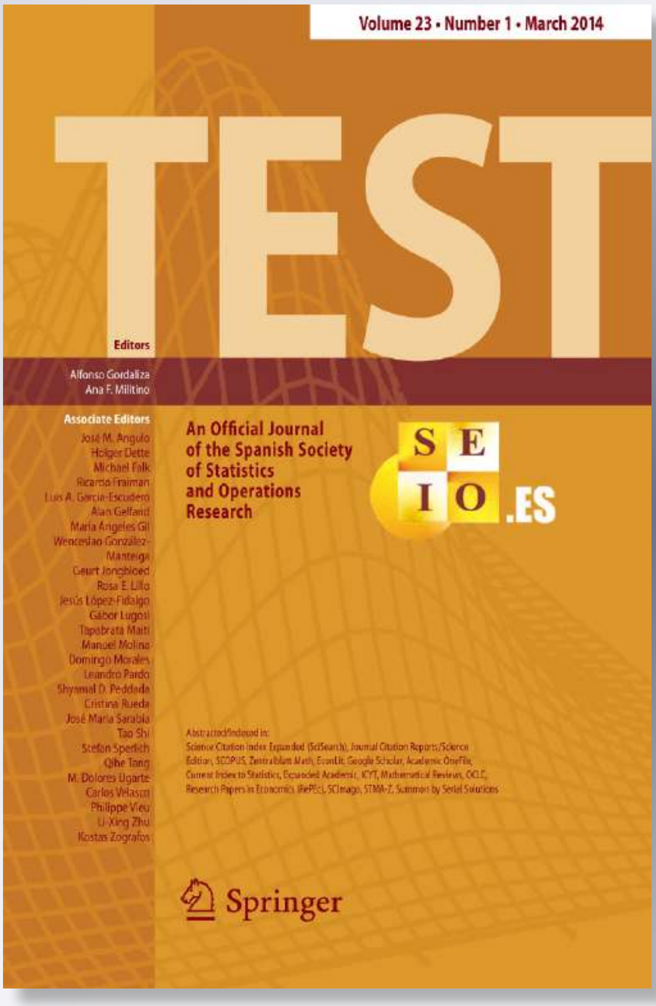
# *Comparison of relative density of two random geometric digraph families in testing spatial clustering*

**Elvan Ceyhan**

**TEST**
An Official Journal of the Spanish Society of Statistics and Operations Research

ISSN 1133-0686
Volume 23
Number 1

TEST (2014) 23:100-134
DOI 10.1007/s11749-013-0344-4


Volume 23 • Number 1 • March 2014
TEST
Editors
Alfonso Gordaliza
Ana F. Militino
Associate Editors
José M. Angulo
Holger Dette
Michael Falk
Ricardo Fraiman
Luis A. García-Escudero
Alan Gelfand
María Angeles Gil
Wenceslao González-Manteiga
Geurt Jongbloed
Rosa E. Lillo
Jesús López-Fidalgo
Gábor Lugosi
Tapabrata Maiti
Manuel Molina
Domingo Morales
Leandro Pardo
Shyamal D. Peddada
Cristina Rueda
José María Sarabia
Tao Shi
Stefan Sperlich
Qihe Tang
M. Dolores Ugarte
Carlos Velasco
Philippe Vieu
Li-Xing Zhu
Kostas Zografos
An Official Journal of the Spanish Society of Statistics and Operations Research
S E I O .ES
Abstracted/Indexed in: Science Citation Index Expanded (SciSearch), Journal Citation Reports/Science Edition, SCOPUS, Zentralblatt Math, EconLit, Google Scholar, Academic OneFile, Current Index to Statistics, Expanded Academic, ICYT, Mathematical Reviews, OCLC, Research Papers in Economics (RePEc), SCImago, STMA-Z, Summon by Serial Solutions
Springer

Springer

Test (2014) 23:100–134
DOI 10.1007/s11749-013-0344-4

ORIGINAL PAPER

# Comparison of relative density of two random geometric digraph families in testing spatial clustering

**Elvan Ceyhan**

Received: 22 February 2012 / Accepted: 23 September 2013 / Published online: 18 October 2013

**Abstract** We compare the performance of relative densities of two parameterized random geometric digraph families called *proximity catch digraphs* (PCDs) in testing bivariate spatial patterns. These PCD families are proportional edge (PE) and central similarity (CS) PCDs and are defined with proximity regions based on relative positions of data points from two classes. The relative densities of these PCDs were previously used as statistics for testing segregation and association patterns against complete spatial randomness. The relative density of a digraph, $D$, with $n$ vertices (i.e., with order $n$) represents the ratio of the number of arcs in $D$ to the number of arcs in the complete symmetric digraph of the same order. When scaled properly, the relative density of a PCD is a $U$-statistic; hence, it has asymptotic normality by the standard central limit theory of $U$-statistics. The PE- and CS-PCDs are defined with an expansion parameter that determines the size or measure of the associated proximity regions. In this article, we extend the distribution of the relative density of CS-PCDs for expansion parameter being larger than one, and compare finite sample performance of the tests by Monte Carlo simulations and asymptotic performance by Pitman asymptotic efficiency. We find the optimal expansion parameters of the PCDs for testing each alternative in finite samples and in the limit as the sample size tending to infinity. As a result of our comparisons, we demonstrate that in terms of empirical power (i.e., for finite samples) relative density of CS-PCD has better performance (which occurs for expansion parameter values larger than one) for the segregation alternative, while relative density of PE-PCD has better performance for the association alternative. The methods are also illustrated in a real-life data set from plant ecology.

**Keywords** Association · Complete spatial randomness · Delaunay triangulation · Pitman asymptotic efficiency · Proximity catch digraphs · Segregation · $U$-statistic

E. Ceyhan (✉)
Department of Mathematics, College of Sciences, Koç University, 34450 Sarıyer, Istanbul, Turkey
e-mail: elceyhan@ku.edu.tr

**Mathematics Subject Classification (2010)** 60D05 · 05C80 · 05C20 · 62G10 · 62M30

**List of Abbreviations**

| | |
|---|---|
| PCD | proximity catch digraph |
| CCCD | class cover catch digraph |
| CS-PCD | central similarity PCD |
| PE-PCDs | proportional edge PCD |
| PAE | Pitman asymptotic efficiency |
| CSR | complete spatial randomness |
| NNCT | nearest neighbor contingency table |
| NN | nearest neighbor |

**Symbols and Notation**

$D(V, A)$ Vertex random PCD with vertex set $V$ and arc set $A$. See Sect. 2, paragraph 1.

$N(\cdot)$ and $N(x)$ Proximity map and the proximity region associated with a point $x$. See Sect. 2, paragraph 5.

$\mathcal{C}_H(\mathcal{Y}_m)$ Convex hull of $\mathcal{Y}_m$. See Sect. 2, paragraph 7.

$T(\mathcal{Y}_3) = T(\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3)$ The triangle with vertices $\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3$. See Sect. 2, paragraph 7.

$\mathcal{S}(F)$ Support of the distribution $F$ and $\mathcal{U}(T(\mathcal{Y}_3))$: Uniform distribution on $T(\mathcal{Y}_3)$. See Sect. 2, paragraph 7.

$N_{\mathrm{PE}}(x, r)$ The proportional edge proximity map with expansion parameter $r$. See Sect. 2, paragraph 8.

$R_V(\mathsf{y}_1)$, $R_V(\mathsf{y}_2)$, and $R_V(\mathsf{y}_3)$ The vertex regions for vertices $\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3$. See Sect. 2, paragraph 8.

$N_{\mathrm{CS}}(x, \tau)$ The central similarity proximity map with expansion parameter $\tau$. See Sect. 2, paragraph 9.

$R_E(e_1)$, $R_E(e_2)$, $R_E(e_3)$ The edge regions for edges $e_1, e_2, e_3$ opposite to the vertices $\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3$ See Sect. 2, paragraph 9.

$\rho(D)$ The relative density of the digraph $D$. See Sect. 3, paragraph 1.

$\rho_{\mathrm{PE}}(n, r)$ and $\rho_{\mathrm{CS}}(n, \tau)$ The relative densities of PE-PCDs and CS-PCDs, respectively. See Theorem 3.1.

$\mu_{\mathrm{PE}}(r)$ and $\nu_{\mathrm{PE}}(r)$ The arc probability (or the asymptotic mean) and asymptotic variance of relative density of PE-PCDs. See Sect. 3.1, paragraph 4.

$\mu_{\mathrm{CS}}(r)$ and $\nu_{C_s}(r)$ The arc probability (or the asymptotic mean) and asymptotic variance of relative density of CS-PCDs. See Sect. 3.1, paragraph 4.

$\widetilde{\rho}_{\mathrm{PE}}(n, m, r)$ and $\widetilde{\rho}_{\mathrm{CS}}(n, m, \tau)$ The relative density for the PE-PCD and CS-PCD in the multiple triangle case. See Sect. 3.2, paragraph 2.

$\widetilde{\mu}_{\mathrm{PE}}(m, r)$ and $\widetilde{\nu}_{\mathrm{PE}}(m, r)$ The asymptotic mean and asymptotic variance of relative density of PE-PCDs in the multiple triangle case. See Corollary 3.4.

$R_{\mathrm{PE}}(r)$ The standardized test statistic based on the relative density of PE-PCD in the one-triangle case. See (13).

$\widetilde{R}_{\mathrm{PE}}(r)$ The standardized test statistic based on the relative density of PE-PCD in the multi-triangle case. See Sect. 4.1, paragraph 3.

$\mathrm{PAE}_{\mathrm{PE}}(r)$ and $\mathrm{PAE}_{\mathrm{CS}}(\tau)$ Pitman asymptotic efficiency score for relative density of PE-PCD and CS-PCD, respectively. See Sect. 7, paragraph 2.

$\pi_{\mathrm{out}}$ and $\widehat{\pi}_{\mathrm{out}}$ Proportion of class 1 points outside the convex hull of class 2 points and its estimate. See Sect. 8, paragraphs 1 & 2, respectively.

$C_{\mathrm{ch}}$ The correction coefficient for the class 1 points outside the convex hull of class 2 points. See (14).

$\widetilde{R}^{\mathrm{ch}}_{\mathrm{PE}}(r)$ and $\widetilde{R}^{\mathrm{ch}}_{\mathrm{CS}}(\tau)$ The convex hull corrected versions of the standardized test statistics based on the relative density of PE- and CS-PCDs. See (15).

## 1 Introduction

Spatial clustering has received considerable attention in the statistical literature (Cressie 1993 and Diggle 2003). Recently, the use of mathematical graphs has gained popularity in spatial analysis (Roberts et al. 2000) although it potentially reduces the benefit of other geo-spatial information, since in general a graph ignores the geographic reference. However, graph-theoretic tools have been discussed in spatial pattern analysis providing the means to go beyond the usual Euclidean metrics for spatial analysis. Graph-theoretic applications in computer vision and pattern recognition have been useful to automate efficient searches for structure in spatial data (Roberts et al. 2000). For example, graph-based approaches have been proposed to determine paths among habitats at various scales and dispersal movement distances, and balance data requirements with information content (Fall et al. 2007). Furthermore, graphs are potentially useful to ecological applications concerned with connectivity or movement (Minor and Urban 2007). Many concepts in spatial ecology depend on the idea of spatial adjacency which requires information on the close vicinity of an object (Keitt 2007). Graph theory conveniently can be adapted to express and communicate adjacency information allowing to compute meaningful quantities related to a spatial point pattern. See Ceyhan (2011) and references therein.

In recent years, a new clustering approach has been developed. This approach uses vertex random digraphs called proximity catch digraphs (PCDs) and is based on the relative positions of the data points from various classes. Priebe et al. (2001) introduced the class cover catch digraphs (CCCDs) which is a special type of PCDs and gave the exact and the asymptotic distribution of the domination number of the CCCD in $\mathbb{R}$. The CCCD approach is extended to multiple dimensions by DeVinney et al. (2002), Marchette and Priebe (2003), and Priebe et al. (2003), who demonstrated relatively good performance of it in classification. The proportional edge (PE) and central similarity (CS) proximity maps are introduced based on the appealing properties of CCCDs for one dimensional uniform data (see Ceyhan 2010b and references therein). One of the graph invariants used as a statistic is the relative density which is the proportion of number of arcs (i.e., directed edges) which are present in the digraph to the total number of arcs possible in a digraph with the same number of vertices.

In the literature, for undirected simple graphs, the graph density is defined as the ratio of the number of edges in the graph to the total number of edges possible in the graph. Hence the maximal density is 1 (for complete graphs) and minimal density is 0 (for graphs with no edges) (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. Hence for an undirected simple graph, $G = (V, E)$, with vertex set $V$ and edge set $E$, graph density is $2|E|/(|V|(|V| - 1))$; on the other hand, if $D = (V, A)$ is a directed graph (i.e., digraph) with vertex set $V$ and arc set $A$, graph density is $|A|/(|V|(|V| - 1))$. The latter quantity is referred to as *relative density* henceforth in this article. A related concept is the average degree of the graph $G$, which is twice the number of edges over number of vertices $2|E|/|V|$, and average degree of the digraph $D$ is $|A|/|V|$. For a given graph, an important problem in practice is finding a subgraph with maximum density. A fast algorithm is introduced for this purpose in Goldberg (1984) by reducing the problem to a minimum capacity cut computation steps, which can be performed with network flow methods. The density of a graph $G = (V, E)$ can also be extended to graphs with edge capacities $\{c(e) : e \in E\}$ as $c(E)/|V|$. A major topic in graph theory algorithms is finding the densest components, e.g., finding a maximum density subgraph which can be performed in polynomial time. However, finding a densest subgraph with exactly $k$ vertices is an NP-hard problem (Leibovich 2009). Various definitions and extensions of the concept of graph density is discussed by Faragó (2008). Graph density is also extended to weighted graphs with positive weights on vertices and nonnegative weights on edges with the introduction of the concept of $w$-density for the graphs (Shenggui et al. 2002). Graph density is also defined as the number of edges divided by number of vertices (Goldberg 1984), but we will stick with the more common definition referred to as relative density above.

In this article, we extend the definition of central similarity PCDs (CS-PCDs) for expansion parameter values larger than one; whereas previously it was defined only for the range of expansion parameter $(0, 1]$ (Ceyhan et al. 2007). Furthermore, we compare various aspects of the relative density of two parameterized PCD families, namely proportional edge PCDs (PE-PCDs) and CS-PCDs in testing bivariate spatial patterns. In particular, we compare the finite sample performance of the relative density of these two PCD families by empirical size and power analysis based on extensive Monte Carlo simulations. We also compare the asymptotic distributions and asymptotic power performance of the tests under the alternatives using Pitman asymptotic efficiency (PAE). For two classes of points labeled as class 1 an 2, respectively, PCDs are constructed with vertices from class 1 while points from class 2 are used to determine the underlying binary relation to determine the occurrence of an arc between two points from class 1. This binary relation is based on the Delaunay triangulation of class 2 points. Hence, we first consider the case of one triangle based on three non-collinear points from class 2, followed by the case of multiple triangles (based on the Delaunay triangulation of four or more class 2 points in general position). We also propose a correction for the proportion of class 1 points lying outside the convex hull of class 2 points. Our Monte Carlo and PAE analysis indicates that relative density of CS-PCDs has better performance for segregation, while that of PE-PCDs has better performance for association. Segregation is the pattern in which classes tend to repel each other in the sense that class 1 and class 2 points tend to be clustered around points from the same class, while association is the pattern in which points from one class tend to cluster around points from the other class. The extension of CS-PCDs for expansion parameter greater than 1 turned out to be useful, as the relative density of CS-PCDs has better performance in terms of empirical size and power in this new range of the expansion parameter. Without such an extension, we would only use CS-PCDs with expansion parameter less than or equal to 1, and thus reach less satisfactory results in our comparison. In particular, PE-PCDs with the corresponding optimal expansion parameter would outperform CS-PCDs with expansion parameter restricted to $(0, 1]$ in terms of size and power for both alternatives.

We describe the two particular PCD families in Sect. 2, provide the asymptotic distribution of relative density of the PCDs for uniform data in Sect. 3, describe the alternative patterns of segregation and association, propose tests based on relative density of PCDs for testing segregation/association, and provide the asymptotic normality and consistency of the tests under the alternatives in Sect. 4. We present the empirical size of the PCD tests in Sect. 5, empirical power under the alternatives in Sect. 6, and asymptotic efficiency in Sect. 7. We propose a correction method for the class 1 points outside the convex hull of class 2 points in Sect. 8 and illustrate the use of the tests in an ecological data set in Sect. 9. We present discussion and conclusions in Sect. 10.

