \;=\; \frac{2}{n(n-1)} \sum_{i<j} h_{ij}$$

where $h_{ij} := (\mathbb{I}_{ij} + \mathbb{I}_{ji})/2$ Since the digraph is based on a relation that is not symmetric, h_{ij} is defined as half of the number of arcs between X_i and X_j in order to produce a symmetric kernel with finite variance ([Lehmann, 2004]). Notice that

$$\mathbf{E}\big[\rho(D_n)\big] \;=\; \mathbf{E}\left[h_{12}\right] \;=\; p_a$$

and

$$0 \le \mathbf{Var}\big[\rho(D_n)\big] = \frac{2}{n(n-1)}\mathbf{Var}\left[h_{12}\right] + \frac{4(n-2)}{n(n-1)}\mathbf{Cov}\left[h_{12}, h_{13}\right] \le 1/4$$

where

$$\mathbf{Var}[h_{ij}] = \mathbf{Var}\left[h_{12}\right] = \frac{1}{4}\mathbf{Var}\big[\mathbb{I}_{12} + \mathbb{I}_{21}\big] = (p_a + p_{sa})/2 - (1-p_a)^2 \ ,$$

where $p_{sa} = P\big(\big\{(X_i, X_j), (X_j, X_i)\big\} \subset \mathcal{A}\big)$ is the symmetric arc probability and

$$\mathbf{Cov}\left[h_{12}, h_{13}\right] = \mathbf{E}\left[h_{12}h_{13}\right] - p_a^2 \ ,$$

with

$$\begin{aligned} 4\,\mathbf{E}\left[h_{12}h_{13}\right] &= 4\,\mathbf{E}\big[(\mathbb{I}_{12} + \mathbb{I}_{21})(\mathbb{I}_{13} + \mathbb{I}_{31})\big] \\ &= P\Big(\{X_2, X_3\} \subset N(X_1)\Big) + 2\,P\Big(X_2 \in N(X_1), X_3 \in \Gamma_1(X_1, N)\Big) \\ &\quad + P\Big(\{X_2, X_3\} \subset \Gamma_1(X_1, N)\Big) \ . \end{aligned}$$

See [Ceyhan, 2012] for the derivations. Since $\rho(D_n)$, is a one-sample U-statistic of degree 2 and is an unbiased estimator of the arc probability p_a, a CLT for U-statistics ([Lehmann, 2004]) yields $\sqrt{n}\left[\rho(D_n) - p_a\right] \xrightarrow{\mathcal{L}} \mathbb{N}(0, 4\,\nu)$ as $n \to \infty$, where $\xrightarrow{\mathcal{L}}$ stands for convergence in law and $\mathbb{N}(\mu, \sigma^2)$ stands for the normal distribution with mean μ and variance σ^2 provided $\nu = \mathbf{Cov}\left[h_{ij}, h_{ik}\right] > 0$ for all $i \ne j \ne k,\ i, j, k \in \{1, 2, ..., n\}$.

Since $\mathbf{E}\left[|h_{ij}|^3\right] \le 1$, for $\nu > 0$, the sharpest rate of convergence in the asymptotic normality of $\rho(D_n)$ is

$$\begin{aligned} \sup_{t \in \mathbb{R}} \left| P\left(\frac{\sqrt{n}\left(\rho(D_n) - p_a\right)}{\sqrt{4\,\nu}} \le t\right) - \Phi(t)\right| &\le 8\,K\,p_a\,(4\nu)^{-3/2}\,n^{-1/2} \\ &= K\,\frac{p_a}{\sqrt{n\,\nu^3}} \ , \end{aligned} \tag{3.1}$$

where K is a constant and $\Phi(t)$ is the distribution function for the standard normal distribution ([Callaert and Janssen, 1978]).

3.1. Distribution of the arc density of CS-ICDs

We consider CS-ICDs for which $\mathcal{X}_n$ and $\mathcal{Y}_m$ are random samples from F_X and F_Y, respectively, so that the joint distribution of X, Y is $F_{X,Y} \in \mathcal{F}(\mathbb{R})$ where

$$\mathcal{F}(\mathbb{R}) := \Big\{F_{X,Y} \text{ on } \mathbb{R} \ \text{ with } P(X = Y) = 0 \\ \text{and the marginals, } F_X \text{ and } F_Y, \text{ are non-atomic}\Big\} \ .$$

Then the order statistics of $\mathcal{X}_n$ and $\mathcal{Y}_m$ are distinct with probability one. We denote such digraphs as $\mathbf{D}_{n,m}(F,\tau,c)$ and focus on the random variable $\rho_{n,m}(F,\tau,c) := \rho\big(\mathbf{D}_{n,m}(F,\tau,c)\big)$. Clearly $0 \leq \rho_{n,m}(F,\tau,c) \leq 1$, and $\rho_{n,m}(F,\tau,c) > 0$ for nontrivial digraphs.

We first partition the real line based on $\mathcal{Y}$ points. Along this line, we let $\mathcal{Y}_{[i]} := \big\{Y_{(i-1)}, Y_{(i)}\big\}$, $\mathcal{I}_i := \big(Y_{(i-1)}, Y_{(i)}\big)$, and $\mathcal{X}_{[i]} := \mathcal{X}_n \cap \mathcal{I}_i$ for $i = 1, 2, ..., (m+1)$. Let $\mathbf{D}_{[i]}(F,\tau,c)$ be the component of the random CS-ICD induced by the vertices in $\mathcal{X}_{[i]}$ (and based on $\mathcal{Y}_{[i]}$). Then we have a disconnected digraph with subdigraphs, each might be null or itself disconnected and denoted as $\mathbf{D}_{[i]}(F,\tau,c)$ for $i = 1, 2, ..., (m+1)$. Let $\mathcal{A}_{[i]}$ be the arc set of $\mathbf{D}_{[i]}(F,\tau,c)$, and $\rho_{[i]}(F,\tau,c)$ denote the arc density of $\mathbf{D}_{[i]}(F,\tau,c)$; $n_i := \big|\mathcal{X}_{[i]}\big|$, and F_i be the distribution F_X restricted to $\mathcal{I}_i$ for $i \in \{1, 2, ..., m+1\}$. Furthermore, let $M_{c,i} \in \mathcal{I}_i$ be the point so that it divides the interval $\mathcal{I}_i$ in ratios c and $1-c$. Since we have at most $m+1$ subdigraphs $\mathbf{D}_{[i]}(F,\tau,c)$ each of which having at most $n_i(n_i-1)$ arcs, it follows that we can have at most $n_{_T} := \sum_{i=1}^{m+1} n_i(n_i-1)$ arcs in the digraph $\mathbf{D}_{n,m}(F,\tau,c)$. We adjust the arc density for the entire digraph as

$$\widetilde{\rho}_{n,m}(F,\tau,c) := \frac{|\mathcal{A}|}{n_{_T}} = \frac{\sum_{i=1}^{m+1} |\mathcal{A}_{[i]}|}{n_{_T}} = \frac{1}{n_{_T}} \sum_{i=1}^{m+1} \big(n_i(n_i-1)\big)\, \rho_{[i]}(F,\tau,c)\,. \tag{3.2}$$

Hence, $\widetilde{\rho}_{n,m}(F,\tau,c)$ is called as the *adjusted arc density* and is a mixture of the $\rho_{[i]}(F,\tau,c)$ values, since $\frac{n_i(n_i-1)}{n_{_T}} \geq 0$ for each i and $\sum_{i=1}^{m+1} \frac{n_i(n_i-1)}{n_{_T}} = 1$. We first focus on the simpler random variable $\rho_{[i]}(F,\tau,c)$. The almost sure (a.s.) results follow from the marginal distributions F_X and F_Y being non-atomic in the rest of this section.

Lemma 3.1. *For $i \in \{1, (m+1)\}$ (i.e., in the end intervals) if $n_i \leq 1$, then $\rho_{[i]}(F,\tau,c) = 0$ for all $\tau > 0$. Moreover, if $n_i > 1$, then $\rho_{[i]}(F,\tau,c) \geq 1/2$ a.s. for all $\tau > 1$.*

Proof: By symmetry, distribution of $\rho_{[i]}(F,\tau,c)$ is same for $i = 1, m+1$. So we only consider $i = m+1$ (i.e., the right end interval). If $n_{m+1} \leq 1$, then by definition $\rho_{[m+1]}(\tau,c) = 0$. So, assume $n_{m+1} > 1$ and let $\mathcal{X}_{[m+1]} = \{Z_1, Z_2, ..., Z_{n_{m+1}}\}$ with $Z_{(j)}$ being the corresponding order statistics. Then there is an arc from $Z_{(j)}$ to each $Z_{(k)}$ for $k < j$, with $j, k \in \{1, 2, ..., n_{m+1}\}$ (and possibly to some other Z_l) for all $\tau > 1$, since $N_e\big(Z_{(j)}, \tau\big) = \big(Y_{(m)}, Z_{(j)} + \tau\,\big(Z_{(j)} - Y_{(m)}\big)\big)$ and so $Z_{(k)} \in N_e\big(Z_{(j)}, \tau\big)$. This implies that there are at least $0 + 1 + 2 + \cdots + n_{m+1} - 1 = n_{m+1}(n_{m+1}-1)/2$ arcs in $D_{[m+1]}(\tau,c)$. Then $\rho_{[m+1]}(\tau,c) \geq \big(n_{m+1}(n_{m+1}-1)/2\big)/\big(n_{m+1}(n_{m+1}-1)\big) = 1/2$. □

Let $\mathbf{D}_{n,m}(F,\tau,c)$ be a CS-ICD with $n > 0$ and $m > 0$. Then we obtain the following lower bound for $\rho_{n,m}(F,\tau,c)$ with $\tau > 1$.

Theorem 3.1. *Let k_1 and k_2 be two natural numbers defined as $k_1 := \sum_{i=2}^{m}\big(n_{i,\ell}(n_{i,\ell}-1)/2 + n_{i,r}(n_{i,r}-1)/2\big)$ and $k_2 := \sum_{i\in\{1,m+1\}} n_i(n_i-1)/2$, where $n_{i,\ell} := \big|\mathcal{X}_n \cap \big(Y_{(i-1)}, M_{c,i}\big)\big|$ and $n_{i,r} := \big|\mathcal{X}_n \cap \big(M_{c,i}, Y_{(i)}\big)\big|$. Then for $\tau > 1$, we have $(k_1+k_2)/n_{_T} \le \rho_{n,m}(F,\tau,c) \le 1$ a.s.*

Proof: We have k_2 as in Lemma 3.1 for the end intervals (i.e., for $i \in \{1,(m+1)\}$). In the middle intervals, i.e., for $i \in \{2,3,...,m\}$, let $\mathcal{X}_{i,\ell} := \mathcal{X}_{[i]} \cap \big(Y_{(i-1)}, M_{c,i}\big) = \{U_1, U_2, ..., U_{n_{i,\ell}}\}$, and $\mathcal{X}_{i,r} := \mathcal{X}_{[i]} \cap \big(M_{c,i}, Y_{(i)}\big) = \{V_1, V_2, ..., V_{n_{i,r}}\}$. Furthermore, let $U_{(j)}$ and $V_{(k)}$ be the corresponding order statistics. For $\tau > 1$, there is an arc from $U_{(j)}$ to $U_{(k)}$ and possibly to some other U_l for $k < j$ with $j,k,l \in \{1,2,...,n_{i,\ell}\}$, and similarly there is an arc from $V_{(j)}$ to $V_{(k)}$ and possibly to some other V_l for $k > j$ with $j,k,l \in \{1,2,...,n_{i,r}\}$. Therefore, we have $\rho_{n,m}(F,\tau,c) \ge (k_1+k_2)/n_{_T}$, since there are at least $n_{i,\ell}(n_{i,\ell}-1)/2 + n_{i,r}(n_{i,r}-1)/2$ arcs in $\mathbf{D}_{[i]}(F,\tau,c)$. □

Theorem 3.2. *When the expansion parameter is infinity (i.e., $\tau = \infty$), we have $\rho_{_{[i]}}(\tau=\infty,c) = \mathbb{I}(n_i > 1)$ and $\rho_{n,m}(\tau=\infty,c) = 1$ a.s. for $i = 1,2,3,...,m+1$ and $n_i > 1$.*

Proof: For $\tau = \infty$, if $n_i \le 1$, then $\rho_{_{[i]}}(\tau=\infty,c) = 0$. So we assume $n_i > 1$ and let $i = m+1$. Then $N_e(x,\infty) = \big(Y_{(m)}, \infty\big)$ for all $x \in \big(Y_{(m)}, \infty\big)$. Hence $D_{[m+1]}(\infty,c)$ is a complete symmetric digraph of order n_{m+1}, which implies $\rho_{_{[m+1]}}(\tau=\infty,c) = 1$. By symmetry, the same holds for $i = 1$. For $i \in \{2,3,...,m\}$ and $n_i > 1$, we have $N(x,\infty,c) = \mathcal{I}_i$ for all $x \in \mathcal{I}_i$, hence $D_{[i]}(\infty,c)$ is a complete symmetric digraph of order n_i, which implies $\rho_{_{[i]}}(\infty,c) = 1$. Then $\rho_{n,m}(\infty,c) = \sum_{i=1}^{m+1} \frac{n_i(n_i-1)\,\rho_{_{[i]}}(\infty,c)}{n_{_T}} = 1$, since when $n_i \le 1$, n_i has no contribution to $n_{_T}$, and when $n_i > 1$, we have $\rho_{_{[i]}}(\infty,c) = 1$. □

4. DISTRIBUTION OF THE ARC DENSITY OF CS-ICDS FOR UNIFORM DATA

Let $\mathcal{X}_n = \{X_1, X_2, ..., X_n\}$ be a random sample from $F_X = \mathcal{U}(\delta_1,\delta_2)$, the uniform distribution on the bounded interval (δ_1,δ_2), and let $\mathcal{Y}_m$ be a random sample from non-atomic F_Y with support $\mathcal{S}(F_Y) \subseteq (\delta_1,\delta_2)$. Then $F_{X,Y} \in \mathcal{F}(\mathbb{R})$. Suppose we have a realization of $\mathcal{Y}_m$ as $\mathcal{Y}_m = \{y_1, y_2, ..., y_m\}$ with the order statistics satisfying $\delta_1 < y_{(1)} < y_{(2)} < \cdots < y_{(m)} < \delta_2$, with the extension that $y_{(0)} := \delta_1$ and $y_{(m+1)} := \delta_2$. Then the distribution of X_i restricted to $\mathcal{I}_i$ is $F_X|_{\mathcal{I}_i} = \mathcal{U}(\mathcal{I}_i)$. We provide the distribution of the arc density of $\mathbf{D}_{n,m}(\tau,c)$ for the whole range of the parameters τ and c. The following "scale invariance" for CS-ICDs will

allow us to consider the special case of the unit interval $(0,1)$ as the support of $\mathcal{X}$ points, thereby simplifying the computations in our subsequent analysis.

