The Canadian Journal of Statistics
Vol. 35, No. 1, 2007, Pages 27–50
La revue canadienne de statistique

A new family of random graphs for testing spatial segregation

Elvan CEYHAN, Carey E. PRIEBE and David J. MARCHETTE

Key words and phrases: Association; complete spatial randomness; Delaunay triangulation; proximity catch digraph; random graph; relative density; segregation.

MSC 2000: Primary 05C20; secondary 62M30.

Abstract: The authors discuss a graph-based approach for testing spatial point patterns. This approach falls under the category of data-random graphs, which have been introduced and used for statistical pattern recognition in recent years. The authors address specifically the problem of testing complete spatial randomness against spatial patterns of segregation or association between two or more classes of points on the plane. To this end, they use a particular type of parameterized random digraph called a proximity catch digraph (PCD) which is based on relative positions of the data points from various classes. The statistic employed is the relative density of the PCD, which is a U-statistic when scaled properly. The authors derive the limiting distribution of the relative density, using the standard asymptotic theory of U-statistics. They evaluate the finite-sample performance of their test statistic by Monte Carlo simulations and assess its asymptotic performance via Pitman's asymptotic efficiency, thereby yielding the optimal parameters for testing. They further stress that their methodology remains valid for data in higher dimensions.

Une nouvelle famille de graphes aléatoires utile pour tester la ségrégation spatiale

Résumé : Les auteurs montrent comment on peut détecter des configurations de points dans l'espace à l'aide de graphes. Leur approche s'appuie sur la notion de graphe aléatoire observé, récemment introduite et utilisée en statistique pour la reconnaissance de formes. Les auteurs cherchent plus précisément à détecter la présence de ségrégation ou d'association entre deux ou plusieurs ensembles de points du plan en testant l'hypothèse d'absence complète de structure. Dans ce but, ils font appel à une classe paramétrique particulière de digraphes aléatoires appelés digraphes "à captation proximale" (DCP) qui tiennent compte de la disposition relative des éléments des diverses classes. Le test s'appuie sur la densité relative du DCP qui, une fois proprement normalisée, est une U-statistique. Les auteurs en déterminent la loi limite en invoquant la théorie asymptotique des U-statistiques. Ils en évaluent la performance à taille finie au moyen de simulations de Monte-Carlo et en étudient aussi le comportement limite sous l'angle de l'efficacité asymptotique de Pitman, dont découlent des choix optimaux de paramètres aux fins de test. Ils soulignent de plus que leur méthodologie reste valide en dimensions supérieures.

1. INTRODUCTION

In this article, a graph-based approach for testing spatial point patterns is discussed. In the statistical literature, the analysis of spatial point patterns in natural populations has been extensively studied and has important implications in epidemiology, population biology, and ecology. The pattern of points from one class with respect to points from other classes, rather than the pattern of points from one class with respect to the ground, is investigated. The spatial relationships among two or more classes have important implications especially for plant species. See, for example, Pielou (1961) and Dixon (1994, 2002).

The goal of this article is to test the spatial pattern of complete spatial randomness against spatial segregation or association. Complete spatial randomness (CSR) is roughly defined as the lack of spatial interaction between the points in a given study area. Segregation is the pattern in which points of one class tend to cluster together, i.e., form one-class clumps. In association, the points of one class tend to occur more frequently around points from the other class. For convenience and generality, we call the different types of points "classes", but the class can be replaced by any characteristic of an observation at a particular location. For example, the pat-

tern of spatial segregation has been investigated for species (Diggle 1983), age classes of plants (Hamill & Wright 1986) and sexes of dioecious plants (Nanami, Kawaguchi & Yamakura 1999).

Data random digraphs are directed graphs in which each vertex corresponds to a data point, and directed edges (arcs) are defined in terms of some bivariate function on the data. For example, nearest neighbour graphs are defined by placing an arc between each vertex and its nearest neighbour. Priebe, DeVinney & Marchette (2001) introduced a data random digraph (called class cover catch digraphs (CCCD)) in $\mathbb{R}$ and extended it to multiple dimensions. In this model, the vertices correspond to data from a single class $\mathcal{X}$ and the definition of the arcs utilizes the other class $\mathcal{Y}$. For each $x_i \in \mathcal{X}$ a radius is defined by $r_i = \min d(x_i, \mathrm{y})$ where the minimum is taken over all $\mathrm{y} \in \mathcal{Y}$. There is an arc from x_i to x_j if $d(x_i, x_j) < r_i$, that is, if the sphere of radius r_i centered at x_i "catches" x_j. DeVinney, Priebe, Marchette & Socolinsky (2002), Marchette & Priebe (2003), Priebe, Marchette, DeVinney & Socolinsky (2003), and Priebe, Solka, Marchette & Clark (2003) demonstrated relatively good performance of CCCD's classification.

We define a new class of random digraphs (proximity catch digraphs or PCDs) and apply it in testing against segregation or association. By construction, in our PCDs, the farther an $\mathcal{X}$ point is from $\mathcal{Y}$, the more arcs to other $\mathcal{X}$ points it will be likely to have. We will use the relative density (number of arcs divided by the total number of possible arcs) as a statistic for testing against segregation or association.

To illustrate our methods, we provide three artificial data sets, one for each pattern. These data sets are plotted in Figure 1, where $\mathcal{Y}$ points are at the vertices of the triangles, and $\mathcal{X}$ points are depicted as squares. The triangles are from the Delaunay triangulation of the $\mathcal{Y}$ points. These triangles will be used to define the proximity function that will in turn define the PCD. Under the segregation pattern (left) the relative density of the PCD will be larger compared to the CSR pattern (middle), while under the association pattern (right) the relative density will be smaller compared to the CSR case.

The statistical tool utilized is the asymptotic theory of U-statistics. When the relative density of our PCDs is properly scaled, we demonstrate that it is a U-statistic, which has asymptotic normality by the general central limit theory of U-statistics. For the digraphs introduced by Priebe, DeVinney & Marchette (2001), whose relative density is also of the U-statistic form, the asymptotic mean and variance of the relative density are not analytically tractable in multiple dimensions, due to geometric difficulties encountered. However, the PCD we introduce is a parameterized family of random digraphs, whose relative density has tractable asymptotic mean and variance.

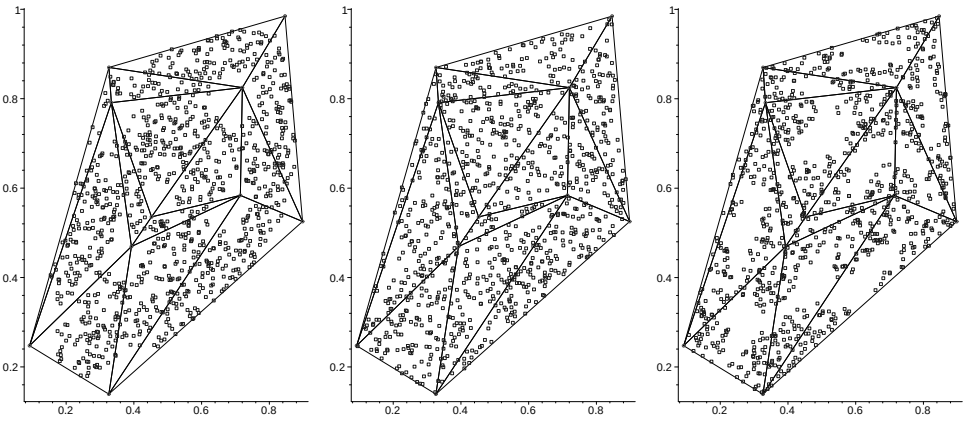

FIGURE 1: Realizations of segregation (left), CSR (middle), and association (right) patterns for $|\mathcal{Y}| = 10$ and $|\mathcal{X}| = 1000$. The $\mathcal{Y}$ points are at the vertices of the triangles and the $\mathcal{X}$ points are squares.

Ceyhan & Priebe (2003) introduced an (unparameterized) version of the PCD we discuss in

this article; Ceyhan & Priebe (2005) also introduced another parameterized family of PCDs and used the domination number (which is another statistic based on the number of arcs from the vertices) of this latter parameterized family for testing segregation and association. The domination number approach is appropriate when both classes are comparably large. Ceyhan, Priebe & Wierman (2006) used the relative density of the same PCD for testing the spatial patterns. Our new parameterized family of PCDs has more geometric appeal, is simpler in distributional parameters in the asymptotics, and the range of the parameters is bounded.

Using the Delaunay triangulation of the $\mathcal{Y}$ observations, in Section 3.1 we will define the parameterized version of the proximity maps of Ceyhan & Priebe (2003) for which the calculations (regarding the distribution of the relative density) are tractable. We then can use the relative density of the digraph to construct a test of complete spatial randomness against the alternatives of segregation or association which are defined explicitly in Sections 2 and 4.1. We will calculate the asymptotic distribution of the relative density for these digraphs, under both the null and alternative patterns in Sections 4.2 and 4.3, respectively. This procedure results in a consistent test, as will be shown in Section 5.1. The finite sample performance (in terms of power) is analyzed using Monte Carlo simulations in Section 5.2. The Pitman asymptotic efficiency is analyzed in Section 5.3. The multiple-triangle case and the extension to higher dimensions are presented in Sections 5.4 and 5.5, respectively. All proofs are provided in the Appendix.

2. SPATIAL POINT PATTERNS

For simplicity, we describe the spatial point patterns for two-class populations. The null hypothesis for spatial patterns has been a controversial topic in ecology from the early days (Gotelli & Graves 1996). But in general, the null hypothesis consists of two random pattern types: complete spatial randomness or random labelling.

Under *complete spatial randomness* (CSR) for a spatial point pattern $\{X_i(D),\ i = 1, \ldots, n : D \subset \mathbb{R}^2\}$, where $X_i(D)$ is the Bernoulli random variable denoting the event that point i is in region D, we have

(i) given n points in domain D, the points are an independent random sample from the uniform distribution on D;

(ii) there is no spatial interaction, i.e., the locations of these points have no influence on one another.

Note that condition (ii) is implied by (i). Furthermore, when the reference region D is large, the number of points in any planar region with area $A(D)$ follows (approximately) a Poisson distribution with intensity λ and mean $\lambda \cdot A(D)$.

Given a fixed set of points in a region, under random labelling, class labels are assigned to these fixed points randomly so that the labels are independent of the locations. Thus, random labelling is less restrictive than CSR. We only consider a special case of CSR as our null hypothesis. More specifically, only $\mathcal{X}$ points are assumed to be uniformly distributed over the convex hull of $\mathcal{Y}$ points.

The alternative patterns fall under two major categories called *association* and *segregation*. Association occurs if the points from the two classes together form clumps or clusters. That is, association occurs when members of one class have a tendency to attract members of the other class, as in symbiotic species, so that the X_i will tend to cluster around the members of $\mathcal{Y}$. For example, in plant biology, $\mathcal{X}$ points might be the geometric coordinates of parasitic plants exploiting another plant whose coordinates are $\mathcal{Y}$ points. As another example, $\mathcal{X}$ and $\mathcal{Y}$ points might represent the coordinates of mutualistic plant species, so they depend on each other to survive. In epidemiology, $\mathcal{Y}$ points might be the geographic coordinates of contaminant sources, such as a nuclear reactor, or a factory emitting toxic waste, and $\mathcal{X}$ points might be the coordinates

of the residences of cases (incidences) of certain diseases caused by the contaminant, e.g., some type of cancer.

Segregation occurs if the members of the same class tend to be clumped or clustered together (see, e.g., Pielou 1961). Many different forms of segregation are possible. Our methods will be useful only for the segregation patterns in which the two classes more or less share the same support (habitat), and members of one class have a tendency to repel members of the other class. For instance, it may be the case that one type of plant does not grow well in the vicinity of another type of plant, and vice versa. This implies, in our notation, that the X_i are unlikely to be located near any elements of $\mathcal{Y}$. See, for instance, (Dixon 1994; Coomes, Rees & Turnbull 1999). In plant biology, $\mathcal{Y}$ points might represent the coordinates of trees from a species with large canopy, so that other plants (whose coordinates are $\mathcal{X}$ points) that need light cannot grow around these trees. As another interesting but contrived example, consider the arsonist who wishes to start fires with maximum duration time (hence maximum damage), so that he starts the fires at the furthest points possible from fire houses in a city. Then $\mathcal{Y}$ points could be the geographic coordinates of the fire houses, while $\mathcal{X}$ points will be the coordinates of the locations of the arson cases.

We consider completely mapped data, i.e., the locations of all events in a defined space are observed rather than sparsely sampled data (i.e., only a random subset of locations is observed).

3. DATA-RANDOM PROXIMITY CATCH DIGRAPHS

In general, in a *random digraph*, there is an arc between two vertices, with a fixed probability, independent of other arcs and vertex pairs. However, in our approach, arcs with a shared vertex will be dependent. Hence the name *data-random digraphs*.