## 2 The proximity map families and the associated PCDs

In the classical sense, random graphs were introduced by Erdős and Rényi (1959). In the classical model, the vertices are fixed in the sense that they only serve as end points of the edges, but edges are independently chosen and attached to the vertices. Similarly, for the random digraphs, arcs (i.e., directed edges) are independently inserted. More specifically, let $n$ be a positive integer and $[n] = \{1, 2, \dots, n\}$, and let $\mathscr{G}_n$ ($\mathscr{D}_n$) denote the set of all simple graphs $G = (V, E)$ (digraphs $D = (V, A)$) with vertex set $V = [n]$ and edge set $E$ (arc set $A$). A random graph (digraph) is a probability space of the form $(\mathscr{G}, P)$ $((\mathscr{D}, P))$ where $P$ is a probability measure defined on $\mathscr{G}_n$ ($\mathscr{D}_n$). In Erdős–Rényi graphs, each of the $\binom{n}{2}$ edges appear independently of others with probability $p$. Hence $P(G) = p^{|E|}(1-p)^{\binom{n}{2}-|E|}$ for $G \in \mathscr{G}_n$. The simplest digraph equivalent of Erdős–Rényi graphs is obtained if each of the $n(n-1)$ arcs appear independently of others with probability $p$. Hence $P(D) = p^{|A|}(1-p)^{n(n-1)-|A|}$ for $D \in \mathscr{D}_n$.

Since their introduction, random graphs have been extended in various directions. For more detail, see Beer et al. (2010) who also classify random graphs as edge random graphs, vertex random graphs, and edge-vertex random graphs based on which component contains randomness. They call Erdős–Rényi graphs as edge random graphs and its digraph counterpart can be labeled as edge random digraph. Edge random graphs can further be extended, if the probability of an edge appearing between vertices $i$ and $j$ equals $p(i, j)$ which is not necessarily constant. A similar extension can be provided for digraphs in a straightforward manner.

Vertex random graphs (digraphs) can be defined as follows. Let $(\Omega, \mu)$ be a probability space, $\varphi : \Omega \times \Omega \to \{0, 1\}$ be a symmetric function and $\nu : \Omega \times \Omega \to \{0, 1\}$ be a function which is not necessarily symmetric. The vertex random graph $G(n, \Omega, \mu, \varphi)$ (or digraph $D(n, \Omega, \mu, \nu)$) is the random graph $(\mathscr{G}, P)$ (or digraph $(\mathscr{D}, P)$) satisfying $P(G) = \int \mathbf{I}(G(x, \varphi) = G)\mu(d\mathbf{x})$, $G \in \mathscr{G}_n$ (or $P(D) =$

$\int \mathbf{I}(D(x,\nu)=D)\mu(d\mathbf{x})$, $D \in \mathscr{D}_n)$ where $\mathbf{I}(\cdot)$ stands for the indicator of a set, $\mu(d\mathbf{x})$ stands for the product integrator $\mu^n(d\mathbf{x}) = \mu(dx_1)\cdots\mu(dx_n)$ on $\Omega^n$, $G(x,\varphi)$ is the graph (or $D(x,\nu)$ is the digraph) with vertex set $[n]$ such that for all $i, j \in [n]$, $i \neq j$, $ij \in E$ iff $\varphi(i,j)=1$ (or $(i,j) \in A$ iff $\nu(i,j)=1$). Here actually $G(\cdot,\varphi)$ (or $D(\cdot,\nu)$) is a graph-valued (or digraph-valued) random variable on $\Omega^n$ with probability assignment being done as the vertex random graph (or digraph) taking the value $G$ (or $D$).

If we let $\Omega$ be the set of real intervals and $\varphi(i,j) = \mathbf{I}(I_i \cap I_j \neq \emptyset)$, we obtain the random interval graph. Also, letting $\Omega = \mathbb{R}^k$, if we take $n$ points iid from some probability distribution on $\Omega$ which is also equipped with a metric $d$, then taking $\varphi(x,y) = \mathbf{I}(d(x,y) \leq t)$ for some threshold $t > 0$, we obtain the random geometric graphs. Notice that for random geometric graphs we are assigning edges to the vertices deterministically, if the distance between the vertices are lower than a certain threshold (i.e., $ij \in E$ iff $d(x_i,x_j) \leq t$) but the vertices are randomly generated or chosen in a metric space (see Penrose 2003 for an extensive treatment).

Our PCDs are vertex random digraphs where the vertices are randomly generated or selected from a probability space and arcs are deterministically inserted between vertices based on proximity regions around the vertices. We first define proximity maps, regions and PCDs in a fairly general setting. For PCDs, the regions around the vertices are based on the proximity maps which are defined as follows. For the probability space $(\Omega,\mu)$ with $\wp(\cdot)$ representing the power set function, the *proximity map* $N(\cdot) : \Omega \to \wp(\Omega)$ defines a *proximity region* $N(x) \subseteq \Omega$ for each point $x \in \Omega$. In our discussion hereinafter, we will have points from two classes, namely classes 1 and 2. Let $\mathcal{X}_n \subseteq \Omega$ be $n$ points from class 1 and $\mathcal{Y}_m \subseteq \Omega$ be $m$ points from class 2. The region $N(x)$ will be defined based on the dissimilarity between $x$ and $\mathcal{Y}_m$. Hence given $\mathcal{Y}_m$, we define the vertex random PCD, $D(V,A)$, with vertex set $V = \mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$ and arc set $A$ by $(X_i, X_j) \in A$ iff $X_j \in N(X_i)$. In the above vertex random digraph setting, we would have $\nu(i,j) = \mathbf{I}(X_j \in N(X_i))$. The term "proximity" comes from thinking of the region $N(x)$ as representing those points in $\Omega$ "close" to $x$ and "catch" comes from $N(u)$ *catching* $v$ whenever $v \in N(u)$. An extensive treatment of the proximity graphs is provided by Toussaint (1980) and Jaromczyk and Toussaint (1992). In the CCCD approach, we have $\Omega = \mathbb{R}^k$ and the points correspond to observations from class 1 and the sets (i.e., proximity regions) are defined to be (open) balls centered at the points with maximal radius (relative to the other class, i.e., class 2): $N(x) = B(x, r(x))$, where $r(x) = d(x, \mathcal{Y}_m)$ is the minimum Euclidean distance between the observation $x \in \mathcal{X}_n$ and the observations, $\mathcal{Y}_m$, from the other class (Priebe et al. 2001). Hence the proximity region for CCCD is also called *spherical proximity region*, and CCCD itself is also called *spherical PCD* (Ceyhan 2010b). In the vertex random graph setting, we would have $\nu(i,j) = \mathbf{I}(X_j \in B(X_i, r(X_i)))$. Notice that PCDs can be viewed as an extension of random interval graphs to higher dimensions with proximity regions replacing intervals. Moreover, PCDs can also be viewed as an extension of random geometric graphs with varying regions around the vertices or by replacing the distance with a dissimilarity, and the defining function being no longer symmetric (hence we have arcs instead of edges).

Next, we briefly define PE and CS proximity maps and the associated PCDs. Here the points correspond to observations from class 1 and the proximity regions are defined to be (closed) regions based on class 1 and class 2 points; and the regions increase in size as the distance of a class 1 point from the set of class 2 points increases. The space is partitioned by the Delaunay tessellation of class 2 points which is a triangulation in $\mathbb{R}^2$. In each triangle, a family of PCDs is constructed based on the relative positions of the class 1 points with respect to each other and to class 2 points. These proximity maps have the advantage that the calculations to obtain the asymptotic distribution of the relative density are analytically tractable (Ceyhan et al. 2006, 2007).

For $\Omega = \mathbb{R}^d$, let $\mathcal{Y}_m = \{\mathsf{y}_1, \mathsf{y}_2, \ldots, \mathsf{y}_m\}$ be $m$ given points from class 2 in general position in $\mathbb{R}^d$. The space, $\mathbb{R}^d$, is partitioned by the Delaunay tessellation of class 2 points. Then, let $T_i$ be the $i$th Delaunay cell for $i = 1, 2, \ldots, J_m$. Let $\mathcal{X}_n$ be a set of iid random variables from distribution $F$ and constitute class 1 points in $\mathbb{R}^d$ with support $\mathcal{S}(F) \subseteq \mathcal{C}_H(\mathcal{Y}_m)$ where $\mathcal{C}_H(\mathcal{Y}_m)$ stands for the convex hull of $\mathcal{Y}_m$. In particular, for illustrative purposes, we focus on $\mathbb{R}^2$ where a Delaunay tessellation is a triangulation provided that no more than three points in $\mathcal{Y}_m$ are cocircular (i.e., lie on the same circle). For simplicity, we consider the one triangle case first. Let $\mathcal{Y}_3 = \{\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3\}$ be three non-collinear points in $\mathbb{R}^2$ and $T(\mathcal{Y}_3) = T(\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3)$ be the triangle with vertices $\mathcal{Y}_3$. Let $\mathcal{X}_n$ be a set of iid random variables from $F$ with support $\mathcal{S}(F) \subseteq T(\mathcal{Y}_3)$ and $\mathcal{U}(T(\mathcal{Y}_3))$ be the uniform distribution on $T(\mathcal{Y}_3)$. We adopt the convention that random variables are represented with capital letters, while fixed quantities are represented with lower case letters. Hence, in our setup, we assume $\mathcal{Y}_m$ is given, i.e., it is a set of fixed class 2 points, while $\mathcal{X}_n$ is a set of random points from class 1.

The *PE proximity maps* are defined in detail in Ceyhan et al. (2006); we provide the definition briefly here for the sake of completeness. For the expansion parameter $r \in [1, \infty]$, we define the PE proximity map with expansion parameter $r$, denoted $N_{\mathrm{PE}}(x, r)$ as follows; see also Fig. 1 (left). Using line segments from the center of mass of $T(\mathcal{Y}_3)$ to the midpoints of its edges, we partition $T(\mathcal{Y}_3)$ into "vertex regions" $R_V(\mathsf{y}_1)$, $R_V(\mathsf{y}_2)$, and $R_V(\mathsf{y}_3)$. For $x \in T(\mathcal{Y}_3) \setminus \mathcal{Y}_3$, let $v(x) \in \mathcal{Y}_3$ be the vertex in whose region $x$ falls, so $x \in R_V(v(x))$. If $x$ falls on the boundary of two vertex regions, we assign $v(x)$ arbitrarily to one of the adjacent regions. Let $e(x)$ be the edge of $T(\mathcal{Y}_3)$ opposite $v(x)$. Let $\ell(x)$ be the line parallel to $e(x)$ through $x$. Let $d(v(x), \ell(x))$ be the Euclidean distance from $v(x)$ to $\ell(x)$. For $r \in [1, \infty)$, let $\ell_r(x)$ be the line parallel to $e(x)$ such that $d(v(x), \ell_r(x)) = r d(v(x), \ell(x))$ and $d(\ell(x), \ell_r(x)) < d(v(x), \ell_r(x))$. Let $T_{\mathrm{PE}}(x, r)$ be the triangle similar to and with the same orientation as $T(\mathcal{Y}_3)$ having $v(x)$ as a vertex and $\ell_r(x)$ as the opposite edge. Then the PE proximity region $N_{\mathrm{PE}}(x, r)$ is defined to be $T_{\mathrm{PE}}(x, r) \cap T(\mathcal{Y}_3)$. Notice that $r \geq 1$ implies $x \in N_{\mathrm{PE}}(x, r)$. Note also that $\lim_{r \to \infty} N_{\mathrm{PE}}(x, r) = T(\mathcal{Y}_3)$ for all $x \in T(\mathcal{Y}_3) \setminus \mathcal{Y}_3$, so we define $N_{\mathrm{PE}}(x, \infty) = T(\mathcal{Y}_3)$ for all such $x$. For $x \in \mathcal{Y}_3$, we define $N_{\mathrm{PE}}(x, r) = \{x\}$ for all $r \in [1, \infty]$.

Define $N_{\mathrm{CS}}(x, \tau)$ to be the CS proximity map with expansion parameter $\tau$ as follows; see also Fig. 1 (right). The *CS proximity maps* were previously defined with expansion parameter $\tau \leq 1$ (Ceyhan et al. 2007). Below, we provide a definition for a much wider range of the expansion parameter $\tau \in (0, \infty]$. Let $e_j$ be the edge opposite vertex $\mathsf{y}_j$ for $j = 1, 2, 3$, and let "edge regions" $R_E(e_1)$, $R_E(e_2)$, $R_E(e_3)$ partition $T(\mathcal{Y}_3)$ using line segments from the center of mass of $T(\mathcal{Y}_3)$ to the vertices. For

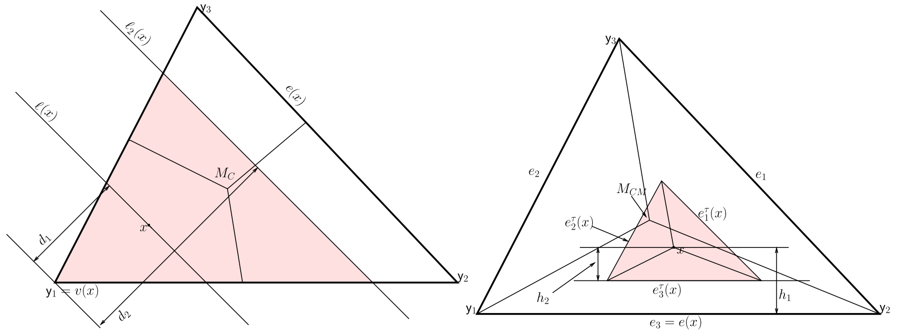


**Fig. 1** Plotted *on the left* is the illustration of the construction of PE proximity region, $N_{\text{PE}}(x, r = 2)$ (*shaded region*) for an $x \in R_V(\mathsf{y}_1)$ where $d_1 = d(v(x), \ell(v(x), x))$ and $d_2 = d(v(x), \ell_2(x)) = 2d(v(x), \ell(x))$; and *on the right* is the illustration of the construction of CS proximity region, $N_{\text{CS}}(x, \tau = 1/2)$ (*shaded region*) for an $x \in R_E(e_3)$ where $h_2 = d(x, e_3^\tau(x)) = \frac{1}{2}d(x, e(x))$ and $h_1 = d(x, e(x))$

$x \in (T(\mathcal{Y}_3))^o$, let $e(x)$ be the edge in whose region $x$ falls; $x \in R_E(e(x))$. If $x$ falls on the boundary of two edge regions we assign $e(x)$ arbitrarily. For $\tau > 0$, the CS proximity region $N_{\text{CS}}(x, \tau)$ is defined to be the triangle $T_{\text{CS}}(x, \tau) \cap T(\mathcal{Y}_3)$ with the following properties:

(i) For $\tau \in (0, 1]$, the triangle $T_{\text{CS}}(x, \tau)$ has an edge $e_\tau(x)$ parallel to $e(x)$ such that $d(x, e_\tau(x)) = \tau\, d(x, e(x))$ and $d(e_\tau(x), e(x)) \le d(x, e(x))$ and for $\tau > 1$, $d(e_\tau(x), e(x)) < d(x, e_\tau(x))$ where $d(x, e(x))$ is the Euclidean distance from $x$ to $e(x)$;
(ii) The triangle $T_{\text{CS}}(x, \tau)$ has the same orientation as and is similar to $T(\mathcal{Y}_3)$; and
(iii) The point $x$ is at the center of mass of $T_{\text{CS}}(x, \tau)$.

Notice that $\tau > 0$ implies that $x \in N_{\text{CS}}(x, \tau)$ and, by construction, we have $N_{\text{CS}}(x, \tau) \subseteq T(\mathcal{Y}_3)$ for all $x \in T(\mathcal{Y}_3)$. Let $\partial(\cdot)$ stand for the boundary of a given region. Then, for $x \in \partial(T(\mathcal{Y}_3))$ and $\tau \in (0, \infty]$, we define $N_{\text{CS}}(x, \tau) = \{x\}$. For all $x \in T(\mathcal{Y}_3)^o$ the edges $e_\tau(x)$ and $e(x)$ are coincident iff $\tau = 1$. Note also that $\lim_{\tau \to \infty} N_{\text{CS}}(x, \tau) = T(\mathcal{Y}_3)$ for all $x \in (T(\mathcal{Y}_3))^o$, so we define $N_{\text{CS}}(x, \infty) = T(\mathcal{Y}_3)$ for all such $x$.

*Remark 2.1* Notice that $X_i \overset{\text{iid}}{\sim} F$, with the additional assumption that the non-degenerate two-dimensional probability density function $f$ exists with support in $T(\mathcal{Y}_3)$, implies that the special case in the construction of $N_{\text{PE}}(\cdot, r)$—$X$ falls on the boundary of two vertex regions—occurs with probability zero; similarly, the special case in the construction of $N_{\text{CS}}(\cdot, \tau)$—$X$ falls on the boundary of two edge regions—occurs with probability zero also.