Theorem 4.1 (Scale Invariance Property). *Let $\mathcal{Y}_m$ be a set of m distinct $\mathcal{Y}$ points in a bounded interval (δ_1,δ_2) and $\mathcal{X}_n$ be random sample from $\mathcal{U}(\delta_1,\delta_2)$. Then the distribution of $\rho_{[i]}(\tau,c)$ is independent of $\mathcal{Y}_{[i]}$ (and hence independent of the restricted support interval $\mathcal{I}_i$) for all $i\in\{1,2,...,m+1\}$, $\tau>0$, and $c\in(0,1)$.*

Proof: Let δ_1 and δ_2 and $\mathcal{Y}_m$ be as in the hypothesis. Any $\mathcal{U}(\delta_1,\delta_2)$ random variable can be transformed into a $\mathcal{U}(0,1)$ random variable by $\phi(x)=(x-\delta_1)/(\delta_2-\delta_1)$, which maps intervals $(t_1,t_2)\subseteq(\delta_1,\delta_2)$ to intervals $\big(\phi(t_1),\phi(t_2)\big)\subseteq(0,1)$. That is, if $X\sim\mathcal{U}(\delta_1,\delta_2)$, then we have $\phi(X)\sim\mathcal{U}(0,1)$ and $P\big(X\in(t_1,t_2)\big)=P\big(\phi(X)\in\big(\phi(t_1),\phi(t_2)\big)\big)$ for all $(t_1,t_2)\subseteq(\delta_1,\delta_2)$. The distribution of $\rho_{[i]}(\tau,c)$ is obtained by calculating such probabilities. So, without loss of generality, we can assume $\mathcal{X}_{[i]}$ is a set of iid (independent identically distributed) random variables from the $\mathcal{U}(0,1)$ distribution. That is, the distribution of $\rho_{[i]}(\tau,c)$ does not depend on $\mathcal{Y}_{[i]}$ and hence does not depend on the restricted support interval $\mathcal{I}_i$. □

For $\tau=\infty$, we have $\rho_{[i]}(\tau=\infty,c)=1$ a.s. for any non-atomic F_X with support in (δ_1,δ_2), hence the scale invariance of $\rho_{[i]}(\tau=\infty,c)$ holds for all $\mathcal{X}_n$ from any such F_X. Based on Theorem 4.1, we may assume each $\mathcal{I}_i$ as the unit interval $(0,1)$ for uniform data. If $x\in\mathcal{I}_i$ for $i\in\{2,...,m\}$ (i.e., in the middle intervals), when transformed to $(0,1)$, the central similarity proximity region for $x\in(0,1)$ with parameters $c\in(0,1)$ and $\tau>0$ is

$$N(x,\tau,c)=\begin{cases}\Big((1-\tau)x,\Big(1+\frac{(1-c)}{c}\tau\Big)x\Big)\cap(0,1) & \text{if } x\in(0,c)\,,\\ \Big(x-\frac{c\,\tau}{(1-c)}(1-x),\,x+(1-x)\tau\Big)\cap(0,1) & \text{if } x\in(c,1)\,.\end{cases}\tag{4.1}$$

In particular, for $\tau\in(0,1]$, we have

$$N(x,\tau,c)=\begin{cases}\Big((1-\tau)x,\Big(1+\frac{(1-c)}{c}\tau\Big)x\Big) & \text{if } x\in(0,c)\,,\\ \Big(x-\frac{c\,\tau}{(1-c)}(1-x),\,x+(1-x)\tau\Big) & \text{if } x\in(c,1)\,,\end{cases}\tag{4.2}$$

and for $\tau>1$, we have

$$N(x,\tau,c)=\begin{cases}\Big(0,\Big(1+\frac{(1-c)}{c}\tau\Big)x\Big) & \text{if } x\in\Big(0,\frac{c}{c+(1-c)\,\tau}\Big),\\ (0,1) & \text{if } x\in\Big(\frac{c}{c+(1-c)\,\tau},\frac{c\,\tau}{1-c+c\,\tau}\Big),\\ \Big(x-\frac{c\,\tau}{(1-c)}(1-x),1\Big) & \text{if } x\in\Big(\frac{c\,\tau}{1-c+c\,\tau},1\Big),\end{cases}\tag{4.3}$$

and $N(x=c,\tau,c)$ is arbitrarily taken to be one of the two defining intervals above. But the case of "$X=c$" happens with probability zero for uniform X.

Furthermore, when transformed to $(0,1)$, if x is in the left end interval (i.e., $x \in \mathcal{I}_1$), we have $N_e(x,\tau) = \big(\max\big(0, x-\tau(1-x)\big), \min\big(1, x+\tau(1-x)\big)\big)$; and if x is in the right end interval (i.e., $x \in \mathcal{I}_{m+1}$), we have $N_e(x,\tau) = \big(\max\big(0, (1-\tau)\,x\big), \min\big(1, (1+\tau)\,x\big)\big)$.

Each subdigraph $D_{[i]}(\tau,c)$ is itself a random CS-ICD (for brevity of notation, we suppress the dependence on the uniform distribution). The distribution of the arc density of $D_{[i]}(\tau,c)$ is given in the following theorem.

Theorem 4.2. *Let $\rho_{[i]}(\tau,c)$ be the arc density of subdigraph $D_{[i]}(\tau,c)$ of the CS-ICD based on $\mathcal{U}(\delta_1,\delta_2)$ data and $\mathcal{Y}_m$ be a set of m distinct $\mathcal{Y}$ points in (δ_1,δ_2). Then, as $n_i \to \infty$, for $\tau \in (0,\infty)$ we have,*

(**i**) $\sqrt{n_i}\left[\rho_{[i]}(\tau,c) - p_a(\tau,c)\right] \xrightarrow{\mathcal{L}} \mathbb{N}(0, 4\,\nu(\tau,c))$, *where* $p_a(\tau,c) = \mathbf{E}\Big[\rho_{[i]}(\tau,c)\Big]$ *is the arc probability and* $\nu(\tau,c) = \mathbf{Cov}[h_{12}, h_{13}]$ *in the middle intervals (i.e., for* $i \in \{2,...,m\}$*), and*

(**ii**) $\sqrt{n_i}\left[\rho_{[i]}(\tau,c) - p_a^e(\tau,c)\right] \xrightarrow{\mathcal{L}} \mathbb{N}(0, 4\,\nu_e(\tau))$, *where* $p_a^e(\tau,c) = \mathbf{E}\Big[\rho_{[i]}(\tau,c)\Big]$ *is the arc probability and* $\nu_e(\tau) = \mathbf{Cov}[h_{12}, h_{13}]$ *in the end intervals (i.e., for* $i \in \{1, m+1\}$*).*

Proof: By Theorem 1 of [Ceyhan, 2012], arc density of CS-ICDs is a U-statistic, and hence the proofs follow by the asymptotic normality of U-statistics provided the asymptotic variance is positive. In particular, in (i) by the scale invariance for uniform data (see Theorem 4.1), a middle interval can be assumed to be the unit interval $(0,1)$. Then

$$\mathbf{E}\Big[\rho_{[i]}(\tau,c)\Big] \;=\; \mathbf{E}[h_{12}] \;=\; P\Big(X_2 \in N(X_1,\tau,c)\Big) \;=\; p_a(\tau,c)$$

which is the arc probability. Similarly in (ii) we have $\mathbf{E}\Big[\rho_{[i]}(\tau,c)\Big] = \mathbf{E}[h_{12}] = P\big(X_2 \in N_e(X_1,\tau)\big) = p_a^e(\tau,c)$.

Furthermore, in (i), for $\tau \in (0,\infty)$, h_{12} and h_{13} tend to be high (resp. low) together, if the proximity region $N(X_1,\tau,c)$ is large (resp. small), since $2h_{12} = \mathbb{I}\big(X_2 \in N(X_1,\tau,c)\big) + \mathbb{I}\big(X_1 \in N(X_2,\tau,c)\big)$ is the number of arcs between X_1 and X_2 in the ICDs. Hence the asymptotic variance of $\rho_{[i]}(\tau,c)$, $\mathbf{Cov}[h_{12}, h_{13}] = 4\,\nu(\tau,c) > 0$. The same holds for end intervals in (ii) as well. □

For middle intervals, the asymptotic variance in Theorem 4.2 can be written as

$$\mathbf{Cov}[h_{12}, h_{13}] \;=\; \frac{1}{4}\left(P_{2N} + 2\,P_{NG} + P_{2G}\right) - p_a(\tau,c)^2\,,$$

where

$$P_{2N} \;:=\; P\Big(\{X_2, X_3\} \subset N(X_1,\tau,c)\Big)\,,$$

$$P_{NG} \;:=\; P\Big(X_2 \in N(X_1,\tau,c),\, X_3 \in \Gamma_1(X_1,\tau,c)\Big)\,,$$

and

$$P_{2G} := P\Big(\{X_2, X_3\} \subset \Gamma_1(X_1, \tau, c)\Big).$$

Similarly, for end intervals

$$\mathbf{Cov}[h_{12}, h_{13}] = \frac{1}{4}\left(P_{2N,e} + 2\,P_{NG,e} + P_{2G,e}\right) - p_a^e(\tau, c)^2,$$

where

$$\begin{aligned} P_{2N,e} &:= P\Big(\{X_2, X_3\} \subset N_e(X_1, \tau)\Big), \\ P_{NG,e} &:= P\Big(X_2 \in N_e(X_1, \tau),\, X_3 \in \Gamma_{1,e}(X_1, \tau)\Big), \\ \text{and} \qquad P_{2G,e} &:= P\Big(\{X_2, X_3\} \subset \Gamma_{1,e}(X_1, \tau)\Big), \end{aligned}$$

with $\Gamma_{1,e}(x, \tau)$ being the Γ_1-region corresponding to $N_e(x, \tau)$ in the end intervals. Furthermore, for $\tau = \infty$, $\mathbf{E}\Big[\rho_{[i]}(\infty, c)\Big] = \mathbf{E}[h_{12}] = \mu(\infty, c) = P\big(X_2 \in N(X_1, \infty, c)\big) = P(X_2 \in \mathcal{I}_i) = 1$ and $\nu(\infty, c) = 0$. Thus, $\rho_{[i]}(\tau = \infty, c) = 1$ a.s. and the CLT result does not hold for $\tau = \infty$.

4.1. Distribution of the arc density of $\mathbf{D}_{n,2}(\tau, c)$

In this section, we find the distribution of the arc density of $\mathbf{D}_{n,2}(\tau, c)$ for $\tau > 0$ and $c \in (0, 1)$. For the special case of $m = 2$, we have $\mathcal{Y}_2 = \{y_1, y_2\}$ and $\delta_1 = y_1 < y_2 = \delta_2$, and only one middle interval and the two end intervals are empty. By Theorems 4.1 and 4.2, the asymptotic distribution of any $\rho_{[i]}(\tau, c)$ for the middle intervals with $m > 2$ will be same as the asymptotic distribution of density of the CS-ICD based on $\mathcal{U}(0, 1)$ data.

For $\tau \in (0, 1]$, the proximity region is defined as in Equation (4.2) and for $\tau > 1$, the proximity region is as in Equation (4.3).

Theorem 4.3. *For $\tau \in (0, \infty)$, we have $\sqrt{n}\left[\rho_{n,2}(\tau, c) - p_a(\tau, c)\right] \xrightarrow{\mathcal{L}} \mathbb{N}\big(0, 4\,\nu(\tau, c)\big)$, as $n \to \infty$, where*

$$p_a(\tau, c) = \begin{cases} \frac{\tau}{2} & \text{if } 0 < \tau < 1, \\ \frac{\tau\,(1 + 2\,c\,(\tau - 1)(1 - c))}{2\,(c\,\tau - c + 1)(\tau + c - c\,\tau)} & \text{if } \tau > 1, \end{cases} \tag{4.4}$$

and

$$4\,\nu_1(\tau, c) = \kappa_1(\tau, c)\,\mathbb{I}(0 < \tau < 1) + \kappa_2(\tau, c)\,\mathbb{I}(\tau > 1)$$

where

$$\kappa_1(\tau, c) = \frac{\tau^2\left(c^2\,\tau^3 - 3\,c^2\,\tau^2 - c\,\tau^3 + 2\,c^2\,\tau + 3\,c\,\tau^2 - c^2 - 2\,c\,\tau - \tau^2 + c + \tau\right)}{3\,(c\,\tau - c + 1)\,(c + \tau - c\,\tau)},$$

and

$$\begin{aligned}\kappa_2(\tau,c) \;=\; &\Big[\, c\,(1-c)\,\Big(2\,c^4\tau^5 - 7\,c^4\tau^4 - 4\,c^3\tau^5 + 8\,c^4\tau^3 + 14\,c^3\tau^4 + 3\,c^2\tau^5 \\ &- 2\,c^4\tau^2 - 16\,c^3\tau^3 - 7\,c^2\tau^4 - c\,\tau^5 - 2\,c^4\tau + 4\,c^3\tau^2 + 12\,c^2\tau^3 \\ &+ c^4 + 4\,c^3\tau - 6\,c^2\tau^2 - 4\,c\,\tau^3 - 2\,c^3 - 3\,c^2\tau + 4\,c\,\tau^2 \\ &+ c^2 + c\,\tau - \tau^2\Big)\Big] \Big/ \Big[3\,\big(c\,\tau - c + 1\big)^3\big(c\,\tau - c - \tau\big)^3\Big]\,.\end{aligned}$$

The proof is provided in the Appendix. Notice that $p_a(\tau, c)$ is independent of the centrality parameter c for $\tau \in (0, 1]$. See Figure 2 for the surface plots of $p_a(\tau, c)$ and $4\,\nu(\tau, c)$. Observe that $\lim_{\tau\to\infty} \nu(\tau, c) = 0$, so the CLT result fails for $\tau = \infty$ and $\lim_{\tau\to 0} \nu_1(\tau, c) = 0$, but CS-ICD is not defined for $\tau = 0$.

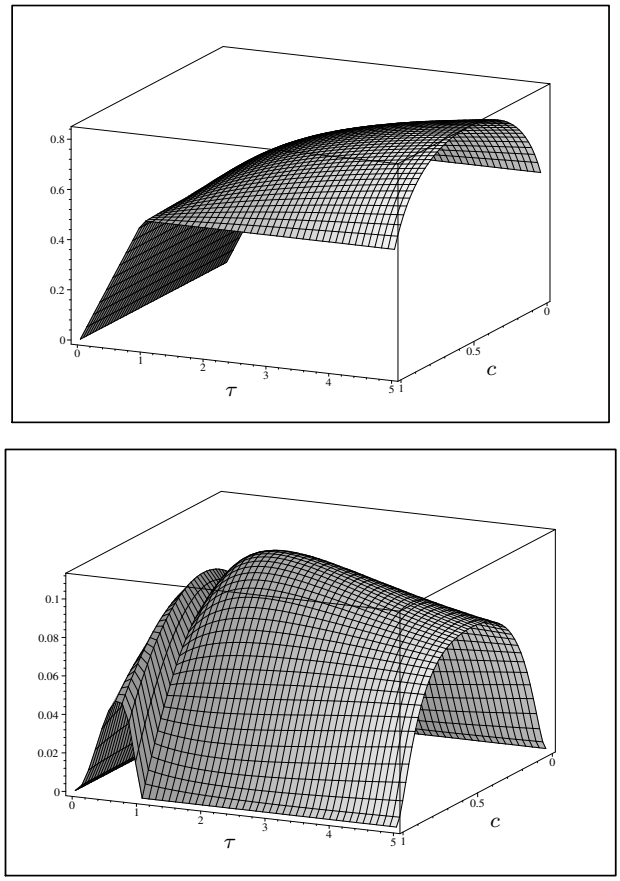

Figure 2: The surface plots of the asymptotic mean $p_a(\tau, c)$ (top) and the variance $4\,\nu(\tau, c)$ (bottom) as a function of τ and c for $\tau \in (0, 5]$ and $c \in (0, 1)$, respectively.

The sharpest rate of convergence in Theorem 4.3 is $K\frac{p_a(\tau,c)}{\sqrt{n\,\nu(\tau,c)^3}}$ (the explicit form not presented) and is minimized at $\tau \approx 1.55$ and $c = 1/2$ which is found by setting

the first order partial derivatives of this convergence rate with respect to τ and c to zero and solving for τ and c numerically and verified by the surface plot. Surface plots for the convergence rates $f^c_{CS}(\tau, c)$ and $f^c_{PE}(\tau, c)$ are presented in Figure 3. At optimal parameters within their entire ranges, the convergence rate for the arc density of CS-ICDs is faster than that of the PE-PCDs.

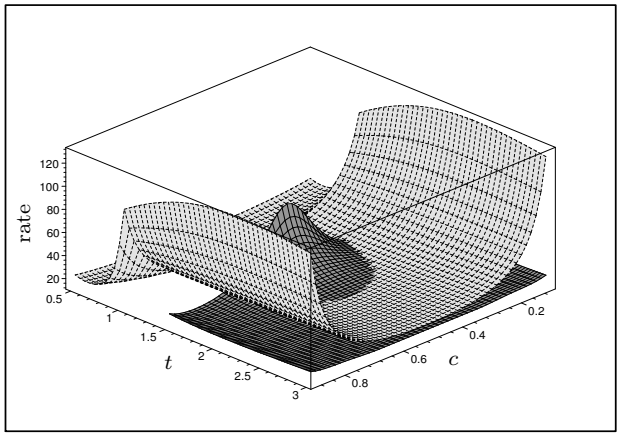

Figure 3: The surface plots of the rates of convergence to normality for PE- and CS-ICDs for the entire ranges of expansion parameter, t, and centrality parameter, c. The rate for CS-ICD is plotted in light gray, while that for PE-PCDs is plotted in dark gray.

Each of the following special cases follows as a corollary of Theorem 4.3.

Special Cases:

Case (i): $\tau > 0$ and $c = 1/2$.

As $n \to \infty$, we have $\sqrt{n}\left[\rho_{n,2}(\tau, 1/2) - p_a(\tau, 1/2)\right] \xrightarrow{\mathcal{L}} \mathbb{N}\big(0, 4\,\nu(\tau, 1/2)\big)$, where

$$p_a(\tau, 1/2) = \begin{cases} \tau/2 & \text{if } 0 < \tau < 1\,, \\ \tau/(\tau+1) & \text{if } \tau > 1\,, \end{cases} \tag{4.5}$$

and

$$4\,\nu(\tau, 1/2) = \begin{cases} \frac{\tau^2\,(1+2\,\tau-\tau^2-\tau^3)}{3\,(\tau+1)^2} & \text{if } 0 < \tau \leq 1\,, \\ \frac{2\,\tau-1}{3\,(\tau+1)^2} & \text{if } \tau > 1\,. \end{cases} \tag{4.6}$$

Case (ii): $\tau = 1$ and $c \in (0, 1)$.

