

空間点パターンのグラフィカルな分類方法に関する研究

高井 勉

博士（統計科学）

総合研究大学院大学

複合科学研究科

統計科学専攻

平成29（2017）年度

空間点パターンのグラフィカルな分類方法に関する研究

高井 勉

Tsutomu Takai

博士（統計科学）

総合研究大学院大学複合科学研究科統計科学専攻

（情報・システム研究機構 統計数理研究所）

平成 29（2017）年度

【目次】

I	序説.....	1
1	研究の目的と論文の構成.....	2
2	空間点パターン.....	6
2.0	はじめに.....	6
2.1	空間点パターンの分類.....	7
2.2	様々な空間点過程.....	8
2.2.1	一様ポアソン点過程.....	9
2.2.2	非一様ポアソン点過程.....	9
2.2.3	SSI 点過程.....	11
2.2.4	マテルンクラスター点過程.....	11
2.3	現実の空間点パターン.....	13
II	空間点パターンのグラフィカルな分類に関連する先行研究.....	15
3	近接グラフ.....	16
3.0	はじめに.....	16
3.1	第1近隣木.....	16
3.2	第 k 近隣木.....	20
3.3	相互最近隣対.....	26
3.4	MST.....	31
3.4.1	MSTの近似としての第1近隣木.....	31
3.4.2	Robertsの近似.....	34
3.4.3	Watanabe (2008)の近似.....	37
3.4.4	MSTの辺長の境界効果.....	37
4	空間点パターンのグラフィカルな分類方法.....	40
4.0	はじめに.....	40
4.1	近接グラフに基づく方法.....	41
4.1.1	G関数を用いた分類.....	41
4.1.2	F関数を用いた分類.....	45
4.1.3	J関数を用いた分類.....	49
4.2	2次特性に基づく方法.....	51

4.2.1	K 関数を用いた分類	51
4.2.1.1	境界効果と境界補正	53
4.2.1.2	K 関数を用いた分類結果と表示の改良	58
4.2.2	L 関数を用いた分類	61
4.2.2.1	L 関数を用いた分類結果	61
4.2.2.2	L*関数を用いた分類結果	63
4.2.3	pair correlation 関数を用いた分類	65
4.3	グラフィカルな分類方法の問題点	69
4.3.1	G 関数の問題点	69
4.3.2	K 関数の問題点	69
5	凝集型階層的クラスター分析	70
5.0	はじめに	70
5.1	凝集型階層的クラスター分析	70
5.2	類似度と非類似度（距離）	71
5.3	AHC 手法の分類アルゴリズム	72
5.3.1	AHC 手法の共通アルゴリズム	73
5.3.2	Lance and Williams の組合せ的な手法	74
5.4	単調性と強メトリック不等式	76
5.4.1	距離が持つ性質	76
5.4.2	単調性の判別	78
5.4.3	補足 1	79
5.5	分類アルゴリズムの具体例	79
5.5.1	最短距離法	79
5.5.2	最長距離法	81
5.5.3	（重み無し）群平均法	83
5.5.4	セントロイド法	83
5.5.5	補足 2	86
5.6	AHC 手法と空間点パターンとの関連性	86
III	本研究で提案する空間点パターンのグラフィカルな分類方法	87
6	AG-curve による空間点パターンの分類	88
6.0	はじめに	88
6.1	凝集型階層的クラスター分析の応用	88
6.1.1	AG-curve の定義	88
6.1.2	AG-curve の作図手順	89

6.1.3	AG-curve の簡単な作図例.....	89
6.2	空間点パターンの分類への応用.....	90
6.2.1	AG-curve を用いた空間点パターンの分類方法.....	91
6.2.2	AG-curve を用いた空間点パターンの分類結果.....	92
6.3	AGsi-curve と既存のグラフィカルな手法との比較.....	96
6.3.1	G 関数との比較.....	96
6.3.2	K 関数との比較.....	100
6.3.3	境界効果に対するロバスト性 (K 関数および G 関数との比較)	102
6.4	AGsi-curve と MST の関係.....	104
6.4.1	AGsi-curve と MST の辺長の経験分布関数 $F_n(x)$ との関係.....	104
6.4.2	AGsi-curve と MST の関係を確認するための簡単な数値例.....	105
6.5	更に踏み込んだ比較.....	107
6.5.1	集中型 (凝集型) 点配置の分類能力 (親子点過程を用いた比較)	107
6.5.2	規則型点配置の分類能力 (SSI 点過程を用いた比較)	111
6.6	【補足】 MST を空間点パターンの分類へ応用する.....	115
6.6.1	μ 関数 $\mu(r)$ の提案.....	115
6.6.2	μ 関数 $\hat{\mu}(r)$ を用いた空間点パターンの分類.....	116
7	AGsi-curve の近似分布の導出.....	119
7.0	はじめに.....	119
7.1	近似分布 U と近似分布 L.....	119
7.2	命題の証明.....	121
7.2.1	命題 7.1 の証明.....	121
7.2.2	命題 7.2 の証明.....	123
7.2.3	命題 7.3 の証明.....	126
8	AGsi-curve の期待値の強度依存性.....	129
8.0	はじめに.....	129
8.1	AGsi-curve の期待値の強度依存性 (2 次元)	129
8.2	MST の辺長の期待値と中央値 (2 次元)	134
8.3	MST の辺長の境界効果 (2 次元)	138
8.4	AGsi-curve の期待値に与える境界効果 (2 次元)	143
8.5	【参考】 3 次元の AGsi-curve.....	144
9	第Ⅲ部のまとめと今後の課題.....	146
9.1	第Ⅲ部のまとめ.....	146
9.1.1	目的①と②に関して.....	146
9.1.2	目的③に関して.....	146

9.1.3	目的④に関して.....	147
9.2	今後の課題.....	147
9.2.1	AGsi-curve および（その基になった）MST の辺長の強度依存性.....	147
9.2.2	AGsi-curve および（その基になった）MST の辺長の境界効果.....	147
9.2.3	AGsi-curve および（その基になった）MST の辺長の近似分布.....	148
9.2.4	AGsi-curve の多次元への対応.....	148
【付録】	149
付録 1	Watanabe (2008) の近似（多次元）.....	150
1.0	はじめに.....	150
1.1	Watanabe の近似.....	150
付録 2	多次元空間に於ける部分球と結合球の体積.....	154
2.1	準備.....	154
2.1.1	領域の定義.....	154
2.1.2	定積分の値.....	155
2.2	部分球の体積.....	155
2.3	結合球の体積.....	157
付録 3	MST の辺長の期待値と中央値（2次元）.....	160
【謝辞】	165
【参考文献】	167

第 I 部

序説

第1章 研究の目的と論文の構成

本論文では、空間点パターンをグラフィカルに分類する方法として AG-curve (AGglomerative-curve) を提案する。これは凝集型階層的クラスター分析 (Agglomerative Hierarchical Clustering Algorithm) を応用した方法である (以降は AHC 手法と呼ぶ)。これまで AHC 手法として多くの方法が提案されており、AG-curve もそれらに対応している。例えば、最短距離法 (single-linkage) を応用した AGsi-curve、最長距離法 (complete-linkage) を応用した AGco-curve、重み無し群平均法 (average-linkage) を応用した AGav-curve 等である。

本研究の目的は、

- AG-curve を用いた空間点パターンの分類方法を提案する。
- 空間点パターンの分類に適した AHC 手法を探索する。
- 既存のグラフィカルな分類方法に対する AG-curve の優位性を検証する。
- AG-curve の数理的性質を明らかにする。

である。

自然科学分野での空間点パターンは2次元の事例が多い。しかし多変量 (多次元) を取り扱う分野での利用を考え、先行研究で提案された手法から本論文で提案する手法まで、各種統計量の導出に当たっては多次元を基本とした。

本論文の構成は以下のようなになる。

第I部 序説

(第1章 研究の目的と論文の構成)

第2章 空間点パターン

第2章では、空間点パターンを研究対象とする分野、および AG-curve による分類対象となる様々な空間点パターンを紹介する。本研究において空間点パターンは、ランダム型・集中型（凝集型）・規則型の3種類に分類される。その中のランダム型を厳密に定義するために用いる完全空間ランダム性（CSR, Complete Spatial Randomness）は、分類の基準となるもので非常に重要である。これは、時系列解析や信号処理における、ホワイト・ノイズ性に相当するものである。また、第2部および第3部の該当章に於いて、分類の対象となる3種類の空間点パターン（2次元）を提示する。これらは全て現実の空間点パターンである。

第II部 空間点パターンのグラフィカルな分類に関連する先行研究

第3章 近接グラフ（Proximity graph）

空間点パターンの分類方法の中には近接グラフと関係の深いものが多い。第3章では先行研究で確立されたグラフィカルな方法および、新たに提案するグラフィカルな方法の基礎をなす近接グラフの辺長の分布を導出する。近接グラフとは、点の近さを表す「近接性」によって構築されるグラフである。本研究に関係する近接グラフは、第1近隣木、第 k 近隣木、相互最近隣対（Reciprocal nearest neighbor pair）、MST（Minimum Spanning Tree）である。領域内に強度 λ で一様ランダムに配置された点過程に対する辺長の分布を対象とする。多次元で導出したのち、具体例として1次元・2次元・3次元の結果（確率密度関数、期待値、標準偏差）を示す。

第4章 空間点パターンのグラフィカルな分類方法

第4章では空間点パターンの探索的データ解析に用いられるグラフィカルな分類方法を取り上げる。これらは近接グラフ（Proximity graph）に基づくものと、2次特性（Second-order characteristics）に基づくものに分けられる。ここでは前者としてG関数・F関数・J関数を、後者としてK関数・L関数・pair correlation 関数を取り上げる。空間点過程の2次特性とは、2点の相互作用、あるいは空間上の関

係が空間内でどのように変化するかを記述する量である。

各分類方法の説明の後に、その分類方法を用いた具体例を示す。分類対象の空間点パターンが、ランダム型・規則型・集中（凝集）型のどのタイプに属するかをグラフィカルに判定する内容である。ここでは分類対象として第1章で示した3種類の現実の点パターンを用いる。

第5章 凝集型階層的クラスター分析

本研究で提案する AG-curve は AHC 手法の併合距離とクラスター数の間の関係を示している。そこで AHC 手法の概略を Lance and Williams (1967) の組合せ的な手法を用いて示す。次に多くの AHC 手法が持つ単調性に焦点をあてる。単調性とはクラスター化の進展とともに併合距離が単調に増加することを意味する。この単調性は AG-curve からの要請である。単調性の説明のため、様々な距離がもつ性質を示したのち、単調性の判別条件となる強メトリック不等式を示す。

最後に、単調性が保証されている AHC 手法として、具体的に最短距離法・最長距離法・（重み無し）群平均法を取り上げ、それぞれの併合距離の定義とクラスター化の特徴を示す。さらに比較のため、単調性が保証されていないセントロイド法も取り上げる。

第Ⅲ部 本研究で提案する空間点パターンのグラフィカルな分類方法

第6章 AG-curve による空間点パターンの分類

第6章は本論文の主要な部分である。提案する AG-curve の定義と作図手順を示したのち、空間点パターンの分類に適した AHC 手法を探索する。AG-curve には単調性が期待されているので、その基礎となる AHC 手法にも単調性が期待される。したがって探索する AHC 手法の候補として、単調性が保証されている、最短距離法・最長距離法・群平均法の3手法を選択する。結果として最短距離法が選択されることになる。最短距離法（single-linkage）を用いた AG-curve を AGsi-curve と呼ぶ。第6章の残りの部分と、第7章および第8章は、AGsi-curve に的を絞って議論を進める。

次に AGsi-curve の分類特性を、既存の手法である G 関数および K 関数と比較する。それらに対する

優位性を示すために計算機実験を活用する。最後に、AGsi-curve は MST の辺長の経験分布関数と等価であることを示す。補足として MST の辺長の経験分布関数を μ 関数とし、この μ 関数を用いた空間点パターンの分類事例を示す。

第 7 章 AGsi-curve の近似分布の導出

第 7 章の内容は、第 6 章で述べた「AGsi-curve と MST の辺長との関係」を発展させたものである。CSR 点過程に対する AGsi-curve の振舞いを MST の辺長の順序統計量を用いて説明する。MST の辺長が従う確率分布は解析的には求められないが、幾つかの近似分布が提案されている。その中から本研究では Roberts (1968) の近似と Watanabe (2008) の近似を取り上げる。Roberts (1968) の近似分布の順序統計量を用いた場合を AGsi-curve の近似分布 U とし、Watanabe (2008) の近似分布の順序統計量を用いた場合を AGsi-curve の近似分布 L とする。これらの近似分布は多次元で導出する。

- ・ AGsi-curve の近似分布 U : MST の辺長に関する Roberts (1968) の近似分布の順序統計量を用いる。
- ・ AGsi-curve の近似分布 L : MST の辺長に関する Watanabe (2008) の近似分布の順序統計量を用いる。

第 8 章 AGsi-curve の期待値の強度依存性

第 8 章では、第 7 章で導出した近似分布 U と近似分布 L の期待値、およびモンテカルロ法を用いて推定した厳密分布の期待値を比較することにより、それらの特性を明らかにする。ここでは 3 つの期待値の強度依存性に着目し、それらに観察された 2 つの特徴を検討する。次に AGsi-curve の近似分布の導出に用いた、MST の辺長の近似分布である Roberts (1968) と Watanabe (2008) についても、期待値の強度依存性に着目し、モンテカルロ法により推定した厳密分布の期待値と比較する。また、MST の辺長に関する境界効果 (edge effect) を、簡易的なモデル化と計算機実験で確認する。最後に、明らかにされた AGsi-curve の特性と MST の辺長の特性を比較し考察する。

第 9 章 まとめと今後の課題

提案する方法 (第 III 部) のまとめと今後の課題を示す。

第2章 空間点パターン

2.0 はじめに

空間データ (Spatial data) には様々な種類があり、Cressie (1993, pp.10-13) はこれらを以下の3クラスに分類している。

- 地球統計データ (Geostatistical data)
- 格子データ (Lattice data)
- 空間点パターン (Point Pattern)

地球統計データとは地球統計学 (Geostatistics) の対象となるものである。地球統計学は最初、鉱山学の手法として誕生した。クリギング (kriging) 補間法を用いて、ボーリング位置で観測されたデータから任意の位置における鉱石の品質を予測したことが始まりである (間瀬, 2010, p.1)。したがって固定された位置で観測された値が対象となる。現在では様々な分野に広がりを見せる。例えば、モニタリングポストで観測された大気中の放射性量、気象庁の観測所 (AMEDAS) から送られてくる降雨量など。地球統計学は地理統計学と呼ばれる場合もある。

格子データとは、規則的または不規則に配置された領域に関連付けられた観測値であり、近傍情報を持っている。規則的な格子データは、地球観測衛星・気象衛星などに搭載されているリモートセンシング機器から送られてきた情報である。地表は小さな矩形に規則的に分割されており、植生の分布・海面温度・穀物生産予測などを調べるために使われる。不規則な格子データは県別の平均寿命や疾病率などである。

本研究では新たに提案する分類方法の対象として、空間点パターンに焦点を当てる。空間点パターンとは、関心のある領域内に何らかのメカニズムによって生成された点の位置座標の集合 (点配置) である。言い換えれば点の位置自体が関心のある変数になっている空間データである。空間点パターンは自然科学・社会科学などの様々な分野で研究対象となっている。例をあげれば、生態学 (Perry *et al.*, 2006)、伝染病学 (Gatrell *et al.*, 1996)、地震学 (Ogata, 1999)、経済学 (Duranton & Overman, 2005)、地理学 (Okabe

& Yamada, 2001)、天文学 (Neyman & Scott, 1958)、時空間解析 (Diggle, 2013) などである。空間という名前がついているが、自然科学の分野では 2 次元を対象とすることが多い。

空間点パターンのデータセットは、点の位置座標だけで構成される場合と、位置に付随した何らかのデータ値 (属性値) を持つ場合がある。後者はマーク付き点パターン (marked point pattern) と呼ばれる。属性値は量的変数でも質的変数でも構わない。森林の樹木を考えた場合、量的変数の例は胸の高さで測った幹の直径になり、質的変数の例は木の種類となる。なお、本研究ではマーク付き点パターンは取り扱わない。

2.1 空間点パターンの分類

空間点パターンを大きく分類すると、構造を持たないものと、構造を持つものとに分けられる。構造を持たないものは、大域的には点間に何も力が働いていないように見えるが、局所的には引力と斥力がランダムに働いているように見える。これはランダム型点配置と呼ばれている。構造を持つものは、点間に引力 (誘引力) が働いているように見える集中型 (凝集型) 点配置と、点間に斥力 (反発力) が働いているように見える規則型点配置に分けられる。集中型 (凝集型) の良く知られた例は、地震 (余震) の震源地の座標である。また規則型としては公立小学校の位置などがあげられる。格子点上に点が配置されると、規則性が強く感じられるようになる。まとめると空間点パターンは大きく以下の 3 種類に分類される。それぞれの型を代表する現実の点パターンを 2.3 節の 図 2.4 に示す (Diggle, 2003)。

- ・ランダム型点配置
- ・規則型点配置
- ・集中型 (凝集型) 点配置

空間点パターンの分類を考える上で最も重要なものは、ランダム型点配置を厳密に定義した完全ランダム型点配置である。Diggle (1979) は完全ランダム型点配置を特徴付ける完全空間ランダム性 (CSR, Complete Spatial Randomness) を定義しているが、ここでは Diggle (1983, p.4) の解説を取り上げ以下に示す。

The hypothesis of complete spatial randomness (CSR) for a spatial point pattern asserts that

- (i) the number of events in any planar region A with area $|A|$ follows a Poisson distribution with mean $\lambda|A|$;
(The constant λ is the intensity, or mean number of events per unit area.)
- (ii) given n events x_i in a region A , the x_i are an independent random sample from the uniform distribution on A .

According to (i) , CSR therefore implies that the intensity of events does not vary over the plane.

According to (ii) , CSR also implies that there are no interactions amongst the events.

Diggle (1983, p.4) の空間点パターンに対する CSR の仮説は、大きく 2 つの内容で構成されている。

(i) に関する内容を箇条書きにすると以下のようになる。

- ・面積が $|A|$ である平面領域 A 内の点 (events) の数は、期待値 $\lambda|A|$ のポアソン分布に従う。
ここで定数 λ は強度、または単位面積当たりの点の平均個数である。
- ・上記より、強度は領域 A 内で変化しない。

(ii) に関する内容を箇条書きにすると以下のようになる。

- ・領域 A 内に n 個の点 x_i が与えられたとすると、 x_i は領域 A 内の一様分布から抽出した独立なランダム標本である。
- ・上記より、点間に相互作用は働かない (引力も斥力も存在しない)。

これらを要約すると、(i) は「点の数はポアソン分布する」であり、(ii) は「点の散らばり方は独立である」となる。この完全空間ランダム性は時系列解析や信号処理の分野におけるホワイトノイズ性に相当するものである。

2.2 様々な空間点過程

ある確率的メカニズム (Stochastic mechanisms) が空間点パターン (点配置) を発生させると考え、その確率的メカニズムの数学モデルが、点過程 (point process) と呼ばれている。空間点パターンを発生させるモデルとして様々な点過程が存在する。

2.2.1 一様ポアソン点過程 (homogeneous Poisson point process)

完全ランダム型点配置を発生させる点過程が一様ポワソン点過程である。

有界集合 B 内の点数を $N(B)$ 、有界集合 B の体積を $v(B)$ とする。集合 B 内に n 個の点が発生する確率 $P(N(B) = n)$ は以下のようになる (Illian *et al.*, 2008, p.67)。

$$P(N(B) = n) = \frac{\lambda^n \cdot (v(B))^n}{n!} \exp(-\lambda v(B)) \quad \text{for } n = 0, 1, 2, \dots$$

$N(B)$ の分布はパラメータ $\lambda v(B)$ のポアソン分布である。したがって、期待値も分散も $\lambda v(B)$ になる。

互いに素な (disjoint) 複数の有界集合 B_1, B_2, \dots, B_k に対しては、 $N(B_1), N(B_2), \dots, N(B_k)$ はパラメータが $\lambda v(B_1), \lambda v(B_2), \dots, \lambda v(B_k)$ の、独立なポアソン分布をするので以下のようになる (Illian *et al.*, 2008, p.67)。

$$\begin{aligned} P(N(B_1) = n_1, N(B_2) = n_2, \dots, N(B_k) = n_k) \\ = \frac{\lambda^{n_1+n_2+\dots+n_k} \cdot v(B_1)^{n_1} v(B_2)^{n_2} \dots v(B_k)^{n_k}}{n_1! \cdot n_2! \cdot \dots \cdot n_k!} \cdot \exp\left(-\sum_{i=1}^k \lambda v(B_i)\right) \end{aligned}$$

一様ポワソン点過程は空間ランダムネスのベンチマークとして使われる。強度 λ も確率変動するので、その期待値が一定となる。ただし本研究では、様々な分類方法の性能比較を容易にするため強度 λ を固定値として扱うことがある。この場合、「点は一様ランダムに分布する」と表記することがある。

一様ポアソン点過程の実現値の一例を図 2.1 に示す。

2.2.2 非一様ポアソン点過程 (inhomogeneous Poisson point process)

非一様ポアソン点過程は、以下の 2 つの性質で特徴づけられる (Illian *et al.*, 2008, p.118)。これは一様ポアソン点過程の強度 λ に関する仮定を変えた (緩和した) ものである。強度 λ は点の位置 x の関数であると考え。点 x を中心とする微小領域 dx に点が 1 つ存在する確率は $\lambda(x)dx$ である。

①任意の有界集合 B 内の点の個数は平均 μ のポアソン分布に従う。

$$\mu = \int_B \lambda(x) dx$$

②各点は独立に分布している (点間に相互作用は働かない)。

非一様ポアソン点過程の実現値の一例を図 2.2 に示す。強度 λ は単位正方形の領域内で

$$\lambda(x, y) = 1000(x^2 + y^2)$$

のように変化する。

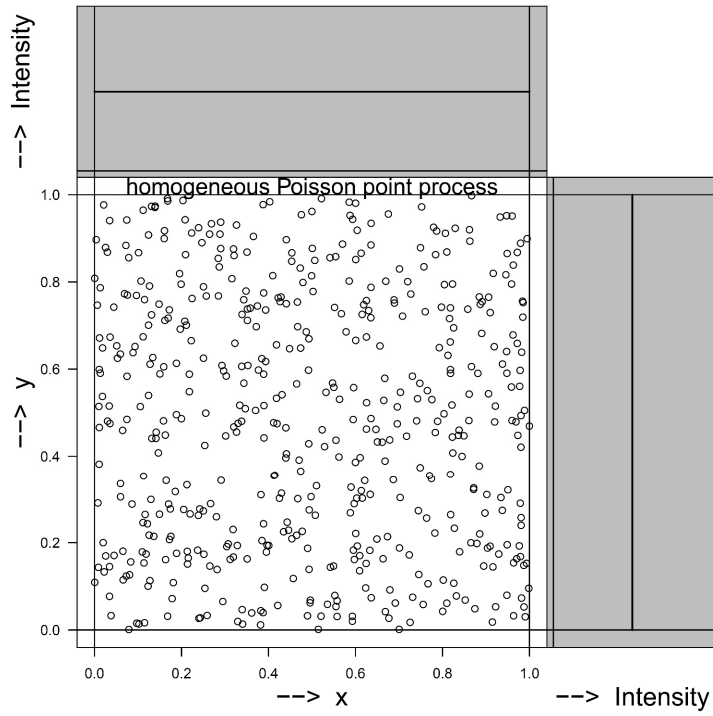


図 2.1 一様ポアソン点過程

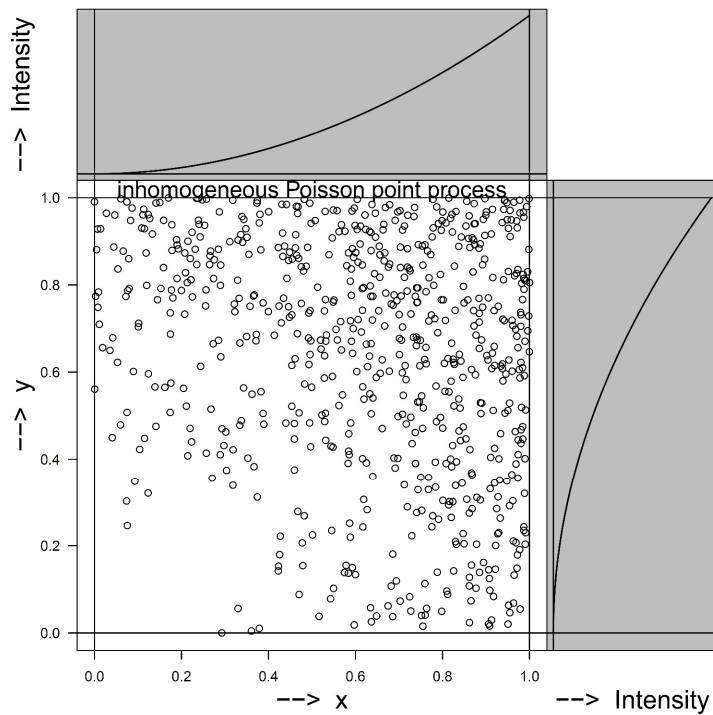


図 2.2 非一様ポアソン点過程

2.2.3 SSI 点過程 (Simple Sequential inhibition point process)

規則型点過程の 1 つとして、第 II 部と第 III 部において、計算機実験に用いる SSI 点過程 (Diggle *et al.*, 1976) をとりあげる。有界集合中に、すでに置かれた点から距離 r 以上離れた部分に一様に次の点を選んでいくことを繰り返して得られる最終点パターンである (間瀬&武田, 2001, p. 43)。この距離 r は禁止半径 (radius of inhibition) とも呼ばれる。禁止半径を大きくし過ぎると、指定した数の点を発生させられない可能性が高くなる。単位正方形内に発生させた SSI 点過程の実現値の一例を図 2.3 の第 1 列 (左図) に示す。強度を 100 に固定し、禁止半径を 3 水準 (0.075, 0.050, 0.025) 変化させた。禁止半径が小さくなるにしたがいランダム型に近づいていく様子が分かる。

2.2.4 マテルンクラスター点過程 (Matérn Cluster point process)

集中型 (凝集型) 点過程の 1 つとして、第 II 部と第 III 部において、計算機実験に用いるマテルンクラスター点過程 (Matérn, 1986) をとりあげる。これは親子点過程とも呼ばれる。親の点を強度 κ のポアソン点過程を用いて発生させる。次に各親の位置を中心とする半径 r の円内に子の点を強度 μ のポアソン点過程を用いて発生させ、親の点と置き換える (子の誕生と同時に親は消滅する)。したがって最終的な強度は $\kappa \times \mu$ になる。単位正方形内に発生させたマテルンクラスター点過程の実現値の一例を図 2.3 の第 2 列 (右図) に示す。親の点数と子の点数は確率変動するので、複数回の試行を行い、親の点数 (クラスター数) を 3 水準 (3, 4, 5) に、領域内の総点数を 100 に揃えた。