## 3 The asymptotic distribution of relative density

The *relative density* of a digraph $D = (V, A)$ of order $|V| = n$, denoted $\rho(D)$, is defined as

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

where $|\cdot|$ stands for set cardinality (Janson et al. 2000). Thus $\rho(D)$ represents the ratio of the number of arcs in the digraph $D$ to the number of arcs in the complete symmetric digraph of order $n$, which is $n(n-1)$. If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$, then the relative density of the associated data-random PCD, denoted $\rho(\mathcal{X}_n; h, N)$, is shown to be a $U$-statistic (Ceyhan et al. 2006, 2007),

$$\rho(\mathcal{X}_n; h, N) = \frac{1}{n(n-1)} \sum\sum_{i<j} h_{ij} \tag{1}$$

where

$$\begin{aligned} h_{ij} &:= h(X_i, X_j; N) = \mathbf{I}\big\{(X_i, X_j) \in A\big\} + \mathbf{I}\big\{(X_j, X_i) \in A\big\} \\ &= \mathbf{I}\big\{X_j \in N(X_i)\big\} + \mathbf{I}\big\{X_i \in N(X_j)\big\}. \end{aligned}$$

Since the digraph is asymmetric, $h_{ij}$ 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). Moreover, by a central limit theorem (CLT) for $U$-statistics (Lehmann 1988), it has been proved that

$$\sqrt{n}\big(\rho_n - \mathbf{E}[\rho_n]\big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \mathbf{Cov}[h_{12}, h_{13}]\big) \tag{2}$$

provided $\mathbf{Cov}[h_{12}, h_{13}] > 0$ where $\mathcal{N}(\mu, \sigma^2)$ stands for the normal distribution with mean $\mu$ and variance $\sigma^2$ and $\mathbf{E}[\rho_n] = \frac{1}{2}\mathbf{E}[h_{12}]$ (Ceyhan et al. 2006, 2007).

### 3.1 The one triangle case

For simplicity, we consider that three non-collinear points $\mathcal{Y}_3 = \{\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3\}$ forming a triangle $T(\mathcal{Y}_3)$ and class 1 points are iid uniform in $T(\mathcal{Y}_3)$. The null hypothesis is a type of *complete spatial randomness* (CSR), that is,

$$H_o : X_i \overset{\text{iid}}{\sim} \mathcal{U}\big(T(\mathcal{Y}_3)\big) \quad \text{for } i = 1, 2, \ldots, n. \tag{3}$$

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

**Theorem 3.1** (Geometry invariance for uniform data) *Let $\mathcal{Y}_3 = \{\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3\} \subset \mathbb{R}^2$ be three non-collinear points. For $i = 1, 2, \ldots, n$, let $X_i \overset{\text{iid}}{\sim} F = \mathcal{U}(T(\mathcal{Y}_3))$. Then*

(i) *For any $r \in [1, \infty]$, the distribution of relative density of PE-PCDs, $\rho_{\text{PE}}(n, r)$, is independent of $\mathcal{Y}_3$, hence the geometry of $T(\mathcal{Y}_3)$.*

(ii) *For any $\tau \in (0, \infty]$, the distribution of relative density of CS-PCDs, $\rho_{\text{CS}}(n, \tau)$, is independent of $\mathcal{Y}_3$, hence the geometry of $T(\mathcal{Y}_3)$.*

The proof is provided in Sect. 1 of the Online Resource 1.

In fact, the geometry invariance of $\rho_{\text{PE}}(n, \infty)$ (or $\rho_{\text{CS}}(n, \infty)$) for data from any continuous distribution on $T(\mathcal{Y}_3)$ follows trivially, since for $r = \infty$ (or $\tau = \infty$), we have $\rho_{\text{PE}}(n, r) = 1$ (or $\rho_{\text{CS}}(n, \tau) = 1$) almost surely (a.s.) (i.e., its distribution is degenerate). Based on the geometry invariance for uniform data, we may assume that $T(\mathcal{Y}_3)$ is a *standard equilateral triangle*, $T_e$, with vertices $\mathcal{Y}_3 = \{(0, 0), (1, 0), (1/2, \sqrt{3}/2)\}$ henceforth.

The CLT for $U$-statistics establishes the asymptotic normality under the uniform null hypothesis. For our proximity maps and uniform null hypothesis, the asymptotic null distribution of $\rho_{\text{PE}}(n, r)$ (or $\rho_{\text{CS}}(n, \tau)$) can be derived as a function of $r$ (or $\tau$). Let $\mu_{\text{PE}}(r) := \mathbf{E}[\rho_{\text{PE}}(n, r)]$ and $\nu_{\text{PE}}(r) := \mathbf{Cov}[h_{12}, h_{13}]$ where $h_{ij}$ is the symmetric kernel in (1) for PE-PCDs. Notice that $\mu_{\text{PE}}(r) = \mathbf{E}[h_{12}]/2 = P(X_2 \in N_{\text{PE}}(X_1, r))$ is the probability of an arc occurring between any pair of vertices, hence is also called the *arc probability*. Similarly, let $\mu_{\text{CS}}(\tau) := \mathbf{E}[\rho_{\text{CS}}(n, \tau)]$, then $\mu_{\text{CS}}(\tau) = P(X_2 \in N_{\text{CS}}(X_1, \tau))$ and let $\nu_{\text{CS}}(\tau) := \mathbf{Cov}[\widetilde{h}_{12}, \widetilde{h}_{13}]$ where $\widetilde{h}_{ij}$ is the symmetric kernel for CS-PCDs.

By detailed geometric probability calculations, the means and the asymptotic variances of the relative density of the PE-PCDs were calculated explicitly and presented below for completeness (details of their derivation is provided in Ceyhan et al. 2006).

**Theorem 3.2** *For* $r \in [1, \infty)$,

$$\frac{\sqrt{n}\,(\rho_{\text{PE}}(n, r) - \mu_{\text{PE}}(r))}{\sqrt{\nu_{\text{PE}}(r)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \tag{4}$$

*where*

$$\mu_{\text{PE}}(r) = \begin{cases} \frac{37}{216} r^2 & \text{for } r \in [1, 3/2), \\ -\frac{1}{8} r^2 + 4 - 8r^{-1} + \frac{9}{2} r^{-2} & \text{for } r \in [3/2, 2), \\ 1 - \frac{3}{2} r^{-2} & \text{for } r \in [2, \infty), \end{cases} \tag{5}$$

*and*

$$\nu_{\text{PE}}(r) = \nu_1(r)\,\mathbf{I}\big(r \in [1, 4/3)\big) + \nu_2(r)\,\mathbf{I}\big(r \in [4/3, 3/2)\big) + \nu_3(r)\,\mathbf{I}\big(r \in [3/2, 2)\big) \\ + \nu_4(r)\,\mathbf{I}\big(r \in [2, \infty]\big) \tag{6}$$

*with*

$$\nu_1(r) = \big[3007\, r^{10} - 13824\, r^9 + 898\, r^8 + 77760\, r^7 - 117953\, r^6 + 48888\, r^5 \\ - 24246\, r^4 + 60480\, r^3 - 38880\, r^2 + 3888\big] \big/ \big[58320\, r^4\big],$$

$$\nu_2(r) = \big[5467\, r^{10} - 37800\, r^9 + 61912\, r^8 + 46588\, r^6 - 191520\, r^5 + 13608\, r^4 \\ + 241920\, r^3 - 155520\, r^2 + 15552\big] \big/ \big[233280\, r^4\big],$$

$$\nu_3(r) = -\big[7\,r^{12} - 72\,r^{11} + 312\,r^{10} - 5332\,r^{8} + 15072\,r^{7} + 13704\,r^{6} - 139264\,r^{5} + 273600\,r^{4} - 242176\,r^{3} + 103232\,r^{2} - 27648\,r + 8640\big] \,/\, \big[960\,r^{6}\big],$$

$$\nu_4(r) = \frac{15\,r^4 - 11\,r^2 - 48\,r + 25}{15\,r^6}.$$

*For $r = \infty$, $\rho_{\text{PE}}(n, r)$ is degenerate.*

The means and asymptotic variances for CS-PCDs were previously calculated for $\tau \in (0, 1]$ (Ceyhan et al. 2007). Here we extend the result for $\tau > 1$ as well (see Sect. 2 of Online Resource 1).

**Theorem 3.3** *For $\tau \in (0, \infty)$,*

$$\frac{\sqrt{n}(\rho_{\text{CS}}(n, \tau) - \mu_{\text{CS}}(\tau))}{\sqrt{\nu_{\text{CS}}(\tau)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \tag{7}$$

*where*

$$\mu_{\text{CS}}(\tau) = \begin{cases} \frac{\tau^2}{6} & \text{for } \tau \in (0, 1], \\ \frac{\tau\,(4\,\tau-1)}{2(1+2\,\tau)(2+\tau)} & \text{for } \tau \in (1, \infty), \end{cases} \tag{8}$$

*and*

$$\nu_{\text{CS}}(\tau) = \begin{cases} \frac{\tau^4(6\,\tau^5-3\,\tau^4-25\,\tau^3+\tau^2+49\,\tau+14)}{45\,(\tau+1)(2\,\tau+1)(\tau+2)} & \text{for } \tau \in (0, 1], \\ \frac{168\,\tau^7+886\,\tau^6+1122\,\tau^5+45\,\tau^4-470\,\tau^3-114\,\tau^2+48\,\tau+16}{5(2\,\tau+1)^4(\tau+2)^4} & \text{for } \tau \in (1, \infty). \end{cases} \tag{9}$$

*For $\tau = 0$, $\rho_{\text{CS}}(n, \tau)$ is degenerate.*

The forms of the mean function are depicted together in Fig. 2 (left). Note that $\mu_{\text{PE}}(r)$ is monotonically increasing in $r$, since $N_{\text{PE}}(x, r)$ increases in size with $r$ for all $x \in R_V(\mathsf{y}_j) \setminus \mathscr{R}_S(N_{\text{PE}}(\cdot, r), M_C)$, where $\mathscr{R}_S(N_{\text{PE}}(\cdot, r), M_C) := \{x \in T(\mathcal{Y}_3) : N_{\text{PE}}(x, r) = T(\mathcal{Y}_3)\}$. In addition, $\mu_{\text{PE}}(r) \to 1$ as $r \to \infty$ at rate $O(r^{-2})$, since the digraph becomes complete in the limit, which explains why $\rho_{\text{PE}}(n, r)$ becomes degenerate, i.e., $\nu_{\text{PE}}(r = \infty) = 0$. $\mu_{\text{PE}}(r)$ is continuous, with the value at $r = 1$, $\mu_{\text{PE}}(1) = 37/216 \approx 0.17$. Note also that $\mu_{\text{CS}}(\tau)$ is monotonically increasing in $\tau$, since $N_{\text{CS}}(x, \tau)$ increases in size with $\tau$ for all $x \in R_E(e_j) \setminus \mathscr{R}_S(N_{\text{CS}}(\cdot, \tau), M_C)$, where $\mathscr{R}_S(N_{\text{CS}}(\cdot, \tau), M_C) := \{x \in T(\mathcal{Y}_3) : N_{\text{CS}}(x, \tau) = T(\mathcal{Y}_3)\}$. Note also that $\mu_{\text{CS}}(\tau)$ is continuous in $\tau$ with $\mu_{\text{CS}}(\tau = 1) = 1/6$ and $\lim_{\tau \to 0} \mu_{\text{CS}}(\tau) = 0$. In addition, $\mu_{\text{CS}}(\tau) \to 1$ as $\tau \to \infty$ at rate $O(\tau^{-1})$, so $\rho_{\text{CS}}(n, \tau)$ becomes degenerate as $\tau \to \infty$. Observe also that $\mu_{\text{PE}}(r) > \mu_{\text{CS}}(\tau)$ for all $r \in [1, \infty)$ and $\tau \in (0, \infty)$.

The asymptotic variance functions are depicted together in Fig. 2 (right). Note that $\nu_{\text{PE}}(r)$ is also continuous in $r$ with $\lim_{r \to \infty} \nu_{\text{PE}}(r) = 0$ and $\nu_{\text{PE}}(1) = 34/58320 \approx 0.0006$ and observe that $\sup_{r \geq 1} \nu_{\text{PE}}(r) \approx 0.13$ which is attained at $r \approx 2.045$. Note also that $\nu_{\text{CS}}(\tau)$ is continuous in $\tau$ with $\lim_{\tau \to \infty} \nu_{\text{CS}}(\tau) = 0$ and $\nu(\tau = 1) = 7/135$

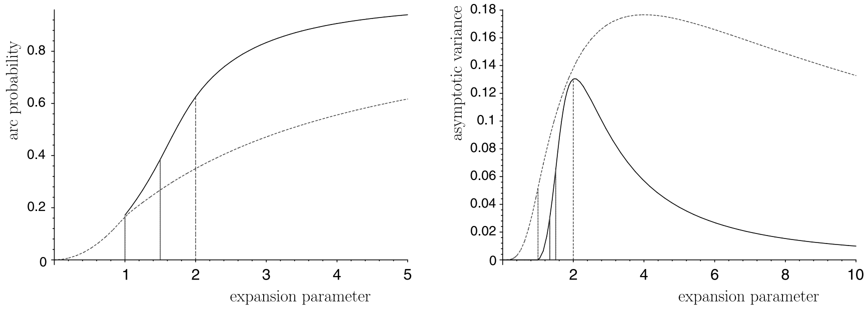


**Fig. 2** Asymptotic null means (i.e., arc probabilities) (*left*) and variances (*right*) as a function of the expansion parameters for relative density of PE-PCDs (*solid line*) and CS-PCDs (*dashed line*). The *vertical lines* indicate the endpoints of the intervals in the piecewise definition of the functions. Notice that the vertical and horizontal axes are differently scaled

and $\lim_{\tau \to 0} \nu_{\mathrm{CS}}(\tau) = 0$—there are no arcs when $\tau = 0$ a.s.—which explains why the limiting distribution of $\rho_{\mathrm{CS}}(n, \tau)$ becomes degenerate as $\tau$ goes to zero. Moreover, $\sup_{\tau > 0} \nu_{\mathrm{CS}}(\tau) \approx 0.18$ which is attained at $\tau \approx 4.0051$. Observe also that $\nu_{\mathrm{CS}}(\tau) > \nu_{\mathrm{PE}}(r)$ for all $r \in [1, \infty)$ and $\tau \in (0, \infty)$.

The finite sample variance and skewness of $\rho_{\mathrm{PE}}(n, r)$ and $\rho_{\mathrm{CS}}(n, \tau)$ can be derived analytically in much the same way as were asymptotic variances. In particular, the variances of $h_{12}$ for PE-PCDs and CS-PCDs are derived and presented in the technical report Ceyhan (2010a).

### 3.2 The multiple triangle case

In this section, we present the asymptotic distribution of the relative density for class 1 points in multiple triangles. Suppose $\mathcal{Y}_m = \{\mathsf{y}_1, \mathsf{y}_2, \ldots, \mathsf{y}_m\} \subset \mathbb{R}^2$ is a set of $m$ points in general position with $m > 3$ and no more than three points are cocircular. As a result of the Delaunay triangulation of $\mathcal{Y}_m$ (Okabe et al. 2000), there are $J_m > 1$ Delaunay triangles denoted as $T_j$, for $j = 1, 2, \ldots, J_m$. The Delaunay triangles partition the convex hull of $\mathcal{Y}_m$. We wish to investigate

$$H_o : X_i \overset{\text{iid}}{\sim} \mathcal{U}\big(C_H(\mathcal{Y}_m)\big) \quad \text{for } i = 1, 2, \ldots, n \tag{10}$$

against segregation and association alternatives (see Sect. 4). Figure 3 presents a realization of 1000 observations independent and identically distributed as $\mathcal{U}(C_H(\mathcal{Y}_m))$ for $m = 10$ and $J_m = 13$.

For $J_m > 1$ (i.e., $m > 3$), as in Sect. 2, let $\widetilde{\rho}_{\mathrm{PE}}(n, m, r) = |A|/(n\,(n-1))$ be the relative density for the PE-PCD in the multiple triangle case. Let $\widetilde{\rho}_{\mathrm{CS}}(n, m, \tau)$ be defined similarly for the CS-PCD. The asymptotic normality of the relative density of PE-PCDs and CS-PCDs (with $\tau \in (0, 1]$) is provided in Ceyhan et al. (2006, 2007), respectively. The result for CS-PCDs with $\tau > 1$ follows similarly (see Ceyhan 2010a).

Let $n_i$ be the number of class 1 points in $T_i$ for $i = 1, 2, \ldots, J_m$. Letting $w_i = A(T_i)/A(C_H(\mathcal{Y}_m))$ with $A(\cdot)$ being the area function and $\mathcal{W} = \{w_1, w_2, \ldots, w_{J_m}\}$, we obtain the following as a corollary to Theorems 3.2 and 3.3.