As $n \to \infty$, we have $\sqrt{n}\left[\rho_{n,2}(1, c) - p_a(1, c)\right] \xrightarrow{\mathcal{L}} \mathbb{N}\big(0, 4\,\nu(1, c)\big)$, where $p_a(1, c) = 1/2$ and $4\,\nu(1, c) = c\,(1-c)/3$.

Case (iii): $\tau = 1$ and $c = 1/2$:

As $n \to \infty$, we have $\sqrt{n}\left[\rho_n(1,1/2) - p_a(1,1/2)\right] \xrightarrow{\mathcal{L}} \mathbb{N}\big(0, 4\,\nu(1,1/2)\big)$, where $p_a(1,1/2) = 1/2$ and $4\,\nu(1,1/2) = 1/12$.

4.2. Arc density in the case of end intervals (for $\mathcal{U}\left(\delta_1, y_{(1)}\right)$ or $\mathcal{U}\left(y_{(m)}, \delta_2\right)$ data)

With $m \geq 1$, we have the end intervals, $\mathcal{I}_1 = \big(\delta_1, y_{(1)}\big)$ and $\mathcal{I}_{m+1} = \big(y_{(m)}, \delta_2\big)$. In these intervals, the proximity and Γ_1-regions are only dependent on x and τ (but not on c). Let $D_{[i]}(1,c)$ be the subdigraph of the CS-ICD based on uniform data in (δ_1, δ_2) where $\delta_1 < \delta_2$ and $\mathcal{Y}_m$ be a set of m distinct $\mathcal{Y}$ points in (δ_1, δ_2). By scale invariance of Theorem 4.1, we can assume that each of the end intervals is $(0,1)$.

For $\tau \in (0,1]$ and x in the right end interval, the proximity region is

$$N_e(x,\tau) = \begin{cases} \big((1-\tau)\,x,\,(1+\tau)\,x\big) & \text{if } x \in \big(0, 1/(1+\tau)\big)\,, \\ \big((1-\tau)\,x, 1\big) & \text{if } x \in \big(1/(1+\tau), 1\big)\,, \end{cases} \tag{4.7}$$

and for $\tau > 1$ and x in the right end interval, the proximity region is

$$N_e(x,\tau) = \begin{cases} \big(0, (1+\tau)\,x\big) & \text{if } x \in \big(0, 1/(1+\tau)\big)\,, \\ (0,1) & \text{if } x \in \big(1/(1+\tau), 1\big)\,. \end{cases} \tag{4.8}$$

Theorem 4.4. *For $i \in \{1, m+1\}$ and $\tau \in (0,\infty)$, as $n_i \to \infty$, we have $\sqrt{n_i}\left[\rho_{[i]}(\tau) - p_a^e(\tau)\right] \xrightarrow{\mathcal{L}} \mathbb{N}(0, 4\,\nu_e(\tau))$, where*

$$p_a^e(\tau, c) = \begin{cases} \frac{\tau\,(\tau+2)}{2\,(\tau+1)} & \text{if } 0 < \tau < 1\,, \\ \frac{1+2\,\tau}{2\,(\tau+1)} & \text{if } \tau > 1\,, \end{cases} \tag{4.9}$$

and

$$4\,\nu_e(\tau) = \begin{cases} \frac{\tau^2\left(4\,\tau+4-2\,\tau^4-4\,\tau^3-\tau^2\right)}{3\,(\tau+1)^3} & \text{if } 0 < \tau < 1\,, \\ \frac{\tau^2}{3\,(\tau+1)^3} & \text{if } \tau > 1\,. \end{cases} \tag{4.10}$$

The proof is provided in the Appendix. See Figure 4 for the plots of $p_a^e(\tau)$ and $4\,\nu_e(\tau)$ with $\tau \in (0, 10]$. Notice that $\lim_{\tau\to\infty} \nu_e(\tau) = 0$, so the CLT result fails for $\tau = \infty$ and $\lim_{\tau\to 0} \nu_e(\tau) = 0$. The sharpest rate of convergence in Theorem 4.4

is $K\frac{p_a^e(\tau)}{\sqrt{n_i\,\nu_e(\tau)^3}}$ (explicit form not presented) for $i \in \{1, m+1\}$ and is minimized at $\tau \approx 0.58$ which is found numerically as before and verified by the plot of $p_a^e(\tau)/\sqrt{\nu_e(\tau)^3}$.

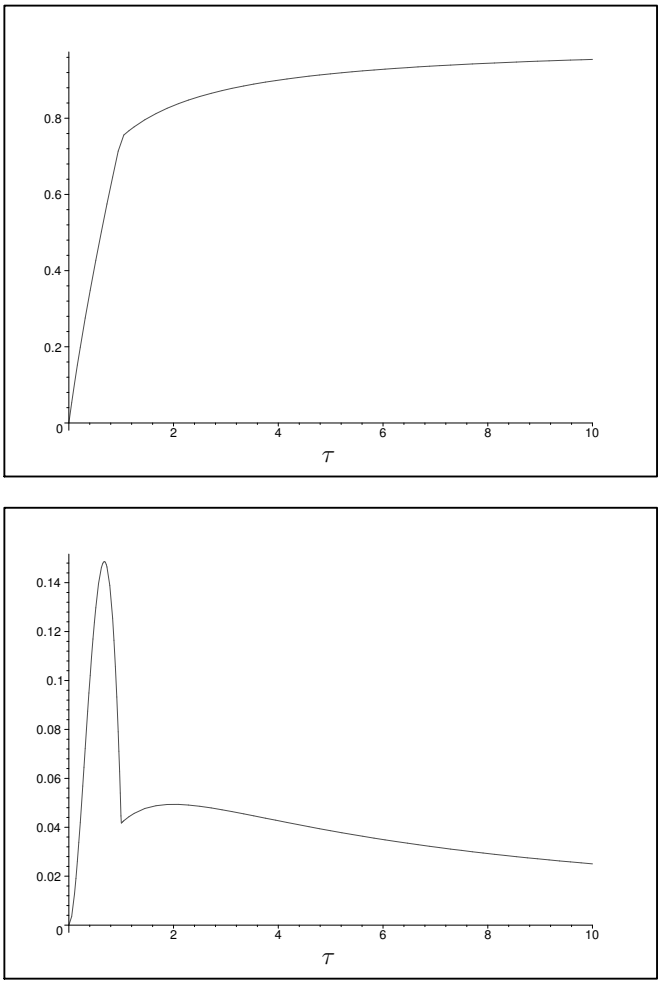

Figure 4: The plots of the asymptotic mean $p_a^e(\tau, c)$ (top) and the variance $4\,\nu_e(\tau)$ (bottom) for the end intervals as a function of τ for $\tau \in (0, 10]$.

The distribution for the special case of $\tau = 1$ follows as a corollary to Theorem 4.4: For x in the right end interval, $N_e(x, 1) = \big(0, \min(1, 2x)\big)$ for $x \in (0, 1)$. For $i \in \{1, m+1\}$ (i.e., in the end intervals), as $n_i \to \infty$, we have $\sqrt{n_i}\Big[\rho_{[i]}(1) - p_a^e(1)\Big] \xrightarrow{\mathcal{L}} \mathbb{N}(0, 4\,\nu_e(1))$, where $p_a^e(1) = 3/4$ and $4\,\nu_e(1) = 1/24$.

Remark 4.1. Multiple Interval Case: In the case of $m > 2$, we have two versions of arc density. One is defined as the (adjusted) arc density as in Equation

(3.2). The asymptotic distribution of $\rho_{n,m}(\tau, c)$ is the same as given in Theorem 11 of [Ceyhan, 2012]. As for the other one, if we consider the entire data set $\mathcal{X}_n$, then we have n vertices. So we can also consider the arc density as $\rho_{n,m}(\tau, c) = |\mathcal{A}| / \big(n\,(n-1)\big)$. The asymptotic distribution for $\rho_{n,m}(\tau, c)$ is as in Theorem 12 of [Ceyhan, 2012].

Remark 4.2. Use of Arc Density for Testing Multi-Class Spatial Interactions: Arc density of CS-ICDs can be employed in testing two-class spatial point patterns of one-dimensional data, as was done in [Ceyhan *et al.*, 2007] for two-dimensional data. In particular, for two classes of points, $\mathcal{X}$ and $\mathcal{Y}$, whose support is in a compact interval in $\mathbb{R}$, we assume some form of *complete spatial randomness* of $\mathcal{X}$ points (i.e., $\mathcal{X}$ points having uniform distribution in the support interval irrespective of the distribution of the $\mathcal{Y}$ points) as our null hypothesis. The alternative patterns of interest are *segregation* of $\mathcal{X}$ from $\mathcal{Y}$ points or *association* of $\mathcal{X}$ points with $\mathcal{Y}$ points. Association is the pattern in which the points from the two classes tend to occur close to each other, while segregation is the pattern in which the points from the same class tend to cluster together. In our context, association implies that $\mathcal{X}$ points are clustered around $\mathcal{Y}$ points, while segregation implies that $\mathcal{X}$ points are clustered away from the $\mathcal{Y}$ points. The use of arc density of CS-ICDs requires number of $\mathcal{X}$ points to be much larger compared to the number of $\mathcal{Y}$ points. Furthermore, the power comparisons are possible for data from large families of distributions to obtain the optimal parameters against segregation and association alternatives.

Remark 4.3. Extension of Central Similarity Proximity Regions to Higher Dimensions: In this article, we discuss the construction of CS-ICDs for one-dimensional data and asymptotic distribution of their arc density (for uniform data). The CS-ICDs in this article can be viewed as the one-dimensional version of the PCDs introduced in [Ceyhan *et al.*, 2007], which also contains the extension to higher dimensions.

5. TESTING UNIFORMITY WITH THE ARC DENSITY OF CS-ICDS

We can employ the arc density of the CS-PCDs for testing uniformity based on its asymptotic normality. Let $X_i \overset{iid}{\sim} F$ for $i = 1, 2, ..., n$ where F has a bounded interval support (a, b) in $\mathbb{R}$. Then our null hypothesis is $H_o\colon F = \mathcal{U}(a, b)$. For testing this hypothesis, we use the arc density $\rho_{n,2}(\tau, c)$ whose asymptotic distribution is provided in Theorem 4.3 for uniform data. By the scale invariance property of the distribution of $\rho_{n,2}(\tau, c)$ (see Theorem 4.1), without loss of generality, we can assume $(a, b) = (0, 1)$. In this approach, for each choice of (τ, c),

we compute the arc density, $\rho_{n,2}(\tau, c)$, and standardize it as

$$R_n(\tau, c) := \sqrt{n}\big(\rho_{n,2}(\tau, c) - p_a(\tau, c)\big)\big/\sqrt{4\,\nu(\tau, c)}$$

and use the standardized version as our test statistic. The critical values for the one- and two-sided alternatives are based on the standard normal distribution, e.g., the level α critical value for the left-sided alternative is z_α, the $\alpha\times 100^{\text{th}}$ percentile of the standard normal distribution.

For comparative purposes, we employ the arc density of PE-ICDs introduced by [Ceyhan, 2012]. In particular, the defining regions for the PE-ICD are

$$N_{PE}(x, r, c) = \begin{cases} (0, r\,x) \cap (0,1) & \text{if } x \in (0, c)\,, \\ \big(1 - r\,(1-x), 1\big) \cap (0,1) & \text{if } x \in (c, 1)\,. \end{cases} \tag{5.1}$$

The asymptotic distribution of the arc density of PE-ICDs for uniform data was provided in [Ceyhan, 2012]. Furthermore, we also employ Kolmogorov–Smirnov (KS) test for uniform distribution and Neyman's smooth test of uniformity since the former is one of the most commonly used tests and latter is recommended for a large family of alternatives for testing uniformity by [Marhuenda *et al.*, 2005].

5.1. Empirical size analysis

We first perform an extensive size analysis to determine for which parameter values the arc density of the ICDs have the appropriate size at specific sample sizes in testing $H_o\colon F = \mathcal{U}(0,1)$. For this purpose, we partition the ranges of τ and c for the CS-ICD as follows. We take $c = .01, .02, ..., .99$ and $\tau = .01, .02, ..., 10.00$, and consider each (τ, c) combination on a 99×1000 grid with $n = 20, 50, 100$. Similarly, we partition the ranges of r and c for the PE-ICD as follows. We use the same partition above for c and take $r = 1.01, ..., 10.00$, and consider each (r, c) combination on a 99×900 grid with $n = 20, 50, 100$. For each (τ, c) (and (r, c)) combination, we generate $N_{mc} = 10000$ samples of size n iid from $\mathcal{U}(0,1)$ distribution. Then for each sample generated, we compute the arc densities and use their standardized versions as approximate test statistics. Empirical size is estimated as the frequency of number of times p-value is significant at $\alpha = .05$ level divided by $N_{mc} = 10000$. We also estimate the empirical sizes for KS and Neyman's smooth tests with $n = 20$ and $N_{mc} = 10000$. With $N_{mc} = 10000$, empirical size estimates larger than .0536 (resp. less than .0464) are deemed liberal (resp. conservative). These bounds are determined using a binomial test for the proportions with $n = 10000$ trials at .05 level of significance. The size estimates for KS and Neyman's smooth test are found to be about the nominal level (i.e., between .0464 and .0536).

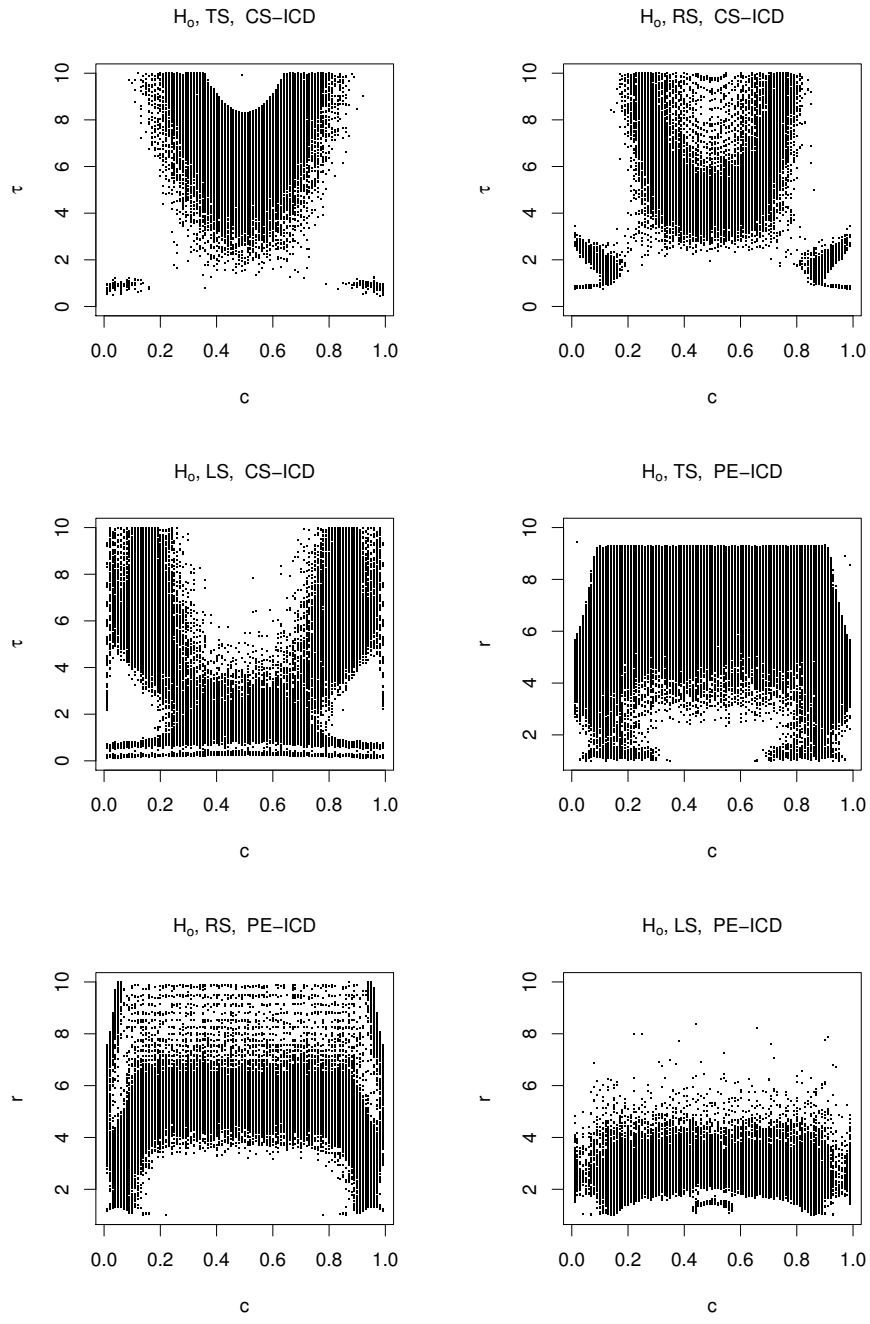

Figure 5: Two-level (i.e., black and white) image plots for the empirical size estimates for the arc density of CS-ICD and PE-PCD based on $n = 20$ and $N_{mc} = 10000$ the two-sided (TS), right-sided (RS) and left-sided (LS) alternatives. The empirical sizes not significantly different from 0.05 are represented with black dots, and others are represented with white dots. For CS-ICD, we take $\tau = .01, .02, ..., 10.00$ and for PE-ICD, we take $r = 1.01, 1.02, ..., 10.00$ and for both ICDs, we take $c = .01, .02, ..., .99$ with $N_{mc} = 10000$ Monte Carlo replications.