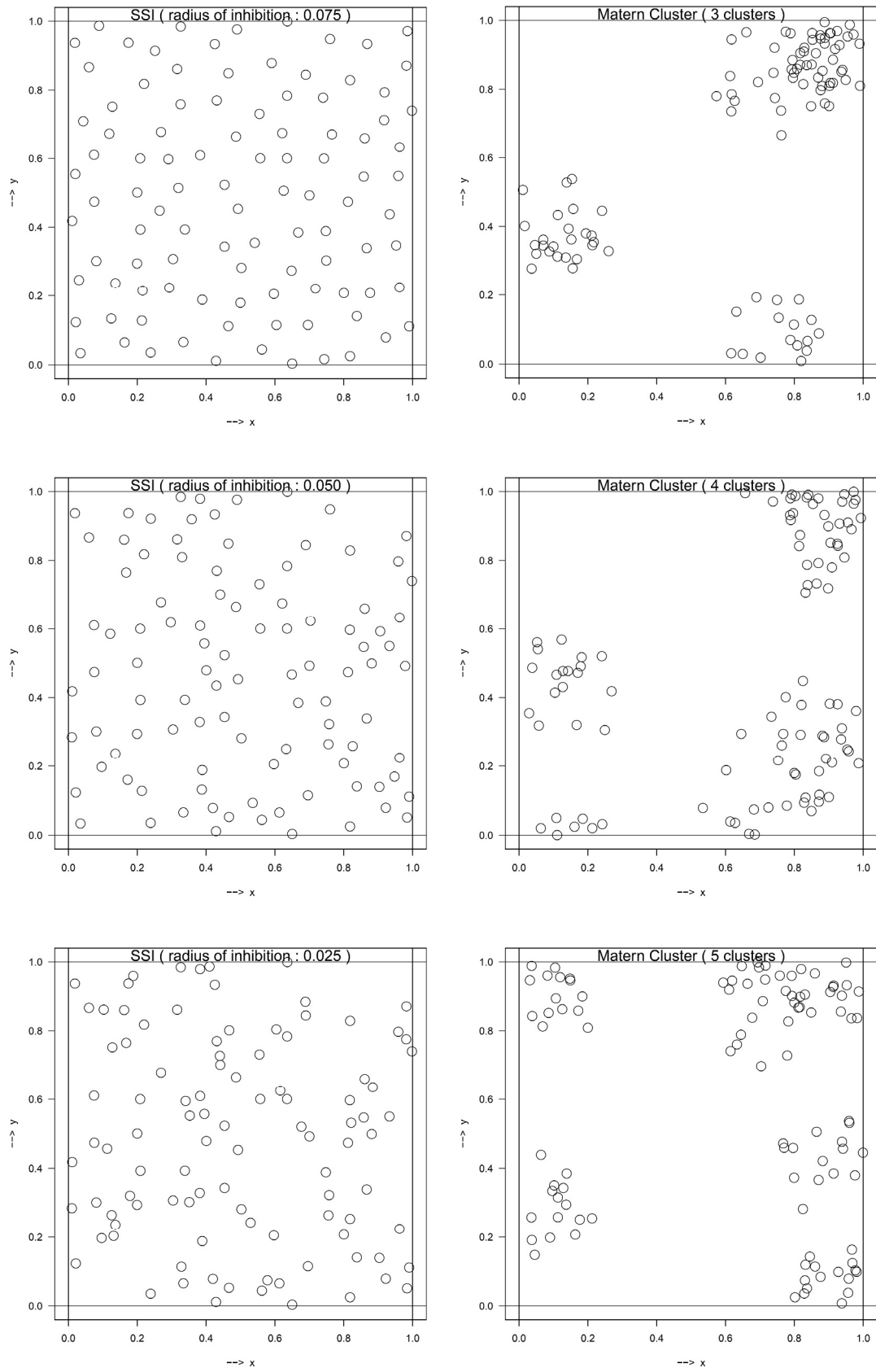


図 2.3 様々な点過程 (左図 : SSI 点過程、右図 : マテルンクラスター点過程)

2.3 現実の空間点パターン

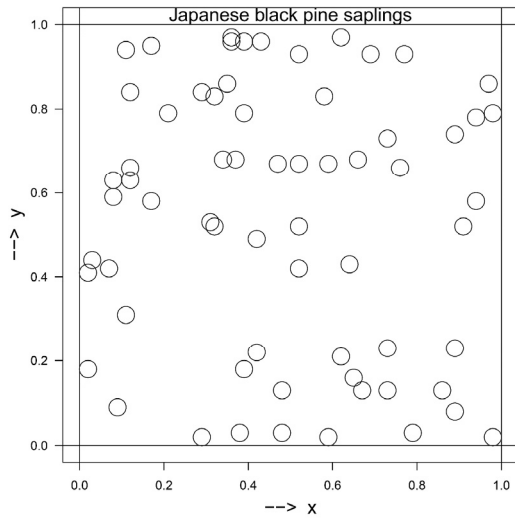
第Ⅱ部と第Ⅲ部において、分類手法の比較に一貫して用いる現実の空間点パターンを図 2.4 に示す。

上から順に

- Japanese black pine sapling ($n = 65$)
- Biological cells ($n = 42$)
- Redwood seedlings ($n = 62$)

となっている。

これらは順に、ランダム型・規則型・集中型（凝集型）であることが、先行研究（Diggle, 2003）により明らかにされている。なお、点の位置座標は全て単位正方形内に収まるように縮尺が変更されている。また n は領域内の点数である。



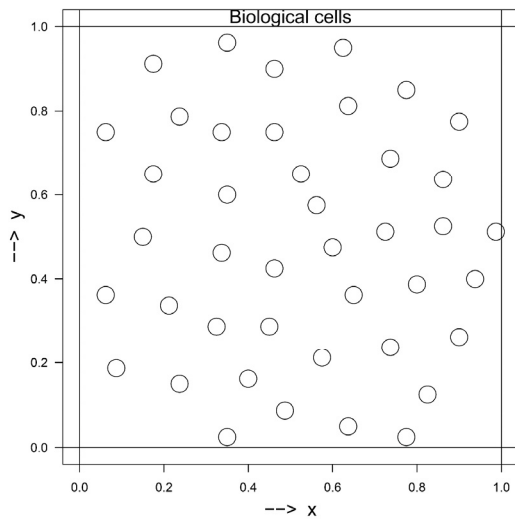
【ランダム型】

Japanese black pine saplings

出典 : Diggle (2003)

千葉県銚子市沿岸部の自然林にある
黒松の若木 (65本) の位置。

Diggle (2003) は、Numata, M. (1961).
Forest vegetation in the vicinity of Choshi.
Coastal flora and vegetation at Choshi, Chiba
Prefecture IV. *Bulletin of Choshi Marine
Laboratory, Chiba University*, vol.3, pp.28-
48 [in Japanese]. のデータを用いている。

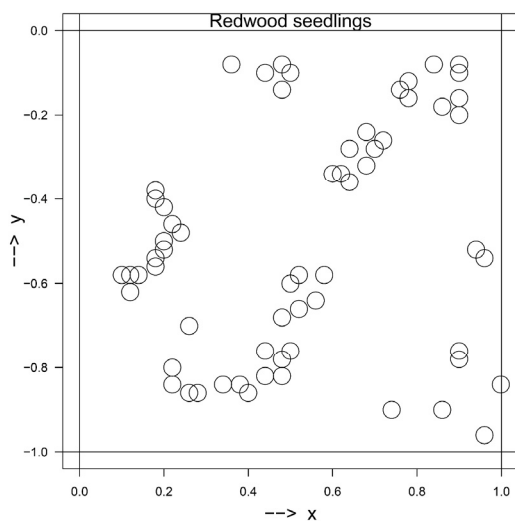


【規則型】

Biological cells

出典 : Ripley (1977)

光学顕微鏡下で観察された
組織切片内にある生体細胞 (42個)
の中心の位置。



【集中型 (凝集型)】

Redwood seedlings

出典 : Strauss (1975)

カリフォルニア州のレッドウッド国立・
州立公園にある樹木 (62本) の位置。

Strauss (1975) のデータセットから
Ripley (1977) が、その 1/4 の領域を
切り出したもの。

図 2.4 分類に用いる現実の空間点パターン (単位正方形内に収まるように縮尺が変更されている)

第Ⅱ部

空間点パターンのグラフィカルな分類に関連する先行研究

第3章 近接グラフ (Proximity graph)

3.0 はじめに

空間点パターンの分類方法の中には近接グラフと関係の深いものが多い。第3章では先行研究で確立されたグラフィカルな方法および、新たに提案するグラフィカルな方法の基礎をなす近接グラフの辺長の分布を導出する。近接グラフとは、点の近さを表す「近接性」によって構築されるグラフである。本研究に関係する近接グラフは、第1近隣木、第 k 近隣木、相互最近隣対 (Reciprocal nearest neighbor pair)、MST (Minimum Spanning Tree) である。領域内に強度 (λ) で一様ランダムに配置された点過程に対する辺長の分布を対象とする。多次元で導出したのち、具体例として1次元・2次元・3次元の確率密度関数・期待値・標準偏差を示す。多次元空間の複雑な領域の体積は付録2で導出したものを用いる。

3.1 第1近隣木

最近傍グラフ (Nearest Neighbor Graph) ・ NNT (Nearest Neighbor Tree) 等とも呼ばれるが、本研究では第1近隣木と呼ぶ。領域内の各点と、その点から最も近い点 (第1近隣点) を結ぶ辺で構成される。したがって辺数は点数と等しい。2次元の辺長の分布は Clark & Evans (1954) と Pollard (1971) によって、多次元の場合は Clark & Evans (1979) によって導出されている。

図3.1は単位正方形内に一様ランダムに10点を配置した第1近隣木の例 ($\lambda = 10$) である。両側に矢印がある場合は2本の線が重なっている。

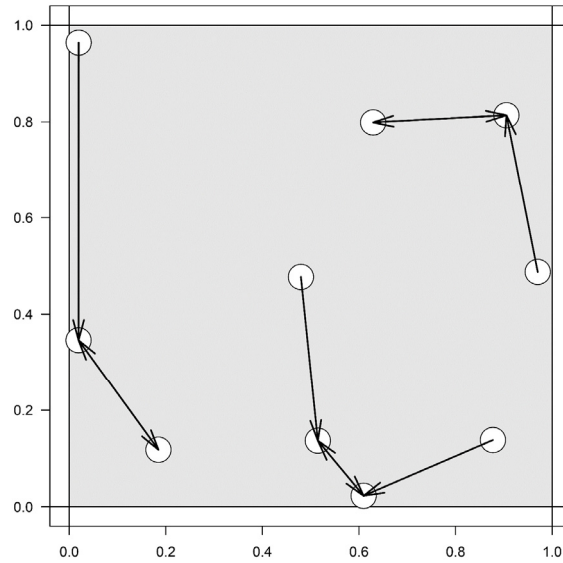


図 3.1 第 1 近隣木 (2 次元の場合)

点 a を中心とした半径 r の多次元球 A を考える。 A の表面上に点 b があり、球内に a 以外の他の点がない時、辺 ab は第 1 近隣木の辺となる。したがって、 u 次元における第 1 近隣木の辺長の分布は、半径 r の u 次元球の内部に a 以外に 1 つも点がなく、且つ u 次元球の表面上に少なくとも 1 つ点があるという確率から導くことができる。図 3.2 は第 1 近隣点 (2 次元の場合) の例である。任意に選んだ点 a (○) と、その第 1 近隣点 b (●) を図中に示す。

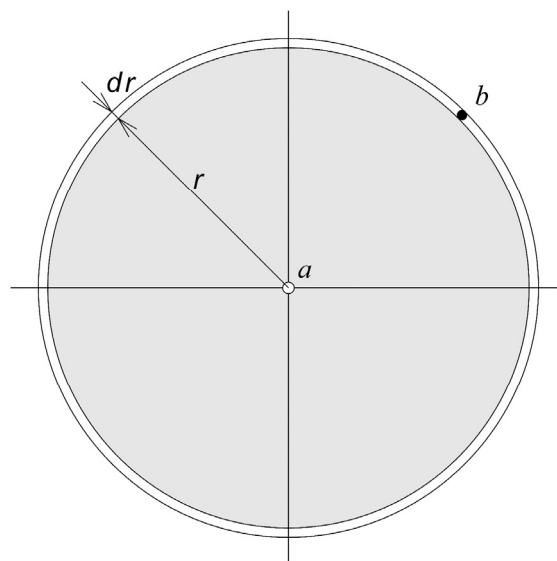


図 3.2 第 1 近隣点 (2 次元の場合)

体積 V の領域内に点が x 個現れる確率 $P(X = x)$ は、一様ポアソン分布の確率より

$$(3.1) \quad P(X = x) = \frac{(\lambda V)^x}{x!} \exp(-\lambda V)$$

である。

半径 r の u 次元球の体積を $V_u^A(r)$ とすると、その内部に a 以外に1つも点が現れない確率 p_1 は

$$p_1 = P(X = 0) = \exp(-\lambda V_u^A(r))$$

である。

一方、 u 次元球の表面上に微小厚み dr をもつ表層領域を考えると、 u 次元球の表面積は、

$$(3.2) \quad \frac{d}{dr} V_u^A(r)$$

であるので、

表層領域の微小体積は、

$$(3.3) \quad \frac{d}{dr} V_u^A(r) \cdot dr$$

となる。

この表層領域の中に b 以外に1つも点が現れない確率は

$$(3.4) \quad P(X = 0) = \exp\left(-\lambda \cdot \frac{d}{dr} V_u^A(r) \cdot dr\right)$$

である。

したがって、この表層領域の中に少なくとも1つ点が現れる確率 p_2 は

$$(3.5) \quad p_2 = 1 - \exp\left(-\lambda \cdot \frac{d}{dr} V_u^A(r) \cdot dr\right)$$

となる。

これら2つの事象は独立なので、同時確率は、

$$p_1 \cdot p_2 = \exp(-\lambda V_u^A(r)) \cdot \left\{1 - \exp\left(-\lambda \cdot \frac{d}{dr} V_u^A(r) \cdot dr\right)\right\}$$

となる。

指数関数の Taylor 展開の1次近似までを使うと、

$$(3.6) \quad \exp\left(-\lambda \cdot \frac{d}{dr} V_u^A(r) \cdot dr\right) \approx 1 - \lambda \cdot \frac{d}{dr} V_u^A(r) \cdot dr$$

と表せるので、同時確率は、

$$\begin{aligned} p_1 \cdot p_2 &\approx \exp(-\lambda V_u^A(r)) \cdot \left(\lambda \cdot \frac{d}{dr} V_u^A(r) \cdot dr \right) \\ &= \lambda \cdot \frac{d}{dr} V_u^A(r) \cdot \exp(-\lambda V_u^A(r)) \cdot dr \end{aligned}$$

となり、確率密度関数は、

$$f_u(r) \approx \lambda \cdot \frac{d}{dr} V_u^A(r) \cdot \exp(-\lambda V_u^A(r))$$

となる。

u 次元球（全球）の体積 $V_u^A(r)$ は、

$$(3.7) \quad V_u^A(r) = \frac{\pi^{\frac{u}{2}}}{\Gamma\left(\frac{u}{2} + 1\right)} r^u$$

なので、 u 次元空間の確率密度関数は、

$$f_u(r) \approx u\lambda \frac{\pi^{\frac{u}{2}}}{\Gamma\left(\frac{u}{2} + 1\right)} r^{u-1} \cdot \exp\left(-\lambda \frac{\pi^{\frac{u}{2}}}{\Gamma\left(\frac{u}{2} + 1\right)} r^u\right)$$

と表すことができる。

また u 次元空間の原点まわりの s 次モーメントは、

$$E_u[r^s] = \int_0^\infty r^s \cdot f_u(r) dr = \frac{\left(\Gamma\left(\frac{u}{2} + 1\right)\right)^{\frac{s}{u}} \Gamma\left(1 + \frac{s}{u}\right)}{\lambda^{\frac{s}{u}} \pi^{\frac{s}{2}}}$$

を用いて求められる。

具体例として1次元・2次元・3次元における第1近隣木の辺長の確率密度関数を求める。確率密度関数の一般式に u の値を代入することにより、

$$f_1(r) = 2\lambda \cdot \exp(-2\lambda r)$$

$$f_2(r) = 2\lambda\pi r \cdot \exp(-\lambda\pi r^2)$$

$$f_3(r) = 4\lambda\pi r^2 \cdot \exp\left(-\frac{4\lambda\pi}{3} r^3\right)$$

と求められる。

次に1次元・2次元・3次元における第1近隣木の辺長の期待値を求める。原点まわりの s 次モーメン

トの一般式に $s = 1$ と u の値を代入することにより、

$$E_1[r^1] = \frac{1}{2\lambda}$$

$$E_2[r^1] = \frac{1}{2\sqrt{\lambda}}$$

$$E_3[r^1] = \frac{\Gamma(1/3)}{3^{2/3} \cdot 4^{1/3} \cdot \lambda^{1/3} \cdot \pi^{1/3}} \approx \frac{0.553960 \dots}{\lambda^{1/3}}$$

と求められる。

最後に 1 次元・2 次元・3 次元における第 1 近隣木の辺長の標準偏差 (SD) を求める。標準偏差 (SD) は、

$$SD_u = \sqrt{E_u(r^2) - \{E_u(r^1)\}^2}$$

を用いて求められる。

また、原点まわりの 2 次モーメントは、

$$E_1[r^2] = \frac{1}{2\lambda^2}$$

$$E_2[r^2] = \frac{1}{\lambda\pi} \approx \frac{0.318309 \dots}{\lambda}$$

$$E_3[r^2] = \frac{2\Gamma(2/3)}{3^{1/3} \cdot 4^{2/3} \cdot \lambda^{2/3} \cdot \pi^{2/3}} \approx \frac{0.347407 \dots}{\lambda^{2/3}}$$

である。

これらを用いて標準偏差 (SD) は、

$$SD_1 = \frac{1}{2\lambda}$$

$$SD_2 = \sqrt{\frac{4 - \pi}{4\lambda\pi}} \approx \frac{0.261361 \dots}{\lambda^{1/2}}$$

$$SD_3 = \frac{\sqrt{2 \cdot 3 \cdot \Gamma(2/3) - \Gamma(1/3)^2}}{3^{2/3} \cdot 4^{1/3} \cdot \lambda^{1/3} \cdot \pi^{1/3}} \approx \frac{0.201334 \dots}{\lambda^{1/3}}$$

と求められる。

3.2 第 k 近隣木

第 k 近隣木は、領域内の各点と、その点から k 番目に近い点（第 k 近隣点）を結ぶ辺で構成される。したがって辺数は点数と等しい。2次元の場合は Thompson (1956) と Pollard (1971) によって、多次元の場合は Yao & Simons (1996) によって導出されている。

図3.3は単位正方形内に一様ランダムに10点を配置した第2近隣木の例 ($\lambda = 10$) である。両側に矢印がある場合は2本の線が重なっている。

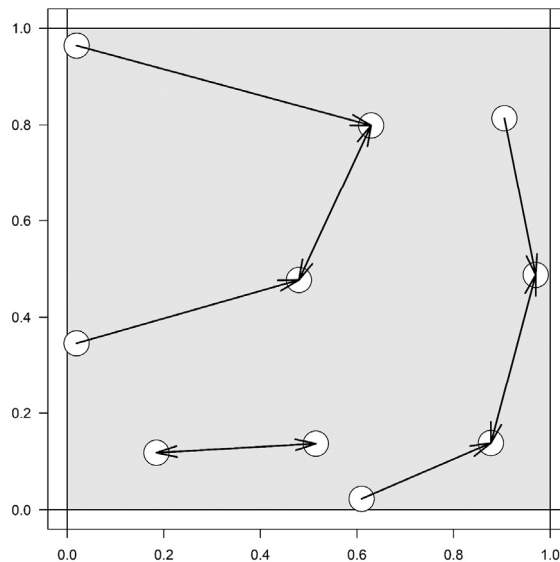


図3.3 第2近隣木 (2次元の場合)

点 a を中心とした半径 r の多次元球 A を考える。 A の表面上に点 b があり、球内に $(k - 1)$ 個の点がある時、辺 ab は第 k 近隣木の辺となる。 u 次元における第 k 近隣木の辺長の分布は、第1近隣木の辺長の導出過程を拡張して、半径 r の u 次元球の内部に a 以外に $(k - 1)$ 個の点があり、且つ u 次元球の表面上に少なくとも1つ点があるという確率から導くことができる。図3.4は第2近隣点 (2次元の場合) の例である。任意に選んだ点 a (○) と、その第1近隣点 (▲) および第2近隣点 b (●) を図中に示す。

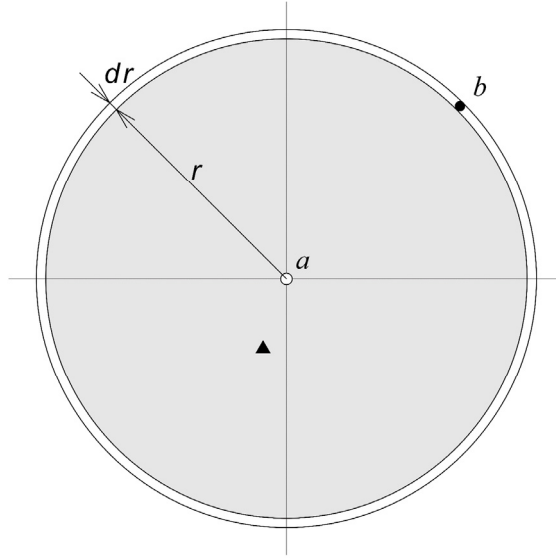


図3.4 第2近隣点 (2次元の場合)

体積 V の領域内に点が x 個現れる確率 $P(X = x)$ は、一様ポアソン分布の確率より式(3.1)となる。

半径 r の u 次元球の体積を $V_u^A(r)$ とすると、その内部に a 以外に $(k-1)$ 個の点が現れる確率 p_1 は

$$p_1 = P(X = k - 1) = \frac{(\lambda V_u^A(r))^{k-1}}{(k-1)!} \exp(-\lambda V_u^A(r))$$

である。

一方、球 A の表面上に微小厚み dr をもつ表層領域を考えると、 u 次元球の表面積は式(3.2)、表層領域の微小体積は式(3.3)である。この表層領域の中に a 以外に1つも点が現れない確率は式(3.4)である。

したがって、この表層領域の中に少なくとも1つ点が現れる確率 p_2 は式(3.5)となる。

これら2つの事象は独立なので、同時確率は、

$$p_1 \cdot p_2 = \frac{(\lambda V_u^A(r))^{k-1}}{(k-1)!} \exp(-\lambda V_u^A(r)) \cdot \left\{ 1 - \exp\left(-\lambda \cdot \frac{d}{dr} V_u^A(r) \cdot dr\right) \right\}$$

となる。

指数関数の Taylor 展開の1次近似である式(3.6)を使うと、同時確率は、

$$\begin{aligned} p_1 \cdot p_2 &\approx \frac{(\lambda V_u^A(r))^{k-1}}{(k-1)!} \exp(-\lambda V_u^A(r)) \cdot \left(\lambda \cdot \frac{d}{dr} V_u^A(r) \cdot dr \right) \\ &= \lambda \cdot \frac{d}{dr} V_u^A(r) \cdot \frac{(\lambda V_u^A(r))^{k-1}}{(k-1)!} \exp(-\lambda V_u^A(r)) \cdot dr \end{aligned}$$

となり、確率密度関数は、

$$\begin{aligned}
f_u(r) &= \lambda \cdot \frac{d}{dr} V_u^A(r) \cdot \frac{(\lambda V_u^A(r))^{k-1}}{(k-1)!} \exp(-\lambda V_u^A(r)) \\
&= \frac{(\lambda V_u^A(r))^{k-1}}{(k-1)!} \cdot \lambda \frac{d}{dr} V_u^A(r) \cdot \exp(-\lambda V_u^A(r)) \\
&= \frac{(\lambda V_u^A(r))^{k-1}}{(k-1)!} \cdot \lambda \frac{u V_u^A(r)}{r} \cdot \exp(-\lambda V_u^A(r)) \\
&= \frac{u (\lambda V_u^A(r))^k}{r (k-1)!} \cdot \exp(-\lambda V_u^A(r))
\end{aligned}$$

となる。

u 次元球(全球)の体積 $V_u^A(r)$ は式(3.7)なので、 u 次元空間の確率密度関数は、

$$f_u(r) = \frac{u \left(\lambda \frac{\pi^{\frac{u}{2}}}{\Gamma\left(\frac{u}{2} + 1\right)} r^u \right)^k}{r (k-1)!} \cdot \exp\left(-\lambda \frac{\pi^{\frac{u}{2}}}{\Gamma\left(\frac{u}{2} + 1\right)} r^u\right)$$

と表すことができる。

また u 次元空間の原点まわりの s 次モーメントは、

$$E_u[r^s] = \int_0^\infty r^s \cdot f_u(r) dr = \frac{1}{(k-1)!} \cdot \frac{\left(\Gamma\left(\frac{u}{2} + 1\right)\right)^{\frac{s}{u}} \Gamma\left(k + \frac{s}{u}\right)}{\lambda^{\frac{s}{u}} \pi^{\frac{s}{2}}}$$

を用いて求められる。

具体例として1次元・2次元・3次元の第2近隣木の辺長の確率密度関数を求める。確率密度関数の一般式に $k=2$ と u の値を代入することにより、

$$f_1(r) = 4\lambda^2 r \cdot \exp(-2\lambda r)$$

$$f_2(r) = 2\lambda^2 \pi^2 r^3 \cdot \exp(-\lambda \pi r^2)$$

$$f_3(r) = \frac{16\lambda^2 \pi^2 \lambda \pi}{3} r^5 \cdot \exp\left(-\frac{4\lambda \pi}{3} r^3\right)$$

と求められる。

次に1次元・2次元・3次元の第2近隣木の辺長の期待値を求める。原点まわりの s 次モーメントの一般式に $k=2$ 、 $s=1$ および u の値を代入することにより、

$$E_1[r^1] = \frac{1}{\lambda}$$

$$E_2[r^1] = \frac{3}{4\sqrt{\lambda}}$$

$$E_3[r^1] = \frac{4^{2/3}\Gamma(1/3)}{3^{5/3}\pi^{1/3}\lambda^{1/3}} \approx \frac{0.738613\dots}{\lambda^{1/3}}$$

と求められる。

最後に 1 次元・2 次元・3 次元の第 2 近隣木の辺長の標準偏差 (SD) を求める。標準偏差は、

$$SD_u = \sqrt{E_u(r^2) - \{E_u(r^1)\}^2}$$

を用いて求められる。

また、原点まわりの 2 次モーメントは、

$$E_1[r^2] = \frac{3}{2\lambda^2}$$

$$E_2[r^2] = \frac{2}{\lambda\pi} \approx \frac{0.636619\dots}{\lambda}$$

$$E_3[r^2] = \frac{10\Gamma(2/3)}{3^{4/3}4^{2/3}\lambda^{2/3}\pi^{2/3}} \approx \frac{0.579012\dots}{\lambda^{2/3}}$$

である。

これらを用いて標準偏差 (SD) は、

$$SD_1 = \frac{1}{\sqrt{2}\lambda} \approx \frac{0.707106\dots}{\lambda}$$

$$SD_2 = \sqrt{\frac{2}{\lambda\pi} - \frac{9}{16\lambda}} \approx \frac{0.272249\dots}{\sqrt{\lambda}}$$

$$SD_3 = \sqrt{\frac{10\Gamma(2/3)}{3^{4/3}4^{2/3}\lambda^{2/3}\pi^{2/3}} - \frac{4^{4/3}\Gamma(1/3)^2}{3^{10/3}\pi^{2/3}\lambda^{2/3}}} \approx \frac{0.182928\dots}{\lambda^{1/3}}$$

と求められる。

$u = 2$ 、 $\lambda = 100$ に固定して、 k を 1 から 10 まで変えた時の第 k 近隣木の確率密度関数を図 3.5 に示す。 k を増やしていくと辺長の期待値が増えると同時に、分布の対称性が増していく。また、ばらつきは若干ではあるが増加していく。

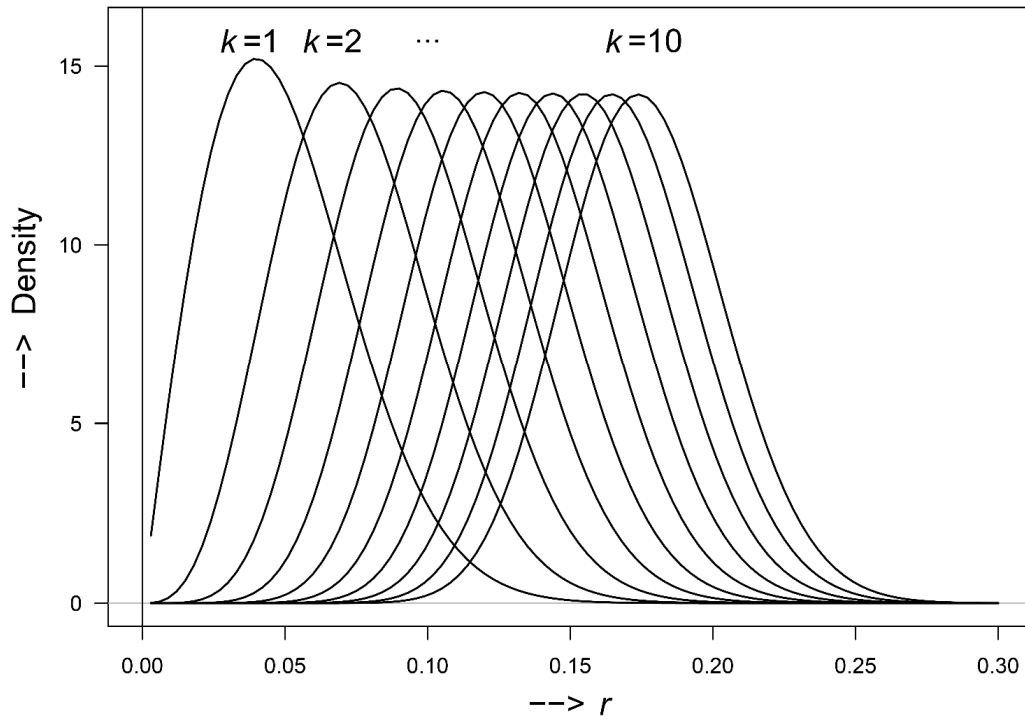


図 3.5 第 k 近隣木の辺長の確率密度関数

3.3 相互最近隣対 (Reciprocal Nearest Neighbor Pairs または Mutual Nearest Neighbor Pairs)