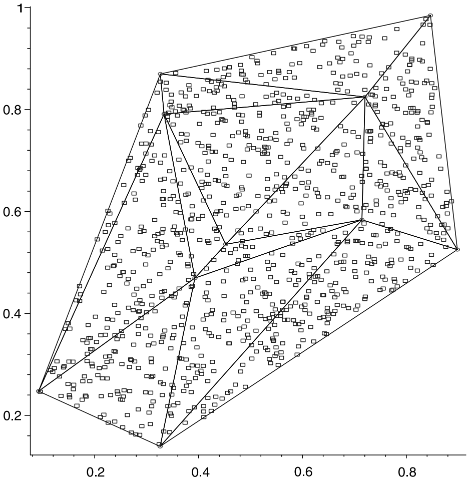


**Fig. 3** A realization of $H_o : X_i \overset{\text{iid}}{\sim} \mathcal{U}(C_H(\mathcal{Y}_m))$ for $i = 1, 2, \ldots, n$ for $|\mathcal{Y}_m| = 10$ points with $n = 1000$ class 1 points generated iid in the convex hull of $\mathcal{Y}_m$

**Corollary 3.4** *For $r \in [1, \infty]$, the asymptotic distribution for $\widetilde{\rho}_{\text{PE}}(n, m, r)$ conditional on $\mathcal{W}$ is given by*

$$\sqrt{n}\big(\widetilde{\rho}_{\text{PE}}(n, m, r) - \widetilde{\mu}_{\text{PE}}(m, r)\big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, 4\,\widetilde{\nu}_{\text{PE}}(m, r)\big), \tag{11}$$

*as $n \to \infty$, where $\widetilde{\mu}_{\text{PE}}(m, r) = \mu_{\text{PE}}(r)(\sum_{i=1}^{J_m} w_i^2)$ and*

$$\widetilde{\nu}_{\text{PE}}(m, r) = \left[\nu_{\text{PE}}(r)\left(\sum_{i=1}^{J_m} w_i^3\right) + \big(\mu_{\text{PE}}(r)\big)^2\left(\sum_{i=1}^{J_m} w_i^3 - \left(\sum_{j=1}^{J_m} w_i^2\right)^2\right)\right]$$

*with $\mu_{\text{PE}}(r)$ and $\nu_{\text{PE}}(r)$ being as in (5) and (6), respectively. The asymptotic distribution of $\widetilde{\rho}_{\text{CS}}(n, m, \tau)$ with $\tau \in (0, \infty]$ is similar.*

See Sect. 3 of the Online Resource 1 for the proof.

By an appropriate application of the Jensen's inequality, we see that $\sum_{i=1}^{J_m} w_i^3 \geq (\sum_{i=1}^{J_m} w_i^2)^2$. So the covariance above is zero iff $\nu_{\text{PE}}(r) = 0$ and $\sum_{i=1}^{J_m} w_i^3 = (\sum_{i=1}^{J_m} w_i^2)^2$, so asymptotic normality may hold even though $\nu_{\text{PE}}(r) = 0$ in the multiple triangle case. That is, $\widetilde{\rho}_{\text{PE}}(n, m, r)$ has the asymptotic normality even for $r = \infty$ provided that $\sum_{i=1}^{J_m} w_i^3 > (\sum_{i=1}^{J_m} w_i^2)^2$. The same holds for $\tau = \infty$ in the central similarity case.

## 4 Parameterization of the alternative patterns: segregation and association

Spatial interaction among species (including segregation and association of species) also has important consequences and potential for applicability in biodiversity theory (Illian and Burslem 2007). Many procedures are suggested for spatial clustering

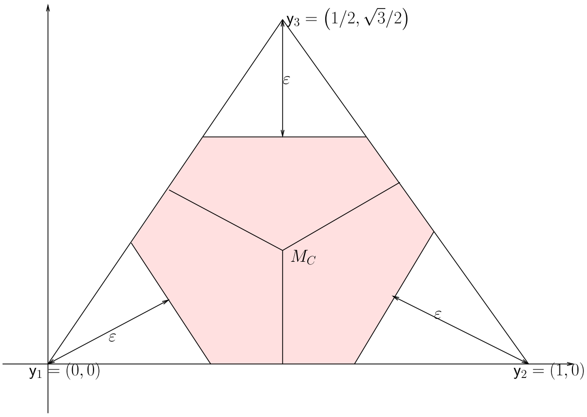


**Fig. 4** An example for the segregation alternative with a particular expansion parameter $\varepsilon$ (*shaded region*), and its complement is for the association alternative with expansion parameter $\sqrt{3}/3 - \varepsilon$ (*unshaded region*) on the standard equilateral triangle

based on count data (see, e.g., Jung and Kulldorff 2007). In a two-class setting, the phenomenon known as *segregation* occurs when 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 $X_i$ are unlikely to be located near elements of $\mathcal{Y}_m$. Alternatively, association occurs when members of one class have a tendency to attract members of the other class, as in symbiotic species, so that $X_i$ will tend to cluster around the elements of $\mathcal{Y}_m$. See, for instance, Dixon (1994) and Coomes et al. (1999).

Under association, the defining proximity regions tend to be small, and hence there should be fewer arcs; while under segregation, the proximity regions tend to be larger and cover many points, resulting in many arcs. Thus, the relative density is a reasonable statistic to employ in this problem. Unfortunately, in the case of the CCCD, it is difficult to make precise calculations in multiple dimensions due to the geometry of the neighborhoods.

In the basic triangle, $T_b$, we define the alternatives $H^S_\varepsilon$ and $H^A_\varepsilon$ with $\varepsilon \in (0, \sqrt{3}/3)$, for segregation and association alternatives, respectively. Under $H^S_\varepsilon$, $4\varepsilon^2/3 \times 100\ \%$ of the area of $T_b$ is chopped off around each vertex so that the class 1 points are restricted to lie in the remaining region. Let $\mathcal{T}_\varepsilon$ be the union of the triangular regions around the vertices as illustrated in Fig. 4. Below, we have the parametrization of the distribution families under the alternatives:

$$\mathscr{U}^S_\varepsilon := \left\{F : F = \mathcal{U}(T_b \setminus \mathcal{T}_\varepsilon)\right\} \quad \text{and} \quad \mathscr{U}^A_\varepsilon := \left\{F : F = \mathcal{U}(\mathcal{T}_{\sqrt{3}/3-\varepsilon})\right\}. \tag{12}$$

See Ceyhan (2010a) for the explicit forms of the support regions under the alternatives. These alternatives $H^S_\varepsilon$ and $H^A_\varepsilon$ with $\varepsilon \in (0, \sqrt{3}/3)$ can be transformed into the equilateral triangle as in Ceyhan et al. (2006, 2007).

For the standard equilateral triangle, we have $\varepsilon_i = \varepsilon$ for $i = 1, 2, 3$ in $T_j(\varepsilon) = \{x \in T_e : d(\mathsf{y}, \ell_j(x)) \le \varepsilon_j\}$. Thus $H^S_\varepsilon$ implies $X_i \overset{\text{iid}}{\sim} \mathcal{U}(T_e \setminus \mathcal{T}_\varepsilon)$ and $H^A_\varepsilon$ be the model under which $X_i \overset{\text{iid}}{\sim} \mathcal{U}(\mathcal{T}_{\sqrt{3}/3-\varepsilon})$. See Fig. 4 for a depiction of the above segregation and the association alternatives in $T_e$.

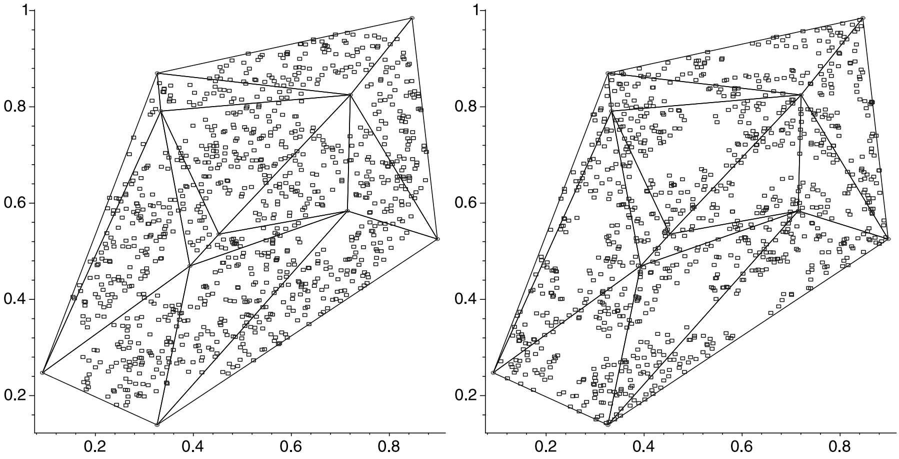


**Fig. 5** Realization of segregation (*left*) and association (*right*) with $n = 1000$ class 1 points for $|\mathcal{Y}_m| = 10$ points as in Fig. 3

The geometry invariance result also holds under the alternatives $H^S_\varepsilon$ and $H^A_\varepsilon$ for both PCD families (see Ceyhan 2010a). In particular, the segregation alternative with $\varepsilon \in (0, \sqrt{3}/4)$ in the standard equilateral triangle corresponds to the case that in an arbitrary triangle, $\kappa \times 100$ % of the area is carved away as forbidden from the vertices using line segments parallel to the opposite edge where $\kappa = 4\varepsilon^2$ (which implies $\kappa \in (0, 3/4)$). This argument is for the segregation alternative with $\varepsilon \in (0, \sqrt{3}/4)$; a similar construction is available for the other cases. In the multiple triangle case, the segregation and association alternatives, $H^S_\varepsilon$ and $H^A_\varepsilon$ with $\varepsilon \in (0, \sqrt{3}/3)$, are defined as in the one-triangle case, in the sense that, when each triangle (together with the data inside it) is transformed to the standard equilateral triangle, we obtain the same alternative pattern described above.

Thus in the case of $J_m > 1$, we have a (conditional) test of $H_o : X_i \overset{\text{iid}}{\sim} \mathcal{U}(C_H(\mathcal{Y}_m))$. The segregation (with $\kappa = 1/16$, i.e., $\varepsilon = \sqrt{3}/8$), and association (with $\kappa = 1/4$, i.e., $\varepsilon = \sqrt{3}/12$) realizations are depicted in Fig. 5 with $n = 1000$.

*Remark 4.1* There are many possible types of parameterizations for the alternatives. The particular parametrization of the alternatives in (12) is chosen so that the distribution of the relative density under the alternatives would also be geometry invariant (i.e., independent of the geometry of the support polygons). The more natural alternatives (i.e., the alternatives that are more likely to be found in practice) can be similar to or might be approximated by our parametrization. Because under a segregation alternative, the class 1 points will tend to be further away from class 2 points and under an association alternative, class 1 points will tend to cluster around the class 2 points. Such patterns can be detected by the test statistics based on the relative density, since under segregation (whether it is parameterized as above or not) we expect them to be larger, and under association (regardless of the parametrization) they tend to be smaller.

### 4.1 Consistency

Asymptotic normality of relative density of the PCDs under both alternative hypotheses of segregation and association were established by the same method as under the null hypothesis. In particular, asymptotic normality of relative density of PE-PCDs under the alternatives is proved in Ceyhan et al. (2006), while that for CS-PCDs for $\tau \in (0, 1]$ is proved in Ceyhan et al. (2007). The proof for $\tau \in (1, \infty)$ follows the same mechanism.

The relative density of the PCD is a test statistic for the segregation/association alternative; rejecting for extreme values of $\rho_{\text{PE}}(n, r)$ is appropriate, since under segregation, we expect $\rho_{\text{PE}}(n, r)$ to be larger, while under association, we expect $\rho_{\text{PE}}(n, r)$ to be smaller compared that under CSR.

In the one triangle case, using the standardized test statistic

$$R_{\text{PE}}(r) = \frac{\sqrt{n}(\rho_{\text{PE}}(r) - \mu_{\text{PE}}(r))}{\sqrt{\nu_{\text{PE}}(r)}}, \tag{13}$$

the asymptotic critical value for the one-sided level $\alpha$ test against segregation is given by $z_\alpha = \Phi^{-1}(1 - \alpha)$ where $\Phi(\cdot)$ is the standard normal distribution function. Against segregation, the test rejecting for $R_{\text{PE}}(r) > z_\alpha$ and against association, the test rejecting for $R_{\text{PE}}(r) < z_{1-\alpha}$ was shown to be consistent (Ceyhan et al. 2006). The same holds for the standardized test statistic in the multiple triangle case, $\widetilde{R}_{\text{PE}}(r) = \frac{\sqrt{n}(\widetilde{\rho}_{\text{PE}}(n,r) - \widetilde{\mu}_{\text{PE}}(r))}{\sqrt{\widetilde{\nu}_{\text{PE}}(r)}}$.

A similar construction is available for $\rho_{\text{CS}}(n, \tau)$ and consistency for $\tau \in (0, 1]$ was established in Ceyhan et al. (2007), consistency for $\tau > 1$ can be proved similarly.

## 5 Empirical size analysis under CSR

In one triangle case, for the null pattern of CSR, we generate $n$ class 1 points iid $\mathcal{U}(T_e)$ where $T_e$ is the standard equilateral triangle. We calculate the relative density of PE-PCDs for $r = 1, 11/10, 6/5, 4/3, \sqrt{2}, 3/2, 2, 3, 5, 10$, and that of CS-PCDs for $\tau = 0.2, 0.4, 0.6, \ldots, 3.0, 3.5, 4.0, \ldots, 20.0$ at each Monte Carlo replicate. We repeat the Monte Carlo procedure $N_{mc} = 10000$ times for each of $n = 10, 50, 100$. Using the critical values based on the normal approximation for the relative density, we calculate the empirical size estimates for both right-sided (i.e., for segregation) and left-sided (i.e., for association) tests as a function of the expansion parameters. Let $R_{\text{PE}}(r, j) := \frac{\sqrt{n}\,(\rho_{\text{PE}}(n,r,j) - \mu_{\text{PE}}(r))}{\sqrt{\nu_{\text{PE}}(r)}}$ be the standardized relative density for PE-PCD for Monte Carlo replicate $j$ with sample size $n$ for $j = 1, 2, \ldots, N_{mc}$. We estimate the empirical size against the segregation alternative as $\frac{1}{N_{mc}} \sum_{j=1}^{N_{mc}} \mathbf{I}(R_{\text{PE}}(r, j) > z_\alpha)$, and against the association alternative as $\frac{1}{N_{mc}} \sum_{j=1}^{N_{mc}} \mathbf{I}(R_{\text{PE}}(r, j) < -z_\alpha)$. For CS-PCDs, the standardized relative density $R_{\text{CS}}(\tau, j)$, asymptotic critical value, and empirical size are defined and calculated similarly. The empirical sizes significantly smaller (larger) than 0.05 are deemed conservative (liberal). The asymptotic normal approximation to proportions is used in determining the significance of the deviations of the empirical sizes from 0.05. For these proportion tests, we also use $\alpha = 0.05$ as the significance level. With $N_{mc} = 10000$, empirical sizes less than 0.0464 are deemed conservative, greater than 0.0536 are deemed liberal at $\alpha = 0.05$ level.

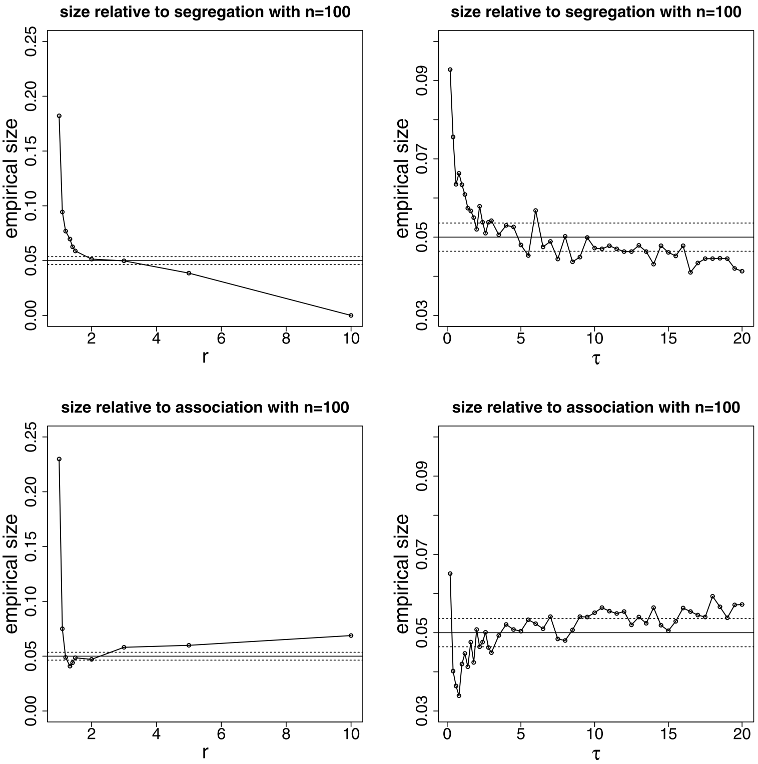


**Fig. 6** The empirical size estimates of the relative density of the PE-PCDs (*left*) and CS-PCDs (*right*) *in the one triangle case* based on 10000 Monte Carlo replicates for the right-sided alternative, (i.e., relative to segregation) (*top*) and the left-sided alternative, (i.e., relative to association) (*bottom*) with $n = 100$ under the CSR pattern (i.e., under $H_o : X_i \overset{\text{iid}}{\sim} \mathcal{U}(T_e)$ for $i = 1, 2, \ldots, n$). The horizontal lines are located at 0.0464 (upper threshold for conservativeness), 0.050 (nominal level), and 0.0536 (lower threshold for liberalness). Notice that the vertical and horizontal axes are differently scaled for the two PCD families

The empirical sizes together with upper and lower bounds of liberalness and conservativeness are plotted in Fig. 6 for $n = 100$. We only present the empirical size results for $n = 100$ in the one triangle case. The results for other $n$ values are deferred to the technical report Ceyhan (2010a).