We present the empirical size estimates in two-level image plots (with empirical sizes not significantly different from 0.05 in black, and others in white) for the two-, right- and left-sided alternatives for the CS-ICD with $n = 20$, $c = .01, .02, ..., .99$ and $\tau = .01, .02, ..., 10.00$ and for the PE-ICD with $n = 20$, $c = .01, .02, ..., .99$ and $r = 1.01, 1.02, ..., 10.00$ in Figure 5. The size estimates for $n = 50$ and 100 have the similar trend with sizes closer to nominal level for more parameter combinations (not presented). Notice that there is symmetry in size estimates around $c = 1/2$. For the one-sided alternatives, the regions at which size estimates are appropriate are somewhat complementary, in the sense that, the sizes are appropriate for the parameter combinations in one region for left-sided alternative and mostly in its complement for the right-sided alternative for each ICD family. Notice also that arc density of PE-ICD has appropriate size for the two-sided alternative for more parameter combinations, and arc density of CS-ICD has appropriate size for the left-sided alternative for more parameter combinations.

5.2. Empirical power analysis

We perform power analysis to determine which tests have higher power under various alternatives against uniformity. For the alternatives, we use three families of non-uniform distributions with support in $(0, 1)$ which are proposed by [Stephens, 1974]:

$$\begin{aligned}
&\textbf{(I)} \quad F_1(x,\delta) = \left(1-(1-x)^{\delta}\right) \mathbb{I}(0 \le x < 1) + \mathbb{I}(x \ge 1)\,,\\
&\textbf{(II)} \quad F_2(x,\delta) = \left(2^{\delta-1}x^{\delta}\right) \mathbb{I}(0 \le x < 1/2) + \left(1-2^{\delta-1}(1-x)^{\delta}\right) \mathbb{I}(1/2 \le x < 1)\\
&\qquad\qquad + \mathbb{I}(x \ge 1)\,,\\
&\textbf{(III)} \quad F_3(x,\delta) = \left(1/2-2^{\delta-1}(1/2-x)^{\delta}\right) \mathbb{I}(0 \le x < 1/2)\\
&\qquad\qquad + \left(1/2+2^{\delta-1}(x-1/2)^{\delta}\right) \mathbb{I}(1/2 \le x < 1) + \mathbb{I}(x \ge 1)\,.
\end{aligned}$$

That is,

$$\begin{aligned}
H_a^I &: \quad F = F_1(x,\delta) \ \text{ with } \delta > 1\,,\\
H_a^{II} &: \quad F = F_2(x,\delta) \ \text{ with } \delta > 1\,,\\
H_a^{III} &: \quad F = F_3(x,\delta) \ \text{ with } \delta > 1\,.
\end{aligned}$$

The corresponding pdfs for the distributions in the alternatives are

$$\begin{aligned}
&\textbf{(I)} \quad f_1(x) = \left(\delta(1-x)^{\delta-1}\right) \mathbb{I}(0<x<1)\,,\\
&\textbf{(II)} \quad f_2(x) = \left(\delta\, 2^{\delta-1}x^{\delta-1}\right) \mathbb{I}(0<x\le 1/2) + \left(\delta\, 2^{\delta-1}(1-x)^{\delta-1}\right) \mathbb{I}(1/2<x<1)\,,\\
&\textbf{(III)} \quad f_3(x) = \left(\delta(1-2x)^{\delta-1}\right) \mathbb{I}(0<x\le 1/2) + \left(\delta(2x-1)^{\delta-1}\right) \mathbb{I}(1/2<x<1)\,.
\end{aligned}$$

See Figure 6 for sample plots of the pdfs with various parameters under types I–III alternatives. Notice that in all the alternatives, $\delta = 1$ corresponds to $\mathcal{U}(0,1)$ distribution. Under type I alternatives, with increasing $\delta > 1$, the pdf of the distribution is more clustered around 0 and less clustered around 1; under type II alternatives, with increasing $\delta > 1$, the pdf of the distribution gets more clustered around 1/2 (and less clustered around the end points, 0 and 1); and under type III alternatives, with increasing $\delta > 1$, the pdf of the distribution is more clustered around the end points, 0 and 1, and less clustered around 1/2. Under the type II and III alternatives, the pdfs are symmetric around 1/2.

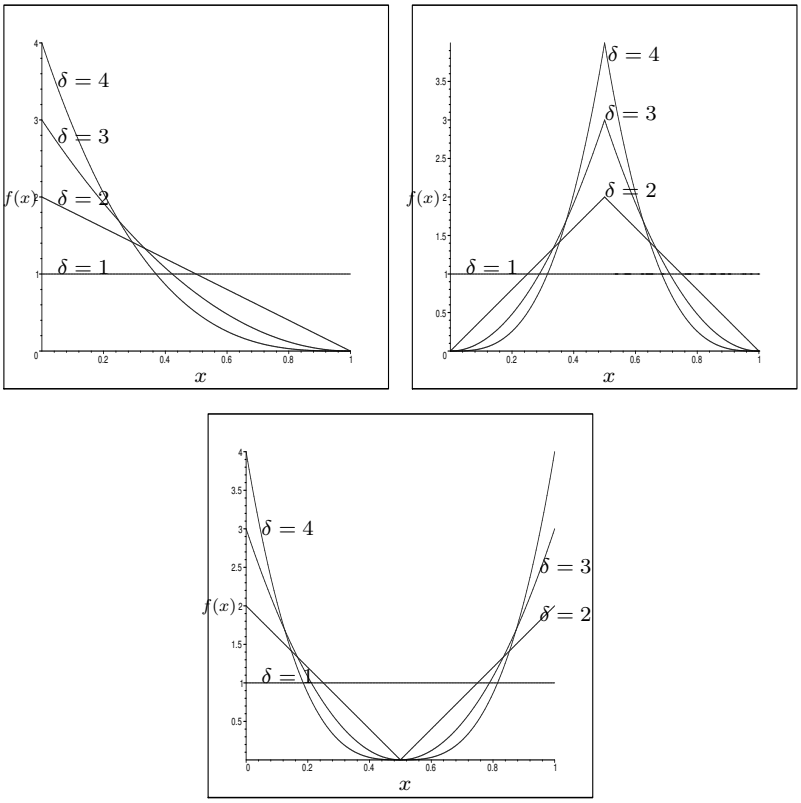

Figure 6: Sample plots for the pdfs of the alternative types I (top left), II (top right), and III (bottom) with $\delta = 2, 3, 4$. The horizontal line at 1 indicates the pdf for $\mathcal{U}(0,1)$ distribution (with $\delta = 1$).

We generate $n = 20$ points according to the specified alternatives with various parameters. In particular, for each of H_a^I–H_a^{III}, we consider $\delta = 2, 3, 4$. With CS-ICDs, we use (τ, c) for $\tau = .01, .02, ..., 10.00$ and $c = .01, .02...., .99$ and with PE-ICDs, we use (r, c) for $r = 1.01, .02, ..., 10.00$ and $c = .01, .02...., .99$. With CS-ICDs, for each (τ, c) and δ combination, and with PE-ICDs, for each (r, c) and δ combination, we replicate the sample generation $N_{mc} = 10000$ times. We compute the power using the asymptotic critical values based on the normal approximation.

Table 1: The maximum power estimates for the one-sided alternatives unadjusted (the first entry) and adjusted (the second entry) for size. In the size adjusted version, only the parameter combinations at which the tests have appropriate level are kept. RS: right-sided, LS: left-sided alternatives.

alternative	CS-ICD		PE-ICD	
	RS	LS	RS	LS
H_a^I	0.86, .73	.65, .30	.93, .75	.41, .41
H_a^{II}	0.93, .90	.29, .00	.91, .90	.60, .00
H_a^{III}	0.41, .18	.81, .81	.27, .13	.81, .81

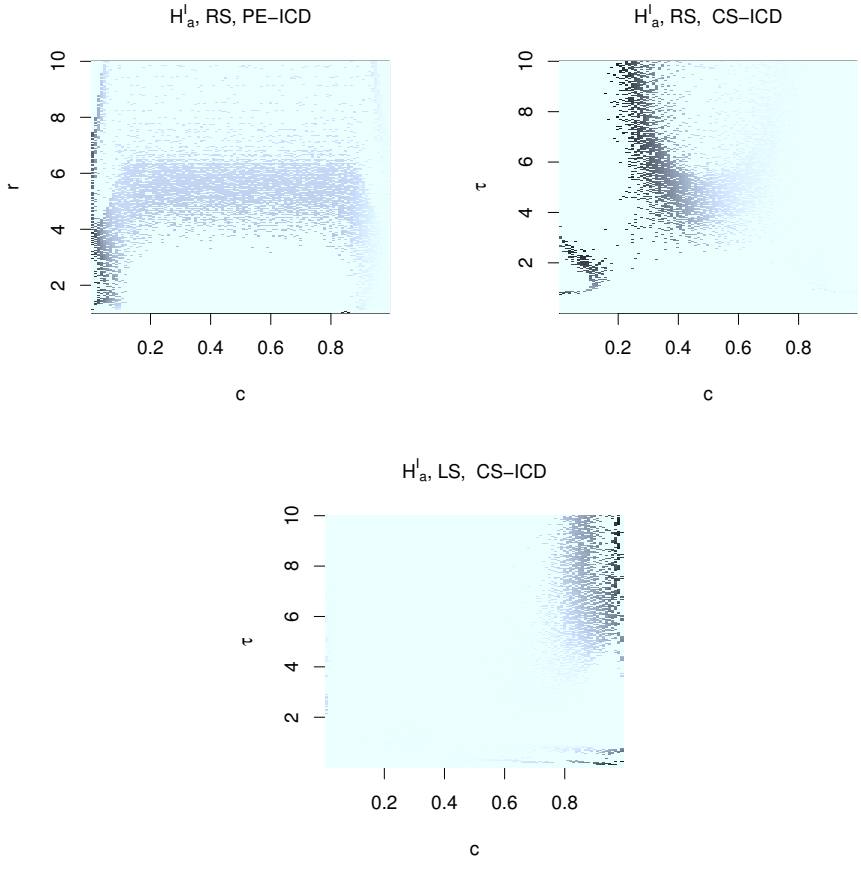

Figure 7: Image plots for the power estimates for PE-ICD with $r \in (1, 10)$ and $c \in (0, 1)$ and CS-ICD with $\tau \in (0, 10)$ and $c \in (0, 1)$ under $H_a^I : \delta = 2$, with $n = 20$, $N_{mc} = 10000$. The intensity of the gray level increases as the power increases, and the same darkness scale is used for each image plot. RS stands for right-sided, LS stands for left-sided alternatives.

We only keep the parameter combinations ((r, c) for PE-ICDs and (τ, c) for CS-ICDs) at which the tests have the appropriate level (of .05), i.e., if the test is conservative or liberal for the one-sided version in question, we ignore that parameter combination in our power estimation, as they would yield unreliable results. We call this procedure the "size adjustment" for the power estimation. The maximum values of the power estimates under the one-sided alternatives adjusted and unadjusted for the correct size are provided in Table 1. Observe that the size adjustment has a substantial effect on the highest power values (and tends to reduce the highest power estimates). Furthermore, under the alternatives H_a^I and H_a^{II}, the ICDs yield higher power for the right-sided alternative, while under H_a^{III} the ICDs yield higher power for the left-sided alternative. In particular, PE-ICDs have high power for the right-sided alternative under H_a^I and H_a^{II}, and left-sided alternative under H_a^{III} with virtually zero power for the opposite direction under these alternatives. On the other hand, CS-ICDs tend to have a

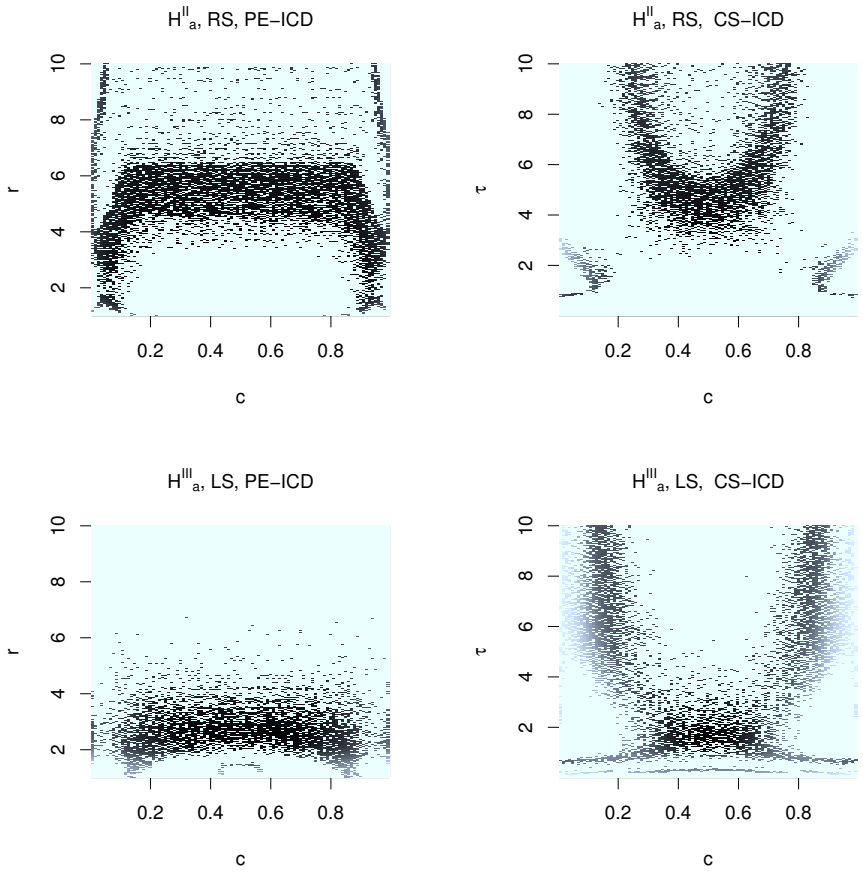

Figure 8: Image plots for the power estimates for PE-ICD with $r \in (1, 10)$ and $c \in (0, 1)$ and CS-ICD with $\tau \in (0, 10)$ and $c \in (0, 1)$ under H_a^{II} and H_a^{III} with $\delta = 2$, $n = 20$, $N_{mc} = 10000$. The gray level intensity and alternative labeling are as in Figure 7.

similar trend, but the power estimates for the direction of the one-sided version depends on the parameters. That is, e.g., under H_a^I, CS-PCD has high power estimates for the right-sided alternative at some (τ, c) combinations, and for the left-sided alternative at some other (τ, c) values. The gray-scale image plots of the power estimates under H_a^I are presented in Figure 7 and under H_a^{II} and H_a^{III} in Figure 8 (with the higher power estimates are represented with darker gray level). Notice that the power estimates are symmetric around $c = 1/2$ under H_a^{II} and H_a^{III}, which is in agreement with the symmetry in the corresponding pdfs (around $c = 1/2$).

The maximum power estimates and at which parameters of the ICDs they occur are presented in Table 2. We also plot the histograms of the power estimates (normalized to have unit area) under the alternatives in Figure 9. Under H_a^I, although the maximum power estimate for the right-sided alternative is attained by PE-ICD test at $(r, c) = (1.02, .78)$, the CS-ICD test tends to have higher power estimates. Among the competitors, the power estimate is .50 for Neyman's smooth test and .82 for KS test (with the right-sided alternative), and the ICD tests have lower power compared to KS-test. Likewise, under H_a^{II}, although the maximum power estimate for the right-sided alternative is attained by CS-ICD at $(\tau, c) = (1.96, .49)$, the PE-ICD test tends to have higher power estimates. Among the competitors, the power estimate is .39 for Neyman's test and .14 for KS test (with the right-sided alternative), and the ICD tests have higher power compared to Neyman's test. Finally, under H_a^{III}, the PE-ICD test tends to have higher power estimates. Among the competitors, the power estimate is .59

Table 2: The maximum power estimates and the parameter combinations at which they occur. RS: right-sided, LS: left-sided alternatives and $\widehat{\beta}$ stands for empirical power estimates.