英語・日本語とも複数の表記法がある。ここでは日本語の表記は大隅・他 (1994) に従う。2点 a と b が相互最近隣対であるとは、 a が b の最近隣になり、かつ b が a の最近隣であるときをいう。多次元の分布の諸元に関しては Pickard (1982) が導出している。図 3.6 は単位正方形内に一様ランダムに 10 点を配置した相互最近隣木の例 ($\lambda = 10$) である。破線の矢印を実線にすれば、図 3.1 に示した第 1 近隣木の例と同じになる。また、図 3.7 に相互最近隣対 (2 次元) となる 2 点の位置関係を示す。

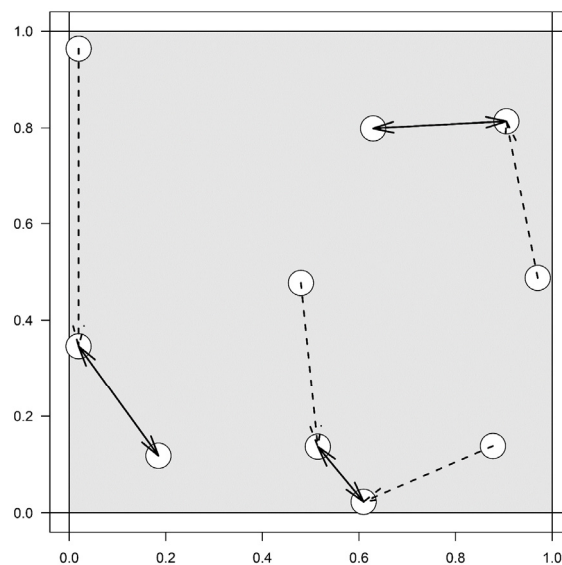


図 3.6 相互最近隣木 (2 次元)

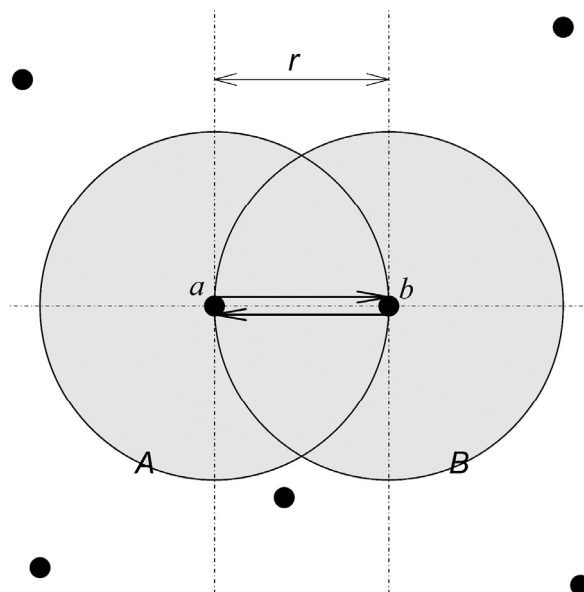


図 3.7 相互最近隣対 (2 次元)

相互最近隣対が発生する事象を場合分けするため、以下のような設定を行う。

u 次元空間に任意の2点 a と b を与える。点 a を中心とする半径 r の u 次元球を A とし、点 b を中心とする半径 r の u 次元球を B とする。 r を少しずつ増加させながら相互最近隣対の発生状況を考える。

ここで、

R : 相互最近隣対の辺長

r' : a, b 間の距離

$\partial(A \cup B)$: 領域 $A \cup B$ の表層領域 (厚みゼロ)

$\text{int}(A \cup B)$: 領域 $A \cup B$ から $\partial(A \cup B)$ を除いた領域

とする。

場合分けの結果を表 3.1 に示す。

表 3.1 相互最近隣対が発生する事象の場合分け

	$R < r$	$R = r$	$R > r$
$r < r'$	$\text{int}(A \cup B)$ に、 a と b 以外に少なくとも1つ点が存在する。	$\partial(A \cup B)$ に、 a と b 以外に少なくとも1つ点が存在する。	$A \cup B = \text{int}(A \cup B) \cup \partial(A \cup B)$ に a と b 以外の点が存在しない。
$r = r'$	同上	a と b が相互最近隣対となる。 または、それ以外に $\partial(A \cup B)$ に点が存在する。	こういう事象は発生しない。 (\emptyset)
$r > r'$	a と b が相互最近隣対となる。 または、 $\text{int}(A \cup B)$ に a または b からの距離が r' 以下の点がある。	こういう事象は発生しない。 (\emptyset)	こういう事象は発生しない。 (\emptyset)

点 a と b からの距離を r としたときに、相互最近隣対の長さが r より大きくなる確率は $P(R > r)$ なので、

$$F_u(r) = P(R \leq r) = 1 - P(R > r)$$

$$\left\{ \begin{array}{ll} = 1 - \exp\{-\lambda(V_u^{A \cup B}(r))\}, & r < r' \\ = 1, & r = r' \\ = 1, & r > r' \end{array} \right.$$

となる。

したがって確率密度関数 $f_u(r)$ は、

$$f_u(r) = \frac{d}{dr} F_u(r) = \lambda \cdot \frac{d}{dr} V_u^{\text{AUB}}(r) \cdot \exp\{-\lambda V_u^{\text{AUB}}(r)\}$$

となる。

ここで偶数次元 ($u = 2t$) の確率密度関数は付録2の式(付2.3)を使って、

$$f_{2t}(r) = 2t\lambda \cdot C_e^{\text{AUB}} \cdot r^{2t-1} \cdot \exp(-\lambda \cdot C_e^{\text{AUB}} \cdot r^{2t})$$

$$\text{ここで、 } C_e^{\text{AUB}} = \frac{V_{2t-1}(r)}{r^{2t-1}} \cdot \left\{ \frac{2\sqrt{\pi}\Gamma(t+\frac{1}{2})}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \right\}$$

$$B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \text{は不完全ベータ関数。}$$

$$\frac{V_{2t-1}(r)}{r^{2t-1}} \text{は次元に依存する定数。}$$

となり、

奇数次元 ($u = 2t + 1$) の確率密度関数は付録2の式(付2.4)を使って、

$$f_{2t+1}(r) = (2t + 1)\lambda \cdot C_o^{\text{AUB}} \cdot r^{2t} \cdot \exp(-\lambda \cdot C_o^{\text{AUB}} \cdot r^{2t+1})$$

$$\text{ここで、 } C_o^{\text{AUB}} = \frac{V_{2t}(r)}{r^{2t}} \cdot \left\{ \frac{2\sqrt{\pi}\Gamma(t+1)}{\Gamma(t+\frac{3}{2})} - B_{\frac{3}{4}}\left(t + 1, \frac{1}{2}\right) \right\}$$

$$B_{\frac{3}{4}}\left(t + 1, \frac{1}{2}\right) \text{は不完全ベータ関数。}$$

$$\frac{V_{2t}(r)}{r^{2t}} \text{は次元に依存する定数。}$$

となる。

また u 次元空間の原点まわりの s 次モーメントは、

$$E_u[r^s] = \int_0^\infty r^s \cdot f_u(r) dr$$

を用いて求められる。

ここで偶数次元 ($u = 2t$) の原点まわりの s 次モーメントは付録2の式(付2.3)を使って、

さらに以下の公式

$$\int_0^\infty r^{b-1} \cdot \exp(-ar^\alpha) dr = \frac{\Gamma\left(\frac{b}{\alpha} + 1\right)}{a^{b/\alpha} \cdot b}$$

を用いて

$$\begin{aligned}
E_{2t}[r^s] &= \int_0^\infty r^s \cdot f_{2t}(r) dr = 2t \lambda C_e^{AUB} \cdot \int_0^\infty r^{s+2t-1} \cdot \exp(-\lambda \cdot C_e^{AUB} \cdot r^{2t}) dr \\
&= 2t \lambda C_e^{AUB} \cdot \frac{\Gamma\left(\frac{s+2t}{2t} + 1\right)}{(\lambda C_e^{AUB})^{\frac{s+2t}{2t}} \cdot (s+2t)} \\
&= 2t \lambda C_e^{AUB} \cdot \frac{\frac{s+2t}{2t} \cdot \Gamma\left(\frac{s+2t}{2t}\right)}{(\lambda C_e^{AUB})^{\frac{s+2t}{2t}} \cdot (s+2t)} \\
&= (\lambda C_e^{AUB})^{1-\frac{s+2t}{2t}} \cdot \Gamma\left(\frac{s+2t}{2t}\right) \\
&= (\lambda C_e^{AUB})^{\frac{-s}{2t}} \cdot \Gamma\left(\frac{s}{2t} + 1\right) \\
&= \frac{s}{2t} \cdot (\lambda C_e^{AUB})^{\frac{-s}{2t}} \cdot \Gamma\left(\frac{s}{2t}\right)
\end{aligned}$$

ここで、 $C_e^{AUB} = \frac{V_{2t-1}(r)}{r^{2t-1}} \cdot \left\{ \frac{2\sqrt{\pi}\Gamma(t+\frac{1}{2})}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \right\}$

$B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right)$ は不完全ベータ関数。

$\frac{V_{2t-1}(r)}{r^{2t-1}}$ は次元に依存する定数。

となり、

奇数次元 ($u = 2t + 1$) の原点まわりの s 次モーメントは付録 2 の式 (付 2.4) を使って、

さらに以下の公式

$$\int_0^\infty r^{b-1} \cdot \exp(-ar^\alpha) dr = \frac{\Gamma\left(\frac{b}{\alpha} + 1\right)}{a^{b/\alpha} \cdot b}$$

を用いて

$$\begin{aligned}
E_{2t+1}[r^s] &= \int_0^\infty r^s \cdot f_{2t+1}(r) dr = (2t+1) \lambda C_o^{AUB} \cdot \int_0^\infty r^{s+2t} \cdot \exp(-\lambda \cdot C_o^{AUB} \cdot r^{2t+1}) dr \\
&= (2t+1) \lambda C_o^{AUB} \cdot \frac{\Gamma\left(\frac{s+2t+1}{2t+1} + 1\right)}{(\lambda C_o^{AUB})^{\frac{s+2t+1}{2t+1}} \cdot (s+2t+1)} \\
&= (2t+1) \lambda C_o^{AUB} \cdot \frac{\frac{s+2t+1}{2t+1} \cdot \Gamma\left(\frac{s+2t+1}{2t+1}\right)}{(\lambda C_o^{AUB})^{\frac{s+2t+1}{2t+1}} \cdot (s+2t+1)} \\
&= (\lambda C_o^{AUB})^{1-\frac{s+2t+1}{2t+1}} \cdot \Gamma\left(\frac{s+2t+1}{2t+1}\right)
\end{aligned}$$

$$\begin{aligned}
&= (\lambda C_o^{\text{AUB}})^{\frac{-s}{2t+1}} \cdot \Gamma\left(\frac{s}{2t+1} + 1\right) \\
&= \frac{s}{2t+1} \cdot (\lambda C_o^{\text{AUB}})^{\frac{-s}{2t+1}} \cdot \Gamma\left(\frac{s}{2t+1}\right)
\end{aligned}$$

ここで、 $C_o^{\text{AUB}} = \frac{V_{2t}(r)}{r^{2t}} \cdot \left\{ \frac{2\sqrt{\pi}\Gamma(t+1)}{\Gamma(t+\frac{3}{2})} - B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right) \right\}$

$B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right)$ は不完全ベータ関数。

$\frac{V_{2t}(r)}{r^{2t}}$ は次元に依存する定数。

となる。

具体例として1次元・2次元・3次元の確率密度関数を求める。

1次元の場合は奇数次元の式で $t = 0$ と置き、2次元の場合は偶数次元の式で $t = 1$ と置き、3次元の場合は奇数次元の式で $t = 1$ と置くことにより、

$$\begin{aligned}
f_1(r) &= 3\lambda \cdot \exp(-3\lambda r) \\
f_2(r) &= 2\lambda \left(\frac{4\pi}{3} + \frac{\sqrt{3}}{2}\right) r \cdot \exp\left\{-\lambda \left(\frac{4\pi}{3} + \frac{\sqrt{3}}{2}\right) r^2\right\} \\
f_3(r) &= 3\lambda \left(\frac{9\pi}{4}\right) r^2 \cdot \exp\left\{-\lambda \left(\frac{9\pi}{4}\right) r^3\right\}
\end{aligned}$$

と求められる。

次に1次元・2次元・3次元の期待値を求める。原点まわりの s 次モーメントの一般式を用いる。1次元の場合は奇数次元の式に $t = 0$ と $s = 1$ を、2次元の場合は偶数次元の式に $t = 1$ と $s = 1$ を、3次元の場合は奇数次元の式に $t = 1$ と $s = 1$ を、それぞれ代入することにより

$$\begin{aligned}
E_1[r^1] &= \frac{1}{3\lambda} \approx \frac{0.333333 \dots}{\lambda} \\
E_2[r^1] &= \frac{1}{2} \sqrt{\frac{\pi}{\lambda \left(\frac{4\pi}{3} + \frac{\sqrt{3}}{2}\right)}} \approx \frac{0.394177 \dots}{\lambda^{1/2}} \\
E_3[r^1] &= \frac{4^{1/3} \cdot \Gamma(1/3)}{3^{5/3} \cdot \lambda^{1/3} \cdot \pi^{1/3}} \approx \frac{0.465297 \dots}{\lambda^{1/3}}
\end{aligned}$$

と求められる。

3.4 MST (Minimum Spanning Tree)

MSTは領域内の全ての点を連結するグラフの中で、辺の総延長距離が最小のものである。Kruskal (1956) や Prim (1957) などにより MST の効率的な探索法が提案されているが、MST の辺長が従う確率分布は解析的に求められていない。したがって、ここでは近似分布について記述する。第 k 近隣木とは異なり、辺数は点数より 1 だけ少ない。図 3.8 は単位正方形内に一様ランダムに 10 点を配置した MST の例 ($\lambda = 10$) である。

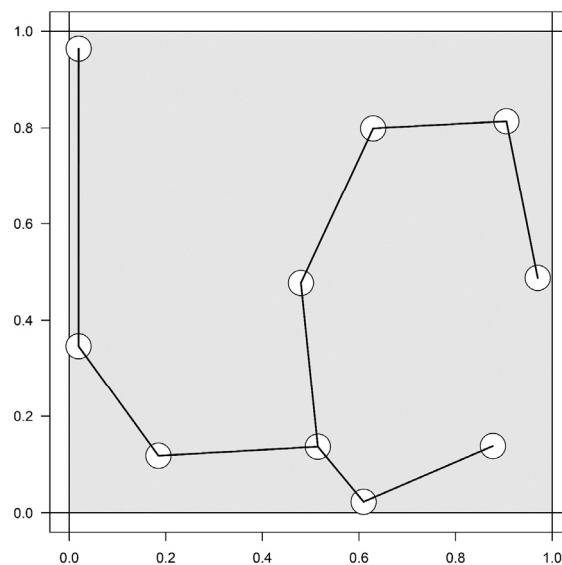


図 3.8 MST (2次元の場合)

3.4.1 MST の近似としての第 1 近隣木

MST の辺長の近似分布として第 1 近隣木が使われる場合がある。ただし近似分布の下限值としてである。第 1 近隣木と MST の関連性に関する文献には、古山 (2003) や Penrose (1997) などがある。古山 (2003) は背理法を用いて第 1 近隣木の全ての辺は MST の辺の構成要素となることを証明している。また Penrose (1997) は、第 1 近隣木の最長辺と MST の最長辺が漸近的に等しいことを証明している。

現実の問題では両者の乖離は大きい。第 1 近隣木には高い確率で相互最近隣対が発生するからである。その結果、第 1 近隣木は、領域内の全ての点をリンクすることができない。2次元の第 1 近隣木に対する相互最近隣対の発生確率を Clark (1955) ・ Pielou (1974) ・ 古山 (2003) が算出している。また多次元

の場合を Pickard (1982) が算出している。ここでは Pielou (1974) の方法を多次元に拡張し Pickard (1982) の結果と一致することを示す。

図 3.9 に MST・第 1 近隣木・相互最近隣対の位置関係を示す。単位正方形内に一様ランダムに 100 点を配置した ($\lambda = 100$) 場合である。灰色の辺が MST を、黒の矢印が第 1 近隣木を、白抜きの辺が相互最近隣対を表す。

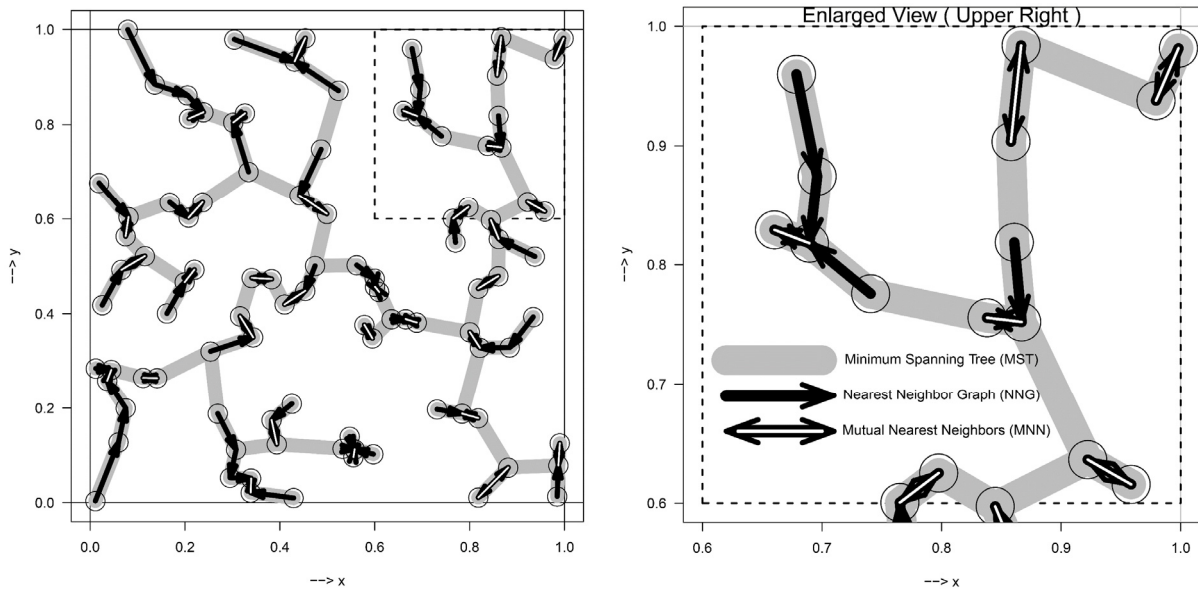


図 3.9 MST・第 1 近隣木・相互最近隣対の位置関係

u 次元空間に距離 r 離れて存在する 2 点 a と b を考える。点 a を中心とする半径 r の u 次元球を A とし、点 b を中心とする半径 r の u 次元球を B とする。また A と B の合併領域 (和集合 $A \cup B$) の体積は $V_u^{A \cup B}(r)$ 、 A と B の共通領域 (積集合 $A \cap B$) の体積は $V_u^{A \cap B}(r)$ 、 A と B の共通領域を、 B から除いた領域の体積 ($V_u^B(r) - V_u^{A \cap B}(r)$) は $V_u^{A^c \cap B}(r)$ と表せる。

強度 λ の一様ランダムな点過程に於いて、第 1 近隣木の辺の両端である点 a と点 b が相互最近隣対になる確率を考える。(a から b を見て) b が a の最近隣個体である確率 p_1 は、 a の最近隣個体が $(r, r + dr)$ の範囲 (超球面上) にある確率なので、

$$p_1 = \lambda \cdot \frac{d}{dr} V_u^A(r) \cdot \exp\{-\lambda V_u^A(r)\} \cdot dr$$

となる。

(b から a を見て) a が b の最近隣個体である確率 p_2 は、領域 $A^c \cap B$ に他の第 3 点が存在しない確率なので、

$$p_2 = \exp\{-\lambda V_u^{A^c \cap B}(r)\}$$

となる。

相互最近接対の発生確率は p_1 と p_2 の同時確率で表せるので、

$$p_1 \cdot p_2 = \lambda \cdot \frac{d}{dr} V_u^A(r) \cdot \exp\{-\lambda V_u^A(r)\} \cdot \exp\{-\lambda V_u^{A^c \cap B}(r)\} \cdot dr$$

となる。

どの個体も相互最近接対の一方となる確率 P は、全ての可能な r の値に亘って上式を積分して求められる。

付録 2 の (付 2.3) 式と (付 2.4) 式を使って、偶数次元 ($u = 2t$) の確率 P_e と、奇数次元 ($u = 2t + 1$) の確率 P_o を求めると、

$$P = \int_0^\infty \lambda \cdot \frac{d}{dr} V_u^A(r) \cdot \exp\{-\lambda V_u^{A \cup B}(r)\} \cdot dr$$

$$P_e = \frac{1}{1 + \frac{\Gamma(t+1)}{\pi^t} C_e^{A^c \cap B}}$$

$$P_o = \frac{1}{1 + \frac{\Gamma\left(t + \frac{3}{2}\right)}{\pi^{t+\frac{1}{2}}} C_o^{A^c \cap B}}$$

$$\text{ここで、 } C_e^{A^c \cap B} = \frac{V_{2t-1}(r)}{r^{2t-1}} \cdot \left\{ \frac{\sqrt{\pi} \Gamma\left(t + \frac{1}{2}\right)}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \right\}$$

$$B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \text{ は不完全ベータ関数。}$$

$$C_o^{A^c \cap B} = \frac{V_{2t}(r)}{r^{2t}} \cdot \left\{ \frac{\sqrt{\pi} \Gamma(t+1)}{\Gamma\left(t + \frac{3}{2}\right)} - B_{\frac{3}{4}}\left(t + 1, \frac{1}{2}\right) \right\}$$

$$B_{\frac{3}{4}}\left(t + 1, \frac{1}{2}\right) \text{ は不完全ベータ関数。}$$

となる。

具体的に 2 次元で $\lambda = 100$ 場合は 62.2% となる。これは第 1 近隣木の辺数に対する割合である。図 3.9 の例では点数が 100 なので、第 1 近隣木の辺数も 100 である。実際に数えるときは、両側に矢印がある辺は 2 本が重なっていることに注意する。この両側に矢印が付いた辺が相互最近隣対で、白抜きで表示されている。したがって期待される相互最近隣対の辺数は、 $100 \times 0.622 \div 2 = 31.1$ となる。実際に数え

てみると白抜きの辺は 31 本ある。

本研究では第 1 近隣木に対する相互最近隣対の発生確率だけを論じている。一方、Clark (1956) は、これを一般化して第 k 近隣木に対する相互最近隣対の発生確率を求めている。また、Cox (1981) は Clark (1956) の第 2 近隣木以降の値を訂正している。第 k 近隣木に対する相互最近隣対の発生確率は、MST の辺長を第 k 近隣木で近似する場合に必要となる。本研究では詳しく取り上げない。

3.4.2 Roberts の近似

Roberts (1968) は、生態学における Gilbert (1965) の *exodic tree* から 2 次元 MST と 3 次元 MST の辺長の近似分布を導出している。この近似分布の多次元への拡張は、Bertsimas and van Ryzin (1990) が行っている。

Gilbert (1965) は *exodic tree* の辺の総延長が MST の辺の総延長よりやや長いことを示している。後戻りしないので相互最近隣対が発生することはない。古山 (2003) と Watanabe (2008) は、Roberts (1968) の近似を MST の辺長の近似分布の上限としている。

Roberts (1968) は 2 次元 MST の辺長の近似分布を、半円の内部には 1 つも点が無く、且つ半円周上 (ただし直線部を除く) に 1 つだけ点があるという確率から導出している。この方法を多次元に拡張し、Bertsimas and van Ryzin (1990) の結果と一致することを示す。 u 次元に於ける Roberts (1968) の近似分布は、半径 r の u 次元半球の内部には 1 つも点がなく、且つ u 次元半球の表面上 (ただし超平面部を除く) に 1 つだけ点があるという確率から導くことができる。

半径 r の u 次元半球の体積を $\frac{1}{2}V_u(r)$ とする、この内部に 1 つも点がない確率 p_1 は、

$$p_1 = \exp\left(-\frac{\lambda V_u(r)}{2}\right)$$

である。

一方、 u 次元半球の表面上 (ただし超平面部を除く) に微小厚み dr をもつ表層領域を考えると、 u 次元半球の表面積は、

$$\frac{d V_u(r)}{dr} \frac{1}{2}$$

なので、表層領域の微小体積は

$$\frac{d V_u(r)}{dr} \cdot \frac{dr}{2}$$

となる。

この表層領域内に1つだけ点があるという確率 p_2 は、

$$p_2 = \lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2} \cdot \exp\left(-\lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2}\right)$$

となる。

これら2つの事象は独立なので、同時確率は、

$$p_1 \cdot p_2 = \exp\left(-\frac{\lambda V_u(r)}{2}\right) \cdot \left\{ \lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2} \cdot \exp\left(-\lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2}\right) \right\}$$

となる。

指数関数の Taylor 展開の1次近似までを使うと

$$\exp\left(-\lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2}\right) \approx 1 - \lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2}$$

と表せるので、同時確率は

$$\begin{aligned} p_1 \cdot p_2 &\approx \exp\left(-\frac{\lambda V_u(r)}{2}\right) \cdot \left\{ \lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2} \cdot \left(1 - \lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2}\right) \right\} \\ &= \exp\left(-\frac{\lambda V_u(r)}{2}\right) \cdot \left\{ \lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2} - \left(\lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2}\right)^2 \right\} \end{aligned}$$

となる。

ここで、 $\left(\lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2}\right)^2$ は高次の無限小であることを考えれば、同時確率は、

$$p_1 \cdot p_2 \approx \lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2} \cdot \exp\left(-\frac{\lambda V_u(r)}{2}\right) \cdot dr$$

と表せる。

したがって、確率密度関数 $f_u(r)$ は、

$$f_u(r) \approx \lambda \cdot \frac{d V_u(r)}{dr} \cdot \frac{dr}{2} \cdot \exp\left(-\frac{\lambda V_u(r)}{2}\right)$$

となる。

u 次元球の体積 $V_u(r) = \frac{\pi^{\frac{u}{2}}}{\Gamma(\frac{u}{2}+1)} r^u$ を代入すると、

$$(3.8) \quad f_u(r) \approx u\lambda \frac{\pi^{\frac{u}{2}}}{2\Gamma\left(\frac{u}{2} + 1\right)} r^{u-1} \cdot \exp\left(-\lambda \frac{\pi^{\frac{u}{2}}}{2\Gamma\left(\frac{u}{2} + 1\right)} r^u\right)$$

となる。

また u 次元空間の原点まわりの s 次モーメントは、

$$E_u[r^s] = \int_0^\infty r^s \cdot f_u(r) dr = \frac{\left(2\Gamma\left(\frac{u}{2} + 1\right)\right)^{\frac{s}{u}} \Gamma\left(1 + \frac{s}{u}\right)}{\lambda^{\frac{s}{u}} \pi^{\frac{s}{2}}}$$