Let $(\Omega, \mathcal{M})$ be a measurable space and consider a function $N: \Omega \times 2^{\Omega} \to 2^{\Omega}$, where 2^{Ω} represents the power set of Ω. Then given $\mathcal{Y} \subseteq \Omega$, the *proximity map* $N_{\mathcal{Y}}(\cdot) = N(\cdot, \mathcal{Y}) : \Omega \to 2^{\Omega}$ associates a *proximity region* $N_{\mathcal{Y}}(x) \subseteq \Omega$ with each point $x \in \Omega$. The region $N_{\mathcal{Y}}(x)$ is defined in terms of the distance between x and $\mathcal{Y}$.

If $\mathcal{X}_n := \{X_1, X_2, \ldots, X_n\}$ is a set of Ω-valued random variables, then the $N_{\mathcal{Y}}(X_i)$, $i = 1, \ldots, n$, are random sets. If the X_i are independent and identically distributed, then so are the random sets, $N_{\mathcal{Y}}(X_i)$.

Define the data-random proximity catch digraph $\mathcal{D}$ with vertex set $\mathcal{V} = \{X_1, \ldots, X_n\}$ and arc set $\mathcal{A}$ by $(X_i, X_j) \in \mathcal{A} \iff X_j \in N_{\mathcal{Y}}(X_i)$ where point X_i catches the point X_j. The random digraph $\mathcal{D}$ depends on the (joint) distribution of the X_i and on the map $N_{\mathcal{Y}}$. The adjective *proximity* (for the catch digraph $\mathcal{D}$ and for the map $N_{\mathcal{Y}}$) comes from thinking of the region $N_{\mathcal{Y}}(x)$ as representing those points in Ω close to x (Toussaint 1980; and Jaromczyk & Toussaint 1992).

The relative density of a digraph $\mathcal{D} = (\mathcal{V}, \mathcal{A})$ of order $|\mathcal{V}| = n$ (i.e., number of vertices is n), denoted $\rho(\mathcal{D})$, is defined as

$$\rho(\mathcal{D}) = \frac{|\mathcal{A}|}{n(n-1)}$$

where $|\cdot|$ stands for the set cardinality (Janson, Łuczak & Ruciński 2000). Thus $\rho(\mathcal{D})$ represents the ratio of the number of arcs in the digraph $\mathcal{D}$ to the number of arcs in the complete symmetric digraph of order n, namely $n(n-1)$.

If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, then the relative density of the associated data-random proximity catch digraph $\mathcal{D}$, denoted $\rho(\mathcal{X}_n; h, N_{\mathcal{Y}})$, is a U-statistic:

$$\rho(\mathcal{X}_n; h, N_{\mathcal{Y}}) = \frac{1}{n(n-1)} \sum_{i<j} \sum h(X_i, X_j; N_{\mathcal{Y}})$$

where

$$\begin{aligned} h(X_i, X_j; N_{\mathcal{Y}}) &= \mathbf{I}\{(X_i, X_j) \in \mathcal{A}\} + \mathbf{I}\{(X_j, X_i) \in \mathcal{A}\} \\ &= \mathbf{I}\{X_j \in N_{\mathcal{Y}}(X_i)\} + \mathbf{I}\{X_i \in N_{\mathcal{Y}}(X_j)\} \end{aligned}$$

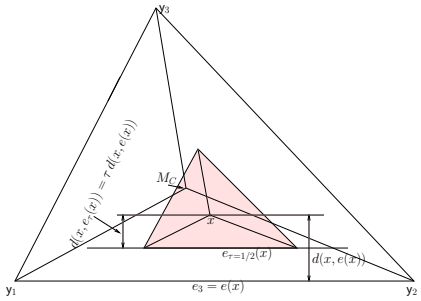

FIGURE 2: Construction of τ-factor central similarity proximity region $N_{\mathrm{CS}}^{1/2}(x)$ (shaded region).

with $\mathbf{I}(\,\cdot\,)$ being the indicator function. We denote $h(X_i, X_j; N_{\mathcal{Y}})$ as h_{ij} henceforth for brevity of notation. Although the digraph is not symmetric (since $(x, y) \in \mathcal{A}$ does not necessarily imply $(y, x) \in \mathcal{A}$), h_{ij} is defined as the number of arcs in D between vertices X_i and X_j, in order to produce a symmetric kernel with finite variance (Lehmann 1988).

The random variable $\rho_n := \rho(\mathcal{X}_n; h, N_{\mathcal{Y}})$ depends on n and $N_{\mathcal{Y}}$ explicitly and on F implicitly. The expectation $\mathbf{E}[\rho_n]$, however, is independent of n and depends only on F and $N_{\mathcal{Y}}$:

$$0 \leq \mathbf{E}[\rho_n] = \frac{1}{2}\mathbf{E}[h_{12}] \leq 1 \quad \text{for all } n \geq 2.$$

The variance $\mathrm{var}[\rho_n]$ simplifies to

$$\mathrm{var}[\rho_n] = \frac{1}{2n(n-1)}\,\mathrm{var}[h_{12}] + \frac{n-2}{n(n-1)}\mathrm{cov}[h_{12}, h_{13}] \leq 1/4.$$

A central limit theorem for U-statistics (Lehmann (1988)) yields

$$\sqrt{n}\,(\rho_n - \mathrm{E}[\rho_n]) \xrightarrow{\mathcal{L}} \mathsf{N}(0, \mathrm{cov}[h_{12}, h_{13}])$$

provided that $\mathrm{cov}[h_{12}, h_{13}] > 0$. The asymptotic variance of ρ_n, namely $\mathrm{cov}[h_{12}, h_{13}]$, depends only on F and $N_{\mathcal{Y}}$. Thus, we need to determine only $\mathrm{E}[h_{12}]$ and $\mathrm{cov}[h_{12}, h_{13}]$ in order to obtain the normal approximation for ρ_n.

3.1. The τ-factor central similarity proximity catch digraphs.

We define the τ-factor central similarity proximity map briefly. Let $\Omega = \mathbb{R}^2$ and let $\mathcal{Y} = \{\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3\} \subset \mathbb{R}^2$ be three non-collinear points. Denote the triangle (including the interior) formed by the points in $\mathcal{Y}$ as $T(\mathcal{Y})$. For $\tau \in [0, 1]$, define $N_{\mathcal{Y}}^{\tau}$ to be the τ-factor central similarity proximity map as follows; see also Figure 2. Let e_j be the edge opposite vertex y_j for $j = 1, 2, 3$, and let "edge regions" $R(e_1)$, $R(e_2)$, $R(e_3)$ partition $T(\mathcal{Y})$ using segments from the centre of mass of $T(\mathcal{Y})$, M_C, to the vertices. For $x \in T(\mathcal{Y}) \setminus \mathcal{Y}$, let $e(x)$ be the edge in whose region x falls; $x \in R(e(x))$. If x falls on the boundary of two edge regions we assign $e(x)$ arbitrarily. For $\tau \in (0, 1]$, the *τ-factor* central similarity proximity region $N_{\mathrm{CS}}^{\tau}(x) = N_{\mathcal{Y}}^{\tau}(x)$ is defined to be the triangle $T_\tau(x)$ with the following properties:

(i) $T_\tau(x)$ has an edge $e_\tau(x)$ parallel to $e(x)$ such that $d(x, e_\tau(x)) = \tau\, d(x, e(x))$ and $d(e_\tau(x), e(x)) \leq d(x, e(x))$ where $d(x, e(x))$ is the Euclidean (perpendicular) distance from x to $e(x)$,

(ii) $T_\tau(x)$ has the same orientation as and is similar to $T(\mathcal{Y})$,

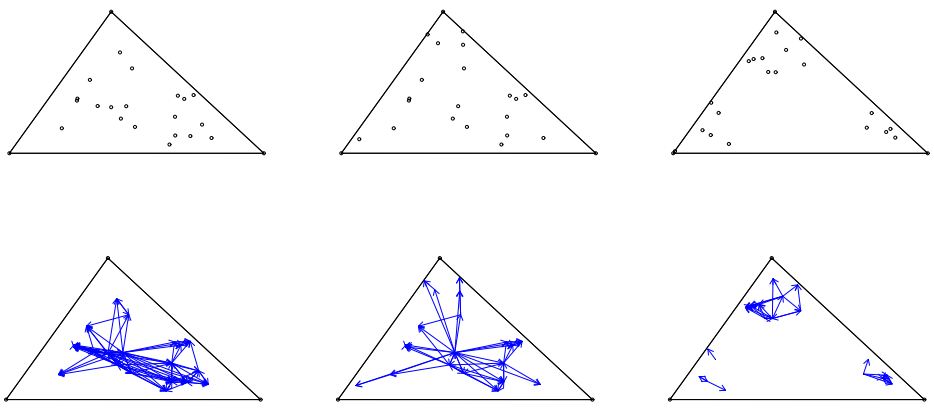

FIGURE 3: Realizations of segregation (left), CSR (middle), and association (right) for $|\mathcal{Y}| = 3$ and $|\mathcal{X}| = 20$. $\mathcal{Y}$ points are at the vertices of the triangle, and $\mathcal{X}$ points are circles.

(iii) x is at the centre of mass of $T_\tau(x)$.

Note that (i) implies the τ-factor, (ii) implies similarity, and (iii) implies central in the name *τ-factor central similarity proximity map*. Notice that $\tau > 0$ implies that $x \in N^\tau_{\mathrm{CS}}(x)$ and $\tau \leq 1$ implies that $N^\tau_{\mathrm{CS}}(x) \subseteq T(\mathcal{Y})$ for all $x \in T(\mathcal{Y})$. For $x \in \partial(T(\mathcal{Y}))$ and $\tau \in [0,1]$, we define $N^\tau_{\mathrm{CS}}(x) = \{x\}$; for $\tau = 0$ and $x \in T(\mathcal{Y})$ we also define $N^\tau_{\mathrm{CS}}(x) = \{x\}$. Let $T(\mathcal{Y})^o$ be the interior of the triangle $T(\mathcal{Y})$. Then for all $x \in T(\mathcal{Y})^o$ the edges $e_\tau(x)$ and $e(x)$ are coincident iff $\tau = 1$. Note that the central similarity proximity map of Ceyhan & Priebe (2003) is $N^\tau_{\mathrm{CS}}(\,\cdot\,)$ with $\tau = 1$. Hence by definition, (x, y) is an arc of the τ-factor central similarity PCD iff $y \in N^\tau_{\mathrm{CS}}(x)$.

Notice that $X_i \overset{\mathrm{iid}}{\sim} F$, with the additional assumption that the non-degenerate two-dimensional probability density function f exists with support in $T(\mathcal{Y})$, implies that the special case (X falls on the boundary of two edge regions) in the construction of N^τ_{CS} occurs with probability zero.

For a fixed $\tau \in (0,1]$, $N^\tau_{\mathrm{CS}}(x)$ gets larger (in area) as x gets farther away from the edges (or equivalently gets closer to the centre of mass M_C) in that $d(x, e(x))$ increases, or equivalently $d(M_C, e_\tau(x))$ decreases. Hence for points in $T(\mathcal{Y})$, the farther the points away from the vertices $\mathcal{Y}$ (or closer the points to M_C as above), the larger the area of $N^\tau_{\mathrm{CS}}(x)$. Hence, it is more likely for such points to catch other points, i.e., have more arcs directed to other points. Therefore, if more $\mathcal{X}$ points are clustered around the centre of mass, then the digraph is more likely to have more arcs, hence larger relative density. So, under segregation, relative density is expected to be larger than that in CSR or association. On the other hand, in the case of association, i.e., when $\mathcal{X}$ points are clustered around $\mathcal{Y}$ points, the regions $N^\tau_{\mathrm{CS}}(x)$ tend to be smaller in area, hence, catch fewer points, thereby resulting in a small number of arcs, or a smaller relative density compared to CSR or segregation. See, for example, Figure 3 with three $\mathcal{Y}$ points, and 20 $\mathcal{X}$ points for segregation (top left), CSR (top middle) and association (top right). The corresponding arcs in the τ-factor central similarity PCD with $\tau = 1$ are plotted in the bottom row in Figure 3. The corresponding relative density values (for $\tau = 1$) are .2579, .1395, and .0974, respectively.

Furthermore, for a fixed $x \in T(\mathcal{Y})^o$, $N^\tau_{\mathrm{CS}}(x)$ gets larger (in area) as τ increases. So as τ increases, it is more likely to have more arcs, hence larger relative density for a given realization of $\mathcal{X}$ points in $T(\mathcal{Y})$.

4. ASYMPTOTIC DISTRIBUTION OF THE RELATIVE DENSITY

There are two major types of asymptotic structures for spatial data (Lahiri 1996). In the first, any two points are required to be at least a fixed distance apart, hence as the number of points increase, the region on which the process (or pattern) is observed eventually becomes unbounded. This type of sampling structure is called *increasing domain asymptotics*. In the second type, the region of interest is a fixed bounded region and more and more points are observed in this region. Hence the minimum distance between data points tends to zero as the sample size tends to infinity. This type of structure is called *infill asymptotics*, due to Cressie (1991).

The sampling structure for our asymptotic analysis is infill, for only the size of the type $\mathcal{X}$ points tends to infinity, while the support, the convex hull $C_H(\mathcal{Y})$ of a given set of points from type $\mathcal{Y}$ points is a fixed bounded region.