With PE-PCDs, for the right-sided tests (i.e., relative to segregation) the size is close to the nominal level for $r \in (2, 3)$, for smaller $r$ values (i.e., $r < 2$), the test seems to be liberal with liberalness increasing as $r$ decreases; and for larger $r$ values (i.e., $r > 3$), the test seems to be conservative with conservativeness increasing as $r$ increases. For the left-sided tests (i.e., relative to association) the size is close to the nominal level for $r \in (1.5, 3)$, for other $r$ values the test seems to be liberal (more liberal for smaller $r$ values). This is due to the fact that very large and small values of $r$ require much larger sample sizes for the normal approximation to hold.

With CS-PCDs, for the right-sided tests, the size is close to the nominal level for $\tau \in (5, 14)$ and closest to 0.05 for $\tau \approx 5$ or $\tau \in (7, 9)$ for all sample sizes; for smaller $\tau$ values (i.e., $\tau \lesssim 4.5$) the test seems to be liberal with liberalness increasing as $\tau$ de-

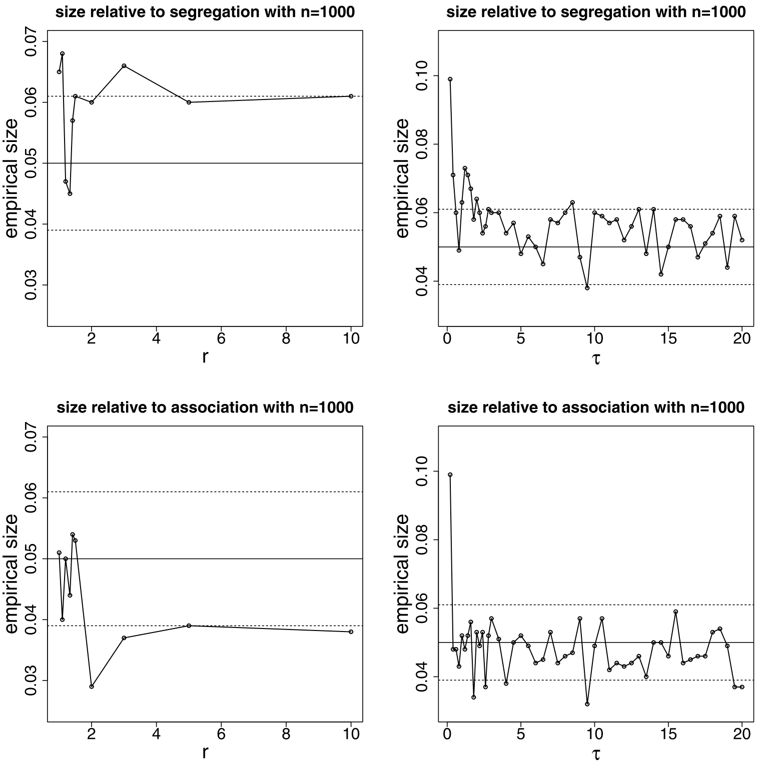


**Fig. 7** The empirical size estimates of the relative density of the PE-PCDs (*left*) and CS-PCDs (*right*) in the *multiple triangle case* based on 1000 Monte Carlo replicates under the segregation alternative (*top*) and the association alternative (*bottom*) with $n = 1000$ under the CSR pattern. The horizontal lines are located at 0.039 (upper threshold for conservativeness), 0.050 (nominal level), and 0.061 (lower threshold for liberalness). The vertical and horizontal axes are differently scaled for the two PCD families

creases; and for $\tau \gtrsim 15$ and the test is slightly conservative for $n = 100$. Considering all sample sizes (see the technical report Ceyhan 2010a), we recommend $\tau \in (5, 10)$ for testing against segregation. For the left-sided alternative, the test has the desired size for $\tau \in (2, 15)$. With all sample sizes, the test seems to be conservative (slightly liberal) for smaller (larger) $\tau$ values. Considering all sample sizes, we recommend $\tau \in (2.5, 5)$ for testing against association. The range of appropriate $\tau$ values gets wider with the increasing sample size and very large and small values of $\tau$ require much larger sample sizes for the normal approximation to hold.

In the multiple triangle case, for the null pattern of CSR, we generate $n$ class 1 points iid $\mathcal{U}(C_H(\mathcal{Y}_{10}))$ where $\mathcal{Y}_{10}$ is the set of the 10 class 2 points given in Fig. 3. With $N_{mc} = 1000$, empirical sizes less than 0.039 are deemed conservative and those greater than 0.061 are deemed liberal at $\alpha = 0.05$ level.

The empirical sizes for the PCDs together with upper and lower bounds of liberalness and conservativeness are plotted in Fig. 7 for $n = 1000$. Observe that in the multiple triangle case (which is more realistic than the one triangle case), the empirical sizes are much closer to the nominal level compared to the one triangle case.

With PE-PCDs, for the right-sided alternative (i.e., against segregation), the size is about the nominal level for $r \in (1.5, 3)$, and for the left-sided alternative (i.e., against association), the size is about the nominal level for $r \in (1.1, 2)$. Furthermore, although the empirical sizes for both right- and left-sided alternatives are about the desired level for $r$ values between 1.5 and 2, it seems that they are not very far from the nominal level for $r \in (1.5, 10)$. The test seems to be liberal for the segregation alternative and conservative for the association alternative, when not at the desired level.

With CS-PCDs, the empirical sizes are much closer to the nominal level compared to the one triangle case also. Furthermore, for the right-sided alternative with $n = 1000$, the test has the desired level for $\tau \geq 2$. Considering all sample sizes, we recommend $\tau \in (2.5, 8)$ for testing against segregation. For the left-sided alternative with $n = 1000$, $\tau \geq 0.5$ seems to yield the appropriate level. Considering all sample sizes, we recommend $\tau \in (0.5, 20)$ for testing against association.

*Remark 5.1 Empirical size comparison for the PCD families:* In the one triangle case, the size estimates for the CS-PCD is close to the nominal level of 0.05 against the segregation alternative for more of the expansion parameter values considered. On the other hand, the size estimates against association are close to the nominal level for both PCD families, but the size estimates for CS-PCD is closer to the nominal level. In the multiple triangle case, the size performance of the two PCD families is similar and the size estimates are close to the nominal level for both one-sided alternatives.

## 6 Empirical power analysis under the alternatives

To compare the power performance of the test statistics under the alternatives, we generate $n$ class 1 points uniformly in the corresponding support sets described in Sect. 4.

### 6.1 Empirical power analysis under the segregation alternative

In the one triangle case, at each Monte Carlo replicate under segregation, $H_\varepsilon^S$, we generate $X_i \overset{\text{iid}}{\sim} \mathcal{U}(T_e \setminus \mathcal{T}_\varepsilon)$, for $i = 1, 2, \ldots, n$ for $n = 10, 50, 100$ and compute the relative density of the PCDs. We consider $r \in \{1, 11/10, 6/5, 4/3, \sqrt{2}, 3/2, 2, 3, 5, 10\}$ for the PE-PCD and $\tau \in \{0.2, 0.4, 0.6, \ldots, 3.0, 3.5, 4.0, \ldots, 20.0\}$ for the CS-PCD. We repeat the above simulation procedure $N_{mc} = 10000$ times. We consider $\varepsilon \in \{\sqrt{3}/8, \sqrt{3}/4, 2\sqrt{3}/7\}$ (which correspond to 18.75 %, 75 %, and $4500/49 \approx$ 91.84 % of the triangle (around the vertices) being unoccupied by the class 1 points, respectively) under the segregation alternatives.

For PE-PCDs, under segregation alternatives with $\varepsilon > 0$, the distribution of $\rho_{\text{PE}}(n, r)$ is degenerate for large values of $r$. For a given $\varepsilon \in (0, \sqrt{3}/4)$, the corresponding digraph is complete a.s. when $r \geq \frac{\sqrt{3}}{2\varepsilon}$, hence $\rho_{\text{PE}}(n, r) = 1$ a.s. For $\varepsilon \in (\sqrt{3}/4, \sqrt{3}/3)$, the corresponding digraph is complete a.s., when $r \geq \frac{\sqrt{3}-2\varepsilon}{\varepsilon}$.

In particular, for $\varepsilon = \sqrt{3}/8$, $\rho_{\mathrm{PE}}(n, r)$ is degenerate when $r \geq 4$, for $\varepsilon = \sqrt{3}/4$, $\rho_{\mathrm{PE}}(n, r)$ is degenerate when $r \geq 2$, and for $\varepsilon = 2\sqrt{3}/7$, $\rho_{\mathrm{PE}}(n, r)$ is degenerate when $r \geq 3/2$. Such a problem does not occur for CS-PCDs.

For a given alternative and sample size, we analyze the empirical power of the test based on $\rho_{\mathrm{PE}}(n, r)$ and $\rho_{\mathrm{CS}}(n, \tau)$—using the asymptotic critical value—as a function of the expansion parameters $r$ and $\tau$, respectively. We estimate the empirical power for PE-PCDs as $\frac{1}{N_{mc}} \sum_{j=1}^{N_{mc}} \mathbf{I}(R_{\mathrm{PE}}(r, j) > z_\alpha)$. The empirical power for CS-PCDs is estimated similarly.

In Fig. 8, we present Monte Carlo power estimates for relative density of the PCDs in the one triangle case as a function of expansion parameters for $n = 10, 50, 100$ against $H^S_{\sqrt{3}/4}$ only. Notice that, for PE-PCDs, Monte Carlo power estimate increases as $r$ gets larger and then decreases, due to the magnitude of $r$ and $n$. Because for small $n$ and large $r$, the critical value is approximately 1 under $H_o$, as we get a complete digraph with high probability. Under moderate segregation (with $\varepsilon = \sqrt{3}/4$), $r$ around 1.5 to 5 yields the highest power (for other $r$ values, the power performance is very poor). Furthermore, under moderate to severe segregation, with $n = 10$ the power estimate seems to be close to 1 for $r \in (1, 4)$, and with $n = 50$ or 100 the power estimate seems to be close to 1 for $r \in (1, 5)$. However, the power estimates are valid only for $r$ within $(2, 3)$, since the test has the desired size for this range of $r$ values against the right-sided alternative. So, for small sample sizes, $r \approx 1.5$ is recommended, and for larger sample sizes, moderate values of $r$ (i.e., $r \in (2, 3)$) are recommended for the segregation alternative as they are more appropriate for normal approximation and they yield the desired significance level.

For CS-PCDs, Monte Carlo power estimate increases as $\tau$ gets larger or $n$ gets larger. With $n = 10$, the power estimates are high for $\tau \in (5, 14)$ and virtually 0 for $\tau \geq 14$. With $n = 50$ or 100, the power values are high for $\tau \geq 1$, with highest power occurring around $\tau \approx 8$. However, for $\tau \geq 6$, the power values are virtually same. Considering the empirical size estimates, we recommend $\tau \approx 8$ for mild segregation, and $\tau \approx 5$ for more severe segregation alternatives.

In the multiple triangle case, we generate the class 1 points uniformly in the support for the segregation alternatives in the triangles based on the 10 class 2 points given in Fig. 3. We use the parameters $\varepsilon \in \{\sqrt{3}/8, \sqrt{3}/4, 2\sqrt{3}/7\}$. The corresponding empirical power estimates as a function of expansion parameters $r$ and $\tau$ (using the normal approximation) are presented in Fig. 9 for $\varepsilon = \sqrt{3}/4$ for $n = 500$ and $n = 1000$. Observe that, for PE-PCDs, the Monte Carlo power estimate increases as $r$ gets larger and then decreases, as in the one triangle case. The empirical power is maximized for $r \in (1.5, 2)$ under mild segregation, and for $r \in (1.5, 3)$ under moderate to severe segregation. Considering the empirical size and power estimates, $r \approx 1.5$ is recommended under mild segregation, while $r \in (2, 3)$ seems to be more appropriate (hence recommended for more severe segregation), since the corresponding test has the desired level with high power.

For CS-PCDs, the Monte Carlo power estimate tends to increase as $\tau$ gets larger. Under mild segregation with $\varepsilon = \sqrt{3}/8$, the empirical power is large for $\tau \geq 2$ with largest being around $\tau \in (4, 8)$. Under moderate to severe segregation, the empirical power is virtually one for $\tau \geq 0.4$. Considering the empirical size and power estimates, $\tau \approx 7$ seems to be more appropriate (hence recommended for segregation), since the corresponding test has the desired level with highest power.

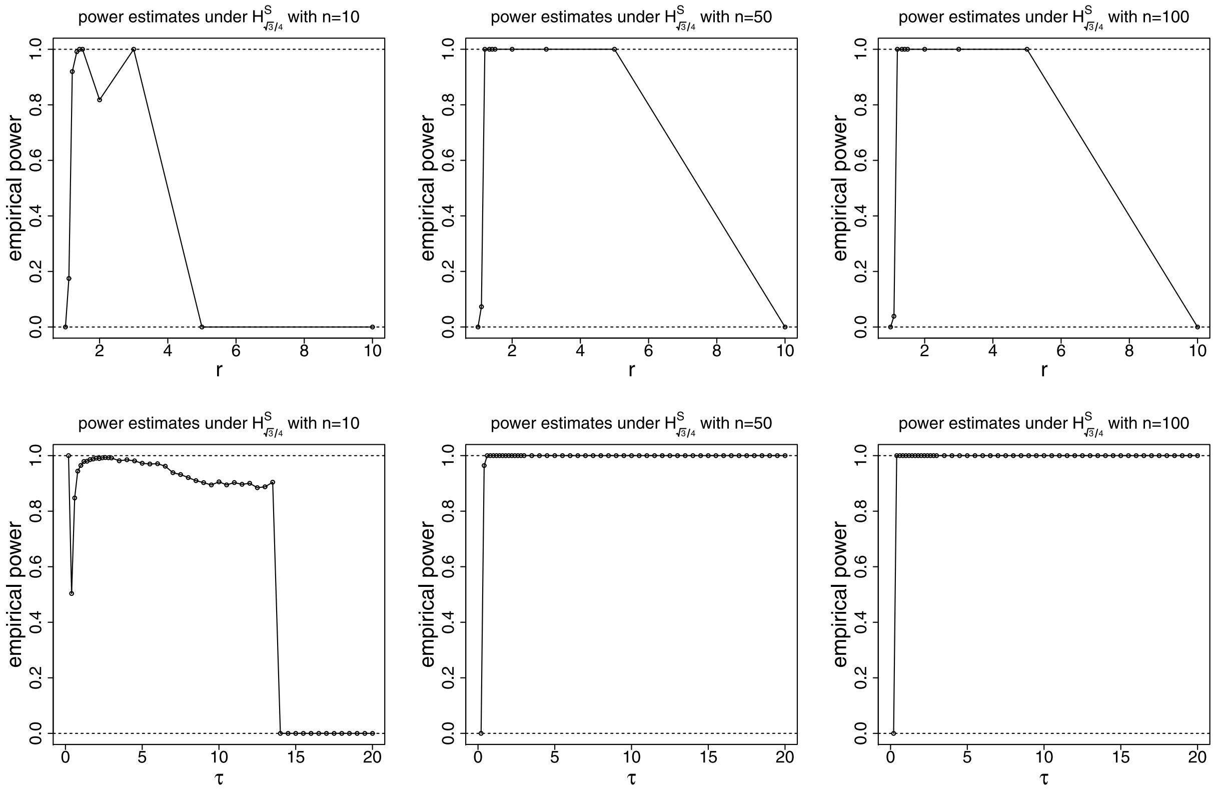


**Fig. 8** Empirical power estimates under segregation in the one triangle case for relative density of PE-PCDs (*top*) and CS-PCDs (*bottom*) using the asymptotic critical value against segregation alternative $H^S_{\sqrt{3}/4}$ as a function of the expansion parameters for $n = 10, 50$ and 100

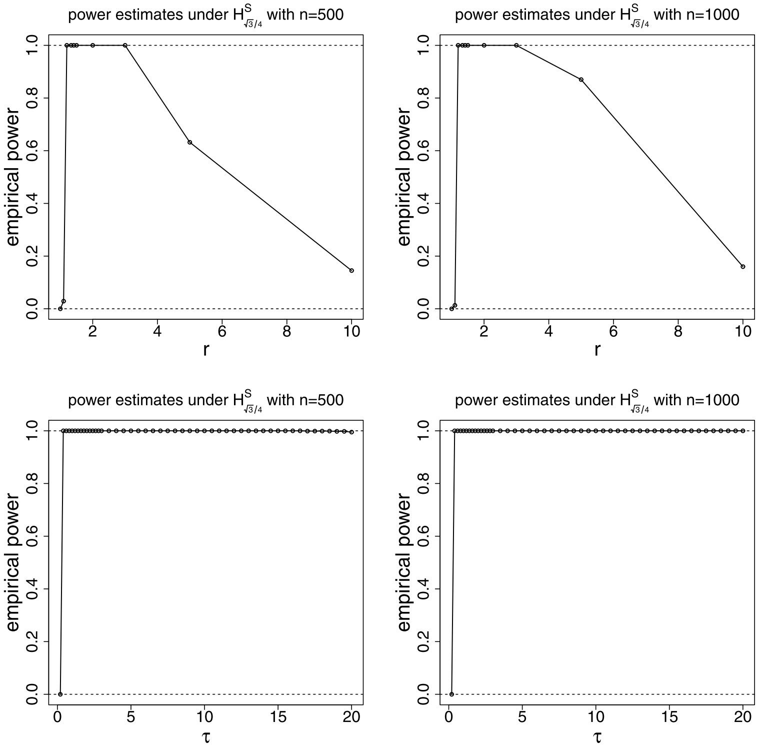


**Fig. 9** Empirical power estimates under segregation in the multiple triangle case for PE-PCDs (*top*) and CS-PCDs (*bottom*) using the asymptotic critical value against segregation alternative $H^S_{\sqrt{3}/4}$ as a function of expansion parameters for $n = 500$ (*left*) and $n = 1000$ (*right*)

## 6.2 Empirical power analysis under the association alternative

In the one triangle case, at each of $N_{mc} = 10000$ Monte Carlo replicates under association, $H^A_\varepsilon$, we generate $X_i \overset{\text{iid}}{\sim} \mathcal{U}(\mathcal{T}_{\sqrt{3}/3-\varepsilon})$, for $i = 1, 2, \ldots, n$ for $n = 10, 50, 100$. Unlike the segregation alternatives, the distribution of $\rho_{\text{PE}}(n, r)$ is non-degenerate for all $\varepsilon \in (0, \sqrt{3}/3)$ and $r \in [1, \infty)$. We consider $\varepsilon \in \{5\sqrt{3}/24, \sqrt{3}/12, \sqrt{3}/21\}$ (which correspond to 18.75 %, 75 %, and $4500/49 \approx 91.84$ % of the triangle being occupied around the class 2 points by the class 1 points, respectively) for the association alternatives.