となる。

具体例として1次元・2次元・3次元の確率密度関数を求める。確率密度関数の一般式に u の値を代入することにより、

$$f_1(r) = \lambda \cdot \exp(-\lambda r)$$

$$f_2(r) = \lambda\pi r \cdot \exp\left(-\frac{\lambda\pi}{2} r^2\right)$$

$$f_3(r) = 2\lambda\pi r^2 \cdot \exp\left(-\frac{2\lambda\pi}{3} r^3\right)$$

と求められる。

次に1次元・2次元・3次元の期待値を求める。原点まわりの s 次モーメントの一般式に $s = 1$ と u の値を代入することにより、

$$E_1[r^1] = \frac{1}{\lambda}$$

$$E_2[r^1] = \frac{1}{\sqrt{2\lambda}} \approx \frac{0.707106\dots}{\lambda^{1/2}}$$

$$E_3[r^1] = \frac{\Gamma(1/3)}{2^{1/3} \cdot 3^{2/3} \cdot \lambda^{1/3} \cdot \pi^{1/3}} \approx \frac{0.697946\dots}{\lambda^{1/3}}$$

と求められる。

最後に1次元・2次元・3次元の標準偏差 (SD) を求める。標準偏差 (SD) は、

$$SD_u = \sqrt{E_u(r^2) - \{E_u(r^1)\}^2}$$

を用いて求められる。

また原点まわりの2次モーメントは、

$$E_1[r^2] = \frac{2}{\lambda^2}$$

$$E_2[r^2] = \frac{2}{\lambda\pi}$$

$$E_3[r^2] = \frac{2^{1/3} \cdot \Gamma(2/3)}{3^{1/3} \cdot \lambda^{2/3} \cdot \pi^{2/3}}$$

である。

これらを用いて標準偏差 (SD) は

$$SD_1 = \frac{1}{\lambda}$$

$$SD_2 = \sqrt{\frac{4 - \pi}{2\lambda\pi}} \approx \frac{0.369621 \dots}{\lambda^{1/2}}$$

$$SD_3 = \frac{\sqrt{2 \cdot 3 \cdot \Gamma(2/3) - \Gamma(1/3)^2}}{2^{1/3} \cdot 3^{2/3} \cdot \lambda^{1/3} \cdot \pi^{1/3}} \approx \frac{0.253666 \dots}{\lambda^{1/3}}$$

と求められる。

3.4.3 Watanabe (2008) の近似

Watanabe (2008) は都市計画の研究において、2次元 MST の辺長の近似分布を導出している。2次元を含めた多次元の導出を付録 1 に記す。

3.4.4 MST の辺長の境界効果

笠原&古山 (1998) は、MST の近似として第 k 近隣木を用い MST の辺長の境界効果を数理的に論じている。そして計算機実験を行い、その導出結果を検証している。本研究と関係が深い計算機実験の結果を以下に記す。

笠原&古山 (1998) は、単位正方形内に一様ランダムに 100 点を配置した状況下で、新たに 1 点を

・点 a : 正方形の中心 ・点 b : 辺の中心 ・点 c : 隅点

に追加し、100 個の点と追加した点を結んでできる 100 本の辺長の総延長距離をモンテカルロ法を用いて求めている。追加した 3 点を図 3.10 に、100 本の辺長の 1 例を図 3.11 に示す。その結果、追加した点が正方形の中心 (点 a) である場合に比べ、辺の中心 (点 b) の場合は約 $\sqrt{2}$ 倍、隅点 (点 c) の場合は約 2

倍長くなることを示している。笠原&古山（1998）は乱数の初期設定を変化させ 100 回の試行を行い、その平均値を求めている。その計算結果と、本研究で行った試行回数 10000 回の追試結果を表 3.2 に示す。

表 3.2 笠原&古山（1998）の実験結果

		<i>a</i>	<i>b</i>	<i>c</i>
笠原・古山(1998) <i>m</i> =100	総延長距離	38.03	57.37	74.011
	期待値	0.3803	0.5737	0.74011
	比率	1	1.509	1.946
本研究の追試 <i>m</i> =10000	期待値	0.3825726	0.593205	0.7648091
	比率	1	1.551	1.999

笠原&古山（1998）は、その数理的検討の中で、近隣点を見つけるための探索方向の範囲が、正方形の中心（点 *a*）の場合が 360 度であるのに対して、辺の中心（点 *b*）の場合は 180 度、隅点（点 *c*）の場合は 90 度に低下することが、この現象の原因であると述べている。境界に近い点は最近隣点ではない点を選択し、求めた MST は最短経路より長くなる可能性が高くなる。

本研究との関連性を考え、最後に以下の考察を追加しておく。

- ・領域内の点数（MST の辺数）が少ないほど、境界効果の影響を受ける点数（MST の辺数）の割合が高くなり、その結果、MST の辺長の期待値は、より長くなってしまふと考えられる。

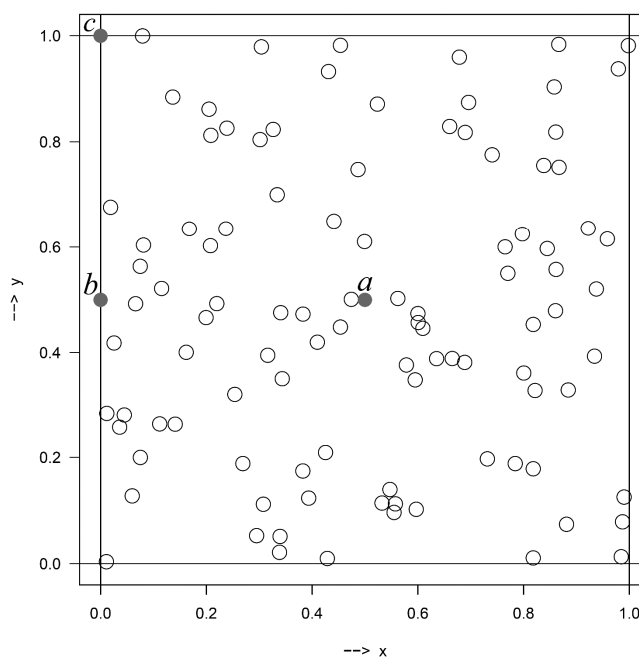


図 3.10 笠原&古山（1998）が追加した 3 点 (*a*, *b*, *c*)

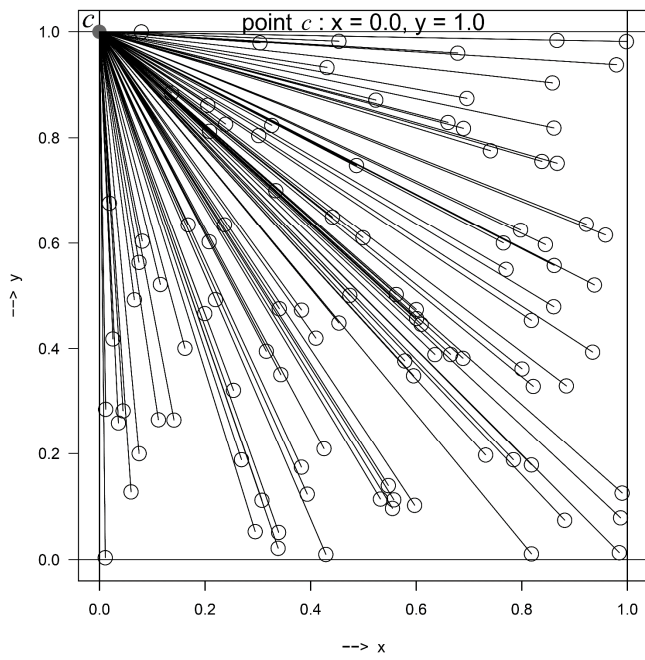
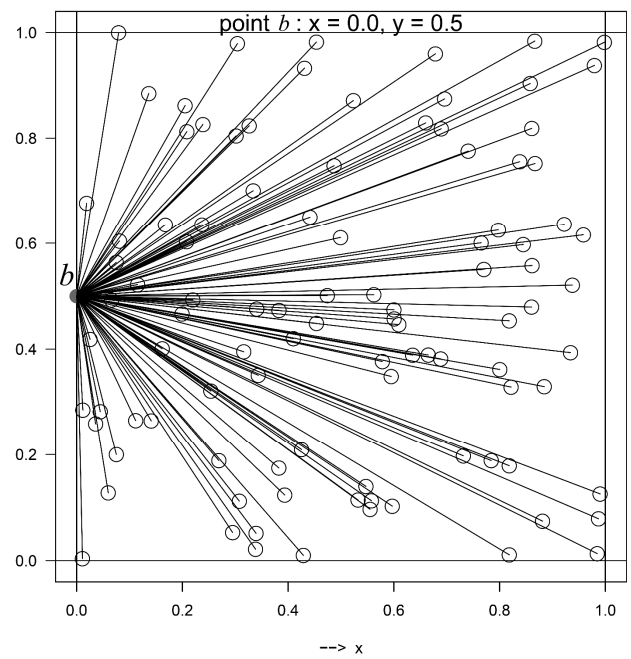
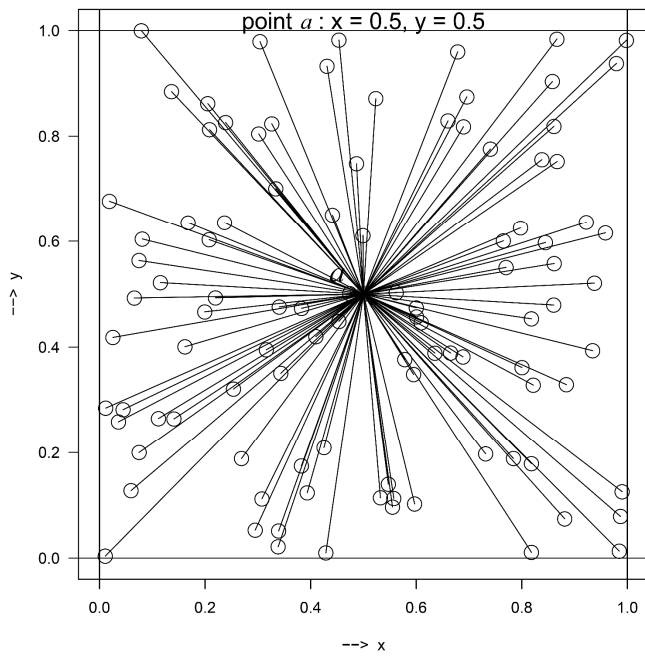


図 3.11 笠原&古山 (1998) の境界効果に関する実験 (上: 点 *a*、中: 点 *b*、下: 点 *c*)

第4章 空間点パターンのグラフィカルな分類方法

4.0 はじめに

本章では空間点パターンの探索的データ解析に用いられるグラフィカルな分類方法を取り上げる。これらは近接グラフ (Proximity graph) に基づくものと、2次特性 (Second-order characteristics) に基づくものに分けられる。ここでは前者としてG関数・F関数・J関数を、後者としてK関数・L関数・pair correlation 関数を取り上げる。空間点過程の2次特性とは、2点の相互作用、あるいは空間上の関係が空間内でどのように変化するのかを記述する量である。

各分類方法の説明の後に、その分類方法を用いた具体例を示す。分類対象の点パターンが、ランダム型・規則型・集中 (凝集) 型のどのタイプに属するのかをグラフィカルに判定する内容である。ここでは分類対象として3種類の現実の点パターンを用いる。これらは第2章にて説明済み (図 2.4) であって、先行研究 (Diggle, 1983, pp.1-2) によって分類先が明らかにされている。第4章に於ける分類事例は全て、上から順にランダム型・規則型・集中 (凝集) 型となっている。

各関数に共通する分類手順を以下に示す。

各関数に共通する分類手順

- Step 1: モンテカルロ法 ($m = 100$ 回) により CSR 点過程に対応する各関数の曲線を100本求める。
- Step 2: 100本の曲線群 (Step1) のエンベロープを作図する。
- Step 3: 分類対象 (点パターン) に対する各関数の曲線を求める。
- Step 4: Step2 のエンベロープの上に Step3 の曲線を重ね描きする。
- Step 5: 分類対象の曲線がエンベロープから逸脱する方向・程度により分類を行う。
この判断基準は各関数により異なり、具体的には以下で説明する。

ここで、各関数に対する境界補正の適用について触れておく。境界効果と、それに対する境界補正は4.2.1.1で説明することになる。

- ・ G 関数, F 関数, J 関数 : 境界補正無し
- ・ K 関数, L 関数, pair correlation 関数 : 境界補正有り

4.1 近接グラフに基づく方法

4.1.1 G関数を用いた分類

Diggle (1983, p.49) は G 関数 $G(r)$ を以下のように定義している（原文は r ではなく y である）。

$$G(r) = P\{\text{distance from an arbitrary event to the nearest other event is at most } r\}$$

一方、間瀬&武田 (2001, p.48) と Illian et al. (2008, p.199) は、 $G(r)$ を $D(r)$ と表記している。

$G(r)$ は領域内のランダムに選んだ点 (an arbitrary event) から、その最近隣点 (the nearest other event) までの距離の分布を考えるものである。これは、第3章で説明した第1近隣木の辺長の分布関数である。 $G(r)$ は CSR 点過程に対してだけ正確に求めることができる。ランダムに選んだ点を中心とした半径 r の円または球体の内部に、ランダムに選んだ点以外に、点が少なくとも1つ有る確率から導出する。2次元、3次元および多次元 (u 次元) の場合を以下に示す。ここで、 λ は CSR 点過程の強度、 $V_u(r)$ は半径 r の u 次元球の体積である。

$$G(r) = 1 - \exp(-\pi\lambda r^2)$$

$$G(r) = 1 - \exp\left(-\frac{4}{3}\pi\lambda r^3\right)$$

$$G(r) = 1 - \exp(-\lambda V_u(r)) \quad \text{ここで、} V_u(r) = \frac{\pi^{\frac{u}{2}}}{\Gamma(\frac{u}{2}+1)} r^u$$

規則型の点過程では点同士が反発するように存在しているので、距離が小さいところで $G(r)$ の値はゼロになり、距離が点間距離の最小値 (hard core distance; Illian et al., p.58) を超えてから $G(r)$ の値は増え始める。この現象を CSR 点過程の変化と比較することにより、規則型の点過程の $G(r)$ は以下の不等式で評価することができる。

$$G(r) < 1 - \exp(-\pi\lambda r^2)$$

$$G(r) < 1 - \exp\left(-\frac{4}{3}\pi\lambda r^3\right)$$

$$G(r) < 1 - \exp(-\lambda V_u(r))$$

集中 (凝集) 型の点過程では点が局所的に集中しているので、距離が小さいところから $G(r)$ の値は増え始める。この現象を CSR 点過程の変化と比較することにより、集中 (凝集) 型の点過程の $G(r)$ は以下の

不等式で評価することができる。

$$G(r) > 1 - \exp(-\pi\lambda r^2)$$

$$G(r) > 1 - \exp\left(-\frac{4}{3}\pi\lambda r^3\right)$$

$$G(r) > 1 - \exp(-\lambda V_u(r))$$

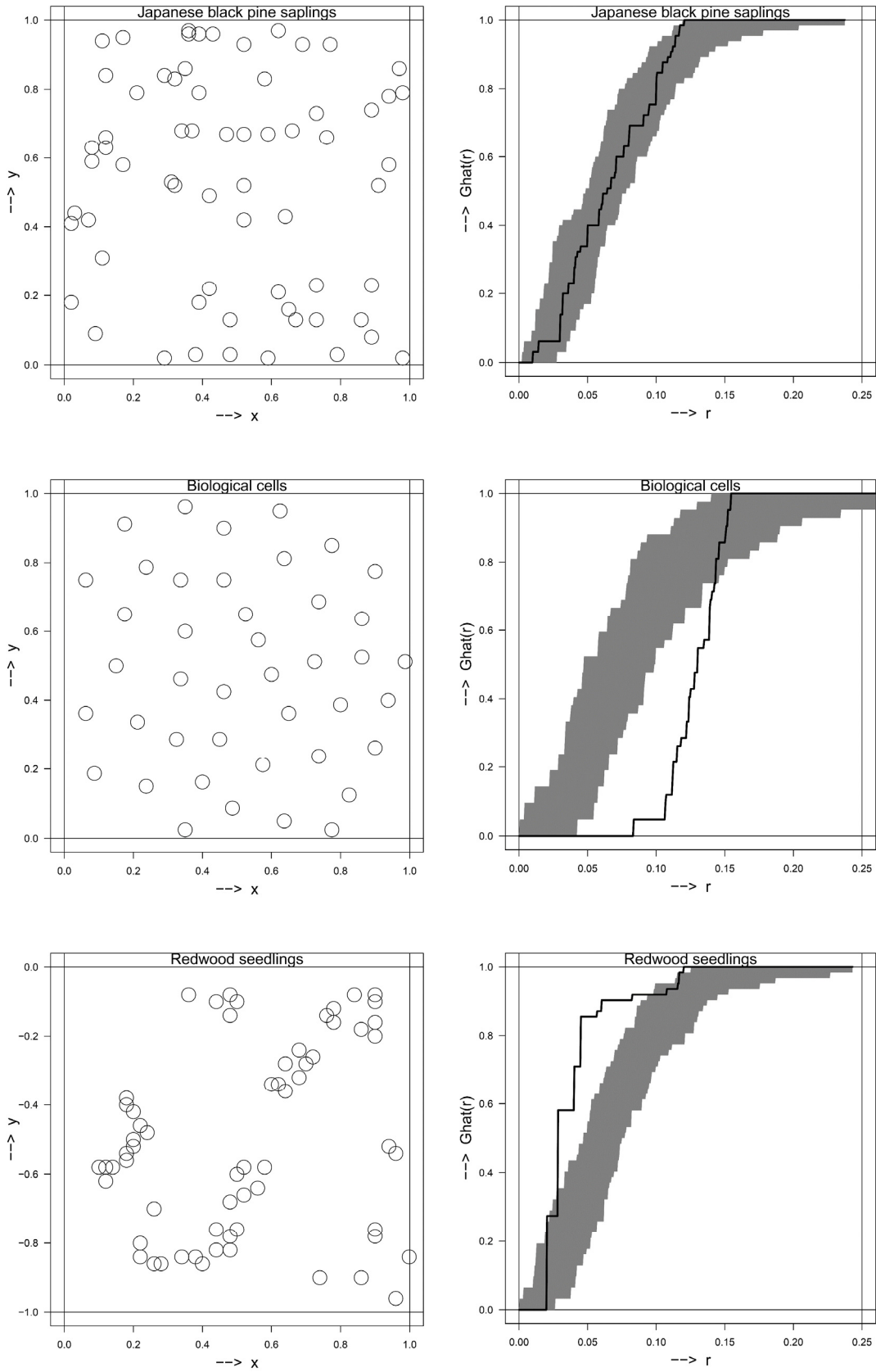
観測データからは点間の第 1 近隣木の辺長（最近隣距離）（nearest neighbor distances） d_i の経験分布関数を用いて $\hat{G}(r)$ を推定することができる（Diggle, 1983, p.16）。

$$\hat{G}(r) = n^{-1}\#(d_i \leq r) = n^{-1} \sum_{d_i \leq r} 1$$

ここで、 n は領域内の点数、 d_i は i 番目の点からその最近隣点までの距離、 $\#$ は個数の意味である。

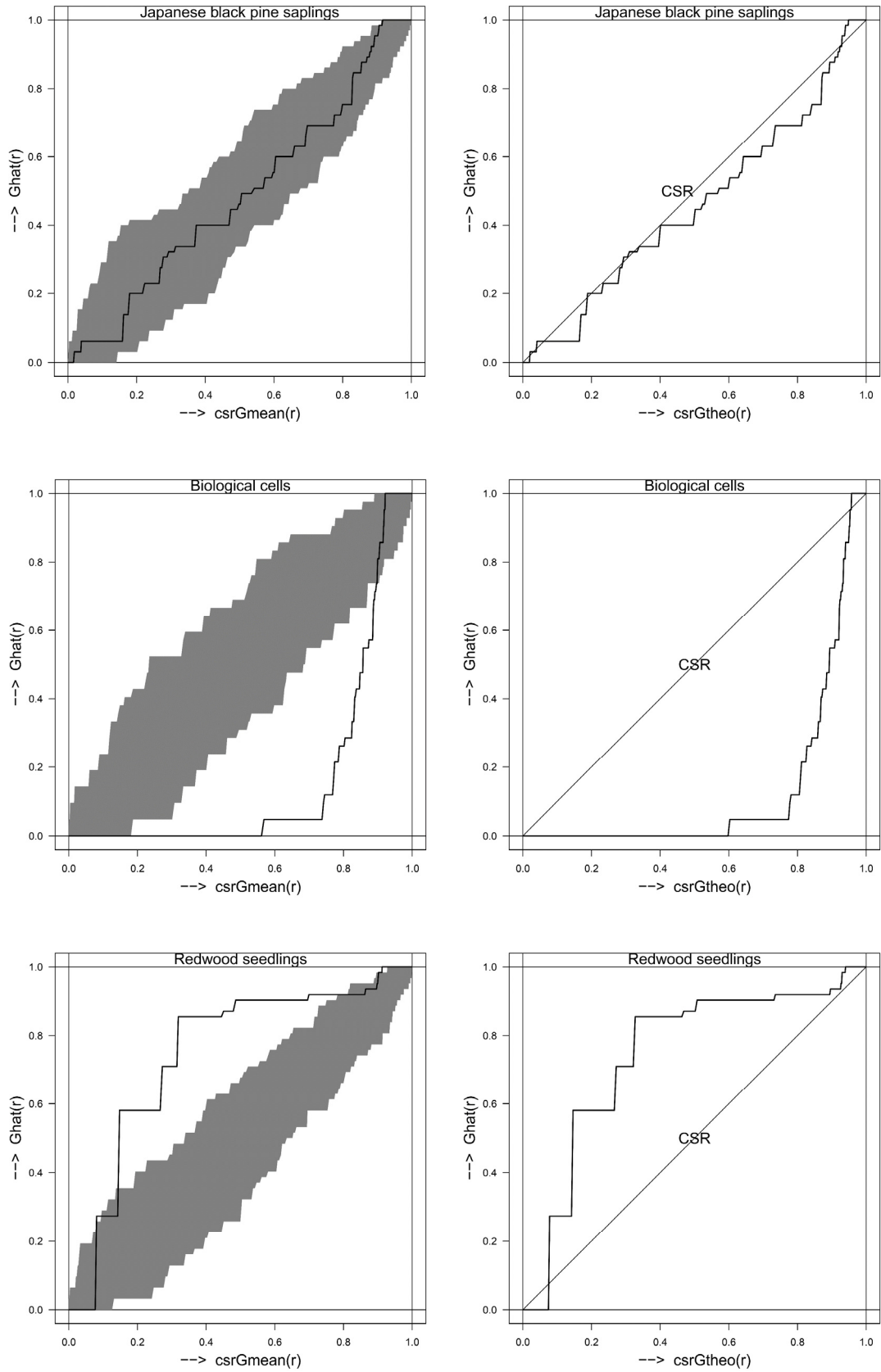
なお等式の最右辺は Diggle (1983, p.16) の表記を変えたものである。

G 関数を用いた空間点パターンの分類事例を図 4.1 に示す。また、G 関数は経験分布関数なので、P-P プロット（P-P plot, Wilk & Gnanadesikan, 1968）形式でも表現することができる。その結果を図 4.2 に示す。図 4.2 の第 1 列はモンテカルロ法で求めた 100 本の曲線群の平均値を CSR 点過程に対する G 関数の推定値として用いた場合である。また第 2 列は CSR 点過程に対する G 関数の理論値を用いた場合である。2次元の場合、G 関数の理論値は $G(r) = 1 - \exp(-\pi\lambda r^2)$ であり、P-P プロット上では原点を通る傾き 1 の直線になる。



注：灰色の帯は CSR 点過程に従う 100 本の曲線群のエンベロップを示す。

図 4.1 G 関数による空間点パターンの分類



注：灰色の帯は CSR 点過程に従う 100 本の曲線群のエンベロップを示す。

図 4.2 G 関数 (P-P プロット) による空間点パターンの分類

4.1.2 F 関数を用いた分類

Diggle (1983, p.49) は F 関数 $F(r)$ を以下のように定義している (原文は r ではなく x である)。

$$F(r) = P\{\text{distance from an arbitrary point to the nearest event is at most } r\}$$

ここで *point* は基準点の位置を示し、*event* は実在する対象物 (例えば森林にある木) の位置を示している。また、Diggle (1983, p.49) は *point* を中心とした半径 r の円内に少なくとも 1 つ *event* があるという解釈を加えている。F 関数には *empty space function* という呼び方もある (Baddeley & Gill, 1997)。

$F(r)$ もまた、CSR 点過程の場合だけ正確に求めることができる。基準点 (*point*) を中心とした半径 r の円または球体の内部に、点 (*event*) が少なくとも 1 つ有る確率から導出する。2 次元、3 次元および多次元 (u 次元) の場合を以下に示す。ここで、 λ は CSR 点過程の強度、 $V_u(r)$ は半径 r の u 次元球の体積である。

$$F(r) = 1 - \exp(-\pi\lambda r^2)$$

$$F(r) = 1 - \exp\left(-\frac{4}{3}\pi\lambda r^3\right)$$

$$F(r) = 1 - \exp(-\lambda V_u(r)) \quad \text{ここで、} V_u(r) = \frac{\pi^{\frac{u}{2}}}{\Gamma(\frac{u}{2}+1)} r^u$$

一方、間瀬&武田 (2001, p.49) は $F(r)$ を球状接触分布関数 (*spherical contact distribution function*) として定義し $H(r)$ と表記している。Illian *et al.* (2008, p.199) もまた同じ定義をして、 $F(r)$ を $H_s(r)$ と表記している。添え字の s は *spherical* の s と考えられる。球状接触分布関数とは、基準点を中心とした半径 r の球を考え、半径 r を増やしていったときに、その球が最初に点 (*event*) に接触した r の値の分布を意味している。理由は述べられていないが、Chiu *et al.* (2013, p.144) は $F(r)$ をパッとしない記号 (*colorless symbol*) と呼んでいる。確率・統計の分野で、 $F(r)$ の F が複数の意味に使われているからだと考えられる。

規則型の点過程では、距離が小さいところで $F(r)$ の値が大きくなるので、これを CSR 点過程の変化と比較することにより、規則型の点過程の $F(r)$ は以下の不等式で評価することができる。

$$F(r) > 1 - \exp(-\pi\lambda r^2)$$

$$F(r) > 1 - \exp\left(-\frac{4}{3}\pi\lambda r^3\right)$$

$$F(r) > 1 - \exp(-\lambda V_u(r))$$

集中（凝集）型の点過程では、距離が大きいところで $F(r)$ の値が大きくなるので、これを CSR 点過程の変化と比較することにより、集中（凝集）型の点過程の $F(r)$ は以下の不等式で評価することができる。

$$F(r) < 1 - \exp(-\pi\lambda r^2)$$

$$F(r) < 1 - \exp(-\frac{4}{3}\pi\lambda r^3)$$

$$F(r) < 1 - \exp(-\lambda V_u(r))$$

つまり、CSR 点過程の場合は $F(r)$ と $G(r)$ は同じだが、規則型と集中（凝集）型では $F(r)$ と $G(r)$ は逆の振舞いをする。

観測データからは基準点（point）と実在の点（event）の間の最近隣距離 e_i の経験分布関数を用いて $\hat{F}(r)$ を推定することができる（Diggle, 1983, p.20）。