Next, we describe the null pattern of CSR and parameterize the alternative patterns of segregation and association briefly, and then provide the asymptotic distribution of the relative density for these patterns.

4.1. Null and alternative patterns.

For statistical testing against segregation and association, the null hypothesis is generally some form of complete spatial randomness; thus we consider

$$\mathcal{H}_0 : X_i \overset{\text{iid}}{\sim} \mathsf{U}(T(\mathcal{Y})).$$

If it is desired to have the sample size be a random variable, we may consider a spatial Poisson point process on $T(\mathcal{Y})$ as our null hypothesis.

We first present a geometry-invariance result that will simplify our calculations by allowing us to consider the special case of the equilateral triangle.

THEOREM 1 (Geometry invariance property). *Let $\mathcal{Y} = \{\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3\} \subset \mathbb{R}^2$ be three non-collinear points. For $i = 1, \ldots, n$, let $\mathbf{X_1} \overset{\text{iid}}{\sim} \mathsf{U}(T(\mathcal{Y}))$, the uniform distribution on the triangle $T(\mathcal{Y})$. Then for any $\tau \in [0, 1]$ the distribution of $\rho_n(\tau) := \rho(\mathcal{X}_n; h, N_{\text{CS}}^{\tau})$ is independent of $\mathcal{Y}$, hence the geometry of $T(\mathcal{Y})$.*

Based on Theorem 1 and our uniform null hypothesis, we may henceforth assume that $T(\mathcal{Y})$ is the standard equilateral triangle with $\mathcal{Y} = \{(0,0), (1,0), (1/2, \sqrt{3}/2)\}$. For our τ-factor central similarity proximity map and uniform null hypothesis, the asymptotic null distribution of $\rho_n(\tau) = \rho(\mathcal{X}_n; h, N_{\text{CS}}^{\tau})$ as a function of τ can be derived. Let $\mu(\tau) := \mathrm{E}[\rho_n]$, then

$$\mu(\tau) = \mathrm{E}[h_{12}]/2 = \mathrm{P}(\mathbf{X_2} \in N_{\text{CS}}^{\tau}(\mathbf{X_1}))$$

is the probability of an arc occurring between any two vertices, and let $\nu(\tau) := \mathrm{cov}[h_{12}, h_{13}]$.

We define two simple classes of alternatives, $\mathcal{H}_{\varepsilon}^{S}$ and $\mathcal{H}_{\varepsilon}^{A}$ with $\varepsilon \in (0, \sqrt{3}/3)$, for segregation and association, respectively. See also Figure 4. For $\mathsf{y} \in \mathcal{Y}$, let $e(\mathsf{y})$ denote the edge of $T(\mathcal{Y})$ opposite vertex y, and for $x \in T(\mathcal{Y})$ let $\ell_{\mathsf{y}}(x)$ denote the $e(\mathsf{y})$ through x. Then define $T(\mathsf{y}, \varepsilon) = \{x \in T(\mathcal{Y}) : d(\mathsf{y}, \ell_{\mathsf{y}}(x)) \leq \varepsilon\}$. Let $\mathcal{H}_{\varepsilon}^{S}$ be the model under which $\mathbf{X_1} \overset{\text{iid}}{\sim} \mathsf{U}\big(T(\mathcal{Y}) \setminus \bigcup_{\mathsf{y} \in \mathcal{Y}} T(\mathsf{y}, \varepsilon)\big)$ and $\mathcal{H}_{\varepsilon}^{A}$ be the model under which $\mathbf{X_1} \overset{\text{iid}}{\sim} \mathsf{U}\big(\bigcup_{\mathsf{y} \in \mathcal{Y}} T\big(\mathsf{y}, \sqrt{3}/3 - \varepsilon\big)\big)$. The shaded region in Figure 4 is the support for segregation for a particular ε value; and its complement is the support for the association alternative with $\sqrt{3}/3 - \varepsilon$. Thus the segregation model excludes the possibility of any $\mathbf{X_1}$ occurring near a y_j and the association model requires that $\mathbf{X_1}$ occur near a y_j. The $\sqrt{3}/3 - \varepsilon$ in the definition of the association alternative is so that $\varepsilon = 0$ yields $\mathcal{H}_0$ under both classes of alternatives. We consider these types of alternatives among many other possibilities, since relative density is geometry invariant for these alternatives as the alternatives are defined with parallel lines to the edges.

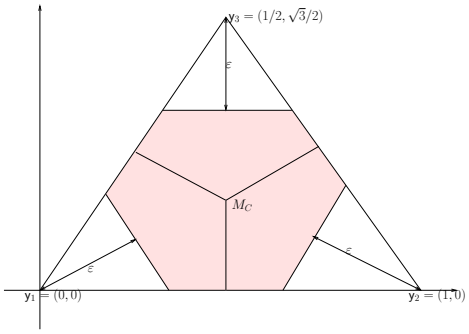

FIGURE 4: An example of the segregation alternative for a particular ε (shaded region); its complement is for the association alternative (unshaded region) on the standard equilateral triangle.

Remark. These definitions of the alternatives are given for the standard equilateral triangle. The geometry-invariance result of Theorem 1 still holds under the alternatives as follows: if, in an arbitrary triangle, a small percentage $\delta \cdot 100\%$ where $\delta \in (0, 4/9)$ of the area is carved away as forbidden from each vertex using line segments parallel to the opposite edge, then under the transformation to the standard equilateral triangle this will result in the alternative $\mathcal{H}^S_{\sqrt{3\,\delta/4}}$. This argument is for segregation with $\delta < 1/4$; a similar construction is available for the other cases.

4.2. Asymptotic normality under the null hypothesis.

By detailed geometric probability calculations provided in the Appendix and in Ceyhan, Priebe & Marchette (2004), the mean and the asymptotic variance of the relative density of our proximity catch digraph can be calculated explicitly. The central limit theorem for U-statistics then establishes the asymptotic normality under the null hypothesis. These results are summarized in the following theorem.

THEOREM 2. *For $\tau \in (0, 1]$, the relative density of the τ-factor central similarity proximity digraph converges in law to the normal distribution; i.e., as $n \to \infty$,*

$$\frac{\sqrt{n}\,(\rho_n(\tau) - \mu(\tau))}{\sqrt{\nu(\tau)}} \xrightarrow{\mathcal{L}} \mathsf{N}(0, 1)$$

where

$$\mu(\tau) = \tau^2/6 \tag{1}$$

and

$$\nu(\tau) = \frac{\tau^4(6\,\tau^5 - 3\,\tau^4 - 25\,\tau^3 + \tau^2 + 49\,\tau + 14)}{45\,(\tau + 1)(2\,\tau + 1)(\tau + 2)}. \tag{2}$$

For $\tau = 0$, $\rho_n(\tau)$ is degenerate for all $n > 1$.

The mean and the variance functions are plotted in Figure 5. Note that $\mu(\tau)$ is strictly increasing in τ, since $N^\tau_{\mathrm{CS}}(x)$ increases with τ for all $x \in T(\mathcal{Y})^o$. Note also that $\mu(\tau)$ is continuous in τ with $\mu(\tau = 1) = 1/6$ and $\mu(\tau = 0) = 0$. Regarding the asymptotic variance, note that $\nu(\tau)$ is strictly increasing and continuous in τ and $\nu(\tau = 1) = 7/135$ and $\nu(\tau = 0) = 0$ (there are no arcs when $\tau = 0$ a.s.) which explains why $\rho_n(\tau = 0)$ is degenerate.

As an example of the limiting distribution, $\tau = 1/2$ yields

$$\frac{\sqrt{n}\,\big(\rho_n(1/2) - \mu(1/2)\big)}{\sqrt{\nu(1/2)}} = \sqrt{\frac{2880\,n}{19}}\,\big(\rho_n(1/2) - 1/24\big) \xrightarrow{\mathcal{L}} \mathsf{N}(0, 1),$$

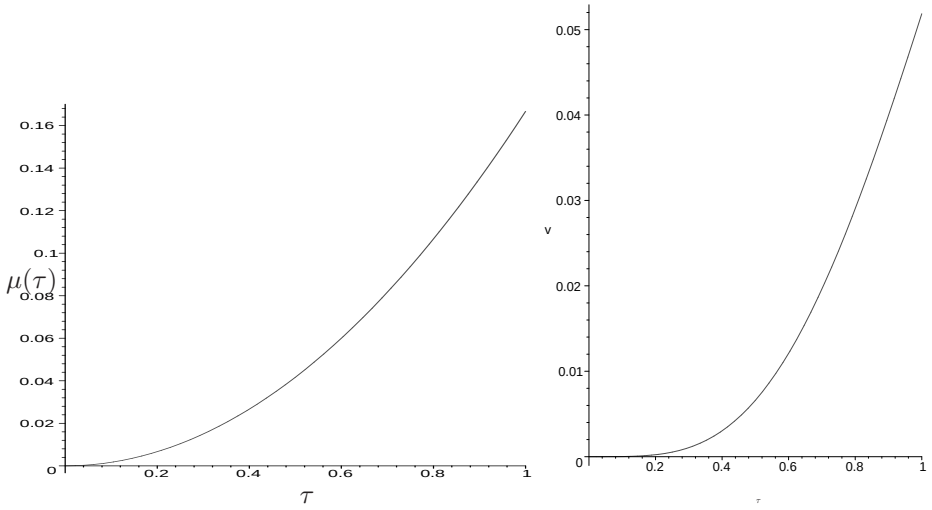

FIGURE 5: Result of Theorem 2: asymptotic null mean $\mu(\tau) = \mu(\tau)$ (left) and variance $\nu(\tau) = \nu(\tau)$ (right), from Equations (1) and (2), respectively.

or equivalently,

$$\rho_n(1/2) \overset{\text{approx}}{\sim} \mathsf{N}\left(\frac{1}{24}, \frac{19}{2880\,n}\right).$$

The finite sample variance may be derived analytically in much the same way as $\operatorname{cov}[h_{12}, h_{13}]$ for the asymptotic variance. In fact, the exact distribution of $\rho_n(\tau)$ is available, in principle, by successively conditioning on the values of the X_i. Alas, while the joint distribution of h_{12}, h_{13} is available, the joint distribution of $\{h_{ij}\}_{1\le i<j\le n}$, and hence the calculation for the exact distribution of $\rho_n(\tau)$, is extraordinarily tedious and lengthy for even small values of n.

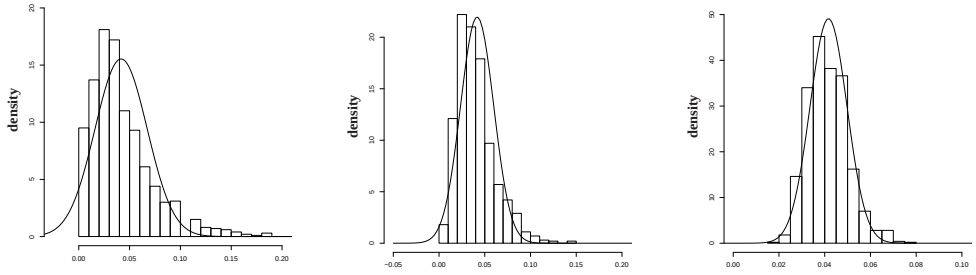

FIGURE 6: Depicted are $\rho_n(1/2) \overset{\text{approx}}{\sim} \mathsf{N}\left(\frac{1}{24}, \frac{19}{2880\,n}\right)$ for $n = 10,\ 20,\ 100$ (left to right). Histograms are based on 1000 Monte Carlo replicates. Solid curves represent the approximating normal densities given in Theorem 2. Note that the axes are differently scaled.

Figure 6 indicates that, for $\tau = 1/2$, the normal approximation is accurate even for small n (although kurtosis and skewness may be indicated for $n = 10,\ 20$). Figure 7 demonstrates, however, that the smaller the value of τ the more severe the skewness of the probability density.

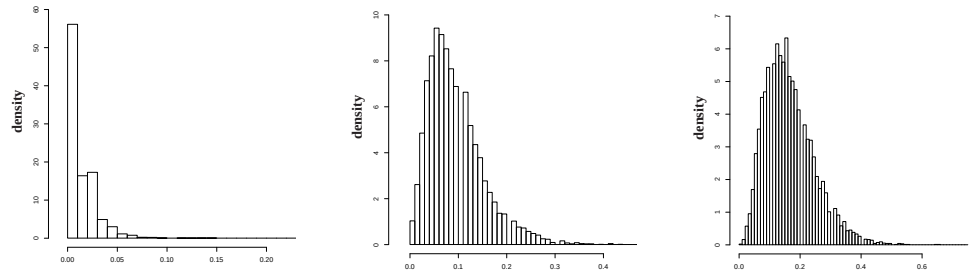

FIGURE 7: Depicted are the histograms for 10000 Monte Carlo replicates of $\rho_{10}(1/4)$ (left), $\rho_{10}(3/4)$ (middle), and $\rho_{10}(1)$ (right) indicating severe small sample skewness for small values of τ.