Under association, for each $r$ value, we estimate the empirical power as $\frac{1}{N_{mc}} \sum_{j=1}^{N_{mc}} \mathbf{I}(R_{\text{PE}}(r, j) < -z_\alpha)$. The asymptotic critical value and empirical power for CS-PCDs are similarly defined. In Fig. 10, we present Monte Carlo power estimates for relative density of the PCDs in the one triangle case against $H^A_{5\sqrt{3}/24}$ as a function of $r$ for $n = 10, 50, 100$. Notice that, for PE-PCDs, Monte Carlo power estimate increases as $r$ gets larger and then decreases, as in the segregation case. Because for small $n$ and large $r$, the critical value is approximately one under $H_o$, as we get a nearly complete digraph with high probability. Highest power is attained for

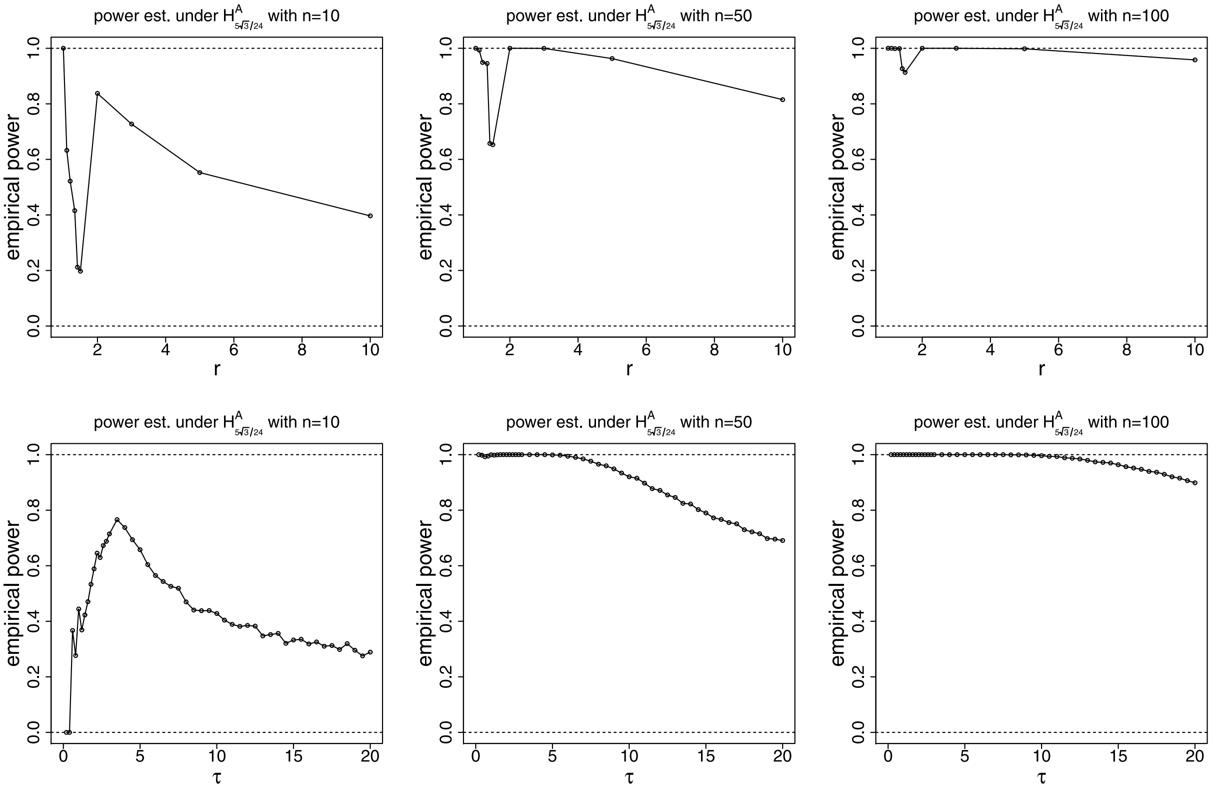


**Fig. 10** Empirical power estimates under association in the one triangle case for relative density of PE-PCDs (*top*) and CS-PCDs (*bottom*) using the asymptotic critical value against association alternative $H^A_{5\sqrt{3}/24}$ as a function of the expansion parameters for $n = 10, 50$ and 100

$r \approx 2$, which is recommended against the association, as it yields the desired level with high power.

For CS-PCDs, under mild association and small $n$, highest power is attained around $\tau \approx 3$, under mild association with large $n$, power increases as $\tau$ increases. For moderate to severe association and large $n$, power is virtually one for all the $\tau$ values we considered. Considering the empirical size and power performances, we recommend $\tau \approx 5$, as it has the desired level and high power.

In the multiple triangle case, we generate the class 1 points uniformly in the support for the association alternatives in the triangles based on the 10 class 2 points given in Fig. 3. We use the parameters $\varepsilon \in \{5\sqrt{3}/24, \sqrt{3}/12, \sqrt{3}/21\}$. The corresponding empirical power estimates as a function of $r$ (using the normal approximation) are presented in Fig. 11 for $\varepsilon = 5\sqrt{3}/24$ for $n = 500$ and $n = 1000$. Observe that, for PE-PCDs, the Monte Carlo power estimate decreases as $r$ gets larger unlike the one triangle case. The empirical power is large (i.e., close to one) for $r \in (1, 5)$. Considering the empirical size estimates, we recommend $r \approx 2$ for association alternative, since the corresponding test has the desired level with high power.

For CS-PCDs, the Monte Carlo power estimate tends to decrease as $\tau$ gets larger. The empirical power is maximized for $\tau \leq 1$. Considering the empirical size and power estimates, we recommend $\tau \approx 1$ for association, since the corresponding test has the desired level with high power.

We only present the empirical power results under some of the alternatives in the one and multiple triangle cases. For the results under other alternatives, see the technical report Ceyhan (2010a).

*Remark 6.1 Empirical power comparison for the two PCD families:* In the one triangle case, under the segregation alternatives, the power estimates of the CS-PCDs tend to be higher than those of the PE-PCDs. Under mild to moderate association alternatives, CS-PCDs have higher power estimates, while under severe association, PE-PCD has higher power estimates. In the multiple triangle case, under segregation, CS-PCDs has higher power estimates; while under association, PE-PCDs has higher power estimates.

## 7 Pitman asymptotic efficiency

Pitman asymptotic efficiency (PAE) provides for an investigation of “local (around $H_o$) asymptotic power”. This involves the limit as $n \to \infty$ as well as the limit as $\varepsilon \to 0$ under the alternatives. A detailed discussion of PAE is available in Kendall and Stuart (1979), van Eeden (1963) and Ceyhan (2010a).

Under segregation or association alternatives, the PAE of $\rho_{\text{PE}}(n, r)$ is given by

$$\text{PAE}_{\text{PE}}(r) = \frac{(\mu^{(k)}(r, \varepsilon = 0))^2}{\nu_{\text{PE}}(r)}$$

where $k$ is the minimum order of the derivative with respect to $\varepsilon$ for which $\mu^{(k)}(r, \varepsilon = 0) \neq 0$. That is, $\mu^{(k)}(r, \varepsilon = 0) \neq 0$ but $\mu^{(l)}(r, \varepsilon = 0) = 0$ for $l = 1, 2, \ldots, k-1$. Sim-

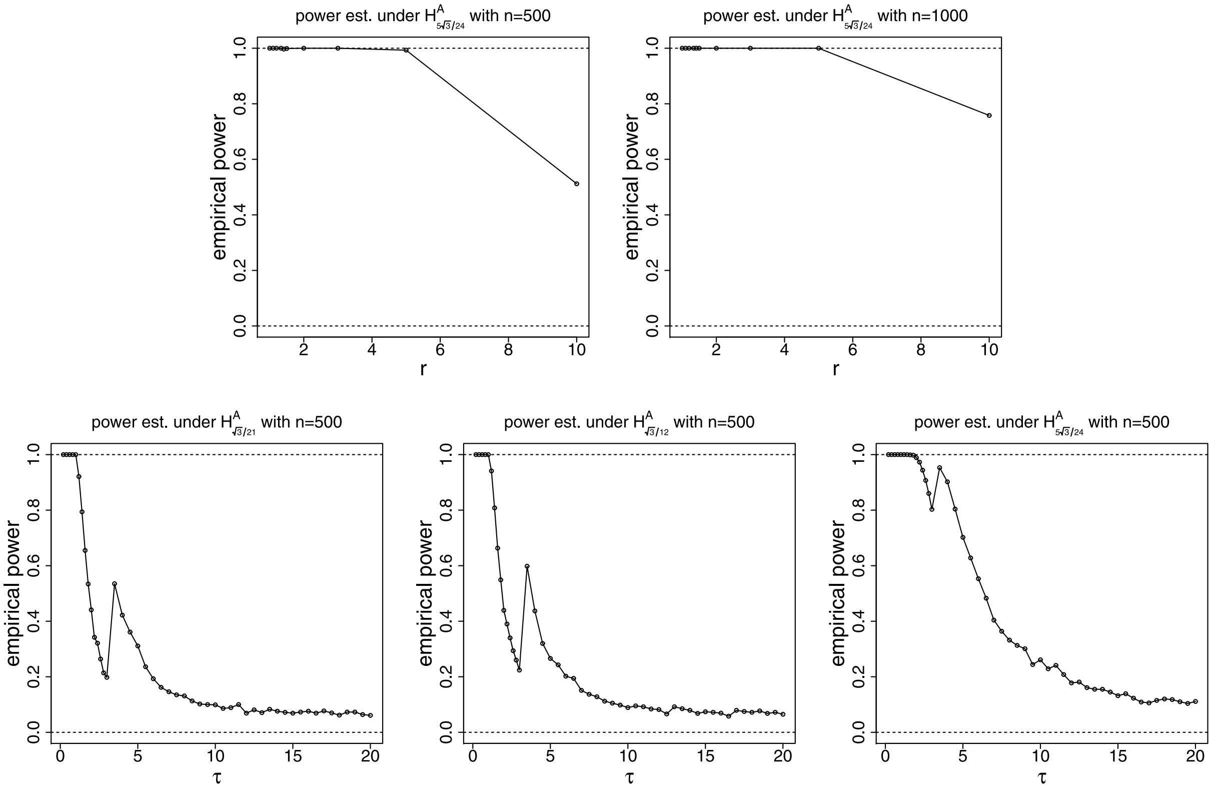


**Fig. 11** Empirical power estimates under association in the multiple triangle case for the relative density of PE-PCDs (*top*) and CS-PCDs (*bottom*) as a function of the expansion parameters

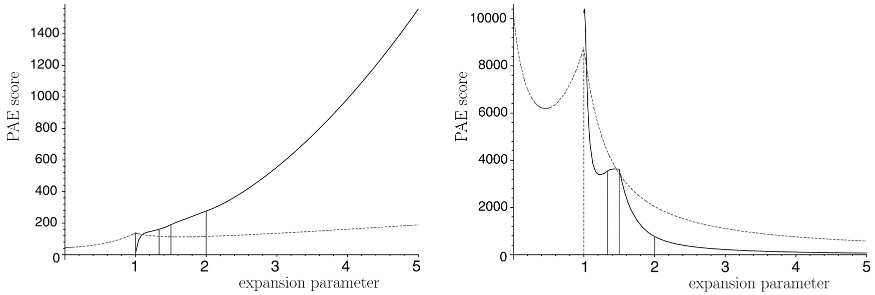


**Fig. 12** Pitman asymptotic efficiency against segregation (*left*) and association (*right*) alternatives as a function of the expansion parameters in the *one triangle case* for the relative density of PE-PCDs (*solid line*) and CS-PCDs (*dashed line*). Notice that the vertical axes are differently scaled

ilarly, the PAE of $\rho_{\text{CS}}(n, \tau)$ is given by

$$\text{PAE}_{\text{CS}}(\tau) = \frac{(\mu^{(k)}(\tau, \varepsilon = 0))^2}{\nu_{\text{CS}}(\tau)}$$

where $k$ is defined as above. For PE-PCDs and CS-PCDs, we need $k = 2$ under both segregation and association alternatives. See Ceyhan (2010a) for details.

### 7.1 PAE analysis in the one-triangle case

The PAE scores for PE-PCDs were calculated in Ceyhan et al. (2006) and for CS-PCDs with $\tau \in (0, 1]$ were calculated in Ceyhan et al. (2007). We extend the PAE calculations for $\tau > 1$ (the details deferred to technical report Ceyhan 2010a).

In Fig. 12 (left), we present the PAE as a function of the expansion parameter for segregation. The corresponding PAE score is denoted with an "$S$" in the superscript. Notice that $\text{PAE}^S_{\text{PE}}(r = 1) = 160/7 \approx 22.86$ and $\lim_{r\to\infty} \text{PAE}^S_{\text{PE}}(r) = \infty$. Furthermore, $\lim_{\tau\to 0} \text{PAE}^S_{\text{CS}}(\tau) = 320/7 \approx 45.71$ and $\lim_{\tau\to\infty} \text{PAE}^S_{\text{CS}}(\tau) = \infty$. Moreover, a local maximum occurs at $\tau = 1$ with $\text{PAE}^S_{\text{CS}}(\tau = 1) = 960/7 \approx 137.14$ and a local minimum occurs at $\tau \approx 1.62$ with PAE score $\approx 112.70$.

Based on the PAE analysis, we suggest, for large $n$ and small $\varepsilon$, choosing expansion parameters large for testing against segregation. However, for small and moderate values of $n$, normal approximation is not appropriate due to the skewness in the density of $\rho_{\text{PE}}(n, r)$ (or $\rho_{\text{CS}}(n, \tau)$) for extreme values of $r$ (or $\tau$). Therefore, for small $n$, we suggest moderate $r$ values for PE-PCDs and moderate $\tau$ values (i.e., $\tau \in [7, 8]$) for CS-PCDs.

Comparing the PAE scores of the relative density of PE-PCDs and CS-PCDs under segregation alternatives, we see that $\text{PAE}^S_{\text{PE}}(t) < \text{PAE}^S_{\text{CS}}(t)$ for $1 \leq t \lesssim 1.09$; and $\text{PAE}^S_{\text{PE}}(t) > \text{PAE}^S_{\text{CS}}(t)$ for $t \gtrsim 1.09$. Therefore, under segregation alternatives, overall, relative density of PE-PCD is asymptotically more efficient compared to the CS-PCD. Furthermore, $\text{PAE}^S_{\text{PE}}(t)$ tends to $\infty$ as $t \to \infty$ at rate $O(t^2)$ while $\text{PAE}^S_{\text{CS}}(t)$ tends to $\infty$ as $t \to \infty$ at rate $O(t)$.