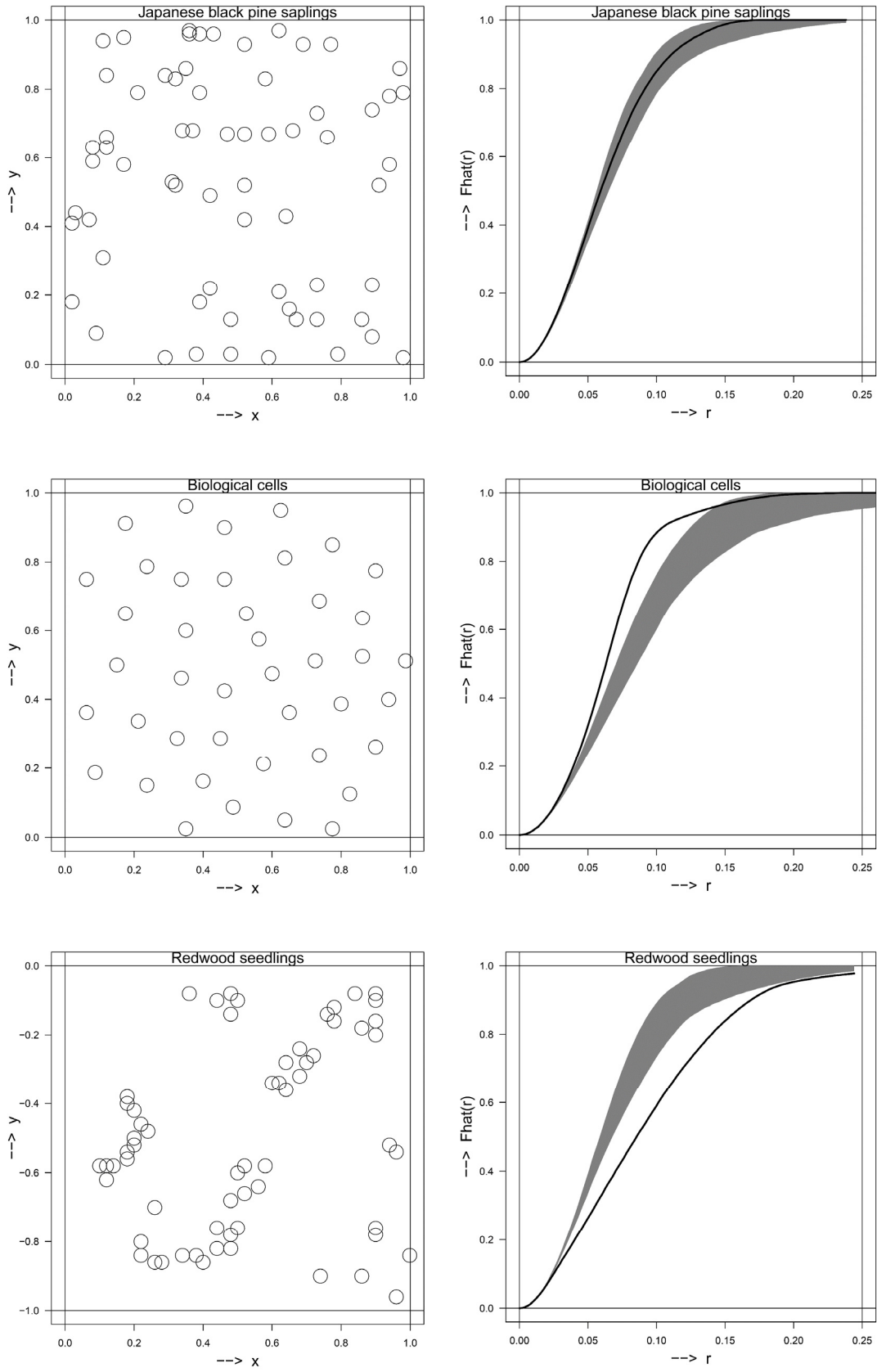
$$\hat{F}(r) = m^{-1}\#(e_i \leq r) = m^{-1} \sum_{e_i \leq r} 1$$

ここで、 m は基準点の個数、 e_i は i 番目の基準点からその最近隣点までの距離、 $\#$ は個数の意味である。この基準点を Kaluzny *et al.* (1996, p.155) は i 番目の原点（the i^{th} origin）と呼んでいる。

なお等式の最右辺は Diggle (1983, p.20) の表記を変えたものである。この式の意味は、 u 次元空間の領域内に k^u のグリッドを重ね（2次元なら $k \times k$ ）、基準点を $m (= k^u)$ 個つくり、その基準点から一番近い点（event）までの距離の分布を考えるものである。

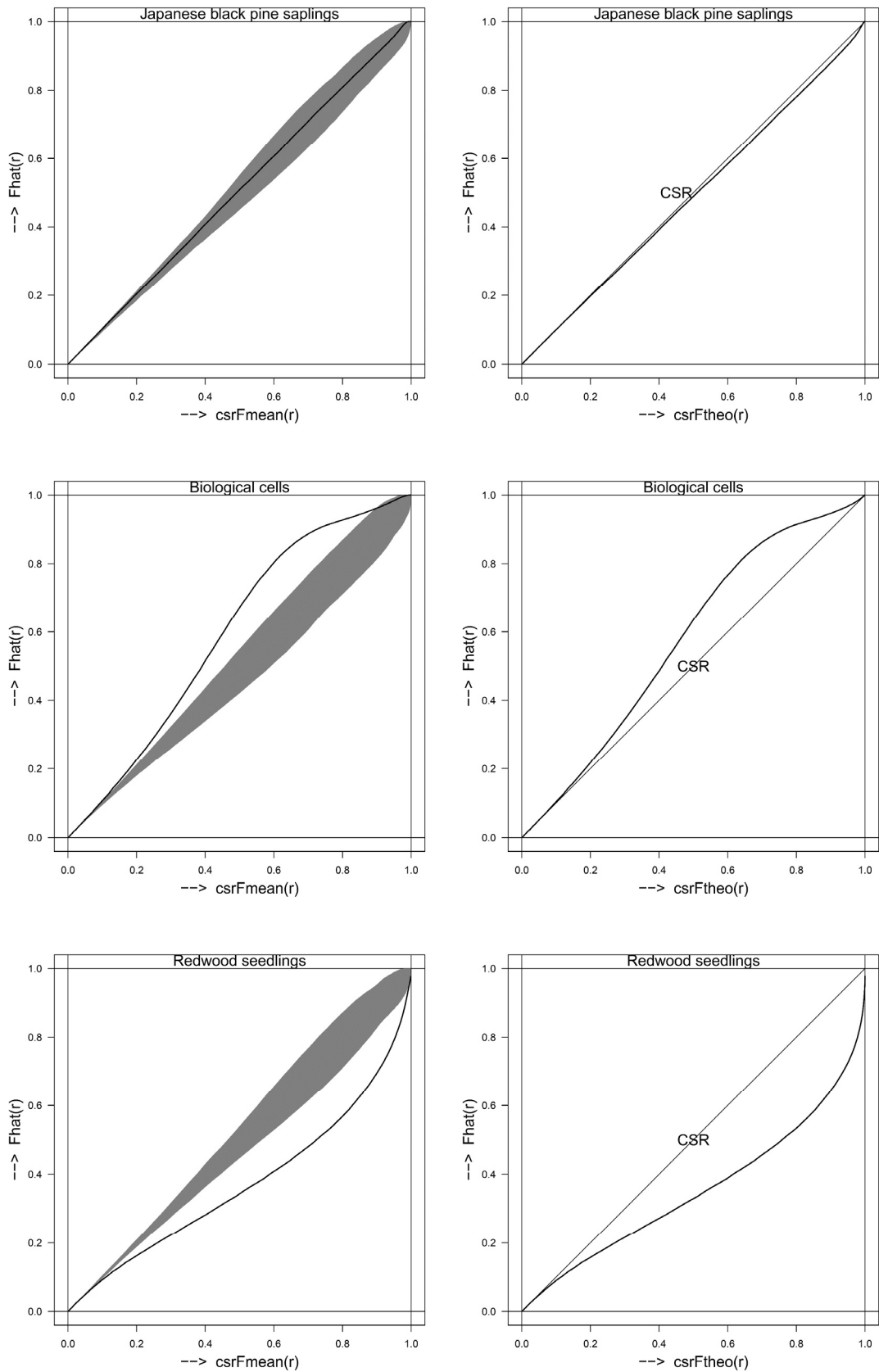
Diggle and Matérn (1981) は、2次元の場合として $k \approx \sqrt{n}$ を推奨している。 k を増やすほど、曲線 $\hat{F}(r)$ は滑らかになっていく。しかし、統計的精度（statistical precision）は、その場合でも依然として、 n （領域内の event 数）の制限を受けることに注意しなければならない（Diggle, 2003, p.21）。なお、多次元（ u 次元）の場合は、 $k \approx n^{1/u}$ と考えられる。

F 関数を用いた空間点パターンの分類事例を図 4.3 に示す。また、F 関数は経験分布関数なので、G 関数と同様に P-P プロット（P-P plot, Wilk & Gnanadesikan, 1968）形式でも表現することができる。その結果を図 4.4 に示す。図 4.4 の第 1 列はモンテカルロ法で求めた 100 本の曲線群の平均値を CSR 点過程に対する F 関数の推定値として用いた場合である。また第 2 列は CSR 点過程に対する F 関数の理論値を用いた場合である。2次元の場合、F 関数の理論値は $F(r) = 1 - \exp(-\pi\lambda r^2)$ であり、P-P プロット上では原点を通る傾き 1 の直線になる。



注：灰色の帯は CSR 点過程に従う 100 本の曲線群のエンベロップを示す。

図 4.3 F 関数による空間点パターンの分類



注：灰色の帯は CSR 点過程に従う 100 本の曲線群のエンベロープを示す。

図 4.4 F 関数 (P-P プロット) による空間点パターンの分類

4.1.3 J 関数を用いた分類

Van Lieshout and Baddeley (1966) は J 関数 $J(r)$ を以下のように定義している。 $G(r)$ と $F(r)$ は既に説明済みのものである。

$$J(r) = \frac{1 - G(r)}{1 - F(r)}, \quad r \geq 0, \quad F(r) < 1$$

Van Lieshout and Baddeley (1966) は J 関数を用いた分類基準を以下のように示しているが、

$$\text{CSR 点過程の場合は } J(r) = 1$$

$$\text{規則型の点過程の場合は } J(r) > 1$$

$$\text{集中（凝集）型の点過程の場合は } J(r) < 1$$

Bedford and Berg (1997) は規則型や集中（凝集）型の点過程に対しても $J(r) = 1$ となる場合があることを示し、分類基準を以下のように修正している。

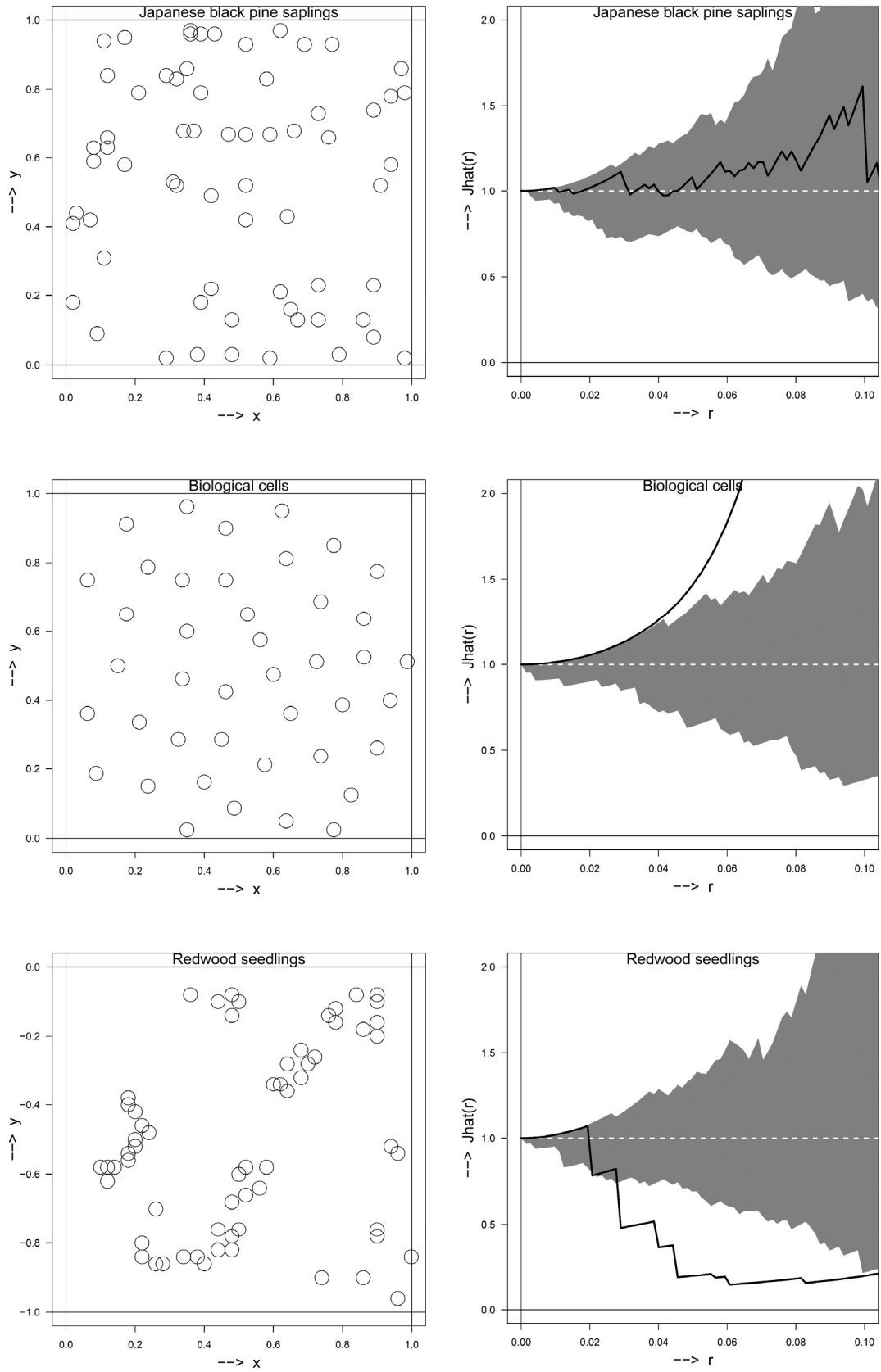
$$\text{CSR 点過程の場合は } J(r) = 1$$

$$\text{規則型の点過程の場合は } J(r) \geq 1$$

$$\text{集中（凝集）型の点過程の場合は } J(r) \leq 1$$

これらの分類基準は、4.1.1 および 4.1.2 の記述内容より、CSR 点過程に対しては $G(r) = F(r)$ 、規則型の点過程に対しては $G(r) \leq F(r)$ 、集中（凝集）型の点過程に対しては $G(r) \geq F(r)$ になることから導出されることである。

J 関数を用いた空間点パターンの分類事例を図 4.5 に示す。



注：灰色の帯は CSR 点過程に従う 100 本の曲線群のエンベロープを示す。

図 4.5 J 関数による空間点パターンの分類

4.2 2次特性 (Second-order characteristics) に基づく方法

4.2.1 K関数を用いた分類

Ripley (1977) は、 λ を強度とした時に、K関数 $K(r)$ を以下のように定義している (原文は r ではなく t である)。

$\lambda K(r)$ is the expected number of further points within r of an arbitrary point of the process.

つまり、 $\lambda K(r)$ は (領域内の) 任意に選んだ点を中心とする半径 r の円 (または u 次元球) に含まれる点の個数の期待値である。この文章の”further”は、中心になった点を個数としてカウントしないということを意味している。

$\lambda K(r)$ もまた CSR 点過程に対してだけ正確に計算することができる。2次元、3次元および多次元 (u 次元) の場合を以下に示す。ここで、 λ は CSR 点過程の強度、 $V_u(r)$ は半径 r の u 次元球の体積である。

$$\lambda K(r) = \lambda \pi r^2$$

$$\lambda K(r) = \frac{4}{3} \lambda \pi r^3$$

$$\lambda K(r) = \lambda V_u(r) \quad \text{ここで、} V_u(r) = \frac{\pi^{\frac{u}{2}}}{\Gamma(\frac{u}{2}+1)} r^u$$

これは CSR 点過程の場合なので、半径 r を増やしていけば、その領域に含まれる点の個数は円の面積 (または u 次元球の体積) に比例して増えていくことを示している。

CSR 点過程に対する K関数は、以下のように更に簡単に表現することができる。

$$K(r) = \pi r^2$$

$$K(r) = \frac{4}{3} \pi r^3$$

$$K(r) = V_u(r)$$

一方、規則型の点過程では点が孤立しているので、半径 r が小さい範囲では点は存在せず、点間距離の最小値 (hard core distance; Illian *et al.*, p.58) より大きくなって初めて点の数が増え始める。この現象を CSR 点過程の変化と比較することにより、規則型の点過程の $K(r)$ は、以下の不等式で評価することができる。

$$K(r) < \pi r^2$$

$$K(r) < \frac{4}{3}\pi r^3$$

$$K(r) < V_u(r)$$

また、集中（凝集）型の点過程では、局所的に点が集中しているため、半径 r が小さい範囲で点の個数が増えることになる。この現象を CSR 点過程の変化と比較することにより、集中（凝集）型の点過程の $K(r)$ は、以下の不等式で評価することができる。

$$K(r) > \pi r^2$$

$$K(r) > \frac{4}{3}\pi r^3$$

$$K(r) > V_u(r)$$

規則型の場合も集中（凝集）型の場合も、CSR 点過程との違いは、半径 r が小さい範囲で顕著になる。

K 関数もまた、近接グラフを用いて表現することができる。Illian *et al.* (2008, p.216) は、近接グラフの1つである第 k 近隣木と $\lambda K(r)$ の関係を以下のように示している。第 k 近隣木は、領域内の各点と、その点から k 番目に近い点（第 k 近隣点）を結ぶ辺で構成されるもので（3.2節の再掲）、この式の右辺もまた半径 r の円または u 次元球に含まれる点の個数を示している。

$$\lambda K(r) = \sum_{k=1}^{\infty} D_k(r)$$

ここで $D_k(r)$ は第 k 近隣木の辺長の分布関数である。

観測データからは以下の式を用いて $\hat{K}(r)$ を推定することができる（Ripley, 1976; Dixon, 2002）。

$$\hat{K}(r) = \hat{\lambda}^{-1} \sum_i \sum_{j \neq i} w(l_i, l_j)^{-1} \frac{I(d_{ij} < r)}{N}$$

ここで、

N は領域内で観測された点数、

$\hat{\lambda}$ は N を領域の体積（または面積）で除したもので強度の推定値、

$d_{i,j}$ は i 番目の点 l_i と j 番目の点 l_j の距離、

$w(l_i, l_j)$ は境界補正（edge correction）に用いる加重関数（weight function）

$I(d_{ij} < r)$ は指標関数（indicator function）であり、 $d_{i,j} < r$ の場合に1、そうでない場合は0となる。

境界補正の加重関数 $w(l_i, l_j)$ については、まず i 番目の点 l_i を中心とした半径 $d_{i,j}$ の円（または u 次元球）を考える。このとき、 $w(l_i, l_j)$ は領域内に入る部分の面積（または体積）の、全体の面積（または体積）に対する比率である。詳しくは次節で取り扱う。

4.2.1.1 境界効果（edge effect）と境界補正（edge correction）

境界の存在しない領域で理論が構築された場合であっても、現実のデータは有限の領域でしか観測できない。したがって分野を問わずデータには境界領域が存在し、データからの推論には何らかの対策が考えられている。例えば信号処理や時系列解析に於いては、有限のデータからスペクトルを推定する場合に、両端の影響を軽減するためにテーパリング（Tapering）を行う（Bloomfield, 2000, pp.66-76）。

K 関数の境界効果を図 4.6 を用いて説明する。左が 2 次元の、右が 3 次元の説明図である。3 次元の場合は、点群と矢印（半径）の表示を省略している。

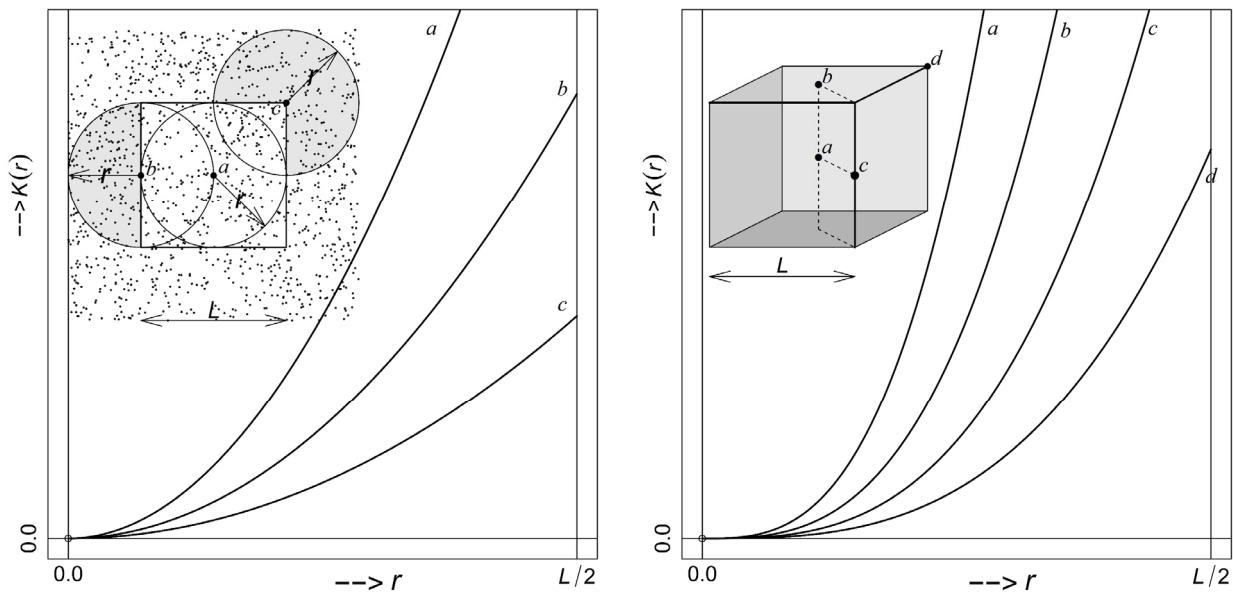


図 4.6 境界効果の説明図（左：2次元、右：3次元）

最初に CSR 点過程（強度 λ ）に対する観測領域が、一辺 L の正方形である場合を考える（図 4.6 左）。説明を単純にするために、円の中心になる点が正方形の以下に示す 3 か所にあると仮定する。

a : (正方形の) 中心, b : 辺の中点, c : 頂点

各点のまわりで半径 r を増やして行った場合に、各円の内部に含まれる点の個数 $\lambda K(r)$ は、点 a の場合は $\lambda\pi r^2$ だが、 b は a の $1/2$ 、 c は a の $1/4$ になる。したがって、領域内で期待値を求めると、 $\hat{K}(r)$ 曲線は下方に偏りをもつことになる。この偏りを補正するのが境界補正である。境界補正では、観測領域の外側も、内側と同じように点が分布していることを前提とする。単純に考えれば、点 b の場合は観測個数を 2 倍し、点 c の場合は観測個数を 4 倍すれば良いことになる。理想的に行われた境界補正の結果を図 4.7 に示す。また具体的な補正方式は、この節の最後に述べる。

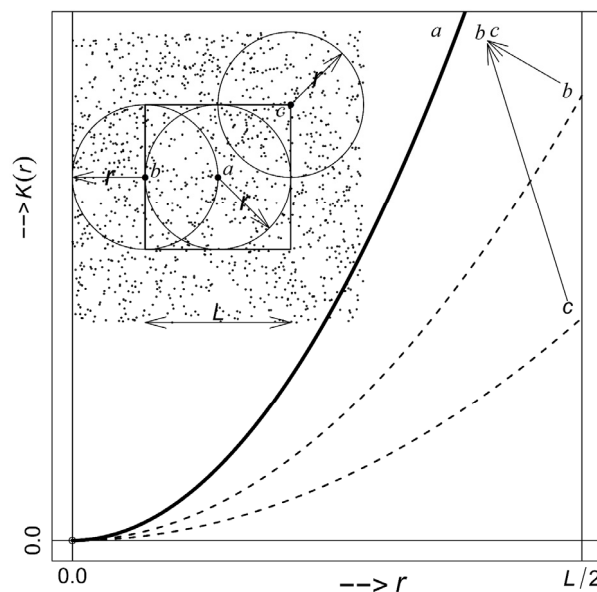


図 4.7 境界補正後 (2 次元)

次に CSR 点過程 (強度 λ) に対する観測領域が、一辺 L の立方体である場合を考える (図 4.6 右)。説明を単純にするために、球の中心になる点が立方体の以下に示す 4 か所にあると仮定する。

a : (立方体の) 中心, b : 面の中心, c : 辺の中点, d : 頂点

各点のまわりで半径 r を増やして行った場合に、各球の内部に含まれる点の個数 $\lambda K(r)$ は、点 a の場合は $\frac{4}{3}\lambda\pi r^3$ だが、 b は a の $1/2$ 、 c は a の $1/4$ 、 d は a の $1/8$ になる。したがって、領域内で期待値を求めると、 $\hat{K}(r)$ 曲線は下方に偏りをもつことになる。

正方形 (2 次元) の事例から、境界効果は領域内の点数が少ない場合と領域の形状が細長い多角形の場合に顕著になることが類推できる。また、正方形 (2 次元) と立方体 (3 次元) の比較から、高次元にな

るほど境界効果の影響は深刻になることが分かる。なお、円または球の中心となる点は領域内の任意の位置にあり、また境界自体も一般的には曲線や曲面である。したがって、実際の境界効果と境界補正を考えるには、より複雑な幾何学を用いなければならない。

K 関数とともに Ripley (1976) が提案した境界補正については、Goreaud & Pelissier (1999) が具体的な解説を行っている。この記述は円の中心となる点が、領域内の何処にあるかによって場合分けを行っている。また境界の形状が曲線の場合を含んでいる。ここでは境界の形状が矩形の場合だけを取り上げ、図 4.8 を用いて説明する。この図には 4 通りの場合分けが示されている。繰り返しになるが、領域の外側の点の分布は、内側と同じであることが前提となっている。

重み関数 $w(l_i, l_j)^{-1}$ は、半径 r の円を使って以下のように表現できる。

$$w(l_i, l_j)^{-1} = (\text{円の全周の長さ}) / (\text{領域内にある円弧の長さの総和})$$

図 4.8 の右上の例では、

$$w(l_i, l_j)^{-1} = \frac{2\pi}{2\pi - \theta}$$

となり、図 4.8 の左下の例では、

$$w(l_i, l_j)^{-1} = \frac{2\pi}{2\pi - (\theta_1 + \theta_2)}$$

となる。

また、図 4.8 の左上の例では、

$$w(l_i, l_j)^{-1} = \frac{2\pi}{2\pi - \frac{3\pi}{2}} = 4$$

となり、図 4.8 の右下の例では、

$$w(l_i, l_j)^{-1} = \frac{2\pi}{2\pi - \pi} = 2$$

となる。

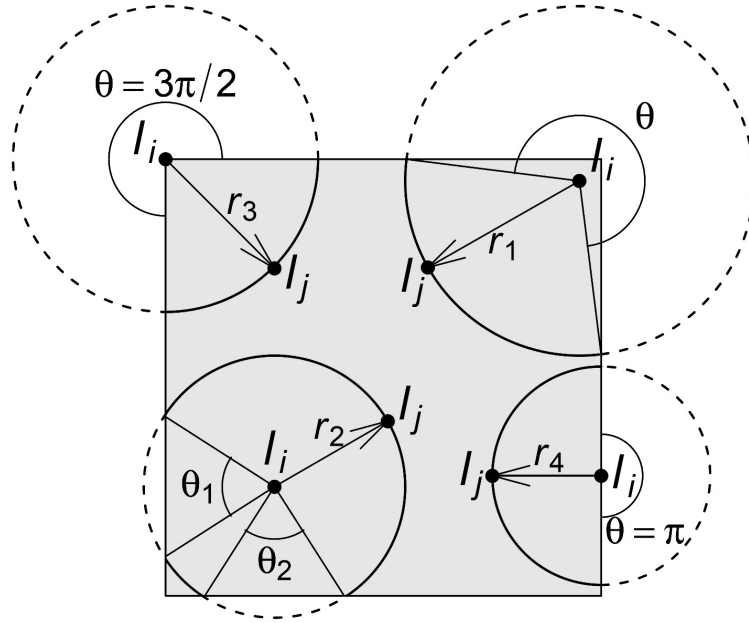


図 4.8 境界補正の説明図 (A)

具体的な境界補正の例を図 4.9 を用いて説明する。円の中心となる点 l_i が観測領域の左下にあり、点 l_i から距離 r_1 離れた位置と距離 r_2 離れた位置に、点が1つずつ観測されている。