For CS-ICDs						
	H_a^I		H_a^{II}		H_a^{III}	
	RS	LS	RS	LS	RS	LS
$\widehat{\beta}$	0.65–.73	.20–.29	.85–.90	—	.15–.18	.75–.80
τ	(7,9)	(6.5,10)	(2.75,4)	—	(2.5,3)	(1,2.5)
c	$\approx .2$	(.96,1)	(.35,.65)	—	$(0, .04) \cup (.96, 1)$	(.4,.6)

For PE-ICDs with RS alternatives			
	H_a^I	H_a^{II}	H_a^{III}
$\widehat{\beta}$	0.65–.75	.88–.89	.80–.81
r	≈ 1	≈ 3.8	≈ 2.5
c	$\approx .86$	(.2,.8)	(.33,.67)

for Neyman's test and .23 for KS test (with the left-sided alternative), and the ICD tests have higher power compared to Neyman's test for most parameter combinations.

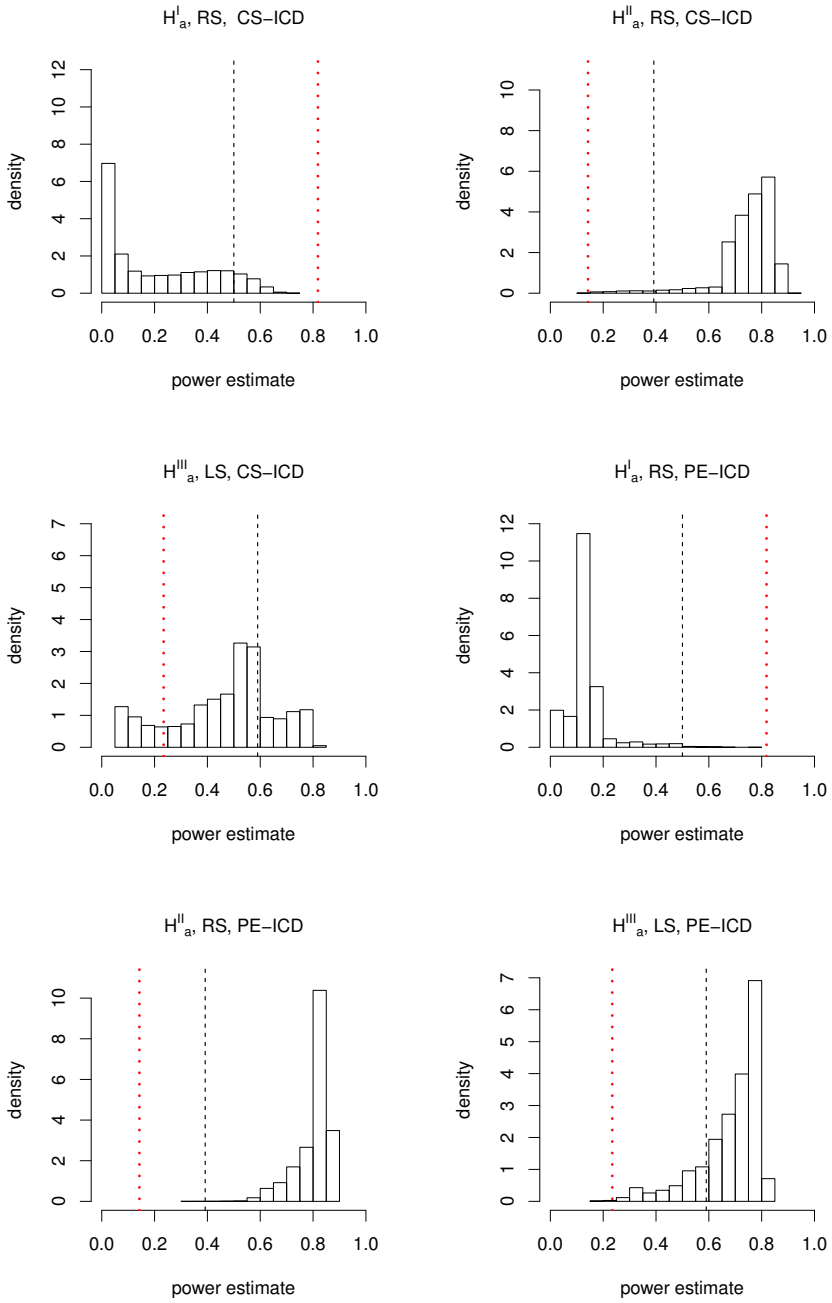

Figure 9: Histograms of the power estimates the alternatives H_a^I–H_a^{III} for the appropriate one-sided alternatives for CS-ICDs and PE-ICDs. The vertical lines are the power estimates for KS (dotted lines) and Neyman's tests (dashed lines). LS: left-sided, RS: right-sided alternatives.

5.3. Consistency of the tests based on the density of ICDs

Any reasonable test should have higher power under the alternatives as the sample size increases, and this property is reflected in the concept of *consistency*. We will prove consistency of the tests under the alternatives based on the arc density of ICDs in a general framework, and then extend the results to our alternative types for certain parameter combinations. Let $H_o: F = \mathcal{U}(0,1)$ and the alternative $H_a: F \neq \mathcal{U}(0,1)$ is parameterized by δ so that $\delta = \delta_o$ corresponds to the null hypothesis and with increasing $\delta > \delta_o$, arc probability tends to increase or decrease.

Theorem 5.1 (Consistency). *Let $\rho_{n,2}(\delta)$ be the arc density of the ICD based on data from F parameterized by δ and $p_a(\delta)$ be the corresponding arc probability. Moreover, suppose $4\nu(\delta)$ be the covariance term $\mathbf{Cov}[h_{12}, h_{13}]$. If the arc probability increases as δ increases (resp. decreases), the test against $H_a: F \neq \mathcal{U}(0,1)$ which rejects for $R_n > z_{1-\alpha}$ (resp. for $R_n < z_\alpha$) are consistent.*

Proof: Under H_o (i.e., for $\mathcal{X}_n$ being a random sample from $\mathcal{U}(0,1)$), the arc density is $\rho_{n,2}(\delta_o)$, and arc probability is $p_a(\delta_o)$ and $\mathbf{Cov}(h_{12}, h_{13})$ is $\nu(\delta_o)$. Similarly, under H_a (i.e., for $\mathcal{X}_n$ being a random sample from F) these quantities are denoted similarly with δ_o being replaced with δ. Suppose arc probability increases as $\delta > \delta_o$ increases. Then $p_a(\delta) > p_a(\delta_o)$ and the asymptotic variances $4\nu(\delta_o)/n$ and $4\nu(\delta)/n$ tends to zero as $n \to \infty$. As standardized arc density, R_n, tends to standard normal distribution or is degenerate with unit mass at $p_a(\delta_o)$ or $p_a(\delta)$ under both null and alternative hypotheses, respectively, the power under H_a tends to 1 as n goes to infinity, and hence consistency follows. The consistency for the alternative under which arc probability increases as δ decreases is similar. □

The alternatives, H_a^I–H_a^{III}, are parameterized with δ so that $\delta_o = 1$. Under H_a^I and H_a^{II} with $F = F_i(x, \delta)$ for $i = 1, 2$ the test based on CS-ICD and PE-ICD which rejects for $R_n > z_{1-\alpha}$ is consistent for most of the parameter combinations. In particular, let $\rho_{n,2}(F, \tau, c)$ be the arc density, $p_a(F, \tau, c)$ and $\nu(F, \tau, c)$ be the arc probability and $\mathbf{Cov}(h_{12}, h_{13})$ for CS-ICD with $\mathcal{X}_n$ being a random sample from F. Then under H_a^I, $p_a(F_1, \tau, c) > p_a(\mathcal{U}, \tau, c)$ for $c \leq 1/2$, since under F_1, X_i are more likely to be in $(0, 1/2)$ and hence more likely to be closer to c and hence the $N(\tau, c)$ regions are more likely to be larger which implies higher arc probability compared to the null case. Moreover, for c closer to 1 and large τ (say $\tau \geq 5$), under F_1, the $N(\tau, c)$ regions are more likely to be smaller which implies lower arc probability compared to the null case. Under H_a^{II}, $p_a(F_1, \tau, c) > p_a(\mathcal{U}, \tau, c)$ for all c away from 0 and 1 and $\tau > 0$, since under F_1, X_i are more likely to be

closer to 1/2 and hence the $N(\tau, c)$ regions are more likely to be larger which implies higher arc probability compared to the null case. Moreover, for c closer to 1 and large τ (say $\tau \geq 5$), under F_1, the $N(\tau, c)$ regions are more likely to be smaller which implies lower arc probability compared to the null case. Similarly, under H_a^{II}, $p_a(F_2, \tau, c) < p_a(\mathcal{U}, \tau, c)$ for all c away from 0 and 1 and $\tau > 0$. Hence consistency follows for these one-sided tests for such parameter combinations. In fact, with careful bookkeeping one can determine the parameter ranges for which consistency holds for each of the one-sided alternatives. For example, under $H_a^I : \delta = 2$ with $c \in (0, 1)$, for $\tau \in (0, 1)$, $p_a(F_1, \tau, c) >$ (resp. $<$) $p_a(\mathcal{U}, \tau, c)$ for $\tau >$ (resp. $<$) $\frac{2c^2-6c+3}{(2c-1)(2c-3)}$, hence consistency for the right-sided (resp. left-sided) alternative follows; likewise, for $\tau > 1$, $p_a(F_1, \tau, c) >$ (resp. $<$) $p_a(\mathcal{U}, \tau, c)$ for $\tau <$ (resp. $>$) $\frac{2c+1-\sqrt{3}}{2c-3+\sqrt{3}}$, hence consistency for the right-sided (resp. left-sided) alternative follows. The corresponding three dimensional figure to illustrate these regions of consistency for the one-sided alternatives are plotted in Figure 10 in the Appendix. Under $H_a^{II} : \delta = 2$, $p_a(F_2, \tau, c) > p_a(\mathcal{U}, \tau, c)$ for all $c \in (0, 1)$ and $\tau > 0$, hence consistency for the right-sided alternative follows; and under $H_a^{III} : \delta = 2$, $p_a(F_3, \tau, c) < p_a(\mathcal{U}, \tau, c)$ for all $c \in (0, 1)$ and $\tau > 0$, hence consistency for the left-sided alternative follows. The actual ranges of (τ, c) for the one-sided alternatives with other specific δ values can also be determined by careful calculations.

Similarly, let $\rho_{n,2}^{PE}(F, r, c)$ be the arc density, $p_a^{PE}(F, r, c)$ and $\nu_{PE}(F, r, c)$ be the arc probability and $\mathbf{Cov}(h_{12}, h_{13})$ for PE-ICD with $\mathcal{X}_n$ being a random sample from F, respectively. Then under H_a^I, $p_a^{PE}(F_1, r, c) > p_a^{PE}(\mathcal{U}, r, c)$ for c close to 0, since for F_1, X_i are more likely to be around 0 and hence the $N_{PE}(r, c)$ regions are more likely to be larger which implies higher arc probability compared to the null case. Under H_a^{II} (resp. H_a^{III}), $p_a^{PE}(F_2, r, c) >$ (resp. $<$) $p_a^{PE}(\mathcal{U}, r, c)$ for c around 1/2, since for F_2 (resp. F_3), X_i are more likely to be closer to 1/2 (resp. 0 and 1) and hence the $N_{PE}(r, c)$ regions are more likely to be larger (resp. smaller) which implies higher (resp. lower) arc probability compared to the null case. Hence consistency follows for the right-sided (resp. left-sided) tests for such parameter combinations. In fact, under $H_a^I : \delta = 2$:

- With $c \in (0, 1/2)$:
  - $p_a^{PE}(F_1, r, c) >$ (resp. $<$) $p_a^{PE}(\mathcal{U}, r, c)$, for $1 < r < 1/(1-c)$ and $r <$ (resp. $>$) $\frac{8c^3-6c^2-6c+3}{6c^4-16c^3+18c^2-12c+3}$;
  - $p_a^{PE}(F_1, r, c) >$ (resp. $<$) $p_a^{PE}(\mathcal{U}, r, c)$, for $1/(1-c) < r < 1/c$ and (r, c) is in (resp. outside) the region bounded by $r = 1/(1-c)$ and the implicit curve $3r^4c^4 - 4r^4c^3 - 8r^3c^3 + 9r^3c^2 + 6r^2c^2 - 6cr^2 + 3r - 3 = 0$;
  - $p_a^{PE}(F_1, r, c) > p_a^{PE}(\mathcal{U}, r, c)$ for $r \geq 1/c$.

  Hence consistency for the right-sided (resp. left-sided) alternative follows for these parameter values.

- With $c \in (1/2, 1)$:
  - $p_a^{PE}(F_1, r, c) > p_a^{PE}(\mathcal{U}, r, c)$ for $1 < r < 1/c$ hence consistency for the right-sided alternative follows;
  - $p_a^{PE}(F_1, r, c) >$ (resp. $<$) $p_a^{PE}(\mathcal{U}, r, c)$, for $1/c < r < 1/(1-c)$ and (r, c) is in (resp. outside) the region bounded by $c = 1$ and the implicit curve $3\,r^4c^4 - 12\,r^4c^3 + 18\,c^2r^4 - 3\,r^3c^2 - 12\,cr^4 - 6\,r^2c^2 + 6\,cr^3 + 3\,r^4 + 6\,cr^2 - 3\,r^3 - r + 1 = 0$;
  - $p_a^{PE}(F_1, r, c) > p_a^{PE}(\mathcal{U}, r, c)$ for $r \geq 1/(1-c)$.

  Hence consistency for the right-sided (resp. left-sided) alternative follows for these parameter values.

These regions of consistency for the one-sided alternatives are plotted in Figure 11 in the Appendix. Under $H_a^{II}: \delta = 2$, $p_a^{PE}(F_2, r, c) > p_a^{PE}(\mathcal{U}, r, c)$ for all $c \in (0, 1)$ and $r > 1$, hence consistency for the right-sided alternative follows; and under $H_a^{III}: \delta = 2$, $p_a^{PE}(F_3, r, c) < p_a^{PE}(\mathcal{U}, r, c)$ for all $c \in (0, 1)$ and $r > 1$ (except (r, c) inside a region that is part of $[1, 1.4] \times ([9.98, 1] \cup [0, .02])$ where the inequality is reversed), hence consistency for the left-sided (resp. right-sided) alternative follows. These regions of consistency for the one-sided alternatives are presented in Figure 12 in the Appendix. The actual ranges of (r, c) for the one-sided alternatives with other specific δ values can also be determined by careful calculations.

5.4. Extension of the methodology to test non-uniform distributions

We can modify the CS-ICD approach to test any distribution in a bounded interval in $\mathbb{R}$. Since any bounded interval (a, b) with $a < b$ can be mapped to $(0, 1)$, we can assume the support for the distribution in question to be $(0, 1)$. First we prove the below result which is instrumental for this purpose.

Proposition 5.1. *Let X_i be iid from an absolutely continuous distribution F with support $(0, 1)$ and let $\mathcal{X}_n = \{X_1, X_2, ..., X_n\}$. Define the proximity map $N_F(x, \tau, c) := F^{-1}\big(N\big(F(x), \tau, c\big)\big)$. More specifically for $\tau \in (0, 1]$,*

$$N_F(x, \tau, c) = \tag{5.2}$$

$$= \begin{cases} \left(F^{-1}\big((1-\tau)F(x)\big), F^{-1}\Big(\Big(1+\frac{(1-c)}{c}\tau\Big)F(x)\Big)\right) & \text{if } x \in \big(0, F^{-1}(c)\big), \\ \left(F^{-1}\Big(F(x) - \frac{c\,\tau}{(1-c)}\,\big(1-F(x)\big)\Big), F^{-1}\Big(F(x) + \big(1-F(x)\big)\,\tau\Big)\right) & \text{if } x \in \big(F^{-1}(c), 1\big), \end{cases}$$

and for $\tau > 1$,

(5.3) $N_F(x,\tau,c) =$

$$= \begin{cases} \left(0, F^{-1}\Big(\Big(1+\frac{(1-c)}{c}\,\tau\Big)F(x)\Big)\right) & \text{if } x \in \left(0, F^{-1}\big(\frac{c}{c+(1-c)\,\tau}\big)\right), \\ (0,1) & \text{if } x \in \left(F^{-1}\big(\frac{c}{c+(1-c)\,\tau}\big), F^{-1}\big(\frac{c\,\tau}{1-c+c\,\tau}\big)\right), \\ \left(F^{-1}\Big(F(x)-\frac{c\,\tau}{(1-c)}\,\big(1-F(x)\big)\Big), 1\right) & \text{if } x \in \left(F^{-1}\big(\frac{c\,\tau}{1-c+c\,\tau}\big), 1\right). \end{cases}$$

Then the arc density of the ICD based on N_F *and* $\mathcal{X}_n$ *has the same distribution as* $\rho_{n,2}(\mathcal{U},\tau,c)$ *(provided in Theorem 4.3).*

Proof: Let $U_i := F(X_i)$ for $i = 1,2,...,n$ and $\mathcal{U}_n := \{U_1, U_2, ..., U_n\}$. Hence, by probability integral transform, $U_i \overset{iid}{\sim} \mathcal{U}(0,1)$. So the image of $N_F(x,r,c)$ under F is $F\big(N_F(x,r,c)\big) = N\big(F(x),r,c\big)$ for (almost) all $x \in (0,1)$. Then $F\big(N_F(X_i,r,c)\big) = N\big(F(X_i),r,c\big) = N(U_i,r,c)$ for $i = 1,2,...,n$. Since $U_i \overset{iid}{\sim} \mathcal{U}(0,1)$, the distribution of the arc density of the ICD based on $N(\cdot,\tau,c)$ and $\mathcal{U}_n$ is given in Theorem 4.3. Observe that for any j, $X_j \in N_F(X_i,\tau,c)$ iff $X_j \in F^{-1}\big(N\big(F(X_i),\tau,c\big)\big)$ iff $F(X_j) \in N\big(F(X_i),\tau,c\big)$ iff $U_j \in N(U_i,\tau,c)$ for $i = 1,2,...,n$. Hence the desired result follows. □

A similar construction is available for the PE-ICDs.