4.3. Asymptotic normality under the alternatives.

Asymptotic normality of the relative density of the proximity catch digraph under the alternative hypotheses of segregation and association can be established by the same method as under the null hypothesis. Let $\mathrm{E}_\varepsilon[\,\cdot\,]$ be the expectation with respect to the uniform distribution under the segregation and association alternatives with $\varepsilon \in \left(0, \sqrt{3}/3\right)$.

THEOREM 3. *Let $\mu_S(\tau,\varepsilon)$ ($\mu_A(\tau,\varepsilon)$) be the mean and let $\nu_S(\tau,\varepsilon)$ ($\nu_A(\tau,\varepsilon)$) be the covariance,* $\mathrm{cov}\,[h_{12}, h_{13}]$ *for $\tau \in (0,1]$ and $\varepsilon \in \left(0, \sqrt{3}/3\right)$ under segregation (association). Then under $\mathcal{H}_\varepsilon^S$,*

$$\sqrt{n}\left(\rho_n(\tau) - \mu_S(\tau,\varepsilon)\right) \xrightarrow{\mathcal{L}} \mathsf{N}\big(0, \nu_S(\tau,\varepsilon)\big)$$

for the values of the pair (τ,ε) for which $\nu_S(\tau,\varepsilon) > 0$. $\rho_n(\tau)$ is degenerate when $\nu_S(\tau,\varepsilon) = 0$. Likewise, under $\mathcal{H}_\varepsilon^A$, $\sqrt{n}\left(\rho_n(\tau) - \mu_A(\tau,\varepsilon)\right) \xrightarrow{\mathcal{L}} \mathsf{N}\big(0, \nu_A(\tau,\varepsilon)\big)$ for the values of the pair (τ,ε) for which $\nu_A(\tau,\varepsilon) > 0$. $\rho_n(\tau)$ is degenerate when $\nu_A(\tau,\varepsilon) = 0$.

5. THE TEST AND ANALYSIS

The relative density of the central similarity proximity catch digraph is a test statistic for the segregation/association alternative; rejecting for extreme values of $\rho_n(\tau)$ is appropriate since under segregation we expect $\rho_n(\tau)$ to be large, while under association we expect $\rho_n(\tau)$ to be small. Using the test statistic

$$R(\tau) = \frac{\sqrt{n}\left(\rho_n(\tau) - \mu(\tau)\right)}{\sqrt{\nu(\tau)}},$$

which is the normalized relative density, the asymptotic critical value for the one-sided level α test against segregation is given by

$$z_\alpha = \Phi^{-1}(1-\alpha).$$

Against segregation, the test rejects for $R(\tau) > z_\alpha$ and against association, the test rejects for $R(\tau) < z_{1-\alpha}$.

5.1. Consistency of the tests under the alternatives.

In this section, we provide the consistency of the tests under segregation and association alternatives.

THEOREM 4. *The test against $\mathcal{H}_\varepsilon^S$ which rejects for $R(\tau) > z_\alpha$ and the test against $\mathcal{H}_\varepsilon^A$ which rejects for $R(\tau) < z_{1-\alpha}$ are consistent for $\tau \in (0,1]$ and $\varepsilon \in \left(0, \sqrt{3}/3\right)$.*

In fact, the analysis of the means under the alternatives reveals more than what is required for consistency. Under segregation, the analysis indicates that $\mu_S(\tau,\varepsilon_1) < \mu_S(\tau,\varepsilon_2)$ for $\varepsilon_1 < \varepsilon_2$. On the other hand, under association, the analysis indicates that $\mu_A(\tau,\varepsilon_1) > \mu_A(\tau,\varepsilon_2)$ for $\varepsilon_1 < \varepsilon_2$.

5.2. Monte Carlo power analysis.

In this section, we asses the finite sample behaviour of the relative density using Monte Carlo simulations for testing CSR against segregation or association. We provide the kernel density estimates, empirical significance levels, and empirical power estimates under the null case and various segregation and association alternatives.

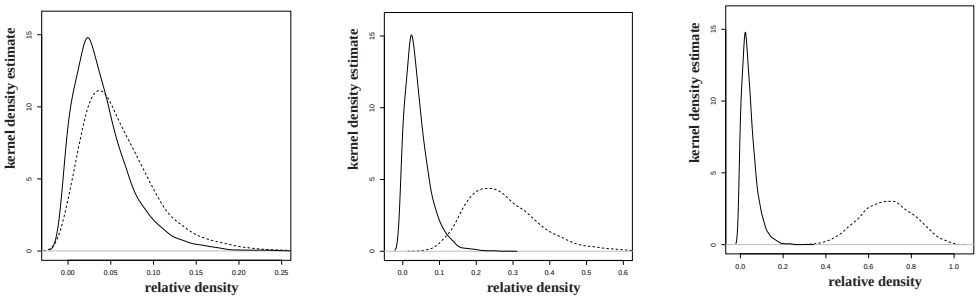

FIGURE 8: Kernel density estimates for the null (solid) and the segregation alternative $\mathcal{H}_\varepsilon^S$ (dashed) with $\tau = 1/2$, $n = 10$, $N = 10000$, and $\varepsilon = \sqrt{3}/8$ (left), $\varepsilon = \sqrt{3}/4$ (middle), and $\varepsilon = 2\sqrt{3}/7$ (right).

5.2.1. Monte Carlo power analysis for segregation alternatives.

In Figure 8, we present the kernel density estimates under $\mathcal{H}_0$ and $\mathcal{H}_\varepsilon^S$ with $\varepsilon = \sqrt{3}/8,\ \sqrt{3}/4,\ 2\sqrt{3}/7$. Observe that with $n = 10$, and $\varepsilon = \sqrt{3}/8$, the density estimates are very similar implying small power; and as ε gets larger, the separation between the null and alternative curves gets larger, hence the power gets larger. With $n = 10$, 10000 Monte Carlo replicates yield power estimates $\hat{\beta}_{\mathrm{mc}}^S(\varepsilon) = .0994,\ .9777,\ 1.000$, respectively. With $n = 100$ (figures not presented), there is more separation between the null and alternative curves at each ε, which implies that power increases as ε or n increases. With $n = 100$, 1000 Monte Carlo replicates yield $\hat{\beta}_{\mathrm{mc}}^S(\varepsilon) = .544,\ 1.000,\ 1.000$.

For a given alternative and sample size, we may consider analyzing the power of the test—using the asymptotic critical value (i.e., the normal approximation)—as a function of τ. The empirical significance levels and power estimates against $\mathcal{H}_{\sqrt{3}/8}^S$, $\mathcal{H}_{\sqrt{3}/4}^S$ as a function of τ for $n = 10$ are presented in Table 1. The empirical significance levels, $\hat{\alpha}_{n=10}$, are all greater than .05 with the smallest being .0868 at $\tau = 1.0$ which have the empirical power $\hat{\beta}_{10}\left(\sqrt{3}/8\right) = .2289$, $\hat{\beta}_{10}\left(\sqrt{3}/4\right) = .9969$. However, the empirical significance levels imply that $n = 10$ is not large enough for normal approximation. Notice that as n gets larger, the empirical significance levels get closer to .05 (except for $\tau = 0.1$), but still are all greater than .05, which indicates that for $n \leq 100$, the test is liberal in rejecting $\mathcal{H}_0$ against segregation. Furthermore, as n increases, for fixed ε the empirical power estimates increase, the empirical significance levels get closer to .05; and for fixed n as τ increases power estimates get larger. Therefore, for segregation, we recommend the use of large τ values ($\tau \lesssim 1.0$).

TABLE 1: The empirical significance levels and empirical power values under H_ε^S for $\varepsilon = \sqrt{3}/8,\ \sqrt{3}/4$ at $\alpha = .05$.

τ	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
$n = 10,\ N = 10000$										
$\hat{\alpha}_S(n)$	.0932	.1916	.1740	.1533	.1101	.0979	.1035	.0945	.0883	.0868
$\hat{\beta}_n^S(\tau, \sqrt{3}/8)$	.1286	.2630	.2917	.2811	.2305	.2342	.2526	.2405	.2334	.2289
$\hat{\beta}_n^S(\tau, \sqrt{3}/4)$	.5821	.9011	.9824	.9945	.9967	.9979	.9990	.9985	.9983	.9969
$n = 100,\ N = 1000$										
$\hat{\alpha}_S(n)$	.155	.101	.080	.077	.075	.066	.065	.063	.066	.069
$\hat{\beta}_n^S(\tau, \sqrt{3}/8)$	.574	.574	.612	.655	.709	.742	.774	.786	.793	.793

5.2.2. Monte Carlo power analysis for association alternatives.

In Figure 9, we present the kernel density estimates under $\mathcal{H}_0$ and $\mathcal{H}_\varepsilon^A$ with $\varepsilon = \sqrt{3}/21,\ \sqrt{3}/12,\ 5\sqrt{3}/24$ and $\tau = 0.5$. Observe that with $n = 10$, the density estimates are very similar for all ε values (with slightly more separation for larger ε) which implies small power. Ten thousand Monte Carlo replicates yield power estimates $\hat{\beta}_{\mathrm{mc}}^A \approx 0$. With $n = 100$ (figures not presented), there is more separation between the null and alternative curves at each ε, which implies that power increases as ε increases. One thousand Monte Carlo replicates yield $\hat{\beta}_{\mathrm{mc}}^A = .324,\ .634,\ .634$, respectively.

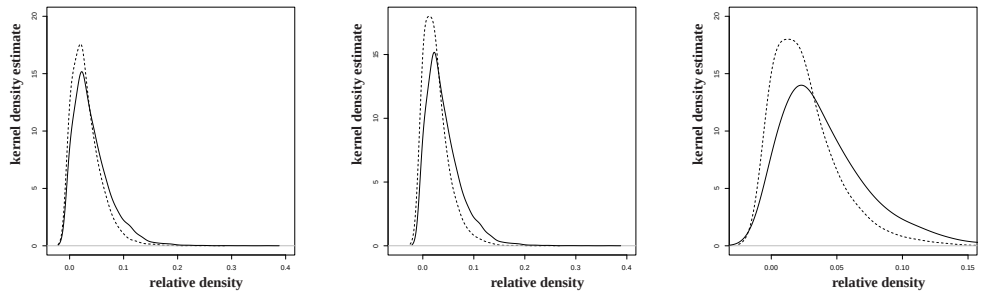

FIGURE 9: Kernel density estimates for the null (solid) and the association alternative $\mathcal{H}_\varepsilon^A$ (dashed) for $\tau = 1/2$ with $n = 10$, $N = 10000$ and $\varepsilon = \sqrt{3}/21$ (left), $\varepsilon = \sqrt{3}/12$ (middle), $\varepsilon = 5\sqrt{3}/24$ (right).

For a given alternative and sample size, we may consider analyzing the power of the test — using the asymptotic critical value—as a function of τ.

The empirical significance levels and power estimates against $\mathcal{H}_\varepsilon^A$, with $\varepsilon = \sqrt{3}/12,\ 5\sqrt{3}/24$ as a function of τ for $n = 10$, are presented in Table 2. The empirical significance level closest to .05 occurs at $\tau = .6$ (much smaller for other τ values), which have the empirical power $\hat{\beta}_{10}\big(\sqrt{3}/12\big) = .1181$, and $\hat{\beta}_{10}\big(5\sqrt{3}/24\big) = .1187$. However, the empirical significance levels imply that $n = 10$ is not large enough for the normal approximation. With $n = 100$, the empirical significance levels are approximately .05 for $\tau \geq .3$ and the highest empirical power is .997 at $\tau = 1.0$. Note that as n increases, the empirical power estimates increase for $\tau \geq .2$ and the empirical significance levels get closer to .05 for $\tau \geq .5$. This analysis indicates that in the one triangle case, the sample size should be large ($n \geq 100$) for the normal approximation to be appropriate. Moreover, the smaller the τ value, the larger the sample needed

for the normal approximation to be appropriate. Therefore, we recommend the use of large τ values ($\tau \lesssim 1.0$) for association.

TABLE 2: The empirical significance level and empirical power values under $\mathcal{H}_\varepsilon^A$ for $\varepsilon = 5\sqrt{3}/24$, $\sqrt{3}/12$, $\sqrt{3}/21$ with $N = 10000$, and $n = 10$ at $\alpha = .05$.