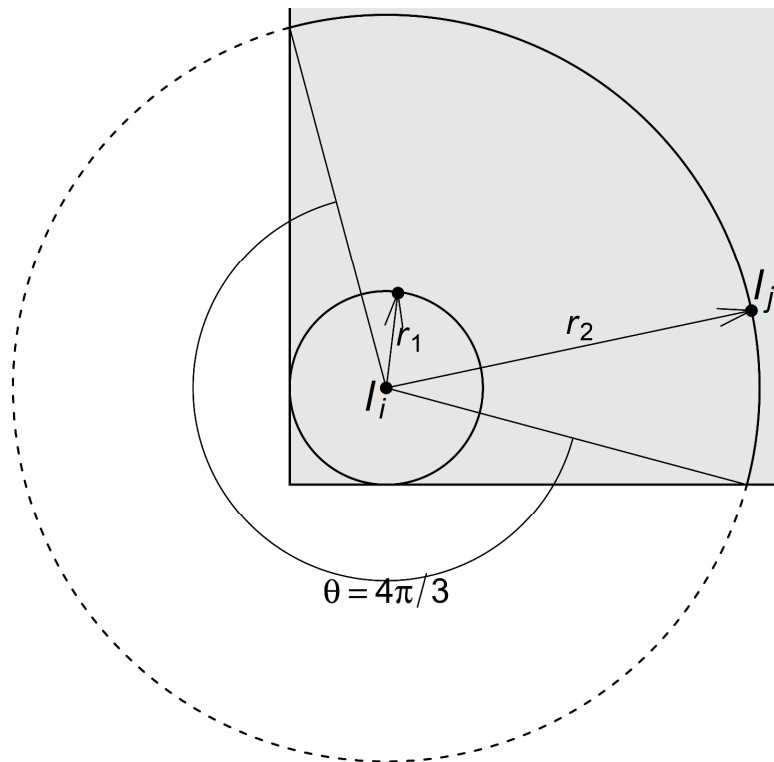


図 4.9 境界補正の説明図 (B)

まず、距離 r_1 離れた点を考える。半径 r_1 の円の全周は観測領域に含まれているので、この場合は境界補正を行う必要はない。次に、距離 r_2 離れた点を考える。半径 r_2 の円は観測領域の外側にはみ出しているので、この場合は境界補正を行う必要がある。重み関数の計算結果は、

$$w(l_i, l_j)^{-1} = \frac{2\pi}{2\pi - \frac{4\pi}{3}} = 3$$

となるので、境界補正により半径 r_2 の円周上には全部で点が3つあることになる。つまり観測領域の外側には点が2つあると推定している。

最後に、間瀬&武田 (2001, pp.48-49) が、最近接距離分布関数 (nearest neighbor distance distribution function) の説明の中で述べている境界効果の対策に関して以下に示す。なお、本研究では最近接距離分布関数を G 関数と呼んでいる。

観測領域外の点が及ぼす影響を境界効果 (edge effect) と称し、空間統計学固有の困難をもたらす。境界効果を除く妙案は当然なく、つぎのようなその場しのぎの処方箋が用いられてきた。

- 単純に無視する。
- 周期的境界条件を用いる (有界観測領域 W をトーラス化する)。
- より小さな領域 W' を考え、 $W \setminus W'$ 中のデータは境界効果対策にだけ使う。
- たとえば、点過程 Φ が定常ポアソン配置であるという仮説下では、推定強度をもとに W 外の点パターンを人工的に発生し代用する。
- 境界近くの点の寄与を割り引くなどの境界補正 (edge correction) を施した統計量を工夫する。

4.2.1.2 K 関数を用いた分類結果と表示の改良

Ripley (1976)の境界補正を行った K 関数の分類結果を図 4.10 に示す。

一方、Diggle (1983, pp. 78-79) は K 関数の表示を以下のように改良している。ここでは、これを K^* 関数 $K^*(r)$ と呼ぶことにする。2次元、3次元および多次元 (u 次元) の場合を以下に示す。ここで、 $V_u(r)$ は半径 r の u 次元球の体積である。

$$K^*(r) = K(r) - \pi r^2$$

$$K^*(r) = K(r) - \frac{4}{3}\pi r^3$$

$$K^*(r) = K(r) - V_u(r) \quad \text{ここで、} V_u(r) = \frac{\pi^{\frac{u}{2}}}{\Gamma(\frac{u}{2}+1)} r^u$$

CSR 点過程に対する K 関数の式より、 K^* 関数 $K^*(r)$ は以下のように簡単に表現することができる。これは次元に依存しない。

$$K^*(r) = 0$$

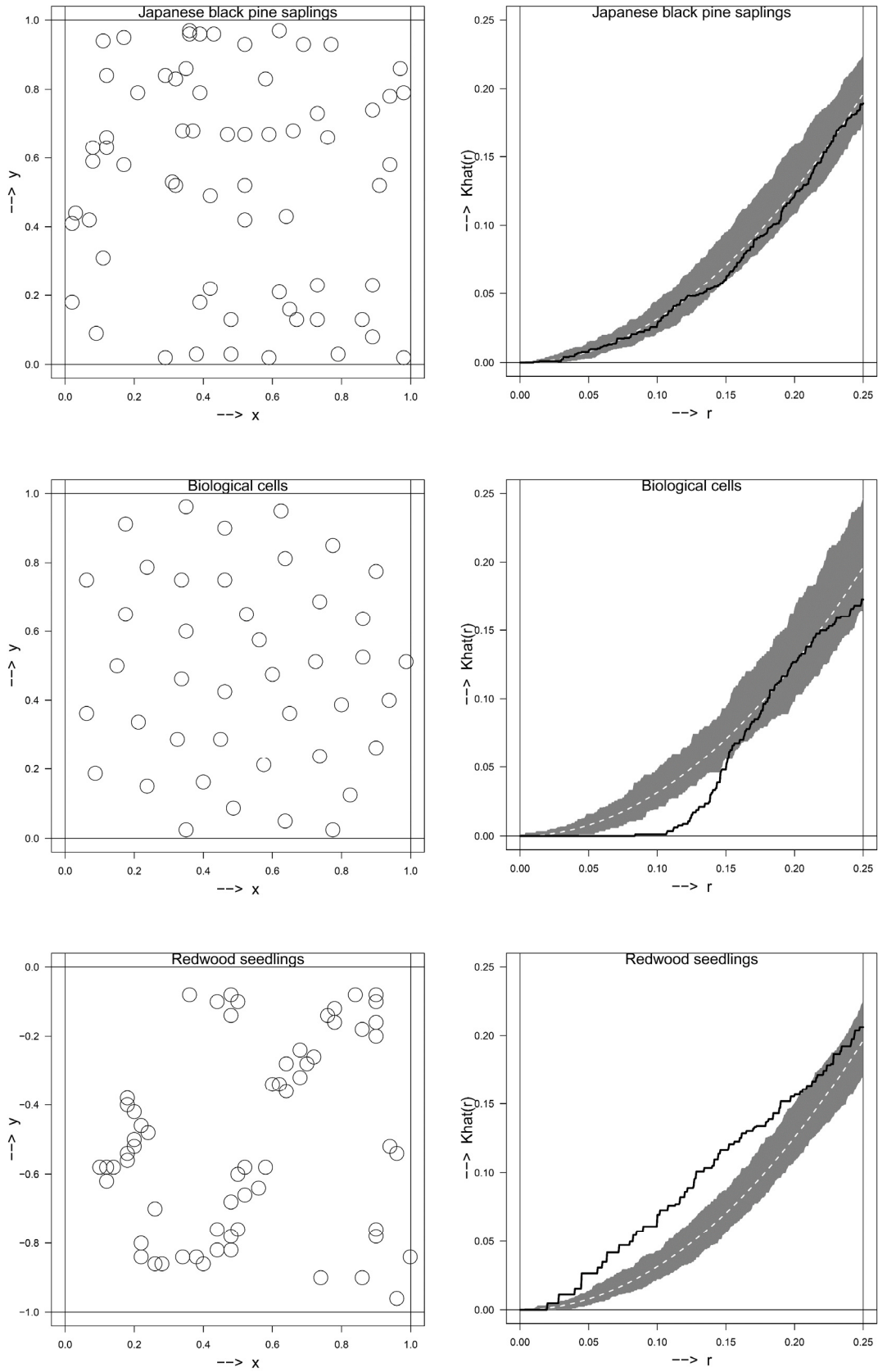
また規則型の点過程に対する K^* 関数 $K^*(r)$ は以下の不等式で評価することができる。

$$K^*(r) < 0$$

さらに集中 (凝集) 型の点過程に対する K^* 関数 $K^*(r)$ は以下の不等式で評価することができる。

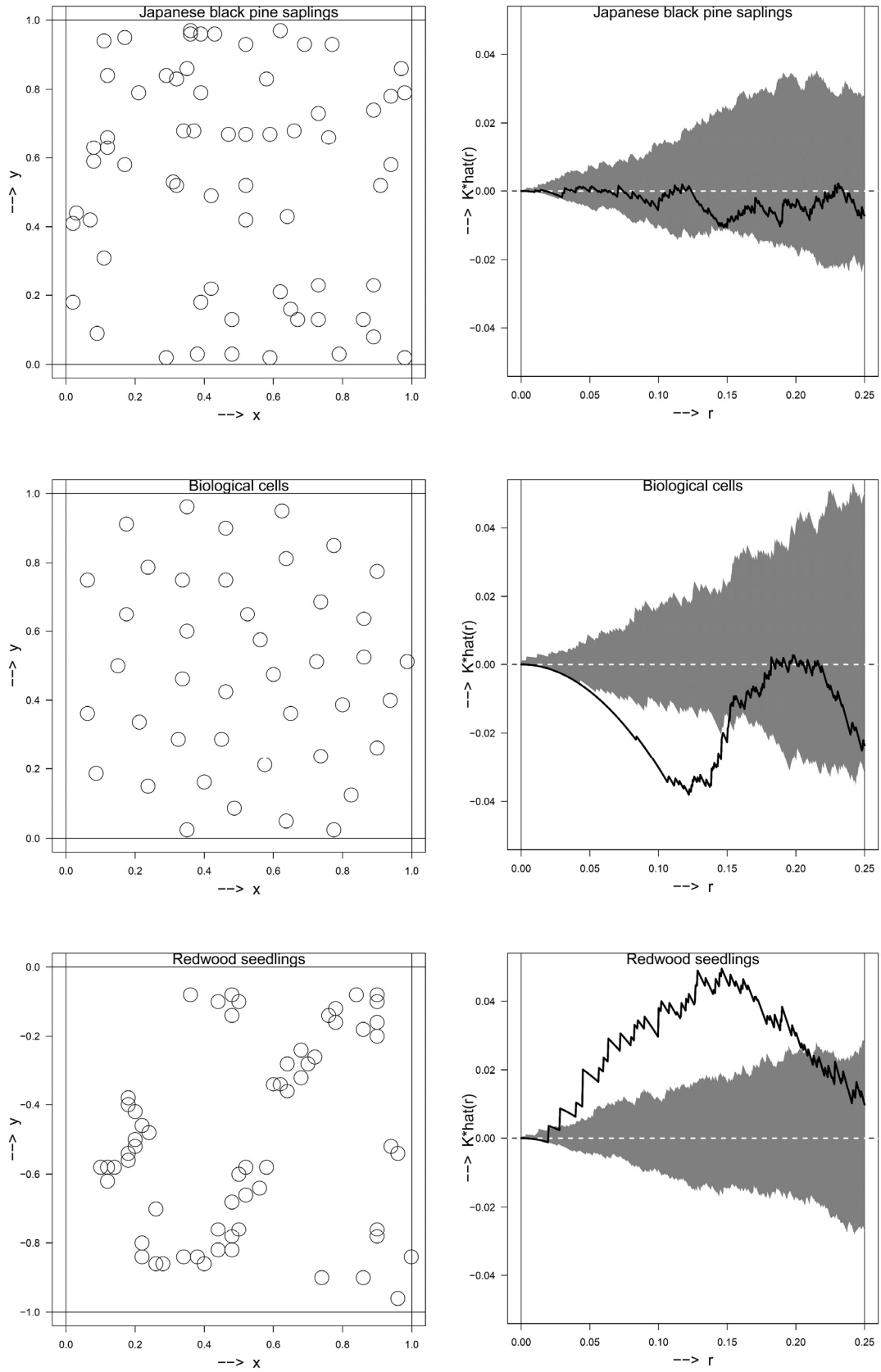
$$K^*(r) > 0$$

K^* 関数 $K^*(r)$ を用いた分類結果を図 4.11 に示す。 r が大きくなるにつれて分散が拡大するのは K 関数と変わらない。これを改良するのが、次節で取り上げる L 関数である。



注：灰色の帯は CSR 点過程に従う 100 本の曲線群のエンベロップを示す。

図 4.10 K 関数による空間点パターンの分類



注：灰色の帯は CSR 点過程に従う 100 本の曲線群のエンベロップを示す。

図 4.11 K^* 関数による空間点パターンの分類

4.2.2 L 関数を用いた分類

L 関数 $L(r)$ は、Besag (1977) により提案された。Besag (1977) は Ripley (1977) の 192 ページ以降にある DISCUSSION ON DR. RIPLEY'S PAPER の内容である。L 関数には規格化の違いにより 2 通りの方式があるので、ここでは L 関数と L^* 関数として区別する。

4.2.2.1 L 関数を用いた分類結果

L 関数 $L(r)$ の 2 次元、3 次元および多次元 (u 次元) の場合の定義を以下に示す。ここで、 $V_u(r)$ は半径 r の u 次元球の体積である。

$$L(r) = \sqrt{\frac{K(r)}{\pi}}$$

$$L(r) = \left(\frac{K(r)}{4\pi/3}\right)^{\frac{1}{3}}$$

$$L(r) = \left(\frac{K(r)}{V_u(r)/r^u}\right)^{\frac{1}{u}} \quad \text{ここで、} V_u(r) = \frac{\pi^{\frac{u}{2}}}{\Gamma(\frac{u}{2}+1)} r^u$$

$L(r)$ は $K(r)$ と同じ内容を意味しているが、以下のように $K(r)$ が曲線になるのに対して、

$$K(r) \propto r^2$$

$$K(r) \propto r^3$$

$$K(r) \propto r^u$$

$L(r)$ は規格化されているので次元によらず常に直線になる

$$L(r) \propto r$$

とりわけ CSR 点過程に対しては傾き 1 の直線になる。

$$L(r) = r$$

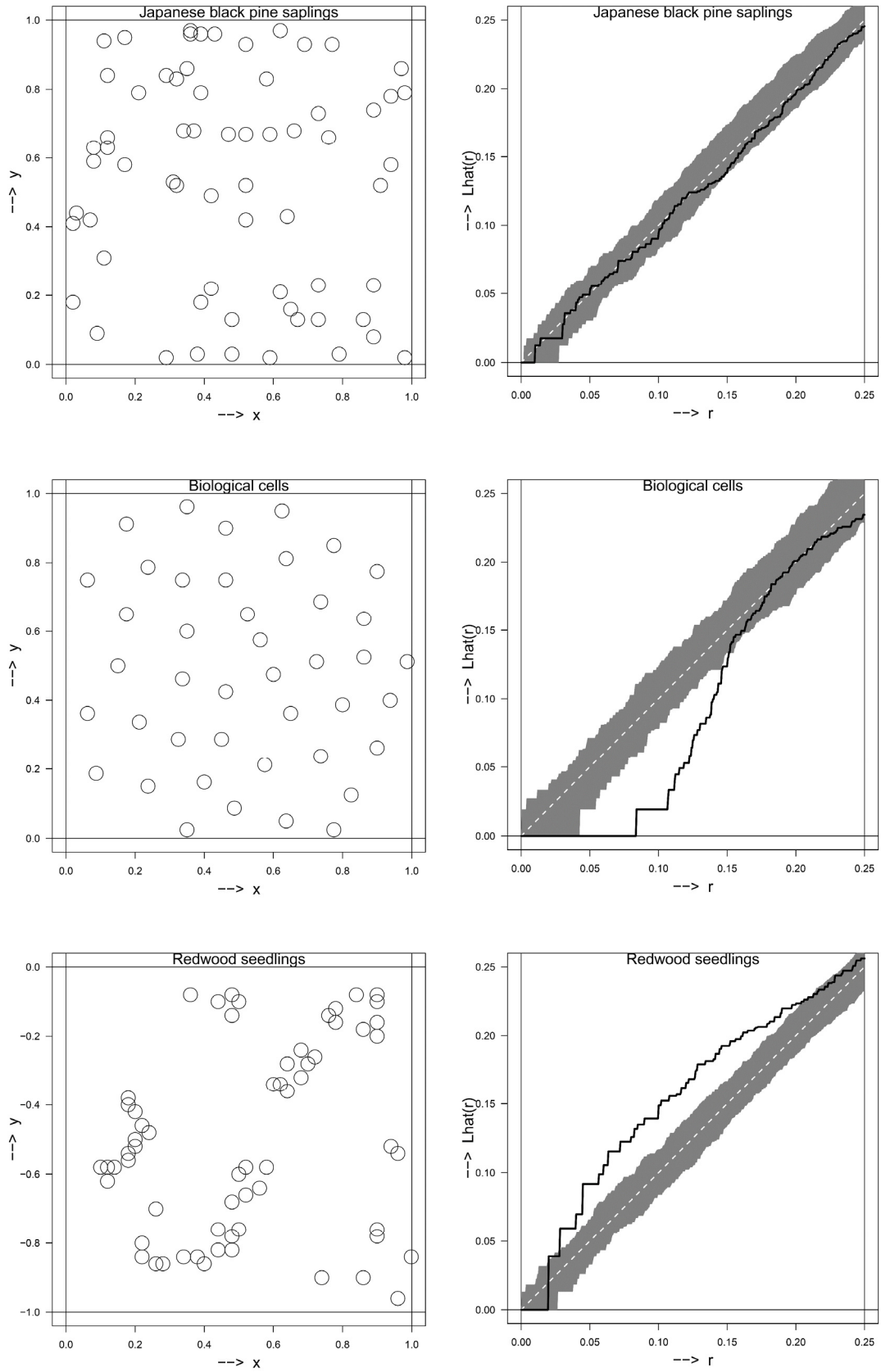
また $K(r)$ の説明を基に考えれば、規則型の点過程の $L(r)$ は以下の不等式で評価することができ、

$$L(r) < r$$

集中（凝集）型の点過程の $L(r)$ は以下の不等式で評価することができる。

$$L(r) > r$$

観測データからは $\hat{K}(r)$ を用いて $\hat{L}(r)$ を推定することができる。分類結果を図 4.12 に示す。



注：灰色の帯は CSR 点過程に従う 100 本の曲線群のエンベロップを示す。

図 4.12 L 関数による空間点パターンの分類

4.2.2.2 L*関数を用いた分類結果

もう一方のL関数である $L^*(r)$ は以下のように定義される。

$$L^*(r) = \sqrt{\frac{K(r)}{\pi}} - r = L(r) - r$$
$$L^*(r) = \left(\frac{K(r)}{V_u(r)/r^u} \right)^{\frac{1}{u}} - r = L(r) - r$$

CSR点過程の場合、 $L(r) = r$ なので、 $L^*(r)$ は次元によらずゼロになる。

$$L^*(r) = 0$$

$L^*(r)$ をグラフに表せば水平線になり $L(r)$ よりも更に見やすくなる。

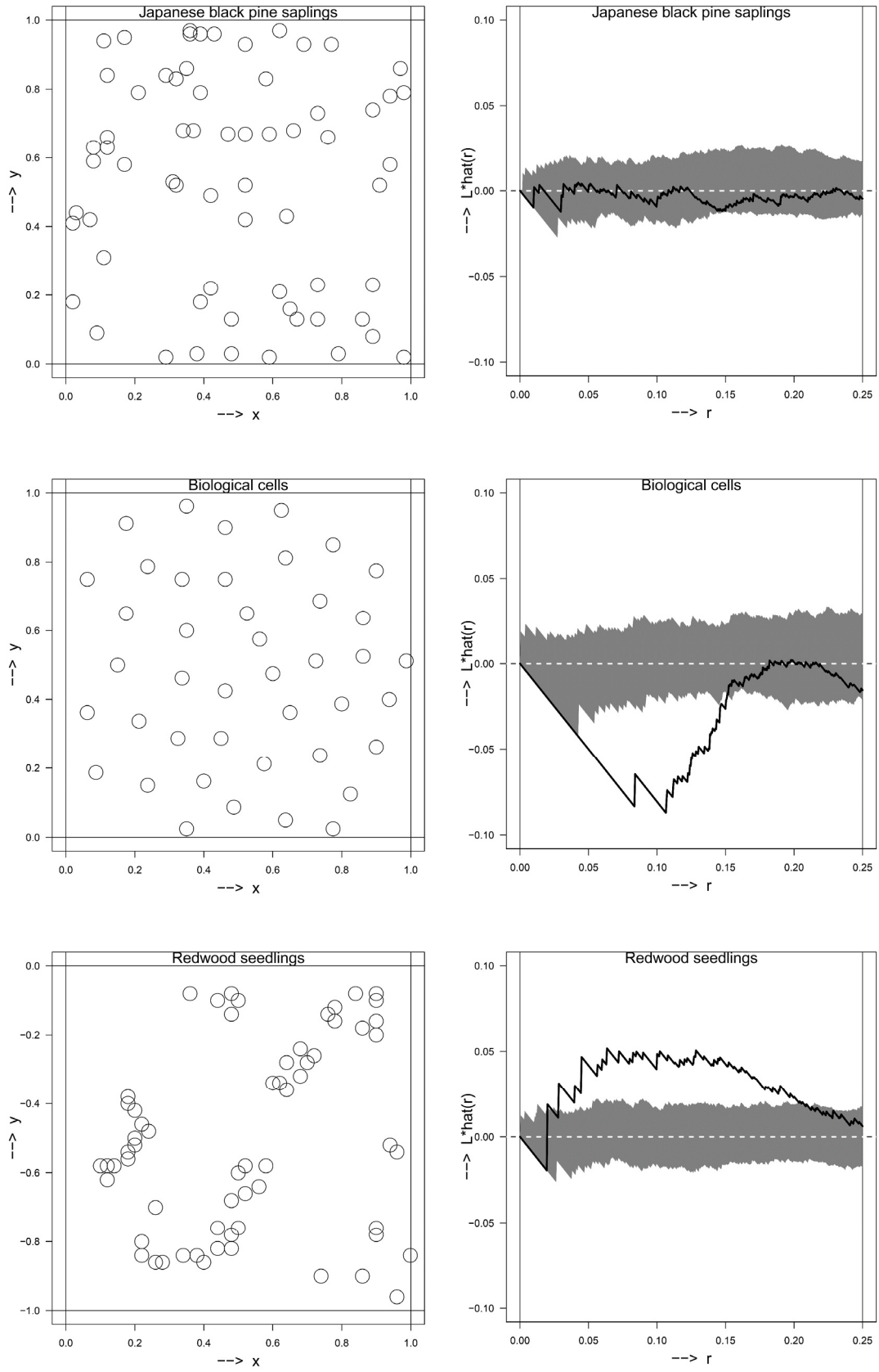
また、これまでの説明を基に考えれば、規則型の点過程の $L^*(r)$ は以下の不等式で評価することができ、

$$L^*(r) < 0$$

集中（凝集）型の点過程の $L^*(r)$ は以下の不等式で評価することができる。

$$L^*(r) > 0$$

観測データからは $\hat{K}(r)$ を用いて $\hat{L}^*(r)$ を推定することができる。分類結果を図 4.13 に示す。



注：灰色の帯は CSR 点過程に従う 100 本の曲線群のエンベロップを示す。

図 4.13 L*関数による空間点パターンの分類

4.2.3 pair correlation 関数を用いた分類

Wiegand & Moloney (2004) は、O-ring 統計量 (O-ring statistic) $O(r)$ から pair correlation 関数 $g(r)$ を導出している。Wiegand & Moloney (2004) の解説に従い、2 変量を 1 変量に変えた $O(r)$ の表現を以下に示す。#は個数の意味である。

$$O(r) = \lambda g(r) = E[\#(\text{points at distance } r \text{ from an arbitrary point})]$$

ここで、 $\lambda K(r)$ が半径 r の円盤内または半径 r の球体内にある点の個数の期待値であったのに対して、 $\lambda g(r)$ は微小幅をもつ半径 r の輪帯内、または微小厚みをもつ半径 r の球殻内にある点の個数の期待値である。

K 関数と pair correlation 関数の比較を図 4.14 に示す。

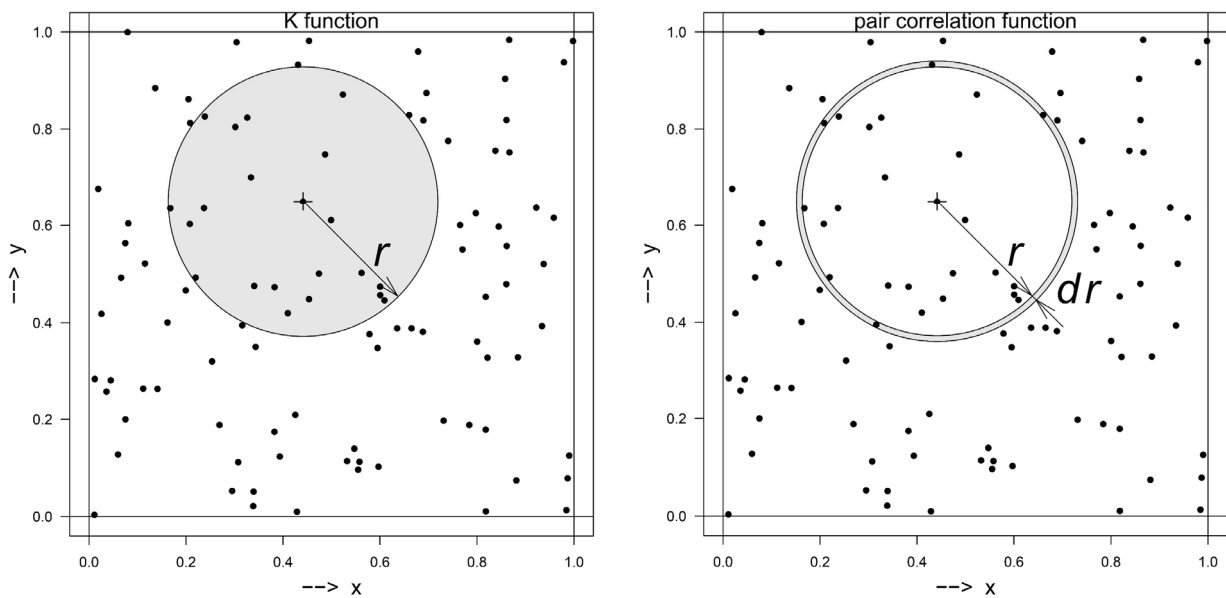


図 4.14 K 関数 (左) と pair correlation 関数 (右) の比較

したがって以下のように $g(r)$ を導出することができる。 $g(r)$ と $K(r)$ の関係は、確率密度関数と分布関数の関係に似ている。2 次元、3 次元および多次元 (u 次元) の場合を以下に示す。ここで、 $V_u(r)$ は半径 r の u 次元球の体積である。

$$g(r) = \frac{\frac{d}{dr} K(r)}{2\pi r}$$

$$g(r) = \frac{\frac{d}{dr} K(r)}{4\pi r^2}$$

$$g(r) = \frac{\frac{d}{dr}K(r)}{\frac{d}{dr}V_u(r)}$$

なお、上記とは異なる導出方法もある (Stoyan & Stoyan, 1994, p.249)。

CSR 点過程に対する pair correlation 関数 $g(r)$ は、CSR 点過程に対する K 関数の値を用いて、以下のよう
に計算することができる。2 次元、3 次元および多次元 (u 次元) の場合を以下に示す。ここで、 $V_u(r)$
は半径 r の u 次元球の体積である。

$$g(r) = \frac{\frac{d}{dr}K(r)}{2\pi r} = \frac{\frac{d}{dr}(\pi r^2)}{2\pi r} = 1$$

$$g(r) = \frac{\frac{d}{dr}K(r)}{4\pi r^2} = \frac{\frac{d}{dr}\left(\frac{4\pi r^3}{3}\right)}{4\pi r^2} = 1$$

$$g(r) = \frac{\frac{d}{dr}K(r)}{\frac{d}{dr}V_u(r)} = \frac{\frac{d}{dr}V_u(r)}{\frac{d}{dr}V_u(r)} = 1$$

pair correlation 関数 $g(r)$ を用いて規則型の点過程を評価する不等式は、規則型の点過程に対する $K(r)$
の不等式から以下のように求められる。

$$g(r) < \frac{\frac{d}{dr}(\pi r^2)}{2\pi r} = 1$$

$$g(r) < \frac{\frac{d}{dr}\left(\frac{4\pi r^3}{3}\right)}{4\pi r^2} = 1$$

$$g(r) < \frac{\frac{d}{dr}V_u(r)}{\frac{d}{dr}V_u(r)} = 1$$

もう少し細かく見ると、規則型の点過程の場合、半径 r が点間の最小距離 r_0 を超えるまでは点が存在しな
いので以下のようなになる。

$$g(r) = 0 \quad \text{for } r \leq r_0$$

また規則型の点過程の場合、半径 r を大きくしていくと複雑な振舞いをする事が報告されている (Illian
et al., p.221, p.239)。

pair correlation 関数 $g(r)$ を用いて集中 (凝集) 型の点過程を評価する不等式は、集中 (凝集) 型の点過
程に対する $K(r)$ の不等式から以下のように求められる。