In Proposition 5.1, we have shown that if the defining proximity region for our ICD is defined as $N_F(x,\tau,c) := F^{-1}\big(N\big(F(x),\tau,c\big)\big)$ where F is an increasing function in (a,b) with $a < b$, the exact (and asymptotic) distribution of the arc density based on the ICD for N_F is the same as $\rho_{n,2}(\mathcal{U},\tau,c)$. Hence we can test whether the distribution of any data set is from F or not with the methodology proposed in this article. For example, to test a "data set is from $F(x) = x^2$ with $\mathcal{S}(F) = (0,1)$" (so the inverse is $F^{-1}(x) = \sqrt{x}$ and the corresponding pdf is $f(x) = 2x\,\mathbb{I}(0 < x < 1)$), we need to compute the arc density for the ICD based on the following proximity region: For $\tau \in (0,1]$,

(5.4)

$$N_F(x,\tau,c) = \begin{cases} \left(x\sqrt{1-\tau},\ x\sqrt{1+\frac{(1-c)}{c}\,\tau}\right) & \text{if } x \in \left(0,\sqrt{c}\right), \\ \left(\sqrt{x^2-\frac{c\,\tau}{(1-c)}\,(1-x^2)},\ \sqrt{x^2+(1-x^2)\,\tau}\right) & \text{if } x \in \left(\sqrt{c},1\right), \end{cases}$$

and for $\tau > 1$,

(5.5)

$$N_F(x,\tau,c) = \begin{cases} \left(0,\ x\sqrt{1+\frac{(1-c)}{c}\,\tau}\right) & \text{if } x \in \left(0,\sqrt{\frac{c}{c+(1-c)\,\tau}}\right), \\ (0,1) & \text{if } x \in \left(\sqrt{\frac{c}{c+(1-c)\,\tau}},\sqrt{\frac{c\,\tau}{1-c+c\,\tau}}\right), \\ \left(\sqrt{x^2-\frac{c\,\tau}{(1-c)}\,(1-x^2)},\ 1\right) & \text{if } x \in \left(\sqrt{\frac{c\,\tau}{1-c+c\,\tau}},1\right). \end{cases}$$

Then the arc density for the ICD based on $N_F(\cdot, \tau, c)$ will have the same distribution as $\rho_{n,2}(\mathcal{U}, \tau, c)$ and hence can be used for testing data is from F or not with the procedure discussed in Section 5.

6. DISCUSSION AND CONCLUSIONS

We consider the central similarity interval catch digraphs (CS-ICDs) based on one dimensional data. The CS-ICDs are defined with two parameters: an expansion parameter $\tau > 0$ and a centrality parameter $c \in (0, 1)$. We study the arc density of CS-ICDs, and using its U-statistics property, we derive its asymptotic (normal) distribution for uniform data for the (interiors of) entire ranges of τ and c. Along this process, we also determine the parameters τ and c for which the rate of convergence to normality is the fastest. We also consider the arc density of proportional-edge ICD (PE-ICD) for comparative purposes. We demonstrate that convergence rate of arc density of CS-ICDs is faster than that of PE-PCDs at their respective optimal parameters, which implies that distribution of arc density of CS-ICDs is closer to normality at smaller sample sizes, compared to the arc density of PE-PCDs.

We use the arc density of the ICDs for testing uniformity (i.e., for testing H_o "data set is a random sample from $\mathcal{U}(0, 1)$"), and show that under type I alternatives in which pdf of the data points is larger around one of the end points (0 or 1) CS-ICD test has higher power compared to PE-ICD test, but under the types II and III alternatives in which pdf is larger around 1/2 or around both end points, then PE-ICD test tends to have higher power compared to CS-ICDs. We also compare the ICD tests with two well known tests in literature (namely, Kolmogorov–Smirnov (KS) test and Neyman's smooth test of uniformity). Under type I alternatives, KS test tends to have higher power compared to the ICD tests and Neyman's smooth test, Neyman's smooth test has higher power compared to PE-ICD test, but lower power compared to CS-ICD tests for some parameter combinations. Under type II (resp. type III) alternatives, ICD tests have higher power than KS and Neyman's smooth test for almost all (resp. most) parameter values which have appropriate size. The recommended parameter combinations for the ICDs are provided in Table 2.

The CS-ICDs for one dimensional data can also be used in testing spatial interaction between multiple classes whose support is one-dimensional (see Remark 4.2). The arc density approach is easily adaptable to testing nonuniform distributions as well (see Section 5.4 for more detail). Furthermore, the study of arc density of CS-ICDs in the one dimensional case will provide insight for and form the foundation of related catch digraph extensions in higher dimensions.

APPENDIX

A. PRELIMINARIES

In the proofs below, we can, without loss of generality, assume that the support of the uniform distribution is $(0, 1)$ based on Theorem 4.1.

A.1. Proof of Theorem 4.3

There are two cases for τ, namely $0 < \tau \leq 1$ and $\tau > 1$.

For $\tau \in (0, 1]$, the proximity region is defined as in Equation (4.2) and the Γ_1-region is

$$\Gamma_1(x, \tau, c) = \begin{cases} \left(\frac{cx}{c+(1-c)\tau}, \frac{x}{1-\tau}\right) & \text{if } x \in \left(0, c(1-\tau)\right], \\ \left(\frac{cx}{c+(1-c)\tau}, \frac{(1-c)x+c\tau}{1-c+c\tau}\right) & \text{if } x \in \left(c(1-\tau), c(1-\tau)+\tau\right], \\ \left(\frac{x-\tau}{1-\tau}, \frac{(1-c)x+c\tau}{1-c+c\tau}\right) & \text{if } x \in \left(c(1-\tau)+\tau, 1\right). \end{cases} \tag{A.1}$$

For $\tau > 1$, the proximity region is as in Equation (4.3) and the Γ_1-region is

$$\Gamma_1(x, \tau, c) = \left(\frac{cx}{c+(1-c)\tau}, \frac{(1-c)x+c\tau}{1-c+c\tau}\right). \tag{A.2}$$

Case 1: $0 < \tau \leq 1$: In this case depending on the location of x_1, the following are the different types of the combinations of $N(x_1, \tau, c)$ and $\Gamma_1(x_1, \tau, c)$. Let

$$a_1 := (1-\tau)x_1, \qquad a_2 := x_1\left(1 + \frac{(1-c)\tau}{c}\right),$$

$$a_3 := x_1 - \frac{c\tau(1-x_1)}{1-c}, \qquad a_4 := x_1 + (1-x_1)\tau,$$

and

$$g_1 := \frac{cx_1}{c+(1-c)\tau}, \qquad g_2 := \frac{x_1}{1-\tau},$$

$$g_3 := \frac{x_1-\tau}{1-\tau}, \qquad g_4 := \frac{x_1(1-c)+c\tau}{1-c+c\tau}.$$

Then

(**i**) for $0 < x_1 \leq c(1-\tau)$, we have $N(x_1, \tau, c) = (a_1, a_2)$ and $\Gamma_1(x_1, \tau, c) = (g_1, g_2)$,

(**ii**) for $c\,(1-\tau) < x_1 \le c$, we have $N(x_1,\tau,c) = (a_1,a_2)$ and $\Gamma_1(x_1,\tau,c) = (g_1,g_4)$,

(**iii**) for $c < x_1 \le c\,(1-\tau)+\tau$, we have $N(x_1,\tau,c) = (a_3,a_4)$ and $\Gamma_1(x_1,\tau,c) = (g_1,g_4)$,

(**iv**) for $c\,(1-\tau)+\tau < x_1 < 1$, we have $N(x_1,\tau,c) = (a_3,a_4)$ and $\Gamma_1(x_1,\tau,c) = (g_3,g_4)$.

Then

$$p_a(\tau,c) \;=\; P\Big(X_2 \in N(X_1,\tau,c)\Big) \;=\; \int_0^c (a_2-a_1)\,dx_1 + \int_c^1 (a_4-a_3)\,dx_1 \;=\; \tau/2\,.$$

For $\mathbf{Cov}(h_{12},h_{13})$, we need to calculate P_{2N}, P_{NG}, and P_{2G}.

$$\begin{aligned} P_{2N} \;&=\; P\Big(\{X_2,X_3\} \subset N(X_1,\tau,c)\Big) \\ &=\; \int_0^c (a_2-a_1)^2\,dx_1 + \int_c^1 (a_4-a_3)^2\,dx_1 \;=\; \tau^2/3\,. \end{aligned}$$

$$\begin{aligned} P_{NG} \;&=\; P\Big(X_2 \in N(X_1,\tau,c),\, X_3 \in \Gamma_1(X_1,\tau,c)\Big) \\ &=\; \int_0^{c\,(1-\tau)} (a_2-a_1)\,(g_2-g_1)\,dx_1 + \int_{c\,(1-\tau)}^{c} (a_2-a_1)\,(g_4-g_1)\,dx_1 \\ &\quad + \int_c^{c\,(1-\tau)+\tau} (a_4-a_3)\,(g_4-g_1)\,dx_1 + \int_{c\,(1-\tau)+\tau}^{1} (a_4-a_3)\,(g_4-g_3)\,dx_1 \\ &=\; \frac{\tau^2\left(c^2\tau^3 - 5\,c^2\tau^2 - c\,\tau^3 + 4\,c^2\tau + 5\,c\,\tau^2 - 2\,c^2 - 4\,c\,\tau - \tau^2 + 2\,c + 2\,\tau\right)}{6\;(c\,\tau - c + 1)\,(c+\tau-c\,\tau)}\,. \end{aligned}$$

Finally,

$$\begin{aligned} P_{2G} \;&=\; P\Big(\{X_2,X_3\} \subset \Gamma_1(X_1,\tau,c)\Big) \\ &=\; \int_0^{c\,(1-\tau)} (g_2-g_1)^2\,dx_1 + \int_{c\,(1-\tau)}^{c\,(1-\tau)+\tau} (g_4-g_1)^2\,dx_1 + \int_{c\,(1-\tau)+\tau}^{1} (g_4-g_3)^2\,dx_1 \\ &=\; \frac{\left(2\,c^2\tau - c^2 - 2\,c\,\tau + c + \tau\right)\tau^2}{3\;(c\,\tau - c + 1)\,(c+\tau-c\,\tau)}\,. \end{aligned}$$

Therefore

$$\begin{aligned} 4\,\mathbf{E}[h_{12}h_{13}] \;&= \\ &=\; P_{2N} + 2\,P_{NG} + P_{2G} \\ &=\; \frac{\tau^2\left(c^2\tau^3 - 6\,c^2\tau^2 - c\,\tau^3 + 8\,c^2\tau + 6\,c\,\tau^2 - 4\,c^2 - 8\,c\,\tau - \tau^2 + 4\,c + 4\,\tau\right)}{3\;(c\,\tau - c + 1)\,(c+\tau-c\,\tau)}\,. \end{aligned}$$

Hence

$$\begin{aligned} 4\,\mathbf{Cov}[h_{12},h_{13}] \;&= \\ &=\; \frac{\tau^2\left(c^2\tau^3 - 3\,c^2\tau^2 - c\,\tau^3 + 2\,c^2\tau + 3\,c\,\tau^2 - c^2 - 2\,c\,\tau - \tau^2 + c + \tau\right)}{3\;(c\,\tau - c + 1)\,(c+\tau-c\,\tau)}\,. \end{aligned}$$

Case 2: $\tau > 1$: In this case depending on the location of x_1, the following are the different types of the combinations of $N(x_1, \tau, c)$ and $\Gamma_1(x_1, \tau, c)$.

(**i**) for $0 < x_1 \leq \frac{c}{c+(1-c)\tau}$, we have $N(x_1, \tau, c) = (0, a_2)$ and $\Gamma_1(x_1, \tau, c) = (g_1, g_4)$,

(**ii**) for $\frac{c}{c+(1-c)\tau} < x_1 \leq \frac{c\tau}{1-c+c\tau}$, we have $N(x_1, \tau, c) = (0, 1)$ and $\Gamma_1(x_1, \tau, c) = (g_1, g_4)$,

(**iii**) for $\frac{c\tau}{1-c+c\tau} < x_1 < 1$, we have $N(x_1, \tau, c) = (a_3, 1)$ and $\Gamma_1(x_1, \tau, c) = (g_1, g_4)$.

Then

$$\begin{aligned} p_a(\tau, c) &= P\Big(X_2 \in N(X_1, \tau, c)\Big) \\ &= \int_0^{\frac{c}{c+(1-c)\tau}} a_2 \, dx_1 + \int_{\frac{c}{c+(1-c)\tau}}^{\frac{c\tau}{1-c+c\tau}} 1 \, dx_1 + \int_{\frac{c\tau}{1-c+c\tau}}^{1} (1 - a_3) \, dx_1 \\ &= \frac{\tau\left(2\,c^2\tau - 2\,c^2 - 2\,c\tau + 2\,c - 1\right)}{2\,(c\tau - c + 1)\,(c\tau - c - \tau)} \, . \end{aligned}$$

Next

$$\begin{aligned} P_{2N} &= P\Big(\{X_2, X_3\} \subset N(X_1, \tau, c)\Big) \\ &= \int_0^{\frac{c}{c+(1-c)\tau}} a_2^2 \, dx_1 + \int_{\frac{c}{c+(1-c)\tau}}^{\frac{c\tau}{1-c+c\tau}} 1 \, dx_1 + \int_{\frac{c\tau}{1-c+c\tau}}^{1} (1 - a_3)^2 \, dx_1 \\ &= \frac{3\,c^2\tau^2 - 2\,c^2\tau - 3\,c\tau^2 - c^2 + 2\,c\tau + c - \tau}{3\,(c\tau - c + 1)\,(c\tau - c - \tau)} \, . \end{aligned}$$

$$\begin{aligned} P_{NG} &= P\Big(X_2 \in N(X_1, \tau, c),\, X_3 \in \Gamma_1(X_1, \tau, c)\Big) \\ &= \int_0^{\frac{c}{c+(1-c)\tau}} a_2\,(g_4 - g_1) \, dx_1 + \int_{\frac{c}{c+(1-c)\tau}}^{\frac{c\tau}{1-c+c\tau}} (g_4 - g_1) \, dx_1 \\ &\quad + \int_{\frac{c\tau}{1-c+c\tau}}^{1} (1 - a_3)\,(g_4 - g_1) \, dx_1 \\ &= \Big[\tau^2\Big(6\,c^6\tau^4 - 24\,c^6\tau^3 - 18\,c^5\tau^4 + 36\,c^6\tau^2 + 72\,c^5\tau^3 \\ &\quad + 18\,c^4\tau^4 - 24\,c^6\tau - 108\,c^5\tau^2 - 84\,c^4\tau^3 - 6\,c^3\tau^4 + 6\,c^6 \\ &\quad + 72\,c^5\tau + 132\,c^4\tau^2 + 48\,c^3\tau^3 - 18\,c^5 - 92\,c^4\tau - 84\,c^3\tau^2 \\ &\quad - 12\,c^2\tau^3 + 26\,c^4 + 64\,c^3\tau + 30\,c^2\tau^2 - 22\,c^3 - 26\,c^2\tau - 6\,c\tau^2 \\ &\quad + 10\,c^2 + 6\,c\tau - 2\,c - \tau\Big)\Big] \Big/ \Big[6\,(c\tau - c + 1)^3\,(c\tau - c - \tau)^3\Big] \, . \end{aligned}$$