τ	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
$n = 10, N = 10000$										
$\hat{\alpha}_A(n)$	0	0	0	0	0	.0465	.0164	.0223	.0209	.0339
$\hat{\beta}_n^A(\tau, \sqrt{3}/12)$	0	0	0	0	0	.1181	.0569	.0831	.0882	.1490
$\hat{\beta}_n^A(\tau, 5\sqrt{3}/24)$	0	0	0	0	0	.1187	.0581	.0863	.0985	.1771
$n = 100, N = 1000$										
$\hat{\alpha}_A(n)$	.169	.075	.053	.047	.049	.044	.040	.044	.049	.049
$\hat{\beta}_n^A(\tau, \sqrt{3}/12)$	.433	.399	.460	.559	.687	.789	.887	.938	.977	.997

5.3. Pitman asymptotic efficiency under the alternatives.

The Pitman asymptotic efficiency (PAE) provides for an investigation of local asymptotic power around $\mathcal{H}_0$. This involves the limit as $n \to \infty$, as well as the limit as $\varepsilon \to 0$. See the proof of Theorem 3 for the ranges of τ and ε for which relative density is continuous as n goes to ∞. A detailed discussion of PAE can be found in Kendall & Stuart (1979) and van Eeden (1963). For segregation or association alternatives the PAE is given by

$$\text{PAE}(\rho_n(\tau)) = \frac{(\mu^{(k)}(\tau, \varepsilon = 0))^2}{\nu(\tau)},$$

where k is the minimum order of the derivative with respect to ε for which $\mu^{(k)}(\tau, \varepsilon = 0) \neq 0$. That is, $\mu^{(k)}(r, \varepsilon = 0) \neq 0$ but $\mu^{(l)}(\tau, \varepsilon = 0) = 0$ for $l = 1, 2, \ldots, k-1$. Then under segregation alternative $\mathcal{H}_\varepsilon^S$ and association alternative $\mathcal{H}_\varepsilon^A$, the PAE of $\rho_n(\tau)$ is given by

$$\text{PAE}^S(\tau) = \frac{(\mu_S''(\tau, \varepsilon = 0))^2}{\nu(\tau)} \quad \text{and} \quad \text{PAE}^A(\tau) = \frac{(\mu_A''(\tau, \varepsilon = 0))^2}{\nu(\tau)},$$

respectively, since $\mu_S'(\tau, \varepsilon = 0) = \mu_A'(\tau, \varepsilon = 0) = 0$. Equation (2) provides the denominator; the numerator requires $\mu_S(\tau, \varepsilon)$ and $\mu_A(\tau, \varepsilon)$ which are provided in Ceyhan, Priebe & Marchette (2004) where we only use the intervals of τ that do not vanish as $\varepsilon \to 0$.

In Figure 10, we present the PAE as a function of τ for both segregation and association. Notice that $\lim_{\tau \to 0} \text{PAE}^S(\tau) = 320/7 \approx 45.7$, $\text{argsup}_{\tau \in (0,1]} \text{PAE}^S(\tau) = 1.0$, and $\text{PAE}^S(\tau = 1) = 960/7 \approx 137.1$. Based on the PAE analysis, we suggest, for large n and small ε, choosing τ large (i.e., $\tau = 1$) for testing against segregation.

Notice that $\lim_{\tau \to 0} \text{PAE}^A(\tau) = 72000/7 \approx 10285.7$, $\text{PAE}^A(\tau = 1) = 61440/7 \approx 8777.1$, $\text{arginf}_{\tau \in (0,1]} \text{PAE}^A(\tau) \approx .46$ with $\text{PAE}^A(\tau \approx .46) \approx 6191.1$. Based on the asymptotic efficiency analysis, we suggest, for large n and small ε, choosing τ small for testing against association. However, for small and moderate values of n the normal approximation is not appropriate due to the skewness in the density of $\rho_n(\tau)$. Therefore, for small and moderate n, we suggest large τ values ($\tau \lesssim 1.0$).

5.4. The case with multiple Delaunay triangles.

Suppose $\mathcal{Y}$ is a finite collection of points in $\mathbb{R}^2$ with $|\mathcal{Y}| \geq 3$. Consider the Delaunay triangulation (assumed to exist) of $\mathcal{Y}$, where T_j denotes the jth Delaunay triangle, J denotes

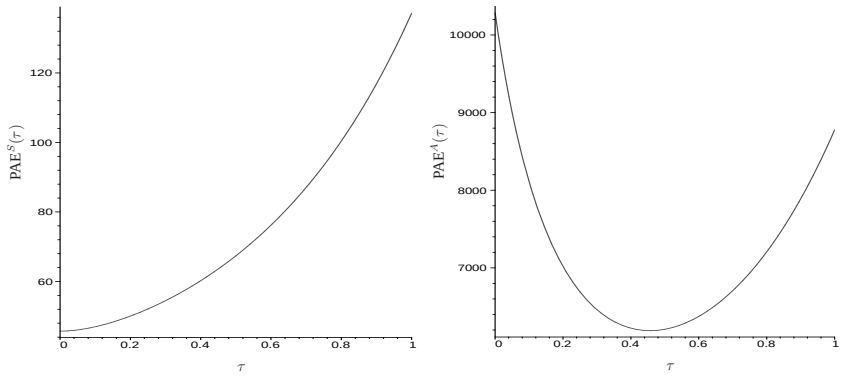

FIGURE 10: Pitman asymptotic efficiency curves against segregation (left) and association (right) as a function of τ. Notice that the axes of the plots are scaled differently.

the number of triangles, and $C_H(\mathcal{Y})$ denotes the convex hull of $\mathcal{Y}$. We wish to investigate $\mathcal{H}_0 : X_i \overset{\text{iid}}{\sim} \mathsf{U}(C_H(\mathcal{Y}))$ against segregation and association alternatives.

Figure 1 is the graph of realizations of $n = 1000$ observations which are independent and identically distributed according to $\mathsf{U}(C_H(\mathcal{Y}))$ for $|\mathcal{Y}| = 10$ and $J = 13$ and under segregation and association for the same $\mathcal{Y}$.

The digraph $\mathcal{D}$ is constructed using $N^\tau_{\text{CS}}(j, \cdot) = N^\tau_{\mathcal{Y}_j}(\cdot)$ as described in Section 3.1, where for $X_i \in T_j$ the three points in $\mathcal{Y}$ defining the Delaunay triangle T_j are used as $\mathcal{Y}_j$. Letting $w_j = A(T_j)/A(C_H(\mathcal{Y}))$ with $A(\cdot)$ being the area functional, we obtain the following as a corollary to Theorem 2.

COROLLARY 1. *The asymptotic null distribution for $\rho_n(\tau, J)$ conditional on $\mathcal{W} = \{w_1, \ldots, w_J\}$ for $\tau \in (0, 1]$ is given by $\mathsf{N}(\mu(\tau, J), \nu(\tau, J)/n)$ provided that $\nu(\tau, J) > 0$ with*

$$\mu(\tau, J) := \mu(\tau) \sum_{j=1}^{J} w_j^2 \quad \textit{and} \quad \nu(\tau, J) := \nu(\tau) \sum_{j=1}^{J} w_j^3 + 4\,\mu(\tau)^2 \left[\sum_{j=1}^{J} w_j^3 - \left(\sum_{j=1}^{J} w_j^2 \right)^2 \right],$$

where $\mu(\tau)$ and $\nu(\tau)$ are given by Equations (1) and (2), respectively.

By an appropriate application of Jensen's inequality, we see that $\sum_{j=1}^{J} w_j^3 \geq \left(\sum_{j=1}^{J} w_j^2\right)^2$. Therefore the covariance $\nu(\tau, J) = 0$ if and only if both $\nu(\tau) = 0$ and $\sum_{j=1}^{J} w_j^3 = \left(\sum_{j=1}^{J} w_j^2\right)^2$ hold.

Similarly, for the segregation (association) alternatives where $4\,\varepsilon^2/3 \times 100\%$ of the area around the vertices of each triangle is forbidden (allowed), we obtain the above asymptotic distribution of $\rho_n(\tau, J)$ with $\mu(\tau, J)$ being replaced by $\mu_S(\tau, J, \varepsilon)$, $\nu(\tau, J)$ by $\nu_S(\tau, J, \varepsilon)$, $\mu(\tau)$ by $\mu_S(\tau, \varepsilon)$, and $\nu(\tau)$ by $\nu_S(\tau, \varepsilon)$. Likewise for association.

The segregation (with $\delta = 1/16$, i.e., $\varepsilon = \sqrt{3}\,/8$), null, and association (with $\delta = 1/4$, i.e., $\varepsilon = \sqrt{3}\,/12$) realizations (from left to right) are depicted in Figure 1 with $n = 1000$. For the null realization, the p-value $p \geq .34$ for all τ values relative to the segregation alternative, also $p \geq .32$ for all τ values relative to the association alternative. For the segregation realization, we obtain $p \leq .021$ for all $\tau \geq .2$. For the association realization, we obtain $p \leq .02$ for all $\tau \geq .2$ and $p = .07$ at $\tau = 0.1$. Note that this is only for one realization of $\mathcal{X}_n$.

We repeat the null and alternative realizations 1000 times with $n = 100$ and $n = 500$ and estimate the significance levels and empirical power. The estimated values are presented in Table 3. With $n = 100$, the empirical significance levels are all greater than .05 and less than

.10 for $\tau \geq .6$ against both alternatives, much larger for other values. This analysis suggests that $n = 100$ is not large enough for normal approximation. With $n = 500$, the empirical significance levels are around .1 for $.3 \leq \tau < .5$ for segregation, and around (but slightly larger than) $.05$ for $\tau \geq .5$. Based on this analysis, we see that, against segregation, our test is liberal (less liberal for larger τ) in rejecting $\mathcal{H}_0$ for small and moderate n, against association it is slightly liberal for small and moderate n, and large τ values. For both alternatives, we suggest the use of large τ values. Observe that the poor performance of relative density in one-triangle case for association does not persist in multiple triangle case. In fact, for the multiple triangle case, $R(\tau)$ gets to be more appropriate for testing against association compared to testing against segregation.

The conditional test presented here is appropriate when $w_j \in \mathcal{W}$ are fixed quantities. An unconditional version requires the joint distribution of the number and relative size of Delaunay triangles when $\mathcal{Y}$ is, for instance, from a Poisson point process. Alas, this joint distribution is not available (Okabe, Boots & Sugihara 2000).

TABLE 3: The empirical significance levels and empirical power values under $H^S_{\sqrt{3}/8}$ and $H^A_{\sqrt{3}/12}$, $N = 1000$, $n = 100$, and $J = 13$, at $\alpha = .05$ for the realization of $\mathcal{Y}$ in Figure 1.

τ	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
					$n = 100$, $N = 1000$, $J = 13$					
$\hat{\alpha}_S(n, J)$	.496	.366	.302	.242	.190	.103	.102	.092	.095	.091
$\hat{\beta}^S_n(\tau, \sqrt{3}/8, J)$	.393	.429	.464	.512	.551	.578	.608	.613	.611	.604
$\hat{\alpha}_A(n, J)$	.726	.452	.322	.310	.194	.097	.081	.072	.063	.067
$\hat{\beta}^A_n(r, \sqrt{3}/12, J)$	.452	.426	.443	.555	.567	.667	.721	.809	.857	.906
					$n = 500$, $N = 1000$, $J = 13$					
$\hat{\alpha}_S(n, J)$	0.246	0.162	0.114	0.103	0.097	0.092	0.095	0.093	0.095	0.090
$\hat{\beta}^S_n(r, \sqrt{3}/8, J)$	0.829	0.947	0.982	0.988	0.995	0.995	0.997	0.998	0.997	0.997
$\hat{\alpha}_A(n, J)$	0.255	0.117	0.077	0.067	0.052	0.059	0.061	0.054	0.056	0.058
$\hat{\beta}^A_n(\tau, \sqrt{3}/12, J)$	0.684	0.872	0.953	0.991	0.999	1.000	1.000	1.000	1.000	1.000

5.4.1. Pitman asymptotic efficiency for multiple triangle case.

The PAE analysis is given for $J = 1$ in Section 5.3. For $J > 1$, the analysis will depend on both the number of triangles as well as the sizes of the triangles. So the optimal τ values with respect to these efficiency criteria for $J = 1$ are not necessarily optimal for $J > 1$, so the analyses need to be updated, conditional on the values of J and $\mathcal{W}$.

Under the segregation alternative $\mathcal{H}^S_\varepsilon$, the PAE of $\rho_n(\tau)$ is given by

$$\text{PAE}^S_J(\tau) = \frac{(\mu''_S(\tau, J, \varepsilon = 0))^2}{\nu(\tau, J)} = \frac{\left(\mu''_S(\tau, \varepsilon = 0) \sum_{j=1}^{J} w_j^2\right)^2}{\nu(\tau) \sum_{j=1}^{J} w_j^3 + 4\,\mu_S(\tau)^2 \left(\sum_{j=1}^{J} w_j^3 - \left(\sum_{j=1}^{J} w_j^2\right)^2\right)}.$$

Under association alternative $\mathcal{H}^A_\varepsilon$ the PAE of $\rho_n(\tau)$ is similar.

The PAE curves for $J = 13$ (as in Figure 1) are similar to the ones for the $J = 1$ case (see Figure 10), hence are omitted. Based on the Pitman asymptotic efficiency analysis, we suggest, for large n and small ε, choosing large τ for testing against segregation and small τ against association. However, for moderate and small n, we suggest large τ values for association due to the skewness of the density of $\rho_n(\tau)$.