$$g(r) > \frac{\frac{d}{dr}(\pi r^2)}{2\pi r} = 1$$

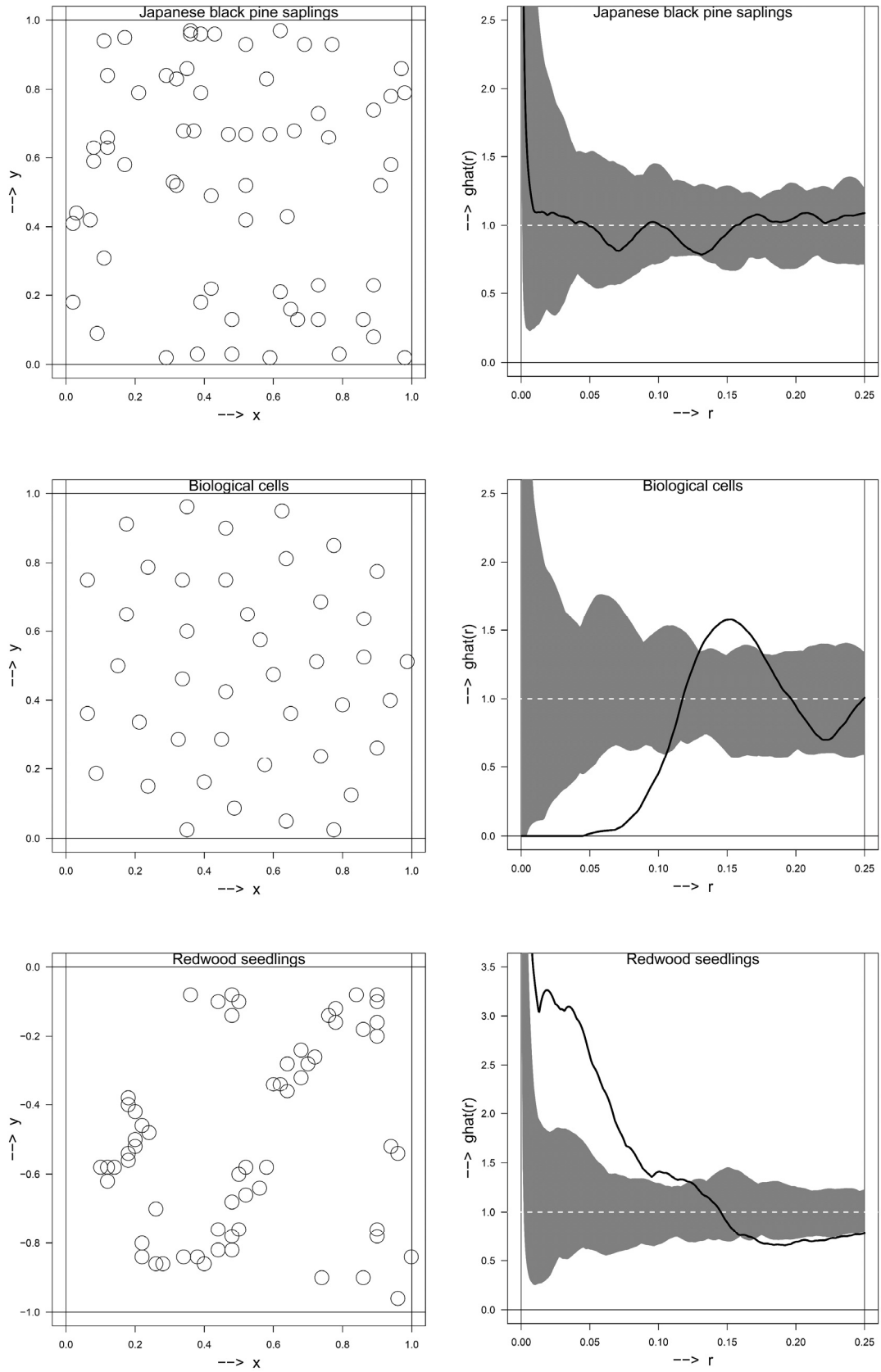
$$g(r) > \frac{\frac{d}{dr} \left(\frac{4\pi r^3}{3} \right)}{4\pi r^2} = 1$$

$$g(r) > \frac{\frac{d}{dr} V_u(r)}{\frac{d}{dr} V_u(r)} = 1$$

もう少し細かく見ると、集中（凝集）型の点過程の場合、点が集中しているので、半径 r が小さい領域で $g(r)$ は大きな値となり、半径 r が大きくなるに従い $g(r)$ は減少する。

観測データから $\hat{g}(r)$ を推定する場合、Illian *et al.* (2008, p.232) は $\hat{K}(r)$ を微分して用いるのではなく、モンテカルロ法を用いて確率密度関数を推定するときのように、カーネル関数の使用を推奨している。例えば、Illian *et al.* (2008, p.482) にあるような Epanechnikov kernel の使用である。pair correlation 関数を用いた分類結果を図 4.15 に示す。

O-ring 統計量 (O-ring statistic) $O(r)$ を用いた分類も可能である (Wiegand & Moloney, 2004)。CSR 点過程に対しては $O(r) = \lambda$ 、規則型の点過程に対しては $O(r) < \lambda$ 、集中（凝集）型の点過程に対しては $O(r) > \lambda$ となる。



注：灰色の帯は CSR 点過程に従う 100 本の曲線群のエンベロップを示す。

図 4.15 pair correlation 関数による空間点パターンの分類

4.3 グラフィカルな分類方法の問題点

G 関数と F 関数は点間の近接距離から構築される方法で、極めてよく似ている。また J 関数は G 関数と F 関数の組み合わせになっている。一方、L 関数と pair correlation 関数は K 関数から派生した（または、K 関数を改良した）ものと考えられる。そこで、ここでは G 関数と K 関数の問題点について述べる。

4.3.1 G 関数の問題点

Chiu *et al.* (2013, p.144) は G 関数の問題点を、近傍の情報しか与えず、その外側で何が起こっているのか分からないので近視眼的 (short-sighted) であると述べている。G 関数は第 1 近隣木の辺長の経験分布関数なので、その問題点は第 3 章「近接グラフ」の 3.4.1 節「MST の近似としての第 1 近隣木」に記述した内容が当てはまる。本研究で提案する AGsi-curve との比較に於いて、具体的な数値例を提示する (第 6 章)。

4.3.2 K 関数の問題点

K 関数の問題点に関して Illian *et al.* (2008, p.216) は、点配置が局所的には規則型 (short-range regularity) だが、大域的には凝集型 (long-range clustering) の場合、K 関数の解釈は理解しにくい (complicated) ものとなると述べている。その理由として K 関数が累積的性質 (cumulative characteristic) を持つからだを指摘している。

本研究では K 関数の問題点として以下の 2 項目を取り上げる。

- ・強度 λ が大きくなるにつれて、規則型の点配置とランダム型の点配置の区別が曖昧になる。
- ・(提案時から境界補正を含んでいることから、) 境界効果に対して極めて敏感である。

本研究で提案する AGsi-curve との比較に於いて、具体的な数値例を提示する (第 6 章)。

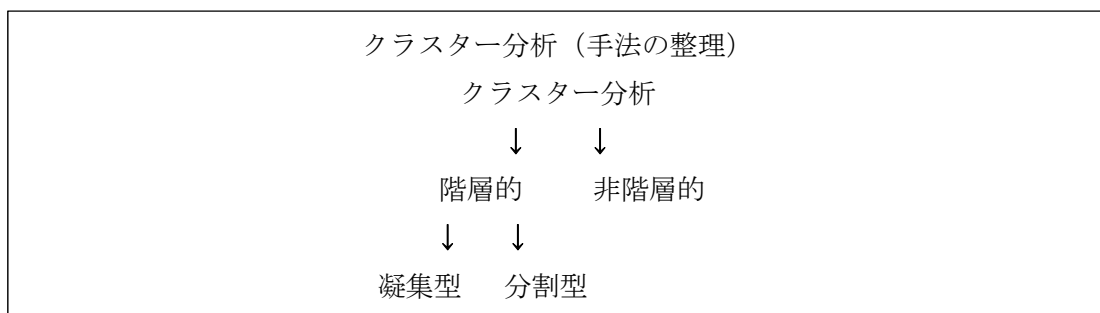
第5章 凝集型階層的クラスター分析

5.0 はじめに

本研究で提案することになる AG-curve は AHC 手法の併合距離とクラスター数間の関係を示している。そこで5章では AHC 手法の概略を示したのち、AHC 手法の単調性に焦点を当てる。単調性とはクラスター化の進展とともに併合距離が単調に増加することを意味する。この単調性は AG-curve からの要請である。そして、単調性の判別条件の基礎となる強メトリック不等式を示す。

5.1 凝集型階層的クラスター分析

クラスター分析は多変量解析やデータマイニングなどで頻繁に使われる教師無しの分類方法であり、概ね以下のように整理される。クラスターとは似たもの同士を集めたグループであり、このグループ分けの手法がクラスター分析と呼ばれている。Sneath & Socol (1973, p.69) は、クラスター分析を行う分類対象を OTU (Operational Taxonomic Unit) と呼んでいる。



クラスター分析は階層的クラスター分析と非階層的クラスター分析に分けられる。後者の非階層的クラスター分析は k 平均法 (k-means algorithm) などが代表的な手法であるが、本研究の対象から外れるので説明を省略する。前者の階層的クラスター分析は凝集型 (Agglomerative) と分割型 (divisive) に分けられる。凝集型はボトムアップ法 (bottom-up method) と呼ばれ、OTU の個体が別々に存在する状態

(OTU が N 個あれば、要素が 1 つしかないクラスターが N 個あると考える) から始まり、ある決まりに基づいて直前のクラスターを連続的に併合して行き、最終的には単一のクラスター (要素を N 個含むクラスターが 1 つできると考える) を形成する方法である。一方、分割型はトップダウン法 (top-down method) と呼ばれ、OTU の個体が全て単一のクラスターに併合されている状態から始まり、ある決まりに基づいて直前のクラスターを連続的に分割して行き、最終的には OTU の個体が別々に存在する状態で終了する。分割型は教師有り分類法の 1 つである決定木 (decision tree) に似ている。事実、Piccarreta and Billari (2007) は、分割型の新しい方法に CART (Breiman *et al.*, 1984) のロジックを利用している。凝集型も分割型も、その結果である階層構造をデンドログラム (dendrogram) という 2 次元の樹形図 (分野によっては系統樹とも呼ばれる) に表現することができる。ここでは本研究と関係が深い凝集型階層的クラスター分析に焦点を当てる。これを AHC 手法 (Agglomerative Hierarchical Clustering Algorithm) と呼ぶことにする。

Gnanadesikan (1997, p.109) は、階層的クラスター分析に現れる 2 つの極端な状態を、弱クラスター化 (weak clustering) と強クラスター化 (strong clustering) と呼んでいる。弱クラスター化とは OTU の個体が別々に存在する状態 (OTU が N 個あれば、要素が 1 つしかないクラスターが N 個あると考える) であり、強クラスター化とは OTU の個体が全て単一のクラスターに併合されている状態である。これらの用語を使うと、凝集型の併合は弱クラスター化から強クラスター化へ向かい、分割型の分解は強クラスター化から弱クラスター化へ向かうと表現できる。

5.2 類似度と非類似度 (距離)

クラスター分析は似たもの同士を集める手法なので、類似性の高い OTU を同一のグループに集める必要がある。そのためにはグループを構成する各 OTU 間の類似性の度合いを数量化しなければならない。類似度の測度としてはピアソンの積率相関係数などが使われる。

一方、クラスター分析では類似度の逆の概念である非類似度を用いることもできる。非類似度を数量化するために様々な距離が使われる。距離の測度としては、ユークリッド (Euclidian) 距離・マンハッ

タン (Manhattan) 距離・ミンコフスキー (Minkovski) 距離・などがある。ミンコフスキー距離は、ユークリッド距離とマンハッタン距離を一般化したものである。奥野・他 (1981, p.395, 表 3.1) は、様々な非類似度 (距離) および類似度の定義を一覧表にして示している。また、アルゴリズム上は、非類似度 (距離) に関する不等式を逆向きにすれば、類似度に関するそれに変えられる (奥野・他, 1981, p.394)。更に非類似度 (距離) としては、それぞれの AHC 手法で定義されるクラスター間の距離がある。これらについては後述する。

5.3 AHC 手法の分類アルゴリズム

AHC 手法には多くのアルゴリズムが存在する。主要なものを以下に示す。

AHC 手法の主なアルゴリズム

- ①最短距離法 (Single-linkage or Nearest-neighbor method)
- ②最長距離法 (Complete-linkage or Furthest-neighbor method)
- ③重み無し群平均法 (Unweighted Pair-Group Method using arithmetic Average)
- ④重み付き群平均法 (Weighted Pair-Group Method using arithmetic Average)
- ⑤セントロイド法 (重心法) (Centroid method)
- ⑥メディアン法 (Median method)
- ⑦ウォード法 (Ward's method)

Sneath & Sokal (1973, p218, TABLE 5-2) は、①～⑥を SAHN clustering methods と呼んでいる。また、これらは特に生物学的物質 (biological material) の中に、クラスターを見つけるために最も頻繁に使われる方法であると述べている (Sneath & Sokal, 1973, p.214)。なお SAHN は、Sequential・Agglomerative・Hierarchic・Nonoverlapping の頭字語 (acronym) である。

Murtagh (1983) は前半の 4 種類をグラフ法 (graph methods)、後半の 3 種類を幾何学法 (geometric methods) として区別している。グラフ法の名称はグラフ理論 (Graph Theory) に由来すると考えられる。

5.3.1 AHC 手法の共通アルゴリズム

それぞれの分類アルゴリズムに共通する具体的な手順 (Step 1 ~ Step 5) を以下に記す。

AHC 手法の共通手順

Step 1 : N 個の OTU を N 個のクラスターとする。

- 要素が 1 個のクラスターが N 個あると考える (弱クラスター化)。

Step 2 : 全てのクラスター間の距離行列を作成する。

- 距離行列は $N \times N$ の対称行列である。
- クラスター間の距離は分類アルゴリズムにより異なる。
- 分類アルゴリズムはクラスター間の距離の定義と言い換えることができる。

Step 3 : 最も距離が近い (似ている) 2 つのクラスターを併合する。

- クラスター C_u と C_v を併合しクラスター $C_u \cup C_v$ をつくる。
- 併合したクラスターの組み合わせと併合距離 $d(C_u, C_v)$ を記録する。
- この併合によりクラスター数が 1 個減る。

Step 4 : 併合後のクラスター間の距離行列を作成する (距離行列の更新)。

- 併合前の距離行列からクラスター C_u と C_v に対応する行と列を削除する。
- 新しいクラスター $C_u \cup C_v$ と残りのクラスター間の距離を与える行と列を追加する。

Step 5 : Step 3 と Step 4 を N-1 回繰り返す。

- (要素を N 個含む) 単一のクラスターができる (強クラスター化)。

注 : クラスターの併合には、以下の 3 通りの組み合わせがある。

- OTU 単体同士の併合 (併合の第 1 段階は全て、この組み合わせになる)
- OTU 単体とクラスターの併合
- クラスター同士の併合 (併合が進展していくにつれ、この組み合わせが多くなる)

併合したクラスターの組み合わせと併合距離 $d(C_u, C_v)$ の記録を用いて、併合の過程 (階層構造) をデンドログラムとして表示することができる。なお、更新される距離行列とその要素は、コーフェネティック行列 (Cophenetic matrix) とコーフェネティック距離 (Cophenetic distance) と呼ばれる場合がある (Sneath & Sokal, 1973, pp.278-280)。また、元の距離行列とコーフェネティック行列の要素間のピアソンの積率相関係数は、コーフェネティック相関係数 (Cophenetic distance) と呼ばれる。

5.3.2 Lance and Williams の組合せ的な手法

Lance and Williams (1967) はクラスター分析を行う計算機用プログラミング作成の観点から、AHC手法の一部に対して組合せ的な手法 (combinatorial method) を提案した。これはクラスターを併合する度に行う距離行列の更新に於いて、併合する前の距離を用いて併合後の距離を求める方法である。特筆すべきことは、複数の手法に対して、4 個のパラメータを持つ 1 つの更新式を使えば良いという点である。つまり、更新式中の 4 個のパラメータの値を変えるだけで、複数の手法に対応できることになり、効率的なプログラミングが可能となる。以下では Nakamura & Ohsumi (1990) の表記法を用いる。また、一般化した併合前後のクラスター間の位置関係を図 5.1 に示す。この作成は、柳井・他 (1986, p.230) を参考にしている。

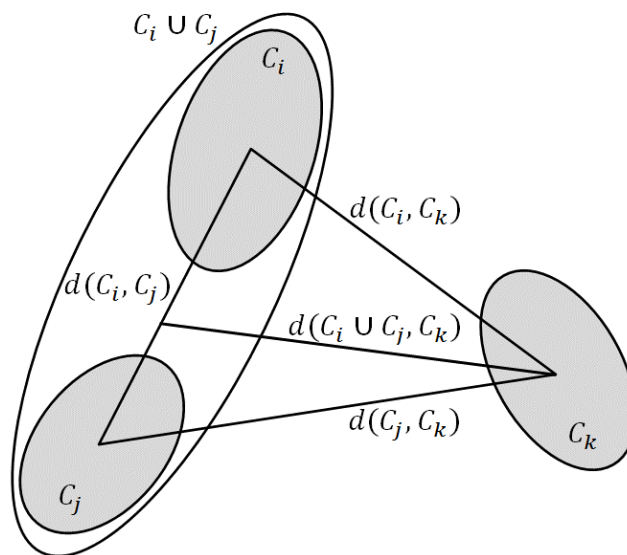


図 5.1 併合前後のクラスター間の位置関係

クラスター C_i とクラスター C_j が併合され新しいクラスター $C_i \cup C_j$ がつくられたと仮定する。この時、クラスター $C_i \cup C_j$ と第 3 のクラスター C_k ($k \neq i, j$) との距離 $d(C_i \cup C_j, C_k)$ は、併合前の各クラスター間の距離 $d(C_i, C_j)$, $d(C_i, C_k)$, $d(C_j, C_k)$ を用いて以下の式で表すことができる。

$$d(C_i \cup C_j, C_k) = \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j) + \gamma |d(C_i, C_k) - d(C_j, C_k)|, \quad (i \neq j, k \neq i, j)$$

各パラメータ $\{\alpha_i, \alpha_j, \beta, \gamma\}$ は、併合前の距離が併合後の距離に与える影響度合いを示していると考えら

れる。分類アルゴリズム別にパラメータの具体的な値を表 5.1 に示す。表 5.1 は Nakamura & Ohsumi (1990) より抜粋したものである。この表は Sneath & Sokal (1973, p218, TABLE 5-2) に、ウォード法 (Ward, 1963) を追加したものである。ウォード法は、後になって Wishart (1969) により Lance & Williams (1967) の組合せ的な手法に組み込まれた。

表 5.1 各パラメータの値 (分類アルゴリズム別)

	パラメータ				$\alpha_i + \alpha_j + \beta$	単調性
	α_i	α_j	β	γ		
最短距離法	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	1	yes
最長距離法	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	1	yes
重み無し群平均法	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0	0	1	yes
重み付き群平均法	$\frac{1}{2}$	$\frac{1}{2}$	0	0	1	yes
ウォード法	$\frac{n_i + n_k}{n_t}$	$\frac{n_j + n_k}{n_t}$	$-\frac{n_k}{n_t}$	0	1	yes
セントロイド法	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$-\alpha_i \cdot \alpha_j$	0	$1 + \beta$	no
メディアン法	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0	$\frac{3}{4}$	no

注 1 : n_i はクラスター C_i のサイズ (要素数) を表す。

注 2 : $n_t = n_i + n_j + n_k$

Lance & Williams (1967) の更新式の発展について少し触れる。Nakamura & Ohsumi (1990) は、パラメータ $\{\gamma\}$ が最短距離法と最長距離法にしか現れないことを手掛かりに、可変法 (flexible method) を改良した一般化可変法 (generalized flexible method) を導入した。そして、Lance and Williams (1967) の更新式の代わりに、 $\{\gamma\}$ を除く 3 個のパラメータ $\{\alpha_i, \alpha_j, \beta\}$ を使った新しい更新式を提案した。これによりパラメータ空間を 3 次元で表現することができ、その空間内に分類アルゴリズムのそれぞれの特徴を明確に位置づけている。本研究では詳しく取り上げない。

5.4 単調性と強メトリック不等式

AHC 手法の結果として得られた併合距離の数列に関して、その単調性を考える。クラスター併合の進展とともに、併合距離が単調に増加するか否かという意味である。これが単調増加または単調非減少（以降は単調性と呼ぶ）になっていることは、併合後のクラスターが併合前のクラスターを常に包含している（入れ子構造になっている）ことと同等である。この場合、分類結果の解釈は簡単になる。しかし、AHC 手法の分類アルゴリズムの全てが、この単調性を保証されているわけではない。単調性とはクラスター併合の各段階に於いて、以下の式が常に成り立っていることである。

$$d(C_i \cup C_j, C_k) \geq d(C_i, C_j)$$

単調性が保証されていない分類アルゴリズムでは、常にではないが以下のような奇異な結果を招くことがある。この現象の発生は局所的な場合が多い。

- 距離の逆転 (inversion または reversal)
- クロスオーバー (crossover)

距離の逆転とは、 $d(C_i \cup C_j, C_k) < d(C_i, C_j)$ となり、ある時点の併合距離が直前の併合距離より短くなってしまうことである。またクロスオーバーとはデンドログラムの縦線と横線が交差する現象である。具体例はセントロイド法の項で提示する。

5.4.1 距離が持つ性質

ここでは距離が持つ性質を考える。2点 a, b 間の距離を $d(a, b)$ と表す。距離は通常、以下の3項目の性質を満たす。

1. $d(x, y) \geq 0$ (正值性)
2. $d(x, y) = 0$ if and only if $x = y$ (同一律)
3. $d(x, y) = d(y, x)$ (対称律)

上記3項目に加え 4a の性質 (三角不等式) を満たすものがメトリック距離 (metric distance) と呼ばれる。

$$4a. \quad d(x, y) \leq d(x, z) + d(y, z)$$

更に 4b の性質を満たすものは、相加的距離 (additive distance) と呼ばれる。

$$4b. \quad d(x, y) + d(z, u) \leq \max\{d(x, z) + d(y, u), d(x, u) + d(y, z)\}$$

この不等式は 4 点条件 (four-point condition) と呼ばれるものである。3 つの合計 $\{d(x, y) + d(z, u)\}$ $\{d(x, z) + d(y, u)\}$ $\{d(x, u) + d(y, z)\}$ のうち、最大の 2 つは互いに等しく、残りの 1 つより小さくないことを意味している (Page & Holmes, 1998, p.26)。相加的距離は系統学 (phylogenetics) 等で使われる。

そのうえ、更に 4c の性質を満たすものは、強メトリック距離 (ultrametric distance) と呼ばれる。

$$4c. \quad d(x, y) \leq \max\{d(x, z), d(y, z)\}$$

これらの包含関係は以下のようにになっており、右にいくほど厳しい条件が課されている。

メトリック距離 \supset 相加的距離 \supset 強メトリック距離

最後の不等式 (4c) は本研究で取り扱う AHC 手法と関係が深い。これは通常三角不等式に制限を加えたもので、強メトリック不等式または超距離不等式 (ultrametric inequality) (Johnson, 1967) と呼ばれる。この強メトリック不等式は、 x と y の間の距離 $d(x, y)$ が、2 つの距離 $d(x, z)$ と $d(y, z)$ の大きい方を超えることはないという意味である。これを幾何学的に言えば、頂角が 60 度以下の鋭角二等辺三角形 (図 5.2) の形成になり、

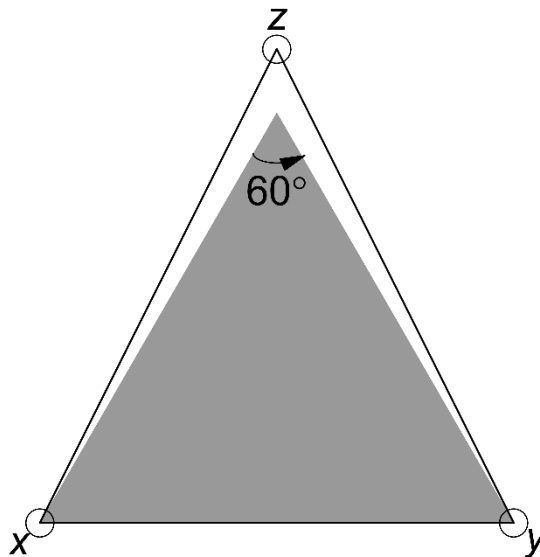


図 5.2 強メトリック不等式の説明図 (二等辺三角形)

併合される x と y の間の距離 $d(x,y)$ が底辺になっている。この観点から大隅・中村（1981）は、強メトリック不等式を二等辺三角不等式と呼んでいる。メトリック距離の代表例はユークリッド距離である。一方、強メトリック距離はAHC手法の1部のアルゴリズムで定義されるクラスター間の距離であり、本研究と関係がある。Johnson（1967）は最短距離法と最長距離法で定義されるクラスター間の距離を強メトリック距離としているが、表 5.1 で取り上げたセントロイド法とメディアン法を除いたものは全て強メトリック距離である。強メトリック距離はデンドログラムを用いてグラフィカルに表現することができる。大隅・他（1994, p329）には、強メトリック距離とデンドログラムの対応図が3種類載せられている。

Lebart et al.,（1984, p121）は、「強メトリック空間に於ける三角形は全て、底辺が最短の二等辺三角形になる」ことを以下のように証明している。

【証明】

x と y が併合されてクラスター t ができると仮定する。
したがって $d(x,z) \geq d(x,y)$ 且つ $d(z,y) \geq d(x,y)$ である。

強メトリック不等式より

$$d(z,x) \leq \max\{d(z,y), d(x,y)\}$$

$d(z,y) \geq d(x,y)$ なので

$$d(z,x) \leq d(z,y) \quad (a)$$

別の強メトリック不等式より

$$d(z,y) \leq \max\{d(z,x), d(x,y)\}$$

$d(x,z) \geq d(x,y)$ なので

$$d(z,y) \leq d(z,x) \quad (b)$$

(a)と(b)より

$$d(z,y) = d(z,x)$$

5.4.2 単調性の判別

Johnson（1967）は最短距離法と最長距離法だけを対象として、単調性を満たすことから強メトリック不等式を導出した。Milligan（1979）は、Lance and Williams（1967）の更新式のパラメータを使い、他の分類アルゴリズムに対して、Johnson（1967）の証明を拡張した。そして、単調性を生み出す分類アルゴリズムは、強メトリック不等式を満たすアルゴリズムでもあると述べている。

Milligan (1979) は Lance and Williams (1967) の更新式の $\{\alpha_i, \alpha_j, \beta, \gamma\}$ が以下の条件の場合、分類アルゴリズムは単調性を満たすことを示した。

$$\gamma \geq 0 \vee (\gamma < 0 \wedge |\gamma| \leq \alpha_1, \alpha_2)$$

$$\min(\alpha_1, \alpha_2) \geq 0.$$

$$\alpha_i + \alpha_j + \beta \geq 1.$$

この条件式には複数の表示方法がある。上記は Batagelj (1981) が整理したもので、表示を改良する前の Milligan (1979) の条件式である。この条件式を用いて表 5.1 に取り上げた分類アルゴリズムの判別を考える。最初の条件式の $\{\gamma\}$ は最短距離法と最長距離法にしか現れず、この条件を持たしている。また表 5.1 の全ての分類アルゴリズムが 2 番目の条件式を満たしている。結局、最後の式だけを用いれば良いことが分かる。その結果を表 5.1 の右端の列に記載する。

5.4.3 補足 1

単調性が保証されていない分類アルゴリズムは、その使用が敬遠されがちである。しかし、この傾向に対して、Williams *et al.* (1971) は、非単調性が分類性能に与える影響は本質的ではないと述べている。