Finally,

$$\begin{aligned}
P_{2G} &= P\Big(\{X_2, X_3\} \subset \Gamma_1(X_1, \tau, c)\Big) \\
&= \int_0^1 (g_4 - g_1)^2 \, dx_1 \\
&= \Big[\tau^2\Big(3\,c^4\tau^2 - 6\,c^4\tau - 6\,c^3\tau^2 + 3\,c^4 + 12\,c^3\tau + 3\,c^2\tau^2 - 6\,c^3 \\
&\quad - 9\,c^2\tau + 7\,c^2 + 3\,c\tau - 4\,c + 1\Big)\Big] \Big/ \Big[3\,(c\tau - c + 1)^2\,(c\tau - c - \tau)^2\Big] .
\end{aligned}$$

Therefore

$$\begin{aligned}
4\,\mathbf{E}[h_{12}h_{13}] &= P_{2N} + 2\,P_{NG} + P_{2G} \\
&= \Big[12\,c^6\tau^6 - 50\,c^6\tau^5 - 36\,c^5\tau^6 + 79\,c^6\tau^4 + 150\,c^5\tau^5 + 36\,c^4\tau^6 \\
&\quad - 56\,c^6\tau^3 - 237\,c^5\tau^4 - 175\,c^4\tau^5 - 12\,c^3\tau^6 + 14\,c^6\tau^2 \\
&\quad + 168\,c^5\tau^3 + 297\,c^4\tau^4 + 100\,c^3\tau^5 + 2\,c^6\tau - 42\,c^5\tau^2 \\
&\quad - 220\,c^4\tau^3 - 199\,c^3\tau^4 - 25\,c^2\tau^5 - c^6 - 6\,c^5\tau + 58\,c^4\tau^2 \\
&\quad + 160\,c^3\tau^3 + 75\,c^2\tau^4 + 3\,c^5 + 7\,c^4\tau - 46\,c^3\tau^2 \\
&\quad - 70\,c^2\tau^3 - 15\,c\tau^4 - 3\,c^4 - 4\,c^3\tau + 20\,c^2\tau^2 + 18\,c\tau^3 \\
&\quad + c^3 + c^2\tau - 4\,c\tau^2 - 3\,\tau^3\Big] \Big/ \Big[3\,(c\tau - c + 1)^3\,(c\tau - c - \tau)^3\Big] .
\end{aligned}$$

Hence

$$\begin{aligned}
4\,\mathbf{Cov}[h_{12}, h_{13}] &= \Big[c\,(1-c)\,\Big(2\,c^4\tau^5 - 7\,c^4\tau^4 - 4\,c^3\tau^5 + 8\,c^4\tau^3 + 14\,c^3\tau^4 \\
&\quad + 3\,c^2\tau^5 - 2\,c^4\tau^2 - 16\,c^3\tau^3 - 7\,c^2\tau^4 - c\tau^5 - 2\,c^4\tau + 4\,c^3\tau^2 \\
&\quad + 12\,c^2\tau^3 + c^4 + 4\,c^3\tau - 6\,c^2\tau^2 - 4\,c\tau^3 - 2\,c^3 - 3\,c^2\tau + 4\,c\tau^2 \\
&\quad + c^2 + c\tau - \tau^2\Big)\Big] \Big/ \Big[3\,(c\tau - c + 1)^3\,(c\tau - c - \tau)^3\Big] . \quad \square
\end{aligned}$$

A.1.1. Special Case (i) $\tau > 0$ and $c = 1/2$

For $x \in (0, 1/2)$, the proximity region for $\tau \in (0, 1]$ is

$$N(x, \tau, 1/2) = \begin{cases} \big((1-\tau)\,x,\ (1+\tau)\,x\big) & \text{if } x \in (0, 1/2)\,, \\ \big(x - (1-x)\,\tau,\ x + (1-x)\,\tau\big) & \text{if } x \in (1/2, 1)\,, \end{cases} \tag{A.3}$$

and for $\tau > 1$

$$N(x, \tau, 1/2) = \begin{cases} \big(0, (1+\tau)\,x\big) & \text{if } x \in \big(0, 1/(1+\tau)\big)\,, \\ (0, 1) & \text{if } x \in \big(1/(1+\tau),\ \tau/(1+\tau)\big)\,, \\ \big(x - (1-x)\,\tau, 1\big) & \text{if } x \in \big(\tau/(1+\tau), 1\big)\,. \end{cases} \tag{A.4}$$

Corollary A.1. *For $\tau \in (0,\infty)$ and $c = 1/2$, we have $\sqrt{n}\big[\rho_{n,2}(\tau,1/2) - p_a(\tau,1/2)\big] \xrightarrow{\mathcal{L}} \mathbb{N}\big(0, 4\,\nu(\tau,1/2)\big)$ as $n \to \infty$, where*

$$p_a(\tau,1/2) = \begin{cases} \tau/2 & \text{if } 0 < \tau < 1\,, \\ \tau/(\tau+1) & \text{if } \tau > 1\,, \end{cases} \tag{A.5}$$

and

$$4\,\nu(\tau,1/2) = \begin{cases} \frac{\tau^2\,(1+2\tau-\tau^2-\tau^3)}{3\,(\tau+1)^2} & \text{if } 0 < \tau \le 1\,, \\ \frac{2\tau-1}{3\,(\tau+1)^2} & \text{if } \tau > 1\,. \end{cases} \tag{A.6}$$

See Figure 10 for the plots of $p_a(\tau,1/2)$ and $4\,\nu(\tau,1/2)$ with $\tau \in (0,5]$.

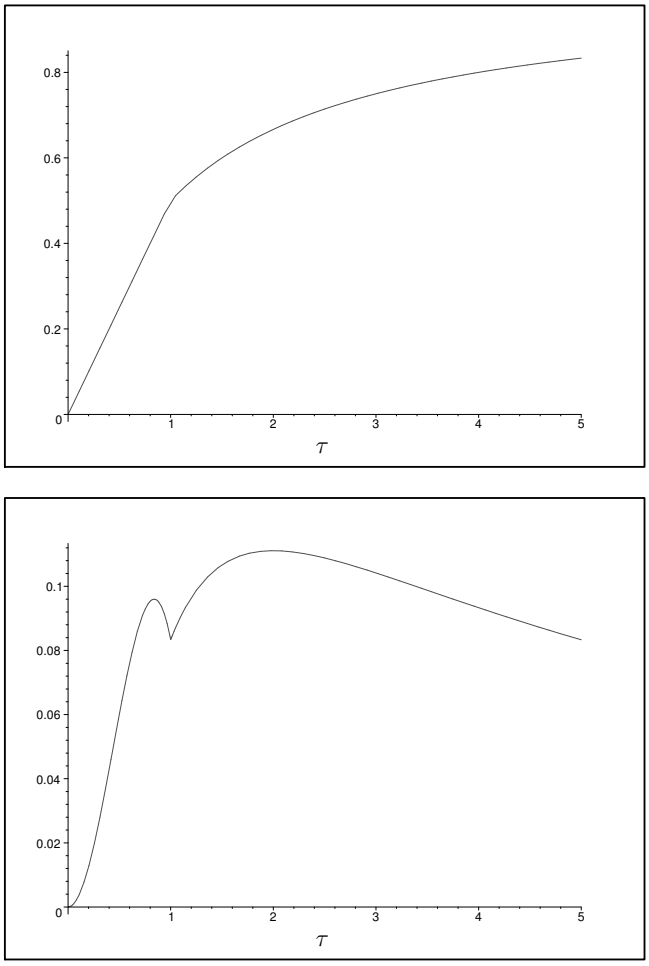

Figure 10: The plots of the asymptotic mean $p_a(\tau,1/2)$ (top) and the variance $4\,\nu(\tau,1/2)$ (bottom) as a function of τ for $\tau \in (0,5]$.

The sharpest rate of convergence in Corollary A.1 is $\frac{K}{\sqrt{n}} f^c_{CS}(\tau, 1/2)$ where

$$f^c_{CS}(\tau, 1/2) = \begin{cases} \frac{27\tau}{2}\left(\frac{(6\tau+3-3\tau^3-3\tau^2)\tau^2}{(\tau+1)^2}\right)^{-3/2} & \text{if } 0 < \tau \leq 1\,, \\ \frac{3\sqrt{3}\tau}{\tau+1}\left(\frac{2\tau-1}{(\tau+1)^2}\right)^{-3/2} & \text{if } \tau > 1\,, \end{cases} \tag{A.7}$$

and is minimized at $\tau \approx .73$ which is found by using simple calculus and numerical methods.

The plot of $p_a(\tau, 1/2)/\sqrt{\nu(\tau, 1/2)^3}$ also indicates that this is where the global minimum occurs. Convergence rates for PE- and CS-ICDs are presented in Figure 11 (bottom) for $c = 1/2$ as a function of expansion parameter. See [Ceyhan, 2012] for the explicit form of $f^c_{PE}(r, 1/2)$. Notice that at the optimal expansion parameters, convergence rate of CS-ICDs is faster with $c = 1/2$.

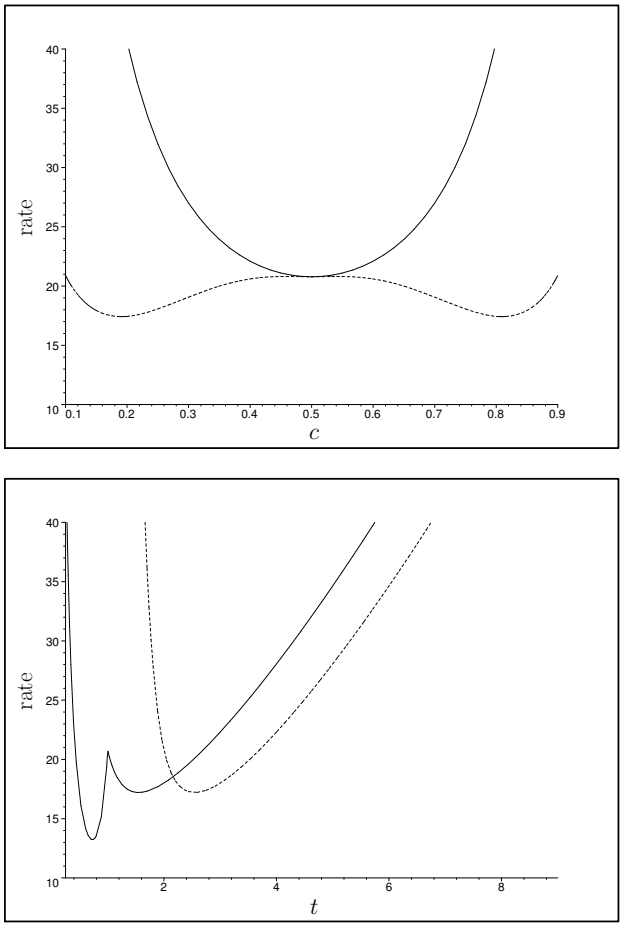

Figure 11: The plots of the rates of convergence to normality for PE- and CS-ICDs. Plotted in the top are $f^c_{CS}(1, c)$ (solid line) and $f^c_{PE}(2, c)$ (dashed line); and in the bottom are $f^c_{CS}(t, 1/2)$ (solid) and $f^c_{PE}(t, 1/2)$ (dashed).

A.1.2. Special Case (ii) $\tau = 1$ and $c \in (0,1)$

For $x \in (0,1)$, the proximity region has the following form:

$$N(x,1,c) = \begin{cases} (0, x/c) & \text{if } x \in (0,c)\,, \\ \big((x-c)/(1-c), 1\big) & \text{if } x \in (c,1)\,. \end{cases} \tag{A.8}$$

Corollary A.2. *As $n \to \infty$, for $c \in (0,1)$, we have $\sqrt{n}\,[\rho_{n,2}(1,c) - p_a(1,c)] \xrightarrow{\mathcal{L}} \mathbb{N}\big(0, 4\,\nu(1,c)\big)$, where $p_a(1,c) = 1/2$ and $4\nu(1,c) = c\,(1-c)/3$.*

Observe that $p_a(1,c)$ is constant (i.e., independent of c) and $\nu(1,c)$ is symmetric around $c = 1/2$ with $\nu(1,c) = \nu(1,1-c)$. Let $\frac{K}{\sqrt{n}}\, f^c_{CS}(\tau,c)$ be the rate of convergence to normality for CS-ICDs. Then the sharpest rate of convergence in Corollary A.2 is $\frac{K}{\sqrt{n}}\, f^c_{CS}(1,c)$ where

$$f^c_{CS}(1,c) \;=\; \frac{3\sqrt{3}}{2\sqrt{c^3\,(1-c)^3}}\,. \tag{A.9}$$

Convergence rate is minimized at $c = 1/2$ (which can be verified by simple calculus). Also, let $\frac{K}{\sqrt{n}}\, f^c_{PE}(r,c)$ be the rate of convergence to normality for PE-ICDs (see [Ceyhan, 2012] for its explicit forms). Then we have $f^c_{PE}(2,c) \le f^c_{CS}(1,c)$ for all $c \in (0,1)$ with equality holding only at $c = 1/2$ (see also Figure 11 (top)). Thus at these specific centrality parameters, convergence rate to normality is faster for PE-PCDs.

A.1.3. Special Case (iii) $\tau = 1$ and $c = 1/2$

In this case we have $N(x,1,1/2) = B(x, r(x))$ where $r(x) = \min(x, 1-x)$ for $x \in (0,1)$. Hence CS-ICD based on $N(x,1,1/2)$ is equivalent to the CCCD of [Priebe *et al.*, 2001] and the PE-ICD with expansion parameter 2 and centrality parameter 1/2 of [Ceyhan, 2012].

Corollary A.3. *As $n \to \infty$, we have $\sqrt{n}\,[\rho_n(1,1/2) - p_a(1,1/2)] \xrightarrow{\mathcal{L}} \mathbb{N}\big(0, 4\,\nu(1,1/2)\big)$, where $p_a(1,1/2) = 1/2$ and $4\nu(1,1/2) = 1/12$ with the sharpest rate of convergence being $K\frac{p_a(1,1/2)}{\sqrt{n\,\nu(1,1/2)^3}} = 12\sqrt{3}\,\frac{K}{\sqrt{n}}$.*

A.2. Proof of Theorem 4.4

There are two cases for τ, namely, $0 < \tau < 1$ and $\tau > 1$.

For $\tau \in (0,1]$ and x in the right end interval, the proximity region is

$$N_e(x,\tau) = \begin{cases} \big((1-\tau)\,x,\ (1+\tau)\,x\big) & \text{if } x \in \big(0, 1/(1+\tau)\big)\,, \\ \big((1-\tau)\,x, 1\big) & \text{if } x \in \big(1/(1+\tau), 1\big)\,, \end{cases} \tag{A.10}$$

and the Γ_1-region is

$$\Gamma_{1,e}(x,\tau) = \begin{cases} \Big(\frac{x}{1+\tau}, \frac{x}{1-\tau}\Big) & \text{if } x \in (0, 1-\tau)\,, \\ \Big(\frac{x}{1+\tau}, 1\Big) & \text{if } x \in (1-\tau, 1)\,. \end{cases} \tag{A.11}$$

For $\tau > 1$ and x in the right end interval, the proximity region is

$$N_e(x,\tau) = \begin{cases} \big(0, (1+\tau)\,x\big) & \text{if } x \in \big(0, 1/(1+\tau)\big)\,, \\ (0,1) & \text{if } x \in \big(1/(1+\tau), 1\big)\,, \end{cases} \tag{A.12}$$

and the Γ_1-region is $\Gamma_{1,e}(x,\tau) = \big(x/(1+\tau), 1\big)$.

Case 1: $0 < \tau \leq 1$: For $x_1 \in (0,1)$, depending on the location of x_1, the following are the different types of the combinations of $N_e(x_1,\tau)$ and $\Gamma_{1,e}(x_1,\tau)$.

- (**i**) for $0 < x_1 \leq 1-\tau$, we have $N_e(x_1,\tau) = \big((1-\tau)\,x_1, (1+\tau)\,x_1\big)$ and $\Gamma_{1,e}(x_1,\tau) = \big(x_1/(1+\tau), x_1/(1-\tau)\big)$,
- (**ii**) for $1-\tau < x_1 \leq 1/(1+\tau)$, we have $N_e(x_1,\tau) = \big((1-\tau)\,x_1, (1+\tau)\,x_1\big)$ and $\Gamma_{1,e}(x_1,\tau) = \big(x_1/(1+\tau), 1\big)$,
- (**iii**) for $1/(1+\tau) < x_1 < 1$, we have $N_e(x_1,\tau) = \big((1-\tau)\,x_1, 1\big)$ and $\Gamma_{1,e}(x_1,\tau) = \big(x_1/(1+\tau), 1\big)$.