5.5. Extension to higher dimensions.

The extension of N_{CS}^{τ} to $\mathbb{R}^d$ for $d > 2$ is straightforward. Let $\mathcal{Y} = \{\mathsf{y}_1, \mathsf{y}_2, \ldots, \mathsf{y}_{d+1}\}$ be $d+1$ points in general position. Denote the simplex formed by these $d+1$ points as $\mathcal{S}(\mathcal{Y})$. (A simplex is the simplest polytope in $\mathbb{R}^d$ having $d+1$ vertices, $d\,(d+1)/2$ edges and $d+1$ faces of dimension $(d-1)$.) For $\tau \in [0,1]$, define the τ-factor central similarity proximity regions as follows. Let φ_j be the face opposite vertex y_j for $j = 1, 2, \ldots, d+1$, and face regions $R(\varphi_1), \ldots, R(\varphi_{d+1})$ partition $\mathcal{S}(\mathcal{Y})$ into $d+1$ regions, namely the $d+1$ polytopes with vertices being the centre of mass together with d vertices chosen from $d+1$ vertices. For $x \in \mathcal{S}(\mathcal{Y}) \setminus \mathcal{Y}$, let $\varphi(x)$ be the face in whose region x falls; $x \in R(\varphi(x))$. (If x falls on the boundary of two face regions, we assign $\varphi(x)$ arbitrarily.) For $\tau \in (0,1]$, the τ-factor central similarity proximity region $N_{\mathrm{CS}}^{\tau}(x) = N_{\mathcal{Y}}^{\tau}(x)$ is defined to be the simplex $\mathcal{S}_\tau(x)$ with the following properties:

(i) $\mathcal{S}_\tau(x)$ has a face $\varphi_\tau(x)$ parallel to $\varphi(x)$ such that $\tau\, d(x, \varphi(x)) = d(\varphi_\tau(x), x)$, where $d(x, \varphi(x))$ is the Euclidean (perpendicular) distance from x to $\varphi(x)$,

(ii) $\mathcal{S}_\tau(x)$ has the same orientation as and is similar to $\mathcal{S}(\mathcal{Y})$,

(iii) x is at the centre of mass of $\mathcal{S}_\tau(x)$. Note that $\tau > 0$ implies that $x \in N_{\mathrm{CS}}^{\tau}(x)$.

For $\tau = 0$, define $N_{\mathrm{CS}}^{\tau}(x) = \{x\}$ for all $x \in \mathcal{S}(\mathcal{Y})$.

Theorem 1 generalizes, so that any simplex $\mathcal{S}$ in $\mathbb{R}^d$ can be transformed into a regular polytope (with edges being equal in length and faces being equal in area) preserving uniformity. Delaunay triangulation becomes Delaunay tessellation in $\mathbb{R}^d$, provided no more than $d+1$ points are cospherical (lying on the boundary of the same sphere). In particular, with $d = 3$, the general simplex is a tetrahedron (4 vertices, 4 triangular faces and 6 edges), which can be mapped into a regular tetrahedron (4 faces are equilateral triangles) with vertices $(0,0,0)\,(1,0,0)\,(1/2, \sqrt{3}\,/2, 0),\ (1/2, \sqrt{3}\,/6, \sqrt{6}\,/3)$.

Asymptotic normality of the U-statistic and consistency of the tests also hold for $d > 2$.

6. DISCUSSION AND CONCLUSIONS

In this article, we investigate the mathematical and statistical properties of a new proximity catch digraph (PCD) and its use in the analysis of spatial point patterns. The mathematical results are the detailed computations of means and variances of the U-statistics under the null and alternative hypotheses. These statistics require keeping good track of the geometry of the relevant neighbourhoods, and the complicated computations of integrals are done in the symbolic computation package MAPLE. The methodology is similar to that given by Ceyhan, Priebe & Wierman (2006). However, the results are simplified by the deliberate choices we make. For example, among many possibilities, the proximity map is defined in such a way that the distribution of the domination number and relative density is geometry invariant for uniform data in triangles, which allows the calculations on the standard equilateral triangle, rather than for each triangle separately.

We develop a technique for testing the patterns of segregation or association. There are many tests available for segregation and association in ecology literature. See (Dixon 1994) for a survey on these tests and relevant references. Two of the most commonly used tests are Pielou's χ^2 test of independence (Pielou 1961) and Ripley's test based on $K(t)$ and $L(t)$ functions (Ripley 1981). However, the test we introduce here is not comparable to either of them. Our test is a conditional test (conditional on a realization of J, the number of Delaunay triangles, and $\mathcal{W}$, the set of relative areas of the Delaunay triangles), and we require that the number of triangles J be fixed and relatively small compared to $n = |\mathcal{X}_n|$. Furthermore, our method deals with a slightly different type of data than most methods for examining spatial patterns. The sample size for one type of point (type $\mathcal{X}$ points) is much larger compared to the other (type $\mathcal{Y}$ points). This implies that in practice, $\mathcal{Y}$ could be stationary or have a much longer life span than members of $\mathcal{X}$. For

example, the geometric coordinates of a special type of fungi might constitute $\mathcal{X}$ points, while the geometric coordinates of trees from a species around which the fungi grow might be viewed as the $\mathcal{Y}$ points.

Based on the asymptotic analysis and finite sample performance of relative density of τ-factor central similarity PCD, we recommend large values of τ ($\tau \lesssim 1$), regardless of the sample size for segregation. For association, we recommend large values of τ ($\tau \lesssim 1$) for small to moderate sample sizes, and small values of τ ($\tau \gtrsim 0$) for large sample sizes. However, in a practical situation, we will not know the pattern in advance. So as an automatic data-based selection of τ to test CSR against segregation or association, one can start with $\tau = 1$, and if the relative density is found to be smaller than that under CSR (which is suggestive of association), use any $\tau \in [.8, 1.0]$ for small to moderate sample sizes ($n \lesssim 200$), and use $\tau \gtrsim 0$ (say $\tau = 0.1$) for large sample sizes $n > 200$. If the relative density is found to be larger than that under CSR (which is suggestive of segregation), then use large τ (any $\tau \in [.8, 1.0]$) regardless of the sample size. However, for large τ values, $\tau = 1$ has more geometric appeal than the rest, so it can be used when large τ is recommended.

Although the statistical analysis and the mathematical properties related to the τ-factor central similarity proximity catch digraph are done in $\mathbb{R}^2$, the extension to $\mathbb{R}^d$ with $d > 2$ is straightforward. Moreover, the geometry invariance, asymptotic normality of the U-statistic and consistency of the tests hold for $d > 2$.

APPENDIX

Proof of Theorem 1. Suppose $X \sim \mathsf{U}(T(\mathcal{Y}))$. A composition of translation, rotation, reflections, and scaling will take any given triangle $T(\mathcal{Y}) = T(\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3)$ to the basic triangle $T_b = T\big((0,0),(1,0),(c_1,c_2)\big)$ with $0 < c_1 \leq 1/2$, $c_2 > 0$ and $(1-c_1)^2 + c_2^2 \leq 1$. Furthermore, when X is also transformed in the same manner, say to X', then X' is uniform on T_b, i.e., $X' \sim \mathsf{U}(T_b)$. The transformation $\phi_e \colon \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\phi_e(u,v) = \left(u + \frac{1-2\,c_1}{\sqrt{3}}\,v, \frac{\sqrt{3}}{2\,c_2}\,v\right)$$

takes T_b to the equilateral triangle $T_e = T\big((0,0),(1,0),\big(1/2,\sqrt{3}\,/2\big)\big)$. Investigation of the Jacobian shows that ϕ_e also preserves uniformity. That is, $\phi_e(X') \sim \mathsf{U}(T_e)$. Furthermore, the composition of ϕ_e with the rigid motion transformations maps the boundary of the original triangle $T(\mathcal{Y})$ to the boundary of the equilateral triangle T_e, the median lines of $T(\mathcal{Y})$ to the median lines of T_e, and lines parallel to the edges of $T(\mathcal{Y})$ to lines parallel to the edges of T_e and straight lines that cross $T(\mathcal{Y})$ to the straight lines that cross T_e. Since the joint distribution of any collection of the h_{ij} involves only probability content of unions and intersections of regions bounded by precisely such lines, and the probability content of such regions is preserved since uniformity is preserved, the desired result follows. □

Derivation of $\mu(\tau)$ and $\nu(\tau)$. Let M_j be the midpoint of edge e_j for $j = 1,2,3$, let M_C be the centre of mass, and $T_s := T(\mathsf{y}_1, M_3, M_C)$. Let $\mathbf{X_i} = (X_i, Y_i)$ for $i = 1,2,3$ be three random points from $\mathsf{U}(T(\mathcal{Y}))$, and let $\mathbf{x_i} = (x_i, y_i)$ be their realizations. Notice that the bivariate variables are denoted in boldface, random variables are denoted in upper case, and realizations of random variables are denoted in lower case characters. By symmetry, $\mu(\tau) = P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1})\big) = 6\,P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}),\, \mathbf{X_1} \in T_s\big)$. Then

$$P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}),\, \mathbf{X_1} \in T_s\big) = \int_0^{1/2} \int_0^{x_1/\sqrt{3}} \frac{A(N_{\mathrm{CS}}^{\tau}(\mathbf{x_1}))}{A(T(\mathcal{Y}))^2}\, dy_1\, dx_1 = \tau^2/36,$$

where $A\big(N_{\mathrm{CS}}^{\tau}(\mathbf{x_1})\big) = 3\sqrt{3}\,\tau^2\,y_1^2$ and $A(T(\mathcal{Y})) = \sqrt{3}\,/4$. Hence $\mu(\tau) = \tau^2/6$.

Next, we find the asymptotic variance. Let

$$\begin{aligned} P_{2N}^{\tau} &:= P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset N_{\mathrm{CS}}^{\tau}(\mathbf{X_1})\big), \\ P_{2G}^{\tau} &:= P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset \Gamma_1^{\tau}(\mathbf{X_1})\big) \quad \text{and} \\ P_{M}^{\tau} &:= P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}), \mathbf{X_3} \in \Gamma_1^{\tau}(\mathbf{X_1})\big), \end{aligned}$$

where $\Gamma_1^{\tau}(x)$ is the Γ_1-*region* of x based on N_{CS}^{τ} and defined as $\Gamma_1^{\tau}(x) := \{y \in T(\mathcal{Y}) : x \subset N_{\mathrm{CS}}^{\tau}(y)\}$. (See Ceyhan, Priebe & Wierman 2006 for more on Γ_1-regions.)

Then $\mathrm{cov}\,[h_{12}, h_{13}] = \mathrm{E}\,[h_{12}\, h_{13}] - \mathrm{E}\,[h_{12}]\mathrm{E}\,[h_{13}]$ where

$$\begin{aligned} \mathrm{E}\,[h_{12}\, h_{13}] &= P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset N_{\mathrm{CS}}^{\tau}(\mathbf{X_1})\big) \\ &\quad + 2\, P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}), \mathbf{X_3} \in \Gamma_1^{\tau}(\mathbf{X_1})\big) \\ &\quad + P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset \Gamma_1^{\tau}(\mathbf{X_1})\big) \\ &= P_{2N}^{\tau} + 2\, P_{M}^{\tau} + P_{2G}^{\tau}. \end{aligned}$$

Hence $\nu(\tau) = \mathrm{cov}\,[h_{12}, h_{13}] = \big(P_{2N}^{\tau} + 2\, P_{M}^{\tau} + P_{2G}^{\tau}\big) - [2\,\mu(\tau)]^2$.

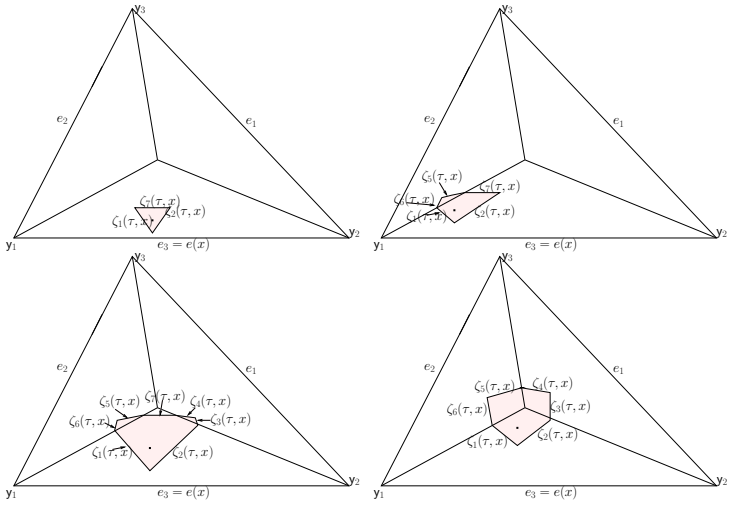

FIGURE 11: The prototypes of the four cases of $\Gamma_1^{\tau}(\mathbf{x_1})$ for $\mathbf{x_1} \in T(\mathrm{y}_1, M_3, M_C)$ with $\tau = 1/2$.