In Fig. 12 (right), we present the PAE as a function of the expansion parameter for association. The corresponding PAE score is denoted with an "$A$" in the superscript. Notice that $\text{PAE}^A_{\text{PE}}(r = 1) = 174240/17 \approx 10249.41$, $\lim_{r\to\infty} \text{PAE}^A_{\text{PE}}(r) = 0$

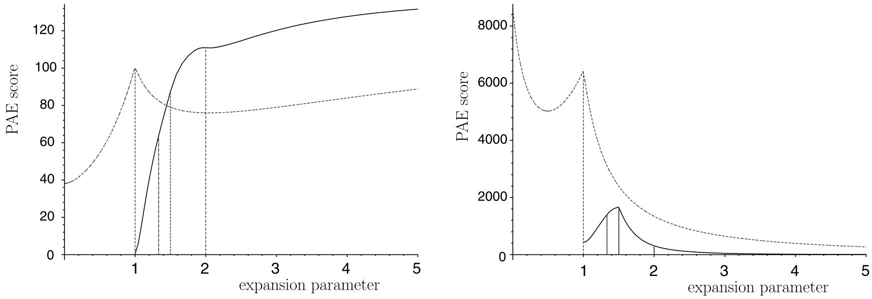


**Fig. 13** Pitman asymptotic efficiency against segregation (*left*) and association (*right*) alternatives as a function of expansion parameters in the *multiple triangle case* with the realization of $\mathcal{Y}_m$ given in Fig. 3 for the relative density of PE-PCDs (*solid line*) and CS-PCDs (*dashed line*). Notice that vertical axes are differently scaled

and $\text{argsup}_{r\in[1,\infty)} \text{PAE}^A_{\text{PE}}(r) \approx 1.01$ with supremum $\approx 10399.77$. $\text{PAE}^A_{\text{PE}}(r)$ has also a local supremum at $r_l \approx 1.44$ with local supremum $\approx 3630.89$. Moreover, we have $\lim_{\tau\to 0} \text{PAE}^A_{\text{CS}}(\tau) = 72000/7 \approx 10285.71$ which is also the global maximum. Moreover, a local minimum of $\text{PAE}^A_{\text{CS}}(\tau)$ occurs at $\tau \approx 0.45$ with PAE score being equal to $\approx 6191.67$ and a local maximum occurs at $\tau = 1$ with $\text{PAE}^A_{\text{CS}}(\tau = 1) = 61440/7 \approx 8777.14$.

Based on the PAE analysis, we suggest, for large $n$ and small $\varepsilon$, choosing $r$ (or $\tau$) small for testing against association. However, for small and moderate values of $n$, normal approximation is not appropriate due to the skewness in the density of $\rho_{\text{PE}}(n, r)$ (or $\rho_{\text{CS}}(n, \tau)$). Therefore, for small $n$, we suggest moderate $r$ values for PE-PCDs and $\tau \approx 1$ for CS-PCDs.

Comparing the PAE scores of the relative density of PE-PCDs and CS-PCDs under association alternatives, we see that $\text{PAE}^A_{\text{PE}}(t) < \text{PAE}^A_{\text{CS}}(t)$ for $1 \leq t \lesssim 1.46$ and for $t \gtrsim 1.52$; and $\text{PAE}^A_{\text{PE}}(t) > \text{PAE}^A_{\text{CS}}(t)$ for $1.46 \lesssim t \lesssim 1.52$. Under association, relative density of CS-PCD is asymptotically more efficient compared to that of the PE-PCD. Furthermore, $\text{PAE}^A_{\text{PE}}(t)$ goes to 0 as $t \to \infty$ at rate $O(t^{-2})$ while $\text{PAE}^A_{\text{CS}}(t)$ goes to 0 as $t \to \infty$ at rate $O(t^{-1})$.

## 7.2 PAE analysis in the multiple triangle case

For $J_m > 1$ (i.e., $m > 3$), in addition to the expansion parameter, PAE analysis depends on the number of triangles as well as the relative sizes of the triangles (i.e., on $\mathcal{Y}_m$). So the optimal expansion parameter values with respect to the PAE criteria in the multiple triangle case might be different than that of the one triangle case. See the technical report Ceyhan (2010a) for explicit form of the PAE scores in the multiple triangle case.

In Fig. 13 (left), we present the PAE scores as a function the expansion parameter under segregation alternative conditional on the realization of $\mathcal{Y}_m$ given in Fig. 3. Let $\text{PAE}^S_{\text{PE}}(m, r)$ be the PAE score for the PE-PCD under the segregation alternative in the multiple triangle case and define the PAE scores under association and those for CS-PCDs in the multiple triangle case similarly. Notice that, unlike the one triangle case, $\mathrm{PAE}^S_{\mathrm{PE}}(m,r)$ is bounded from above by $\lim_{r\to\infty}\mathrm{PAE}^S_{\mathrm{PE}}(m,r)\approx 139.34$. Some values of interest are $\mathrm{PAE}^S_{\mathrm{PE}}(m,r=1)\approx 0.39$, and a local maximum value of $\approx 110.97$ is attained at the $\operatorname{argsup}_{r\in[1,2]}\mathrm{PAE}^S_{\mathrm{PE}}(m,r)\approx 1.97$. On the other hand, the PAE curve for the CS-PCDs in the multiple triangle case is similar to that in the one triangle case (See Fig. 12 (left)). But unlike the one triangle case, $\mathrm{PAE}^S_{\mathrm{CS}}(m,\tau)$ is bounded with $\lim_{\tau\to\infty}\mathrm{PAE}^S_{\mathrm{CS}}(m,\tau)\approx 139.34$. Some values of note are $\lim_{\tau\to 0}\mathrm{PAE}^S_{\mathrm{CS}}(m,\tau)\approx 38.20$; and a local maximum of $\approx 100.77$ is attained at $\tau=1$, and a local minimum of $\approx 75.97$ is attained at $\tau\approx 2.04$. Based on the PAE analysis of the relative density of PE-PCDs, under segregation alternative, larger $r$ values yield larger asymptotic relative efficiency. However, due to the skewness of the pdf of $\rho_{\mathrm{PE}}(m,r)$, moderate $r$ values ($r$ around 1.5 or 2) are recommended. As for the CS-PCDs, larger $\tau$ values have larger asymptotic relative efficiency. However, due to the skewness of the pdf of $\rho_{\mathrm{CS}}(m,\tau)$, moderate $\tau$ values ($\tau$ around 1) are recommended.

Comparing the PAE scores for PE-PCDs and CS-PCDs under the segregation alternative, we see that for $1\le t\lesssim 1.45$ asymptotic relative efficiency of relative density of CS-PCDs is larger, since $\mathrm{PAE}^S_{\mathrm{CS}}(m,t)>\mathrm{PAE}^S_{\mathrm{PE}}(m,t)$, and for $t\gtrsim 1.45$ asymptotic relative efficiency of relative density of PE-PCDs is larger since $\mathrm{PAE}^S_{\mathrm{CS}}(m,t)<\mathrm{PAE}^S_{\mathrm{PE}}(m,t)$. Therefore, PE-PCD tends to be more efficient asymptotically compared to the CS-PCD under segregation.

In Fig. 13 (right), we present the PAE scores as a function the expansion parameter under association alternative conditional on the realization of $\mathcal{Y}_m$ given in Fig. 3. Notice that, as in the one triangle case, $\mathrm{PAE}^A_{\mathrm{PE}}(m,r)$ tends to 0 as $r\to\infty$. Some values of interest are $\mathrm{PAE}^A_{\mathrm{PE}}(m,r=1)\approx 422.96$, and a global maximum value of $\approx 1855.97$ is attained at $r=1.5$. On the other hand, the PAE curve for the CS-PCDs in the multiple triangle case is similar to the one in the one triangle case. (See Fig. 12 (left)). Notice also that $\lim_{\tau\to 0}\mathrm{PAE}^A_{\mathrm{CS}}(m,\tau)\approx 8593.97$; a local maximum value of $\approx 6449.54$ is attained at $\tau=1$; and a local minimum value of $\approx 5024.22$ is attained at $\tau\approx 0.49$. Moreover, $\lim_{\tau\to\infty}\mathrm{PAE}^A_{\mathrm{CS}}(m,\tau)=0$ at rate $O(\tau^{-2})$. Based on the PAE analysis for relative density of PE-PCDs, smaller $\tau$ values tend to yield larger asymptotic relative efficiency. However, we suggest, for large $n$ and small $\varepsilon$, choosing moderate $\tau$ for testing against association due to the skewness of the density of $\rho_{\mathrm{CS}}(n,\tau)$ for very small $\tau$ values.

Comparing the PAE scores for PE-PCDs and CS-PCDs, under the association alternative, we see that asymptotic relative efficiency of relative density of CS-PCDs is larger for $t\ge 1$, since $\mathrm{PAE}^A_{\mathrm{CS}}(m,t)>\mathrm{PAE}^A_{\mathrm{PE}}(m,t)$. Therefore, CS-PCD tends to be more efficient asymptotically compared to the PE-PCD under association.

*Remark 7.1 Empirical power versus PAE for the PCD families:* The finite sample performance (based on the Monte Carlo simulations) and the asymptotic efficiency (based on PAE scores) may seem to give conflicting results. The reason for this is two fold: (i) in the Monte Carlo simulations, we only have a finite number of observations, and the asymptotic normality of the relative density of the PCDs require smaller sample sizes for moderate values of the expansion parameters, and (ii) PAE is designed for infinitesimal deviations from the null hypothesis (i.e., as close as possible to the null case), while in our simulations we use mild to severe but fixed levels of deviations. Hence, if we had extremely large samples, the results of our finite sample and asymptotic comparisons would agree under extremely mild segregation or association alternatives.

Furthermore, when the PAE scores are compared at the optimal expansion parameters, the comparison results agree with that of the Monte Carlo simulation results. In particular, recall that in the one triangle case, the optimal parameters for PE-PCDs were 1.5 and 2 (and for CS-PCDs, they were 8 and 5) against mild segregation and association, respectively. Under segregation, CS-PCD is asymptotically more efficient, while under association, PE-PCD is asymptotically more efficient at these optimal parameters. These agree with the conclusions of empirical power comparisons. In the multiple triangle case, the optimal parameters for PE-PCDs were 1.5 and 2 (and for CS-PCDs, they were 7 and 1) against mild segregation and association, respectively. Under both alternatives, CS-PCD is asymptotically more efficient. In this case, only the segregation results are in agreement. The power estimates under association were virtually same at these optimal values for both PCD families.

An extension of PE proximity regions and CS proximity regions to higher dimensions (hence the corresponding PCDs to data in higher dimensions) are provided in Ceyhan et al. (2006, 2007), respectively.

## 8 Correction for class 1 points outside the convex hull of $\mathcal{Y}_m$

Our null hypothesis in (10) is somewhat restrictive, in the sense that, it might not be realistic to assume the support of class 1 points being $C_H(\mathcal{Y}_m)$ in practice. Up to now, our inference was restricted to the $C_H(\mathcal{Y}_m)$. However, crucial information from the data (hence power) might be lost, since a substantial proportion of class 1 points, denoted $\pi_{\text{out}}$, might fall outside the $C_H(\mathcal{Y}_m)$. A correction is suggested in Ceyhan (2011) to mitigate the effect of $\pi_{\text{out}}$ (or restriction to the $C_H(\mathcal{Y}_m)$) on the use of the domination number for the PE-PCDs. We propose a similar correction for the points outside the $C_H(\mathcal{Y}_m)$ for the relative density in this article.

Along this line, Ceyhan (2011) estimated the $\pi_{\text{out}}$ values for independently generated $\mathcal{X}_n$ and $\mathcal{Y}_m$ as random samples from $\mathcal{U}((0,1)\times(0,1))$. The considered values were $n = 100, 200, \ldots, 900, 1000, 2000, \ldots, 9000, 10000$ for each of $m = 10, 20, \ldots, 50$. The procedure is repeated $N_{mc} = 1000$ times for each $n, m$ combination. Let $\widehat{\pi}_{\text{out}}$ be the estimate of the proportion of class 1 points outside the $C_H(\mathcal{Y}_m)$ which is obtained by averaging the $\pi_{\text{out}}$ values (over $n$) for each $m, n$ combination. The simulation results suggested that $\widehat{\pi}_{\text{out}} \approx 1.7932/m + 1.2229/\sqrt{m}$ (see Ceyhan 2011). Notice that as $m \to \infty$, we have $\widehat{\pi}_{\text{out}} \to 0$.

Based on the Monte Carlo simulation results, we propose a coefficient to adjust for the proportion of class 1 points outside $C_H(\mathcal{Y}_m)$, namely,

$$C_{\text{ch}} := \text{signum}\big(p_{\text{out}} - \mathbf{E}[\widehat{\pi}_{\text{out}}]\big) \times \big(p_{\text{out}} - \mathbf{E}[\widehat{\pi}_{\text{out}}]\big)^2 \tag{14}$$

where $\text{signum}(p_{\text{out}} - \mathbf{E}[\widehat{\pi}_{\text{out}}])$ is the sign of the difference $p_{\text{out}} - \mathbf{E}[\widehat{\pi}_{\text{out}}]$ and $p_{\text{out}}$ is the observed proportion and $\mathbf{E}[\widehat{\pi}_{\text{out}}] \approx 1.7932/m + 1.2229/\sqrt{m}$ is the expected proportion of class 1 points outside $C_H(\mathcal{Y}_m)$. For the test statistics in Sect. 4.1, we suggest

$$\widetilde{R}^{\text{ch}}_{\text{PE}}(r) := \widetilde{R}_{\text{PE}}(r) + C_{\text{ch}}\left|\widetilde{R}_{\text{PE}}(r)\right| \quad \text{and} \quad \widetilde{R}^{\text{ch}}_{\text{CS}}(\tau) := \widetilde{R}_{\text{CS}}(\tau) + C_{\text{ch}}\left|\widetilde{R}_{\text{CS}}(\tau)\right|. \quad (15)$$

The convex hull adjustment slightly affects the empirical size estimates under CSR of class 1 and 2 points in the same rectangular support, since $p_{\text{out}}$ and $\mathbf{E}[\widehat{\pi}_{\text{out}}]$ values would be very similar. On the other hand, under segregation alternatives, we expect $\widetilde{R}^{\text{ch}}_{\text{PE}}(r)$ value and $p_{\text{out}} - \mathbf{E}[\widehat{\pi}_{\text{out}}]$ to be positive, so the convex hull correction increases the value of $\widetilde{R}_{\text{PE}}(r)$ in favor of the right-sided alternative (i.e., segregation). Under association alternatives, we expect $\widetilde{R}^{\text{ch}}_{\text{PE}}(r)$ value and $p_{\text{out}} - \mathbf{E}[\widehat{\pi}_{\text{out}}]$ to be negative, so the convex hull correction decreases the value of $\widetilde{R}_{\text{PE}}(r)$ in favor of the left-sided alternative (i.e., association).

## 9 Example data set

We illustrate the method on an ecological data set, namely, swamp tree data of Dixon (2002b). Good and Whipple (1982) considered the spatial patterns of tree species along the Savannah River, SC, USA. From this data, Dixon (2002b) used a single 50 m × 200 m rectangular plot (denoted as the $(0, 200) \times (0, 50)$ rectangle) to illustrate his nearest neighbor contingency table (NNCT) methods. All live or dead trees with 4.5 cm or more dbh (diameter at breast height) were recorded together with their species labels. The plot contains 13 different tree species, four of which comprising over 90 % of the 734 tree stems. See Ceyhan (2010c) for more detail on the data.

In this article, we only consider the middle 50 m × 55 m rectangular plot from the original study area (i.e., the subset $(95, 150) \times (0, 50)$ of the 50 m × 200 m rectangular plot) and investigate the spatial interaction of all other tree species (i.e., other than bald cypress trees) with bald cypresses (i.e., bald cypresses are taken to be the class 2 points, while all other trees are taken to be the class 1 points; hence Delaunay triangulation is based on the locations of bald cypresses). The study area contains 8 bald cypress trees and 156 other trees. See also Fig. 14 which is suggestive of segregation of other trees from bald cypresses.

For this data, we find that 108 other trees are inside and 48 are outside of the convex hull of bald cypresses. Hence the proportion of other trees outside the convex hull of bald cypresses is $p_{\text{out}} = 0.3077$ and the expected proportion is $\pi_{\text{out}} = 0.6515$. Hence the convex hull correction decreases the magnitude of the raw test statistics. We calculate the standardized test statistics, $R_{\text{PE}}(r)$, for $r = 1, 11/10, 6/5, 4/3, \sqrt{2}, 3/2, 2, 3, 5, 10$ values and, $R_{\text{CS}}(\tau)$, for $\tau = 0.2, 0.4, 0.6. \ldots, 3.0, 3.5, 4.0, \ldots, 20.0$ values and the corresponding convex hull corrected versions. The $p$-values based on the normal approximation are presented in Fig. 15. Observe that, with $R_{\text{PE}}(r)$, the convex hull corrected version is not significant (for both the right- and the left-sided alternatives) at 0.05 level at any of the $r$ values considered (only significant at 0.10 level at $r$ between 1.4 and 2.0 for the right-sided alternative), while the uncorrected version is significant (at 0.05 level) for $r$ values between 1.4 and 2.0. On the other hand, with $R_{\text{CS}}(\tau)$, the convex hull corrected version is significant (for the right-sided alternatives) at 0.05 level at $\tau$ values between 0.2 and 4.0, while the uncorrected version is significant (at 0.05 level) for $\tau$ values between 0.2 and 7. Hence, there is significant evidence for segregation of other trees from bald cypresses. We also perform Monte Carlo randomization tests for this data set (not presented), and see that the Monte Carlo randomized tests are more conservative in this example (see Ceyhan 2010a for more detail).