Then

$$\begin{aligned} p_a^e(\tau,c) &= P\Big(X_2 \in N_e(X_1,\tau)\Big) \\ &= \int_0^{1/(1+\tau)} \big((1+\tau)\,x_1 - (1-\tau)\,x_1\big)\,dx_1 + \int_{1/(1+\tau)}^1 \big(1-(1-\tau)\,x_1\big)\,dx_1 \\ &= \int_0^{1/(1+\tau)} (2\,\tau\,x_1)\,dx_1 + \int_{1/(1+\tau)}^1 (1-x_1+x_1\,\tau)\,dx_1 \;=\; \frac{\tau\,(\tau+2)}{2\,(\tau+1)}\,. \end{aligned}$$

For $\mathbf{Cov}(h_{12}, h_{13})$, we need to calculate $P_{2N,e}$, $P_{NG,e}$, and $P_{2G,e}$.

$$\begin{aligned} P_{2N,e} &= P\Big(\{X_2, X_3\} \subset N_e(X_1,\tau)\Big) \\ &= \int_0^{1/(1+\tau)} (2\,\tau\,x_1)^2\,dx_1 + \int_{1/(1+\tau)}^1 (1-x_1+x_1\,\tau)^2\,dx_1 \\ &= \frac{\tau^2\,(\tau^2+3\,\tau+4)}{3\,(\tau+1)^2}\,. \end{aligned}$$

$$\begin{aligned} P_{NG,e} &= P\Big(X_2 \in N_e(X_1,\tau), X_3 \in \Gamma_{1,e}(X_1,\tau)\Big) \\ &= \int_0^{1-\tau} (2\,\tau\,x_1)\left(\frac{2\,\tau\,x_1}{1-\tau^2}\right) dx_1 + \int_{1-\tau}^{1/(1+\tau)} (2\,\tau\,x_1)\left(1-\frac{x_1}{1+\tau}\right) dx_1 \\ &\quad + \int_{1/(1+\tau)}^{1} \big(1-(1-\tau)\,x_1\big)\left(1-\frac{x_1}{1+\tau}\right) dx_1 \\ &= \frac{\big(7\,\tau^2+14\,\tau+8-2\,\tau^4-2\,\tau^3\big)\,\tau^2}{6\,(\tau+1)^3}\,. \end{aligned}$$

Finally,

$$\begin{aligned} P_{2G,e} &= P\Big(\{X_2,X_3\} \subset \Gamma_{1,e}(X_1,\tau)\Big) \\ &= \int_0^{1-\tau}\left(\frac{2\,\tau\,x_1}{1-\tau^2}\right)^2 dx_1 + \int_{1-\tau}^{1}\left(1-\frac{x_1}{1+\tau}\right)^2 dx_1 \;=\; \frac{\tau^2\,(3\,\tau+4)}{3\,(\tau+1)^2}\,. \end{aligned}$$

Therefore $4\,\mathbf{E}[h_{12}h_{13}] = P_{2N,e} + 2\,P_{NG,e} + P_{2G,e} = \frac{\tau^2\left(2\tau^2+5\tau+4\right)\left(2\tau+4-\tau^2\right)}{3\,(\tau+1)^3}$. Hence

$$4\,\mathbf{Cov}[h_{12},h_{13}] \;=\; \frac{\tau^2\left(4\,\tau+4-2\,\tau^4-4\,\tau^3-\tau^2\right)}{3\,(\tau+1)^3}\,.$$

Case 2: $\tau > 1$: For $x_1 \in (0,1)$, depending on the location of x_1, the following are the different types of the combinations of $N_e(x_1,\tau)$ and $\Gamma_{1,e}(x_1,\tau)$.

- (**i**) for $0 < x_1 \le 1/(1+\tau)$, we have $N_e(x_1,\tau) = \big(0,(1+\tau)\,x_1\big)$ and $\Gamma_{1,e}(x_1,\tau) = \big(x_1/(1+\tau),1\big)$,
- (**ii**) for $1/(1+\tau) < x_1 < 1$, we have $N_e(x_1,\tau) = (0,1)$ and $\Gamma_{1,e}(x_1,\tau) = \big(x_1/(1+\tau),1\big)$.

Then

$$\begin{aligned} p_a^e(\tau,c) &= P\Big(X_2 \in N_e(X_1,\tau)\Big) \\ &= \int_0^{1/(1+\tau)} (1+\tau)\,x_1\,dx_1 + \int_{1/(1+\tau)}^{1} 1\,dx_1 \;=\; \frac{1+2\,\tau}{2\,(\tau+1)}\,. \end{aligned}$$

Next,

$$\begin{aligned} P_{2N,e} &= P\Big(\{X_2,X_3\} \subset N_e(X_1,\tau)\Big) \\ &= \int_0^{1/(1+\tau)} \big((1+\tau)\,x_1\big)^2\,dx_1 + \int_{1/(1+\tau)}^{1} 1\,dx_1 \;=\; \frac{1+3\,\tau}{3\,(\tau+1)}\,, \end{aligned}$$

$$\begin{aligned} P_{NG,e} &= P\Big(X_2 \in N_e(X_1,\tau), X_3 \in \Gamma_{1,e}(X_1,\tau)\Big) \\ &= \int_0^{1/(1+\tau)} \big((1+\tau)\,x_1\big)\left(1-\frac{x_1}{1+\tau}\right) dx_1 + \int_{1/(1+\tau)}^{1}\left(1-\frac{x_1}{1+\tau}\right) dx_1 \\ &= \frac{6\,\tau^3+12\,\tau^2+6\,\tau+1}{6\,(\tau+1)^3}\,. \end{aligned}$$

Finally,

$$P_{2G,e} = P\Big(\{X_2, X_3\} \subset \Gamma_{1,e}(X_1, \tau)\Big)$$
$$= \int_0^1 \left(1 - \frac{x_1}{1+\tau}\right)^2 dx_1 = \frac{3\,\tau^2 + 3\,\tau + 1}{3\,(\tau+1)^2}\,.$$

Therefore $4\,\mathbf{E}[h_{12}h_{13}] = P_{2N,e} + 2\,P_{NG,e} + P_{2G,e} = \frac{12\,\tau^3+25\,\tau^2+15\,\tau+3}{3\,(\tau+1)^3}$. Hence

$$4\,\mathbf{Cov}[h_{12}, h_{13}] = \frac{\tau^2}{3\,(\tau+1)^3}\,. \quad \square$$

A.3. Figures for the consistency results in Section 5.3

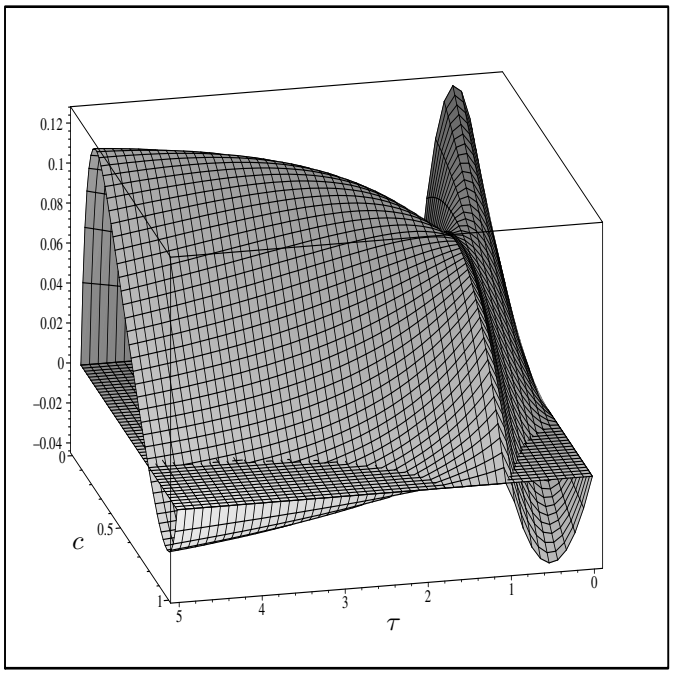

Figure 12: The three dimensional plot of the difference between arc probability of CS-ICD under $H_a^I\colon \delta = 2$ and the null hypothesis $p_a(F_1, \tau, c) - p_a(\mathcal{U}, \tau, c)$ for $c \in (0, 1)$ and $\tau \in (0, 5)$. The horizontal plane is at $z = 0$ and is used to determine the sign changes in the difference.

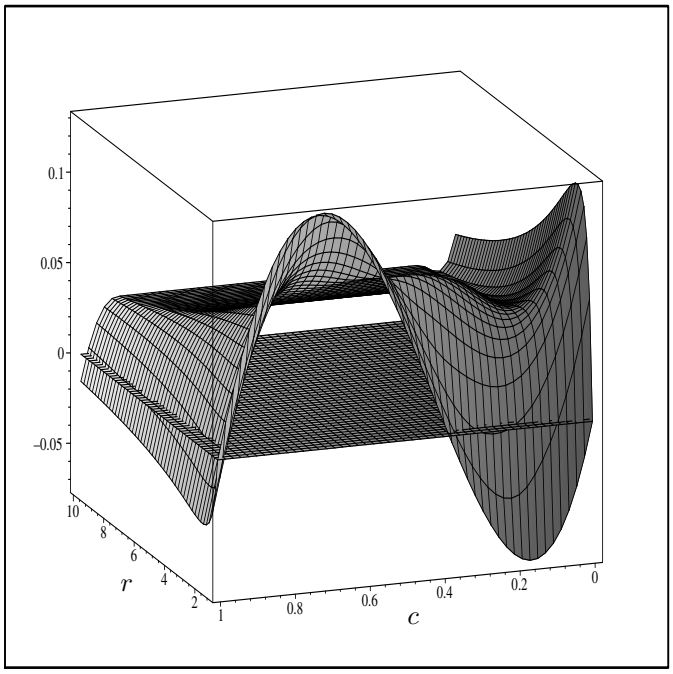

Figure 13: The three dimensional plot of the difference between arc probability of PE-ICD under $H_a^I : \delta = 2$ and the null hypothesis $p_a^{PE}(F_1, r, c) - p_a^{PE}(\mathcal{U}, r, c)$ for $c \in (0, 1)$ and $r \in (1, 10)$. The horizontal plane is at $z = 0$ and is used to determine the sign changes in the difference.

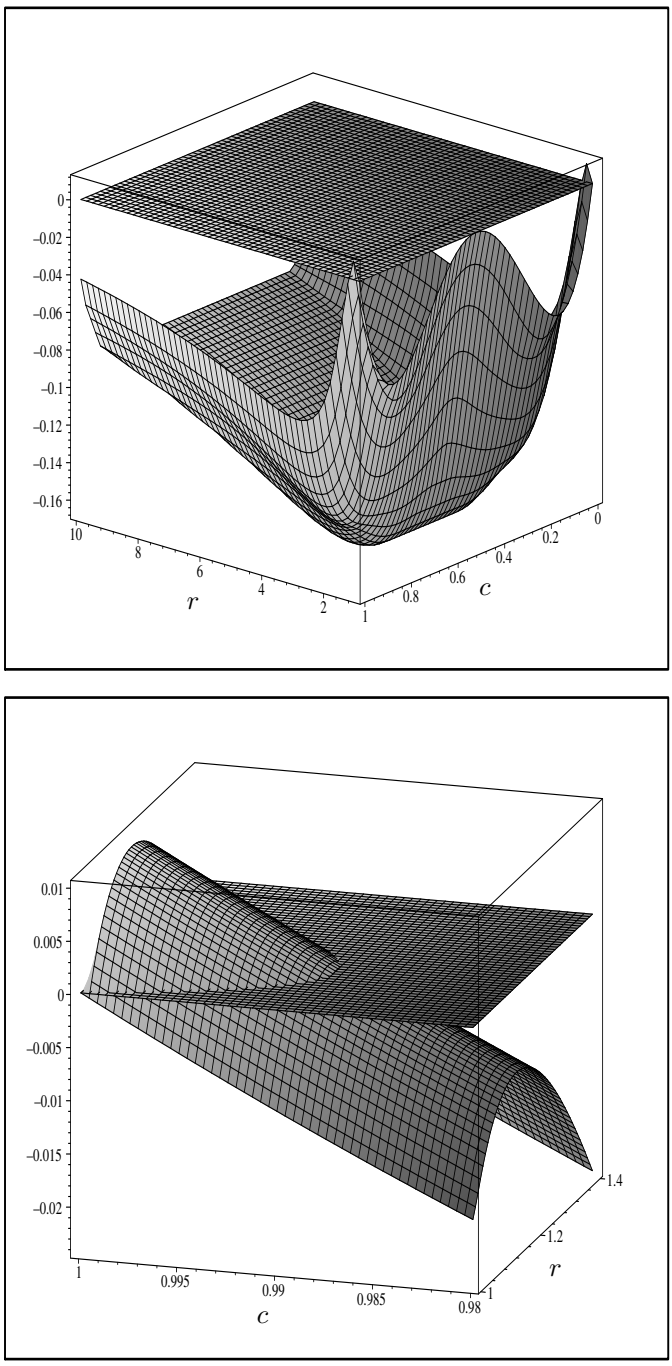

Figure 14: The three dimensional plot of the difference between arc probability of PE-ICD under $H_a^{III}: \delta = 2$ and the null hypothesis $p_a^{PE}(F_3, r, c) - p_a^{PE}(\mathcal{U}, r, c)$. The top plot is with $c \in (0, 1)$ and $r \in (1, 10)$ and the bottom plot is with $c \in (.98, 1)$ and $r \in (1, 1.4)$ (to better visualize the region of positive difference around $(r, c) = (1, 1)$). The horizontal planes at $z = 0$ are used to determine the sign changes in the difference.

ACKNOWLEDGMENTS

I would like to thank the anonymous referee, whose constructive comments and suggestions greatly improved the presentation and flow of this article. This research was supported by the European Commission under the Marie Curie International Outgoing Fellowship Programme via Project # 329370 titled PRinHDD.

REFERENCES

[1] Beineke, L.W. and Zamfirescu, C.M. (1982). Connection digraphs and second order line graphs, *Discrete Mathematics*, **39**, 237–254.

[2] Callaert, H. and Janssen, P. (1978). The Berry–Esseen theorem for U-statistics, *Annals of Statistics*, **6**, 417–421.

[3] Cannon, A. and Cowen, L. (2000). *Approximation algorithms for the class cover problem*. In "Proceedings of the 6th International Symposium on Artificial Intelligence and Mathematics", January 5–7, 2000, Fort Lauderdale, Florida.

[4] Ceyhan, E. (2012). The distribution of the relative arc density of a family of interval catch digraph based on uniform data, *Metrika*, **75**(6), 761–793.

[5] Ceyhan, E. and Priebe, C.E. (2005). The use of domination number of a random proximity catch digraph for testing spatial patterns of segregation and association, *Statistics & Probability Letters*, **73**, 37–50.

[6] Ceyhan, E.; Priebe, C.E. and Marchette, D.J. (2007). A new family of random graphs for testing spatial segregation, *Canadian Journal of Statistics*, **35**(1), 27–50.

[7] Chartrand, G.; Lesniak, L. and Zhang, P. (2010). *Graphs & Digraphs*, Chapman and Hall/CRC 5th Edition, Boca Raton, Florida.

[8] Coleman, T.F. and Moré, J.J. (1983). Estimation of sparse Jacobian matrices and graph coloring problems, *SIAM Journal on Numerical Analysis*, **20**(1), 187–209.

[9] Douglas, B.W. (1996). Short proofs for interval digraphs, *Discrete Math*, **178**, 287–292.

[10] Goldberg, A.V. (1984). Finding a maximum density subgraph, Technical Report UCB/CSD-84-171, EECS Department, University of California, Berkeley.

[11] Grünbaum, B. (1988). The edge-density of 4-critical planar graphs, *Combinatorica*, **8**(1), 137–139.

[12] Jain, A.K.; Xu, X.; Ho, T.K. and Xiao, F. (2002). *Uniformity testing using minimal spanning tree*. In "Proceedings of the 16th International Conference on Pattern Recognition (ICPRŠ02)". 04:40281.

[13] Janson, S.; Łuczak, T. and Ruciński, A. (2000). *Random Graphs*, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, Inc., New York.

[14] Lehmann, E.L. (2004). *Elements of Large Sample Theory*, Springer, New York.

[15] Maehara, H. (1984). A digraph represented by a family of boxes or spheres, *Journal of Graph Theory*, **8**(3), 431–439.

[16] Marhuenda, Y.; Morales, D. and Pardo, M.C. (2005). A comparison of uniformity tests, *Statistics*, **39**(4), 315–327.

[17] Priebe, C.E.; DeVinney, J.G. and Marchette, D.J. (2001). On the distribution of the domination number of random class cover catch digraphs, *Statistics & Probability Letters*, **55**, 239–246.

[18] Prisner, E. (1989). A characterization of interval catch digraphs, *Discrete Mathematics*, **73**, 285–289.

[19] Prisner, E. (1994). Algorithms for interval catch digraphs, *Discrete Applied Mathematics*, **51**, 147–157.

[20] Stephens, M.A. (1974). EDF statistics for goodness of fit and some comparisons, *Journal of American Statistical Association*, **69**, 730–737.

[21] Toussaint, G.T. (1980). The relative neighborhood graph of a finite planar set, *Pattern Recognition*, **12**(4), 261–268.