To find the covariance, we need to find the possible types of $\Gamma_1^{\tau}(\mathbf{x_1})$ and $N_{\mathrm{CS}}^{\tau}(\mathbf{x_1})$ for $\tau \in (0, 1]$. There are four cases regarding $\Gamma_1^{\tau}(\mathbf{x_1})$ and one case for $N_{\mathrm{CS}}^{\tau}(\mathbf{x_1})$. See Figure 11 for the prototypes of these four cases of $\Gamma_1^{\tau}(\mathbf{x_1})$ where, for $\mathbf{x_1} = (x_1, y_1) \in T(\mathcal{Y})$, the explicit forms of $\zeta_j(\tau, x)$ are

$$\begin{aligned} \zeta_1(\tau, x) &= \frac{y_1 + \sqrt{3}\,(x_1 - x)}{(1 + 2\,\tau)}, & \zeta_2(\tau, x) &= \frac{y_1 - \sqrt{3}\,(x_1 - x)}{(1 + 2\,\tau)}, \\ \zeta_3(\tau, x) &= \frac{\sqrt{3}\,x(\tau + 1) + y_1 - \sqrt{3}\,(x_1 + \tau)}{(1 - \tau)}, & \zeta_4(\tau, x) &= \frac{\sqrt{3}\,\tau(x - 1) - 2\,y_1}{2 + \tau}, \\ \zeta_5(\tau, x) &= \frac{\tau\,\sqrt{3}\,x + 2\,y_1}{2 + \tau}, & \zeta_6(\tau, x) &= \frac{\sqrt{3}\,[(x_1 + y_1) - x(1 + \tau)]}{(1 - \tau)}, \\ \zeta_7(\tau, x) &= \frac{y_1}{1 - \tau}. \end{aligned}$$

Each case j corresponds to the region R_j in Figure 12, where

$$q_1(x) = \frac{1 - \tau}{2\,\sqrt{3}}, \quad q_2(x) = \frac{(1 - x)(1 - \tau)}{\sqrt{3}\,(1 + \tau)}, \quad q_3(x) = \frac{(1 - \tau)x}{\sqrt{3}\,(1 + \tau)}, \quad \text{and} \quad s_1 = (1 - \tau)/2.$$

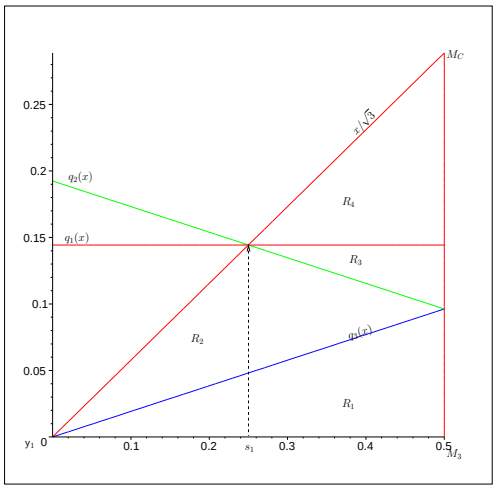

FIGURE 12: The regions corresponding to the prototypes of the four cases with $\tau = 1/2$.

The explicit forms of R_j, $j = 1, \ldots, 4$ are as follows:

$$\begin{aligned} R_1 &= \{(x,y) \in [0,1/2] \times [0, q_3(x)]\}, \\ R_2 &= \{(x,y) \in [0,s_1] \times \left[q_3(x), x/\sqrt{3}\,\right] \cup [s_1,1/2] \times [q_3(x), q_2(x)]\}, \\ R_3 &= \{(x,y) \in [s_1,1/2] \times [q_2(x), q_1(x)]\}, \\ R_4 &= \left\{(x,y) \in [s_1,1/2] \times \left[q_1(x), x/\sqrt{3}\,\right]\right\}. \end{aligned}$$

By symmetry,

$$P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset N_{\mathrm{CS}}^{\tau}(\mathbf{X_1})\big) = 6\, P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}),\ \mathbf{X_1} \in T_s\big),$$

and

$$P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}),\ \mathbf{X_1} \in T_s\big) = \int_0^{1/2} \int_0^{x_1/\sqrt{3}} \frac{A(N_{\mathrm{CS}}^{\tau}(\mathbf{x_1}))^2}{A(T(\mathcal{Y}))^3}\, dy_1\, dx_1 = \tau^4/90,$$

where $A(N_{\mathrm{CS}}^{\tau}(\mathbf{x_1})) = 3\sqrt{3}\,\tau^2\, y_1^2$. Hence,

$$P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset N_{\mathrm{CS}}^{\tau}(\mathbf{X_1})\big) = \tau^4/15.$$

Next, by symmetry,

$$P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset \Gamma_1^{\tau}(\mathbf{X_1})\big) = 6\, P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset \Gamma_1^{\tau}(\mathbf{X_1}), \mathbf{X_1} \in T_s\big),$$

and

$$P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset \Gamma_1^{\tau}(\mathbf{X_1}), \mathbf{X_1} \in T_s\big) = \sum_{j=1}^{4} P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset \Gamma_1^{\tau}(\mathbf{X_1}), \mathbf{X_1} \in R_j\big).$$

For $\mathbf{x_1} = (x_1, y_1) \in R_1$,

$$\begin{aligned} P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset \Gamma_1^{\tau}(\mathbf{X_1}), \mathbf{X_1} \in R_1\big) &= \int_0^{1/2} \int_0^{q_3(x)} \frac{A(\Gamma_1^{\tau}(\mathbf{x_1}))^2}{A(T(\mathcal{Y}))^3}\, dy_1\, dx_1 \\ &= \frac{\tau^4(1-\tau)}{90\,(1+2\,\tau)^2(1+\tau)^5}, \end{aligned}$$

where

$$A\big(\Gamma_1^\tau(\mathbf{x_1})\big) = 3\,\frac{\tau^2\sqrt{3}\,y^2}{(\tau-1)^2(2\,\tau+1)}.$$

For $\mathbf{x_1} = (x_1, y_1) \in R_2$,

$$\begin{aligned} &P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset \Gamma_1^\tau(\mathbf{X_1}),\ \mathbf{X_1} \in R_2\big) \\ &= \int_0^{s_1}\int_{q_3(x_1)}^{x_1/\sqrt{3}} \frac{A(\Gamma_1^\tau(\mathbf{x_1}))^2}{A(T(\mathcal{Y}))^3}\,dy_1\,dx_1 + \int_{s_1}^{1/2}\int_{q_3(x_1)}^{q_2(x_1)} \frac{A(\Gamma_1^\tau(\mathbf{x_1}))^2}{A(T(\mathcal{Y}))^3}\,dy_1\,dx_1 \\ &= \frac{\tau^5(4\,\tau^6+6\,\tau^5-12\,\tau^4-21\,\tau^3+14\,\tau^2+40\,\tau+20)(1-\tau)}{45\,(2\,\tau+1)^2(\tau+2)^2(\tau+1)^5}, \end{aligned}$$

where

$$A\big(\Gamma_1^\tau(\mathbf{x_1})\big) = \frac{3\,\sqrt{3}\,(x_1^2\tau+2\,\sqrt{3}\,x_1y_1\tau-y_1^2\tau-x_1^2+2\sqrt{3}\,x_1y_1-3\,y_1^2)\tau}{4\,(1-\tau)(2\,\tau+1)(\tau+2)}.$$

For $\mathbf{x_1} = (x_1, y_1) \in R_3$,

$$\begin{aligned} P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset \Gamma_1^\tau(\mathbf{X_1}), \mathbf{X_1} \in R_3\big) &= \int_{s_1}^{1/2}\int_{q_2(x_1)}^{q_1(x_1)} \frac{A(\Gamma_1^\tau(\mathbf{x_1}))^2}{A(T(\mathcal{Y}))^3}\,dy_1\,dx_1 \\ &= \frac{\tau^6(1-\tau)(6\,\tau^6-35\,\tau^4+130\,\tau^2+160\,\tau+60)}{90(2\,\tau+1)^2(\tau+2)^2(\tau+1)^5}, \end{aligned}$$

where

$$\begin{aligned} A\big(\Gamma_1^\tau(\mathbf{x_1})\big) &= \frac{-3\,\sqrt{3}\tau(2\,x_1^2\tau^2+2y_1^2\tau^2-4\,x_1^2\tau-2\,x_1\,\tau^2+4\,y_1^2\tau+2\,\sqrt{3}\,y_1\,\tau^2}{4\,(2\,\tau+1)(\tau-1)^2(\tau+2)} \\ &\quad + \frac{2\,x_1^2+4\,x\,\tau+6\,y_1^2+\tau^2-2\,x_1-2\,\sqrt{3}\,y_1-2\,\tau+1)}{4\,(2\,\tau+1)(\tau-1)^2(\tau+2)}. \end{aligned}$$

For $\mathbf{x_1} = (x_1, y_1) \in R_4$,

$$\begin{aligned} P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset \Gamma_1^\tau(\mathbf{X_1}), \mathbf{X_1} \in R_4\big) &= \int_{s_1}^{1/2}\int_{q_1(x_1)}^{x_1/\sqrt{3}} \frac{A(\Gamma_1^\tau(\mathbf{x_1}))^2}{A(T(\mathcal{Y}))^3}\,dy_1\,dx_1 \\ &\quad + \int_{s_4}^{s_5}\int_{q_3(x_1)}^{x_1/\sqrt{3}} \frac{A(\Gamma_1^\tau(\mathbf{x_1}))^2}{A(T(\mathcal{Y}))^3}\,dy_1\,dx_1 \\ &\quad + \int_{s_5}^{1/2}\int_{q_3(x_1)}^{q_{12}(x_1)} \frac{A(\Gamma_1^\tau(\mathbf{x_1}))^2}{A(T(\mathcal{Y}))^3}\,dy_1\,dx_1 \\ &= \frac{\tau^6(\tau^2-5\,\tau+10)}{15\,(2\,\tau+1)^2(\tau+2)^2}, \end{aligned}$$

where

$$A\big(\Gamma_1^\tau(\mathbf{x_1})\big) = \frac{-\sqrt{3}\,\tau(3\,x_1^2+3\,y_1^2-3\,x_1-\sqrt{3}\,y_1-\tau+1)}{2\,(2\,\tau+1)(\tau+2)}.$$

So

$$\begin{aligned} P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset \Gamma_1^\tau(\mathbf{X_1})\big) &= 6\left(\frac{-(\tau^2-7\,\tau-2)\tau^4}{90\,(\tau+1)(2\,\tau+1)(\tau+2)}\right) \\ &= \frac{-(\tau^2-7\,\tau-2)\tau^4}{15\,(\tau+1)(2\,\tau+1)(\tau+2)}. \end{aligned}$$

Furthermore, by symmetry,

$$P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}), \mathbf{X_3} \in \Gamma_1^{\tau}(\mathbf{X_1})\big) = 6\left(\sum_{j=1}^{4} P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}), \mathbf{X_3} \in \Gamma_1^{\tau}(\mathbf{X_1}), \mathbf{X_1} \in R_j\big)\right),$$

where $P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}), \mathbf{X_3} \in \Gamma_1^{\tau}(\mathbf{X_1}), \mathbf{X_1} \in R_j\big)$ can be calculated with the same regions of integration with integrand being replaced by

$$\frac{A(N_{\mathrm{CS}}^{\tau}(\mathbf{x_1}))A(\Gamma_1^{\tau}(\mathbf{x_1}))}{A(T(\mathcal{Y}))^3}.$$

Then

$$\begin{aligned} P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}), \mathbf{X_3} \in \Gamma_1^{\tau}(\mathbf{X_1})\big) &= 6\left(\frac{(2\,\tau^4 - 3\,\tau^3 - 4\,\tau^2 + 10\,\tau + 4)\tau^4}{180\,(2\,\tau + 1)(\tau + 2)}\right) \\ &= \frac{(2\,\tau^4 - 3\,\tau^3 - 4\,\tau^2 + 10\,\tau + 4)\tau^4}{30\,(2\,\tau + 1)(\tau + 2)}. \end{aligned}$$

Hence

$$\mathrm{E}\,[h_{12}\,h_{13}] = \frac{\tau^4(2\,\tau^5 - \tau^4 - 5\,\tau^3 + 12\,\tau^2 + 28\,\tau + 8)}{15\,(\tau + 1)(2\,\tau + 1)(\tau + 2)}.$$

Therefore,

$$\nu(\tau) = \frac{\tau^4(6\,\tau^5 - 3\,\tau^4 - 25\,\tau^3 + \tau^2 + 49\,\tau + 14)}{45\,(\tau + 1)(2\,\tau + 1)(\tau + 2)}.$$

Sketch of the Proof of Theorem 3. Under the alternatives, i.e., $\varepsilon > 0$, $\rho_n(\tau)$ is a U-statistic with the same symmetric kernel h_{ij} as in the null case. The mean $\mu_S(\tau, \varepsilon) = \mathrm{E}_{\,\varepsilon}[\rho_n(\tau)] = \mathrm{E}_{\,\varepsilon}[h_{12}]/2$ (and $\mu_A(\tau, \varepsilon)$), now a function of both τ and ε, is again in $[0, 1]$. $\nu_S(\tau, \varepsilon) = \mathrm{cov}_{\,\varepsilon}[h_{12}, h_{13}]$ (and $\nu_A(\tau, \varepsilon)$), also a function of both τ and ε, is bounded above by $1/4$, as before. Thus asymptotic normality obtains provided that $\nu_S(\tau, \varepsilon) > 0$ ($\nu_A(\tau, \varepsilon) > 0$); otherwise $\rho_n(\tau)$ is degenerate. The explicit forms of $\mu_S(\tau, \varepsilon)$ and $\mu_A(\tau, \varepsilon)$ are given, defined piecewise, in Ceyhan, Priebe & Marchette (2004). Note that under H_{ε}^{S},

$$\nu_S(\tau, \varepsilon) > 0$$

for

$$(\tau, \varepsilon) \in \Big((0, 1] \times \big(0, 3\,\sqrt{3}\,/10\big]\Big) \bigcup \left(\left(\frac{2\big(\sqrt{3} - 3\,\varepsilon\big)}{4\,\varepsilon - \sqrt{3}}, 1\right] \times \big(3\,\sqrt{3}\,/10, \sqrt{3}\,/3\big)\right),$$

and under $\mathcal{H}_{\varepsilon}^{A}$,

$$\nu_A(\tau, \varepsilon) > 0 \quad \text{for } (\tau, \varepsilon) \in (0, 1] \times \big(0, \sqrt{3}\,/3\big).$$

Sketch of Proof of Theorem 4. Since the variance of the asymptotically normal test statistic, under both the null and the alternative cases, converges to 0 as $n \to \infty$ (or is degenerate), it remains to show that the mean under the null, $\mu(\tau) = \mathrm{E}\,[\rho_n(\tau)]$, is less than (greater than) the mean under the alternative, $\mu_S(\tau, \varepsilon) = \mathrm{E}_{\,\varepsilon}[\rho_n(\tau)]$ ($\mu_A(\tau, \varepsilon)$) against segregation (association) for $\varepsilon > 0$. Whence it will follow that power converges to 1 as $n \to \infty$.