We also analyze the same data in a 2 × 2 NNCT with Dixon's overall test of segregation (Dixon 2002a). See Table 1 for the corresponding NNCT and the percentages (observe that the row sum for live trees is 157 instead of 156 due to ties in nearest neighbor (NN) distances). The cell percentages are relative to the row sums (i.e., number of other or bald cypress trees) and marginal percentages are relative to the overall sum. Notice that the table is not suggestive of segregation. Dixon's overall test statistic is $C_D = 0.9735$ ($p = 0.6146$) and Ceyhan's test is $C_N = 0.1825$ ($p = 0.6692$), both of which are suggestive of no significant deviation from CSR independence. So, NNCT-analysis and our relative density approach seem to yield different results about the spatial interaction of other trees with bald cypresses. However, NNCT and

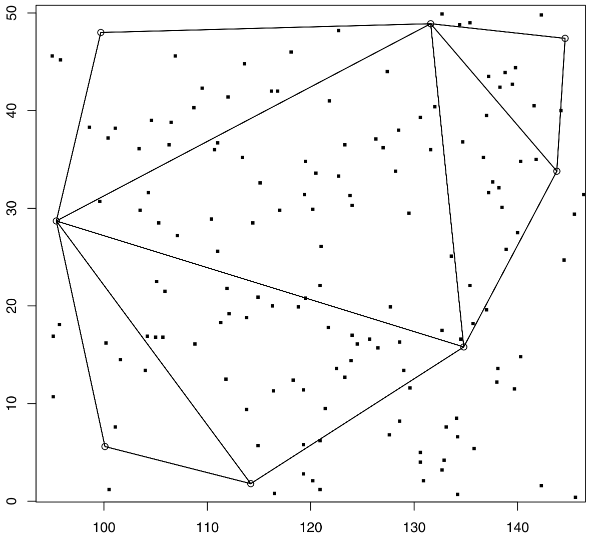


**Fig. 14** The scatter plot of the locations of bald cypresses (*circles* ○) and other trees (*black squares* ▪) in the swamp tree data. The Delaunay triangulation is based on the locations of the bald cypresses

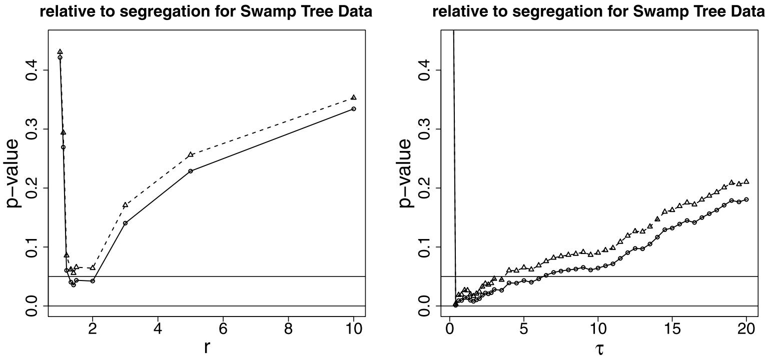


**Fig. 15** The $p$-values based on PE-PCDs (*left*) and CS-PCDs (*right*) with convex hull corrected test statistics (*circles* connected with *solid lines*) and uncorrected test statistics (*triangles* connected with *dashed lines*). The *horizontal lines* are at 0 and 0.05 values. Notice that the horizontal axes are differently scaled

**Table 1** The NNCT for swamp tree data (left) and the corresponding percentages (right). O.T. stands for "other trees" and B.C. for "bald cypresses"

| | | NN | | Sum |
|---|---|---|---|---|
| | | O.T. | B.C. | |
| Base | O.T. | 151 | 6 | 157 |
| | B.C. | 8 | 0 | 8 |
| | Sum | 159 | 6 | 736 |

| | NN | | |
|---|---|---|---|
| | O.T. | B.C. | |
| O.T. | 96 % | 4 % | 95 % |
| B.C. | 100 % | 0 % | 5 % |
| | 96 % | 4 % | 100 % |

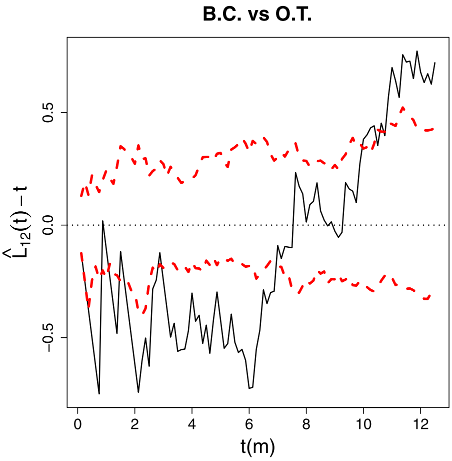


**Fig. 16** Ripley's bivariate $L$-function $\widehat{L}_{12}(t) - t$ for the part of the swamp tree data we considered. *Wide dashed lines* are the upper and lower (pointwise) 95 % confidence bounds for the functions based on Monte Carlo simulations under the CSR independence pattern. B.C. = bald cypresses and O.T. = other trees

our relative density approach answer different questions. More specifically, NNCT-tests are used to detect the spatial interaction between the two tree groups, while the relative density approach only tests the spatial interaction of other trees with bald cypresses, but not vice versa. Furthermore, this situation is an example where relative density is more appropriate, since there is much more other trees compared to bald cypresses. On the other hand, the NNCT tests are more appropriate in the cases where the relative abundance of the two species are similar and cell sizes are larger than 5 (Dixon 2002a and Ceyhan 2010c).

To find out the level of interaction between the tree species at different scales (i.e., distances between the trees), we also present the second-order analysis of the swamp tree data (Diggle 2003) using the functions (or some modified version of them) provided in spatstat package in R (Baddeley and Turner 2005). We use Ripley's bivariate $L$-functions which are modified versions of his $K$-functions. For a rectangular region, to remove the bias in estimating $K(t)$, it is recommended to use distance values up to 1/4 of the smaller side length of the rectangle. So we take the values $t \in [0, 12.5]$ in our analysis, since the rectangular region is 50 m × 55 m.

Ripley's bivariate $L$-function, $L_{ij}(t)$, is symmetric in $i$ and $j$ in theory, that is, $L_{ij}(t) = L_{ji}(t)$ for all $i, j$. In practice although edge corrections will render it slightly asymmetric, i.e., $\widehat{L}_{ij}(t) \neq \widehat{L}_{ji}(t)$ for $i \neq j$. The corresponding estimates are pretty close in our example, so we only present one of them. Ripley's bivariate $L$-function for the bald cypresses and other trees are plotted in Fig. 16, which suggests that bald cypresses and other trees are significantly segregated for distances about 0.5 to 7 meters, and do not significantly deviate from CSR for distances from 7 to 10 meters. This significant finding for segregation is in agreement with the results of the PCD test results.

## 10 Discussion

In this article, we compare the relative density of two proximity catch digraphs (PCDs), namely, proportional edge (PE) and central similarity (CS) PCDs, each of which is defined with an expansion parameter, for testing bivariate spatial patterns of segregation and association against complete spatial randomness (CSR). To the author's knowledge, the PCD-based methods are the only graph-theoretic tools for testing spatial point patterns in literature (Ceyhan et al. 2006, 2007, and Ceyhan 2011).

We extend the expansion parameter, $\tau$, of the CS-PCD to values higher than one (previously it was defined only up to one in Ceyhan et al. 2007). This extension proved to be useful, since the relative density of CS-PCDs has better performance in terms of empirical size and power for expansion parameter values in this new range (i.e., $\tau > 1$) compared to the ones in the previous range (i.e., $\tau \in (0, 1]$). For finite samples, we assess the empirical size and power of the relative density of the PCDs by extensive Monte Carlo simulations. For the PE-PCDs, the optimal expansion parameters (in terms of appropriate empirical size and high power) are about 1.5 under mild segregation and values in $(2, 3)$ under moderate to severe segregation; and about 2 under association. On the other hand, for CS-PCDs, the optimal parameters are about 7 under segregation, and about 1 under association. Furthermore, we have shown that relative density of CS-PCDs has better empirical size performance; and also, it has higher power against the segregation alternatives. On the other hand, relative density of PE-PCDs has higher power against the association alternatives.

For the two samples, $\mathcal{X}_n$ and $\mathcal{Y}_m$, with sizes $n$ and $m$ from classes 1 and 2, respectively, with class 1 points being used as the vertices of the PCDs and class 2 points being used in the construction of Delaunay triangulation, the null hypothesis is assumed to be CSR of class 1 points, i.e., the uniformness of class 1 points in the convex hull of class 2 points, $C_H(\mathcal{Y}_m)$. Although we have two classes here, the null pattern is not the CSR independence, since for finite $m$, we condition on relative areas of the Delaunay triangles based on class 2 points (assumed to have no more than three co-circular points). The relative density of the two PCD families lend themselves for spatial pattern testing conveniently, because of the geometry invariance property for uniform data on triangles (Ceyhan 2010b).

We also compare the asymptotic relative efficiency of the relative densities of the two PCD families. Based on Pitman asymptotic efficiency, we have shown that in general the relative density of PE-PCDs is asymptotically more efficient for segregation, while relative density of CS-PCDs is more efficient for association. However, this result is for $n \to \infty$ under very mild deviations from CSR. Besides, for the above optimal expansion parameter values (optimal with respect to empirical size and power), the asymptotic efficiency and empirical power analysis yield the same ordering in terms of performance.

For the relative density approach to be appropriate, the size of class 1 points (i.e., $n$) should be much larger compared to size of class 2 points (i.e., $m$). This implies that $n$ tends to infinity while $m$ is assumed to be fixed. That is, the imbalance in the relative abundance of the two classes should be large for our method to be appropriate. Such an imbalance usually confounds the results of other spatial interaction tests. Furthermore, by construction, our method uses only the class 1 points in $C_H(\mathcal{Y}_m)$ which might cause substantial data (hence information) loss. To mitigate this, we propose a correction for the proportion of class 1 points lying outside $C_H(\mathcal{Y}_m)$, because the pattern inside $C_H(\mathcal{Y}_m)$ might not be the same as the pattern outside $C_H(\mathcal{Y}_m)$. We suggest a two-stage analysis with our relative density approach: (i) analysis for $C_H(\mathcal{Y}_m)$, which provides inference restricted to class 1 points in $C_H(\mathcal{Y}_m)$, (ii) overall analysis with convex hull correction (i.e., for all class 1 points inside or outside $C_H(\mathcal{Y}_m)$). We recommend the use of normal approximation if $n \approx 10 \times m$ or more, although Monte Carlo simulations suggest smaller $n$ might also work.

**Acknowledgements** I would like to thank an anonymous associate editor and referees whose constructive comments and suggestions greatly improved the presentation and flow of the paper. Most of the Monte Carlo simulations presented in this article were executed at Koç University High Performance Computing Laboratory. This research was supported by the European Commission under the Marie Curie International Outgoing Fellowship Programme via Project # 329370 titled PRinHDD.

## References

Baddeley AJ, Turner R (2005) Spatstat: an R package for analyzing spatial point patterns. J Stat Softw 12(6):1–42

Beer E, Fill JA, Janson S, Scheinerman ER (2010) On vertex, edge, and vertex-edge random graphs. arXiv:0812.1410v2 [math.CO]

Ceyhan E (2010a) A comparison of two proximity catch digraph families in testing spatial clustering. Technical Report # KU-EC-10-3, Koç University, Istanbul, Turkey. arXiv:1010.4436v1 [math.CO]

Ceyhan E (2010b) Extension of one-dimensional proximity regions to higher dimensions. Comput Geom Theor Appl 43(9):721–748

Ceyhan E (2010c) New tests of spatial segregation based on nearest neighbor contingency tables. Scand J Stat 37:147–165

Ceyhan E (2011) Spatial clustering tests based on domination number of a new random digraph family. Commun Stat, Theory Methods 40(8):1363–1395

Ceyhan E, Priebe CE, Wierman JC (2006) Relative density of the random $r$-factor proximity catch digraphs for testing spatial patterns of segregation and association. Comput Stat Data Anal 50(8):1925–1964

Ceyhan E, Priebe CE, Marchette DJ (2007) A new family of random graphs for testing spatial segregation. Can J Stat 35(1):27–50

Coleman TF, Moré JJ (1983) Estimation of sparse Jacobian matrices and graph coloring problems. SIAM J Numer Anal 20(1):187–209

Coomes DA, Rees M, Turnbull L (1999) Identifying aggregation and association in fully mapped spatial data. Ecology 80(2):554–565

Cressie NAC (1993) Statistics for spatial data. Wiley, New York

DeVinney J, Priebe CE, Marchette DJ, Socolinsky D (2002) Random walks and catch digraphs in classification. In: Proceedings of the 34th symposium on the interface: computing science and statistics, vol 34. http://www.galaxy.gmu.edu/interface/I02/I2002Proceedings/DeVinneyJason/DeVinneyJason.paper.pdf

Diggle PJ (2003) Statistical analysis of spatial point patterns. Hodder Arnold Publishers, London

Dixon PM (1994) Testing spatial segregation using a nearest-neighbor contingency table. Ecology 75(7):1940–1948

Dixon PM (2002a) Nearest-neighbor contingency table analysis of spatial segregation for several species. Ecoscience 9(2):142–151

Dixon PM (2002b) Nearest neighbor methods. In: El-Shaarawi AH, Piegorsch WW (eds) Encyclopedia of environmetrics, vol 3. Wiley, New York, pp 1370–1383

Erdős P, Rényi A (1959) On random graphs I. Publ Math (Debr) 6:290–297

Fall A, Fortin MJ, Manseau M, O'Brien D (2007) Ecosystems. Int J Geogr Inf Sci 10(3):448–461

Faragó A (2008) A general tractable density concept for graphs. Math Comput Sci 1(4):689–699

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

Good BJ, Whipple SA (1982) Tree spatial patterns: South Carolina bottomland and swamp forests. Bull Torrey Bot Club 109:529–536

Illian J, Burslem D (2007) Contributions of spatial point process modelling to biodiversity theory. Coexistence 148(148):9–29

Janson S, Łuczak T, Ruciński A (2000) Random graphs. Wiley-Interscience series in discrete mathematics and optimization. Wiley, New York

Jaromczyk JW, Toussaint GT (1992) Relative neighborhood graphs and their relatives. Proc IEEE 80:1502–1517

Jung I, Kulldorff M (2007) Theoretical properties of tests for spatial clustering of count data. Can J Stat 35(3):433–446

Keitt T (2007) Introduction to spatial modeling with networks. Presented at the workshop on networks in ecology and beyond organized by the PRIMES (program in interdisciplinary math, ecology and statistics) at Colorado State University, Fort Collins, Colorado

Kendall M, Stuart A (1979) The advanced theory of statistics, vol 2, 4th edn. Griffin, London

Lehmann EL (1988) Nonparametrics: statistical methods based on ranks. Prentice-Hall, Upper Saddle River

Leibovich E (2009) Approximating graph density problems. PhD thesis, The Open University of Israel, Department of Mathematics and Computer Science

Marchette DJ, Priebe CE (2003) Characterizing the scale dimension of a high dimensional classification problem. Pattern Recognit 36(1):45–60

Minor ES, Urban DL (2007) Graph theory as a proxy for spatially explicit population models in conservation planning. Ecol Appl 17(6):1771–1782

Okabe A, Boots B, Sugihara K, Chiu SN (2000) Spatial tessellations: concepts and applications of Voronoi diagrams. Wiley, New York

Penrose M (2003) Random geometric graphs. Number 5 in Oxford studies in probability. Oxford University Press, London

Priebe CE, DeVinney JG, Marchette DJ (2001) On the distribution of the domination number of random class cover catch digraphs. Stat Probab Lett 55:239–246

Priebe CE, Marchette DJ, DeVinney J, Socolinsky D (2003) Classification using class cover catch digraphs. J Classif 20(1):3–23

Roberts SA, Hall GB, Calamai PH (2000) Analysing forest fragmentation using spatial autocorrelation, graphs and GIS. Int J Geogr Inf Sci 14(2):185–204

Shenggui Z, Hao S, Xueliang L (2002) $w$-density and $w$-balanced property of weighted graphs. Appl Math J Chin Univ 17(3):355–364

Toussaint GT (1980) The relative neighborhood graph of a finite planar set. Pattern Recognit 12(4):261–268

van Eeden C (1963) The relation between Pitman's asymptotic relative efficiency of two tests and the correlation coefficient between their test statistics. Ann Math Stat 34(4):1442–1451