It is possible, albeit tedious, to compute $\mu_S(\tau, \varepsilon)$ and $\mu_A(\tau, \varepsilon)$ under the two alternatives. The calculations are deferred to the technical report by Ceyhan, Priebe & Marchette (2004) due to its extreme length and technicality, and the resulting explicit forms are provided in the Appendix of that report. Detailed analysis of $\mu_S(\tau, \varepsilon)$ and $\mu_A(\tau, \varepsilon)$ indicates that under segregation

$\mu_S(\tau,\varepsilon) > \mu(\tau)$ for all $\varepsilon > 0$ and $\tau \in (0,1]$. Likewise, detailed analysis of $\mu_A(\tau,\varepsilon)$ indicates that under association $\mu_A(\tau,\varepsilon) < \mu(\tau)$ for all $\varepsilon > 0$ and $\tau \in (0,1]$. We direct the reader to the technical report for the details of the calculations. Hence the desired result follows for both alternatives.

Notice that under the association alternatives any $\tau \in (0,1]$ yields asymptotic normality for all $\varepsilon \in \left(0, \sqrt{3}/3\right)$, while under the segregation alternatives only $\tau = 1$ yields this universal asymptotic normality.

Proof of Corollary 1. In the multiple triangle case,

$$\begin{aligned}\mu(\tau,J) &= \mathrm{E}\left[\rho_n(\tau,J)\right] = \frac{1}{n\,(n-1)}\sum\sum_{i<j}\mathrm{E}\left[h_{ij}\right] \\ &= \frac{1}{2}\mathrm{E}\left[h_{12}\right] = \mathrm{E}\left[\mathbf{I}(\mathbf{X_1},\mathbf{X_2}) \in \mathcal{A})\right] \\ &= P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1})\big).\end{aligned}$$

But, by definition of $N_{\mathrm{CS}}^{\tau}(\,\cdot\,)$, $\mathbf{X_2} \notin N_{\mathrm{CS}}^{\tau}(\mathbf{X_1})$ a.s. if $\mathbf{X_1}$ and $\mathbf{X_2}$ are in different triangles. So by the law of total probability

$$\begin{aligned}\mu(\tau,J) &:= P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1})\big) \\ &= \sum_{j=1}^{J} P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}) \,|\, \{\mathbf{X_1},\mathbf{X_2}\} \subset T_j\big)\, P\big(\{\mathbf{X_1},\mathbf{X_2}\} \subset T_j\big) \\ &= \sum_{j=1}^{J} \mu(\tau)\, P\big(\{\mathbf{X_1},\mathbf{X_2}\} \subset T_j\big) \\ &\qquad\qquad \text{(since } P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}) \,|\, \{\mathbf{X_1},\mathbf{X_2}\} \subset T_j\big) = \mu(\tau)\text{)} \\ &= \mu(\tau) \sum_{j=1}^{J} \big(A(T_j)/A(C_H(\mathcal{Y}))\big)^2 \\ &\qquad\qquad \text{(since } P\big(\{\mathbf{X_1},\mathbf{X_2}\} \subset T_j\big) = \big(A(T_j)/A(C_H(\mathcal{Y}))\big)^2\text{).}\end{aligned}$$

Letting $w_j := A(T_j)/A(C_H(\mathcal{Y}))$, we get $\mu(\tau,J) = \mu(\tau)\cdot\left(\sum_{j=1}^{J} w_j^2\right)$ where $\mu(\tau)$ is given by Equation (1).

Furthermore, the asymptotic variance is

$$\begin{aligned}\nu(\tau,J) &= \mathrm{E}\left[h_{12}\,h_{13}\right] - \mathrm{E}\left[h_{12}\right]\mathrm{E}\left[h_{13}\right] \\ &= P\big(\{\mathbf{X_2},\mathbf{X_3}\} \subset N_{\mathrm{CS}}^{\tau}(\mathbf{X_1})\big) \\ &\quad + 2\,P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}), \mathbf{X_3} \in \Gamma_1^{\tau}(\mathbf{X_1})\big) \\ &\quad + P\big(\{\mathbf{X_2},\mathbf{X_3}\} \subset \Gamma_1^{\tau}(\mathbf{X_1})\big) - 4(\mu(\tau,J))^2.\end{aligned}$$

Then for $J > 1$, we have

$$\begin{aligned}&P\big(\{\mathbf{X_2},\mathbf{X_3}\} \subset N_{\mathrm{CS}}^{\tau}(\mathbf{X_1})\big) \\ &= \sum_{j=1}^{J} P\big(\{\mathbf{X_2},\mathbf{X_3}\} \subset N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}) \,|\, \{\mathbf{X_1},\mathbf{X_2},\mathbf{X_3}\} \subset T_j\big)\, P\big(\{\mathbf{X_1},\mathbf{X_2},\mathbf{X_3}\} \subset T_j\big) \\ &= \sum_{j=1}^{J} P_{2N}^{\tau}\left(A(T_j)/A(C_H(\mathcal{Y}))\right)^3 = P_{2N}^{\tau}\left(\sum_{j=1}^{J} w_j^3\right).\end{aligned}$$

Similarly, $P\big(\mathbf{X_2} \in N_{\mathrm{CS}}^{\tau}(\mathbf{X_1}), \mathbf{X_3} \in \Gamma_1^{\tau}(\mathbf{X_1})\big) = P_M^{\tau}\big(\sum_{j=1}^{J} w_j^3\big)$ and $P\big(\{\mathbf{X_2}, \mathbf{X_3}\} \subset \Gamma_1^{\tau}(\mathbf{X_1})\big) = P_{2G}^{\tau}\big(\sum_{j=1}^{J} w_j^3\big)$, hence, $\nu(\tau, J) = \big(P_{2N}^{\tau} + 2\,P_M^{\tau} + P_{2G}^{\tau}\big)\big(\sum_{j=1}^{J} w_j^3\big) - 4\mu(\tau, J)^2 = \nu(\tau)\big(\sum_{j=1}^{J} w_j^3\big) + 4\mu(\tau)^2\big(\sum_{j=1}^{J} w_j^3 - \big(\sum_{j=1}^{J} w_j^2\big)^2\big)$, so conditional on $\mathcal{W}$, if $\nu(\tau, J) > 0$ then $\sqrt{n}\left(\rho_n(\tau) - \tilde{\mu}(\tau)\right) \xrightarrow{\mathcal{L}} \mathsf{N}(0, \nu(\tau, J))$.

ACKNOWLEDGEMENTS

This work was partially supported by a U.S. Office of Naval Research Grant and a U.S. Defence Advanced Research Projects Agency Grant. We also thank anonymous referees, whose constructive comments and suggestions greatly improved the presentation and flow of the article.

REFERENCES

E. Ceyhan & C. E. Priebe (2003). Central similarity proximity maps in Delaunay tessellations. In *ASA Proceedings of the Joint Statistical Meetings*, 840–845.

E. Ceyhan & C. E. Priebe (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.

E. Ceyhan, C. E. Priebe & D. J. Marchette (2004). Relative density of random τ-factor proximity catch digraph for testing spatial patterns of segregation and association. *Technical Report 645, Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, Maryland 21218, USA*.

E. Ceyhan, C. E. Priebe & J. C. Wierman (2006). Relative density of the random r-factor proximity catch digraphs for testing spatial patterns of segregation and association. *Computational Statistics and Data Analysis*, 50, 1925–1964.

D. A. Coomes, M. Rees & L. Turnbull (1999). Identifying aggregation and association in fully mapped spatial data. *Ecology*, 80, 554–565.

N. A. C. Cressie (1991). *Statistics for Spatial Data*. Wiley, New York.

J. DeVinney, C. E. Priebe, D. J. Marchette & D. Socolinsky (2002). Random walks and catch digraphs in classification. In *Computing Science and Statistics, Volume 34: Montréal, Québec* (Edward J. Wegman & Amy Braverman, eds.), Interface Foundation of North America, Fairfax, Virginia; online at http://www.galaxy.gmu.edu/interface/I02/I2002Proceedings/DeVinneyJason/DeVinneyJason.paper.pdf

P. J. Diggle (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, New York.

P. M. Dixon (1994). Testing spatial segregation using a nearest-neighbour contingency table. *Ecology*, 75, 1940–1948.

P. M. Dixon (2002). Nearest-neighbour contingency table analysis of spatial segregation for several species. *Ecoscience*, 9, 142–151.

N. J. Gotelli & G. R. Graves (1996). *Null Models in Ecology*. Smithsonian Institution Press, Washington, DC.

D. M. Hamill & S. J. Wright (1986). Testing the dispersion of juveniles relative to adults: A new analytical method. *Ecology*, 67, 952–957.

S. Janson, T. Łuczak & A. Rucinński (2000). *Random Graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley, New York.

J. W. Jaromczyk & G. T. Toussaint (1992). Relative neighbourhood graphs and their relatives. *Proceedings of IEEE*, 80, 1502–1517.

M. G. Kendall & A. Stuart (1979). *The Advanced Theory of Statistics, Volume 2*, 4th edition. Griffin, London.

S. N. Lahiri (1996). On consistency of estimators based on spatial data under infill asymptotics. *Sankhyā: The Indian Journal of Statistics, Series A*, 58, 403–417.

E. L. Lehmann (1988). *Nonparametrics: Statistical Methods Based on Ranks*. Prentice-Hall, Upper Saddle River, New Jersey.

D. J. Marchette & C. E. Priebe (2003). Characterizing the scale dimension of a high dimensional classification problem. *Pattern Recognition*, 36, 45–60.

S. H. Nanami, H. Kawaguchi & T. Yamakura (1999). Dioecy-induced spatial patterns of two codominant tree species, podocarpus nagi and neolitsea aciculata. *Journal of Ecology*, 87, 678–687.

A. Okabe, B. Boots & K. Sugihara (2000). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Wiley, New York.

E. C. Pielou (1961). Segregation and symmetry in two-species populations as studied by nearest-neighbour relationships. *Journal of Ecology*, 49, 255–269.

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

C. E. Priebe, D. J. Marchette, J. G. DeVinney & D. Socolinsky (2003). Classification using class cover catch digraphs. *Journal of Classification*, 20, 3–23.

C. E. Priebe, J. L. Solka, D. J. Marchette, B. T. Clark (2003). Class cover catch digraphs for latent class discovery in gene expression monitoring by DNA microarrays. *Computational Statistics and Data Analysis on Visualization*, 43, 621–632.

B. D. Ripley (1981). *Spatial Statistics*. Wiley, New York.

G. T. Toussaint (1980). The relative neighbourhood graph of a finite planar set. *Pattern Recognition*, 12, 261–268.

C. van Eeden (1963). The relation between Pitman's asymptotic relative efficiency of two tests and the correlation coefficient between their test statistics. *The Annals of Mathematical Statistics*, 34, 1442–1451.

Received 19 June 2005
Accepted 28 August 2006

Elvan CEYHAN: elceyhan@ku.edu.tr
Department of Mathematics
Koç University, Sarıyer
TR-34450 Istanbul, Turkey

Carey E. PRIEBE: cep@jhu.edu
David J. MARCHETTE: marchettedj@nswc.navy.mil
Department of Applied Mathematics and Statistics
The Johns Hopkins University
Baltimore, MD 21218, USA