クラスターの併合による空間の歪は、単調性と深くかかわっており、Nakamura & Ohsumi (1990) が両者の関係を論じている。本研究では詳しく取り上げない。

5.5 分類アルゴリズムの具体例

AHC 手法の分類アルゴリズムは多いが、ここでは本研究と関係が深い最短距離法、最長距離法、(重み無し) 群平均法、および比較の観点からセントロイド法を取り上げる。

5.5.1 最短距離法

最短距離法に於けるクラスター間の距離の定義は、2 つのクラスターに含まれる要素 (OTU 単体) 間の距離のうち最小の値である。Lance and Williams (1967) の一般式に、 $\alpha_i = \alpha_j = \frac{1}{2}$, $\beta = 0$, $\gamma = -\frac{1}{2}$ (表

5.1 の値) を代入すれば、最短距離法の更新式が得られる。

$$\begin{aligned}
 d(C_i \cup C_j, C_k) &= \frac{1}{2}d(C_i, C_k) + \frac{1}{2}d(C_j, C_k) - \frac{1}{2}|d(C_i, C_k) - d(C_j, C_k)| \\
 &= d(C_j, C_k), \quad \text{if } d(C_i, C_k) \geq d(C_j, C_k); \\
 &= d(C_i, C_k), \quad \text{if } d(C_i, C_k) < d(C_j, C_k).
 \end{aligned}$$

この結果は表現を変えれば以下のようになり、

$$d(C_i \cup C_j, C_k) = \min(d(C_i, C_k), d(C_j, C_k))$$

これが手法の名前の由来ともなっている。また、併合後の距離は併合前の距離行列の要素の 1 つと等しくなることを意味している。つまり、距離行列の更新に於いて、出発時の距離行列の要素の値を変えずに、最後まで使い切ることになる。またパラメータ $\{\alpha_i, \alpha_j, \beta, \gamma\}$ は、単調性の条件式 (不等式) を満たしている。

$$\begin{aligned}
 |\gamma| &= \frac{1}{2} \leq \frac{1}{2} \\
 \min(\alpha_1, \alpha_2) &= \frac{1}{2} \geq 0 \\
 \alpha_i + \alpha_j + \beta &= 1 \geq 1
 \end{aligned}$$

併合前後のクラスター間の位置関係 (最短距離法) を図 5.3 に示す。

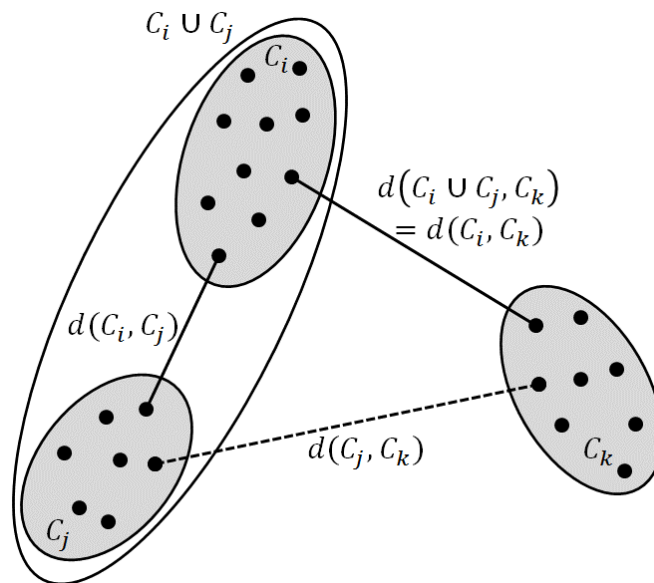


図 5.3 併合前後のクラスター間の位置関係 (最短距離法)

最短距離法のアルゴリズムに関してはグラフ理論 (Graph Theory) を用いたアプローチが存在する。

Gower & Ross (1969) は、MST を求める複数のアルゴリズムを紹介し、以下のように述べている。

- ・最短距離法の併合距離と MST の辺長が等価である。(このことは、本研究に於いて新たに提案する AGsi-curve の基礎を成す。)
- ・データセットが巨大になり、最短距離法の距離行列の計算コストが高くなり過ぎた場合には、MST は補助的技術 (ancillary technique) として役に立つ。

Lebart et al., (1984, p125) は、最短距離法の階層構造と MST の関係を図を用いて説明している。

最短距離法の特徴として連鎖効果 (chain effects) と空間の凝縮 (space contraction) を取り上げる。

連鎖効果とは、相対的に大きなクラスターと OTU 単体との間で併合が頻繁に発生する現象である。1 つのクラスターが近傍にある OTU 単体を次々に取り込んでいき、大きなクラスターに成長する。その結果、大きさの近い複数のクラスターが形成されにくい。これはデータマイニングでは欠点とされるが、分野が異なれば長所となる場合がある。本研究に於いては、この特徴 (連鎖効果) を積極的に活用しているとも言える。連鎖効果はデンドログラムより、MST の方が分かり易い。Zahn (1971) には、様々な広がりを見せる点群に対して、MST が奇異な図形を描く様子が多く示されている。

空間の凝縮とは、クラスター併合の進展とともに空間が凝縮していくように見える現象である。これは単調性と深くかかわっており、Nakamura & Ohsumi (1990) が両者の関係を論じている。本研究では詳しく取り上げない。

5.5.2 最長距離法

最長距離法に於けるクラスター間の距離の定義は、2 つのクラスターに含まれる要素 (OTU 単体) 間の距離のうち最大の値である。Lance and Williams (1967) の一般式に、 $\alpha_i = \alpha_j = \frac{1}{2}$, $\beta = 0$, $\gamma = \frac{1}{2}$ (表 5.1 の値) を代入すれば、最長距離法の更新式が得られる。

$$\begin{aligned}d(C_i \cup C_j, C_k) &= \frac{1}{2}d(C_i, C_k) + \frac{1}{2}d(C_j, C_k) + \frac{1}{2}|d(C_i, C_k) - d(C_j, C_k)| \\ &= d(C_i, C_k), \quad d(C_i, C_k) \geq d(C_j, C_k);\end{aligned}$$

$$= d(C_j, C_k), \quad d(C_i, C_k) < d(C_j, C_k).$$

この結果は表現を変えれば以下のようになり、

$$d(C_i \cup C_j, C_k) = \max(d(C_i, C_k), d(C_j, C_k))$$

これが手法の名前の由来ともなっている。また距離 $d(C_i \cup C_j, C_k)$ は、併合後の3クラスターを囲む、最小の(超)球面の直径となる。最短距離法の場合と同様に、併合後の距離は併合前の距離行列の要素の1つと等しくなる。つまり、出発時の距離行列の要素の値を変えずに、最後まで使い切ることになる。また最長距離法のパラメータ $\{\alpha_i, \alpha_j, \beta, \gamma\}$ は、単調性の条件式(不等式)を満たしている。

$$|\gamma| = \frac{1}{2} \leq \frac{1}{2}$$

$$\min(\alpha_1, \alpha_2) = \frac{1}{2} \geq 0$$

$$\alpha_i + \alpha_j + \beta = 1 \geq 1$$

併合前後のクラスター間の位置関係(最長距離法)を図5.4に示す。

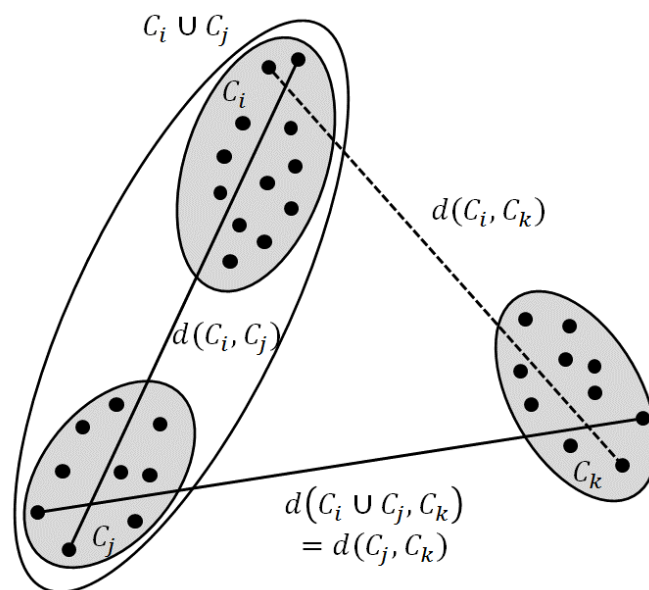


図 5.4 併合前後のクラスター間の位置関係(最長距離法)

最長距離法のアルゴリズムに関してもグラフ理論を用いたアプローチが存在する。Baker & Hubert (1976) はノード彩色 (node colorings) を用いた方法を提案しているが、最長距離法とノード彩色の関係は、最短

距離法と MST の関係ほど単純ではないと述べている。

最長距離法の主な特徴は以下の 2 項目である。

- 大きさの近い複数のクラスターが形成される。
- 空間の拡大 (space dilation) : これは、最短距離法とは逆に、クラスター併合の進展とともに空間が拡大していくように見える現象である。これは単調性と深くかかわっており、Nakamura & Ohsumi (1990) が両者の関係を論じている。本研究では詳しく取り上げない。

5.5.3 (重み無し) 群平均法

最短距離法と最長距離法は、クラスター間の距離を定義するために、クラスター内の 1 点を用いている。それに対して群平均法は、一方のクラスター内の個体と、もう一方のクラスター内の個体間の全ての距離の平均値を用いる。

Lance and Williams (1967) の一般式に、 $\alpha_i = \frac{n_i}{n_i+n_j}$, $\alpha_j = \frac{n_j}{n_i+n_j}$, $\beta = \gamma = 0$ (表 5.1 の値) を代入すれば、(重み無し) 群平均法の更新式が得られる。

$$d(C_i \cup C_j, C_k) = \frac{n_i}{n_i + n_j} d(C_i, C_k) + \frac{n_j}{n_i + n_j} d(C_j, C_k)$$

パラメータ $\{\alpha_i, \alpha_j, \beta, \gamma\}$ は、単調性の条件式 (不等式) を満たしている。

$$|\gamma| = 0 \leq \frac{1}{2}$$

$$\min(\alpha_1, \alpha_2) \geq 0$$

$$\alpha_i + \alpha_j + \beta = 1 \geq 1$$

群平均法の適用例としては、距離を用いる場合 (Everitt et al, 2011, p.71, Table 4.1) と、距離の平方 (奥野・他, 1986, p.324, 表 1) を用いる場合の 2 通りが存在する。本研究では最短距離法と最長距離法との比較を行うため距離を用いる。最短距離法では空間は凝縮され、最長距離法では空間は拡大されたが、群平均法では空間は保存 (space conservative) される。

5.5.4 セントロイド法

本研究では、提案する AG-curve の基礎を成す AHC 手法としてセントロイド法を用いていない。その理由は単調性が保証されていないからである。言い換えれば、セントロイド法では距離の逆転が発生する可能性がある。これについては、Anderberg (1973, pp.140-142) が、セントロイド法の距離の逆転現象 (Centroid reversal phenomenon) として解説している。パラメータ $\{\alpha_i, \alpha_j, \beta, \gamma\}$ (表 5.1 の値) は、単調性の条件式 (不等式) を満たしていない。

$$|\gamma| = 0 \leq \frac{1}{2}$$

$$\min(\alpha_1, \alpha_2) \geq 0$$

$$\alpha_i + \alpha_j + \beta = 1 - \alpha_i \cdot \alpha_j = 1 - \frac{n_i n_j}{(n_i + n_j)^2} < 1$$

最短距離法・最長距離法・群平均法はクラスター間の距離として点間の距離を用いている。これに対してセントロイド法のクラスター間の距離は、その名前が示すように重心間の距離の平方である。以下では簡単な例を用いて距離の逆転が発生する条件を示す。第 1 回目のクラスター化に於いては、セントロイド法といえども、2 点間の距離 (の平方) が基準となることに注意する。説明図の作成は Legendre & Legendre (1998, pp.341-343) を参考にしている。

2 変量 (2 次元) の 3 つの OTU である P・Q・R が未分類の状態 (弱クラスター化) で存在し、それらに対する第 1 回目と第 2 回目の併合を考える。ここでは第 1 回目の併合が P と Q の間で発生することを仮定する。

$$\text{OTU 単体 : } P(-0.5, 0.0), Q(0.5, 0.0), R(0.0, y)$$

3 点 P・Q・R の位置関係を図 5.5 に示す。左図が距離の逆転現象が発生する場合で、右図が発生しない場合を表している。

第 1 回目の併合が P と Q の間で発生するためには

$$y > \sqrt{3}/2 \quad (y^2 > 3/4)$$

である必要がある。このとき、 $\{P, Q\}$ の併合距離は $d^2(P, Q) = 1$ で、併合後の重心位置は C_{PQ} となる。

第 2 回目の併合は $\{P, Q\}$ と R の間で発生するので、その併合距離は $d^2(C_{PQ}, R) = y^2$ である。

距離の逆転現象とは $d^2(C_{PQ}, R) < d^2(P, Q)$ なので、

$$(3/4 <) y^2 < 1$$

となる。したがって、この事例では

$$(\sqrt{3}/2 <) y < 1$$

のときに、距離の逆転が発生し、

$$y \geq 1$$

では、距離の逆転は発生しない。

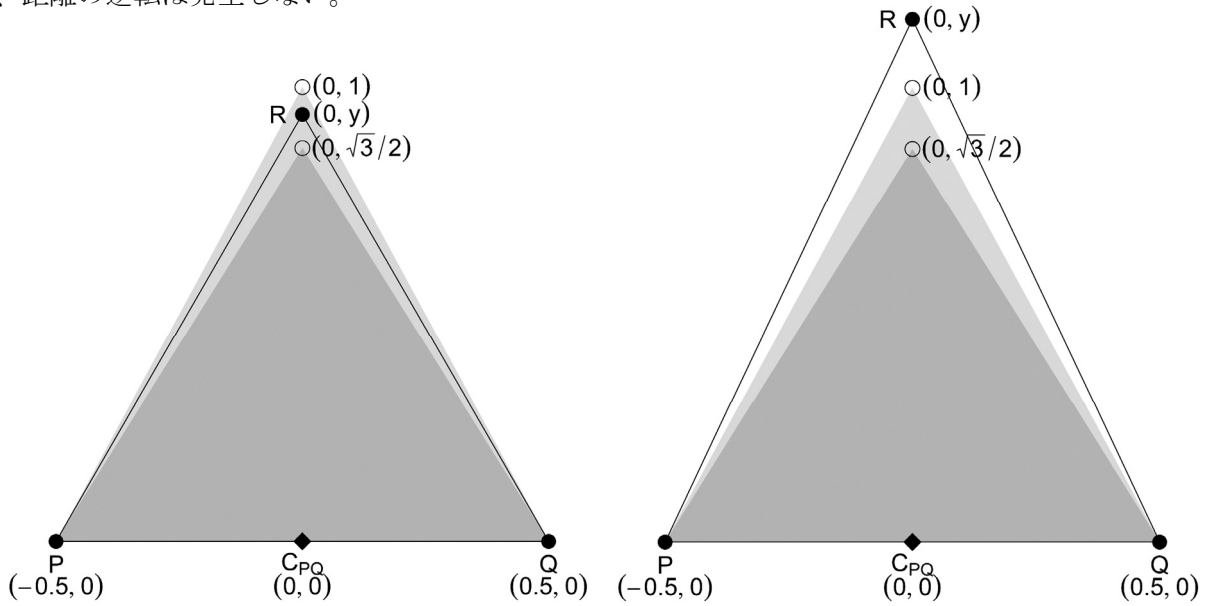


図 5.5 距離の逆転現象（左が発生し、右は発生しない。）

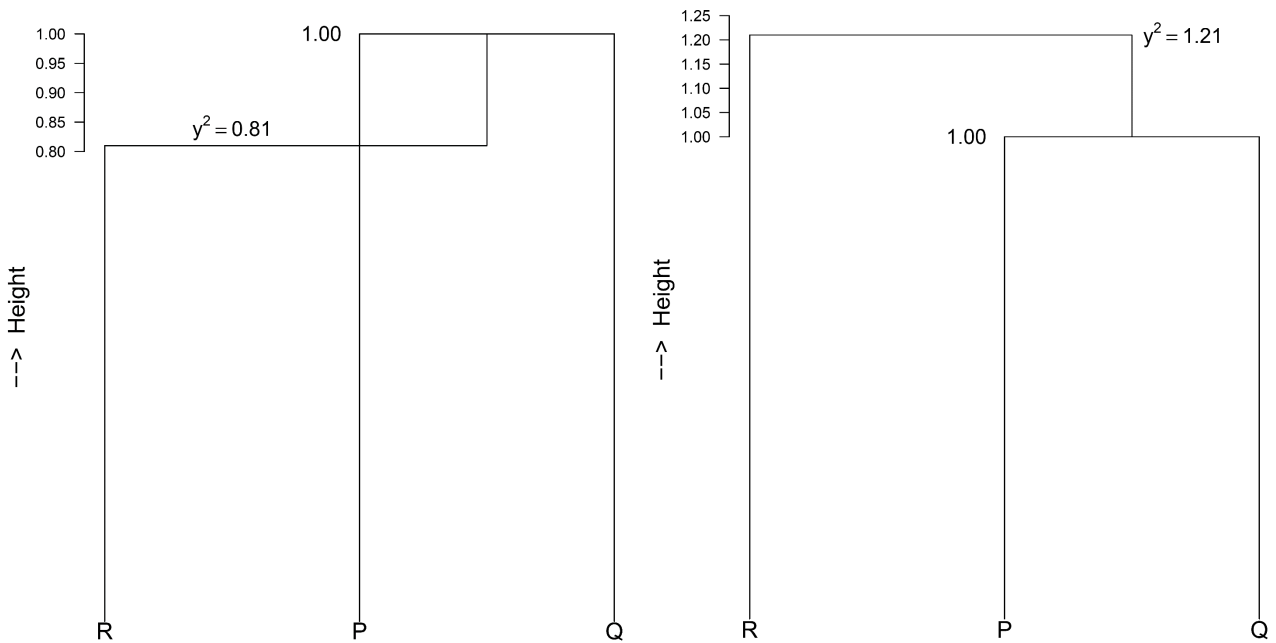


図 5.6 デンドログラム（セントロイド法）

具体例として、 $y = 0.9$ と $y = 1.1$ の場合のデンドログラムを作成し図 5.6 に示す。左図が $y = 0.9$ の場合で、距離の逆転現象により、デンドログラムに交差 (crossover) が発生している。

5.5.5 補足 2

5.3 節で取り上げた AHC 手法のうち、最短距離法・最長距離法・群平均法は、全ての非類似度 (距離) および類似度に関して用いることができるが、他の手法はユークリッド距離以外を用いることはできない点に注意が必要である (柳井・他, 1986, p.233)。

5.6 AHC 手法と空間点パターンとの関連性

Sneath (1966) は最短距離法・最長距離法・群平均法の分類アルゴリズムを比較している。使ったデータは 2 次元の点パターンで、正方形内に一様ランダム (randomly-spaced) に 20 点を配置したものである。Sneath (1966) の内容には空間統計学への直接的な貢献はなく、2 変量のデータを使って純粋にアルゴリズムの比較を行う研究であったと考えられる。2 次元の点パターンを使ったのはクラスター分析の結果を視覚的に表現するのに都合が良いからである。また、Sneath & Sokal (1973, pp.220,224,225,236-239) は Sneath (1966) の内容を他の AHC 手法に拡張している。

本研究で提案する AG-curve は、AHC 手法を空間点パターンの分類へ応用するものである。Sneath (1966) および Sneath & Sokal (1973, pp.220,224,225,236-239) は、クラスターの併合に対応して、2 次元マップ上の点群が順次リンクされていく様子を示している。これは本研究を進めていく上で参考になる。

第Ⅲ部

本研究で提案する空間点パターンのグラフィカルな分類方法

第6章 AG-curve (AGglomerative-curve) を用いた空間点パターンの分類

6.0 はじめに

第6章の内容は Takai, Tamura & Motoyama (2017) に基づいている。提案する方法の定義・空間点パターンの分類への応用などを述べたのち、最後に数理的背景に触れる。

6.1 凝集型階層的クラスタ分析の応用

Takai, Tamura & Motoyama (2017) は、凝集型階層的クラスタ分析 (Agglomerative Hierarchical Cluster Algorithm) の結果を利用した、空間点パターンのグラフィカルな分類方法を提案した。第5章と同様に凝集型階層的クラスタ分析を AHC 手法と呼ぶ。AHC 手法の結果はデンドログラム (dendrogram) と呼ばれる2次元のグラフに表示することができる。提案する方法の名称は AG-curve (AGglomerative-curve) であり、このデンドログラムが持つ情報を利用するものである。第5章で示したように AHC 手法には様々なアルゴリズムがあり、AG-curve もそれらに対応している。例えば、AGsi-curve、AGco-curve、AGav-curve、等である。これらは順番に、最短距離法 (single-linkage) を利用した AG-curve、最長距離法 (complete-linkage) を利用した AG-curve、群平均法 (average-linkage) を利用した AG-curve を意味している。

6.1.1 AG-curve の定義

AHC 手法に於いて、 $(n - k)$ 番目の併合時のクラスター間の距離を、 k 個のクラスター間で計算される強メトリック距離 (ultrametric distance) (Johnson, 1967) (大隅・他, 1994) の最大値

$$h(k) \quad (k = 1, 2, \dots, n - 1)$$

として定義する。ここで n は領域内の点数を表す。 $h(k)$ は最初の併合時 ($k = n - 1$) に最小となり、最後の併合時 ($k = 1$) に最大となる。したがって、 $h(k)$ は単調減少し以下のように表すことができる。

$$h(n - 1) < h(n - 2) < \dots < h(k) < \dots < h(2) < h(1)$$

ここで、点群 $(k, h(k))$, $k = 1, 2, \dots, n - 1$ を直線で補完したグラフを AG-curve と定義する。すなわち AG-curve は AHC 手法の併合距離とクラスター数間の関係を示している。

第 5 章で AHC 手法の併合距離は併合の推移とともに単調増加すると述べた。しかし、併合が進んで行けばクラスター数は減少するので、クラスター数を基準に考える $h(k)$ は単調減少することになる。この単調性は AHC 手法の全てのアルゴリズムで保証されているわけではないことに注意が必要である(5.4 節)。

6.1.2 AG-curve の作図手順

AG-curve の作図手順を整理して以下に示す。本研究では AHC 手法の非類似度としてユークリッド距離を用いる。空間点パターンの分類を行う上では自然な選択である。

AG-curve の作図手順 (まとめ)

- Step 1 : 領域内の点に対する距離行列を計算する。
 - Step 2 : Step 1 の距離行列に対して AHC 手法を適用する。
 - Step 3 : 各段階の併合距離 $h(k)$ を求める。
 - Step 4 : Step 3 に対応したクラスター数 k を求める。
 - Step 5 : k を横軸に $h(k)$ を縦軸にして点群 $(k, h(k))$ をプロットし直線で補完する。
- 注 : AHC 手法の手順は第 5 章で説明済みである。

6.1.3 AG-curve の簡単な作図例

単位正方形内に 10 点を一様ランダムに配置した点パターンに対して、AGsi-curve (AHC 手法 : 最短距離法) を適用した結果である。点パターンを図 6.1 に、そのデンドログラムを図 6.2 に示す。なお点パターンには、クラスター分析による分類結果を楕円で表示した。また、点間をリンクしている直線は後で登場する MST である。デンドログラムからクラスター数 k と併合距離 $h(k)$ を読み取り、点群 $(k, h(k))$ をプロットしたものが図 6.3 である。

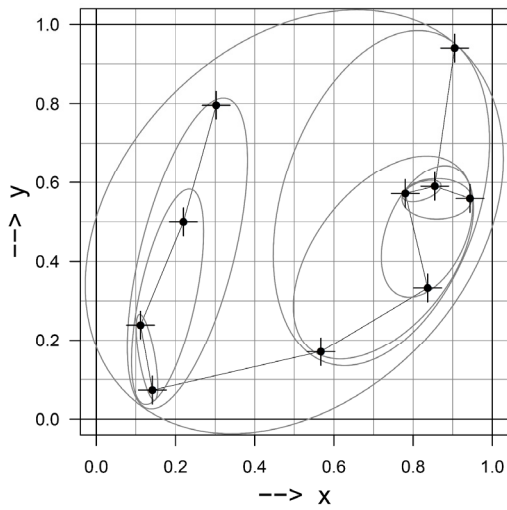


図 6.1 点パターン

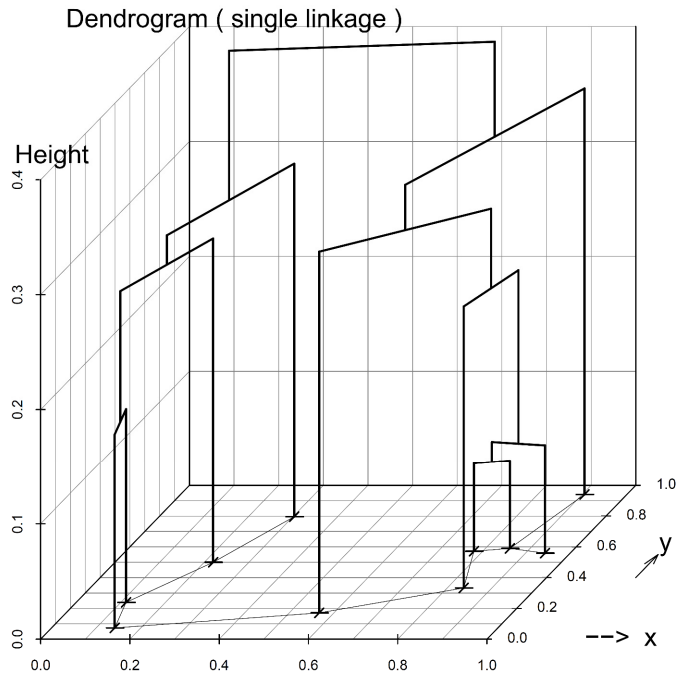


図 6.2 デンドログラム

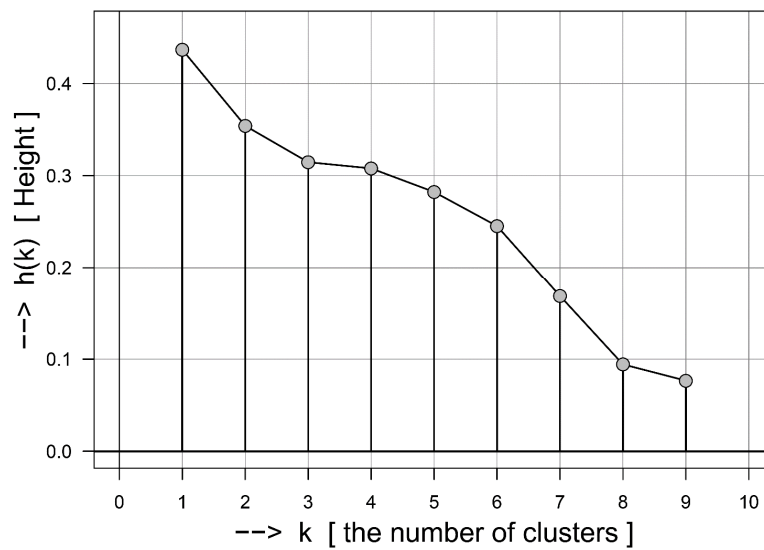


図 6.3 AGsi-curve

6.2 空間点パターンの分類への応用

第2章で示したように空間点パターンは大きく、ランダム型点配置・規則型点配置・集中型（凝集型）点配置の3種類に分類される。本研究ではグラフィカルな方法である AG-curve を用いて、観測された点

パターンを3種類のどれか1つに分類する。AG-curveを適用するのは第2章で紹介した以下の3データである。

- Japanese black pine saplings ⇒ ランダム型点配置
- Biological cells ⇒ 規則型点配置
- Redwood seedlings ⇒ 集中型（凝集型）点配置

これらは先行研究（Diggle, 1983, pp.1-2）により分類先が明らかにされている。

ここでは、以下に示す3種類のAG-curveを用いて空間点パターンを分類し、その結果を比較する。（）内は対応するAHC手法の名称である。最終的に最短距離法を利用したAGsi-curveの分類感度が3つの中で一番良好であることを示す。

- AGsi-curve（AHC手法：最短距離法）
- AGav-curve（AHC手法：群平均法）
- AGco-curve（AHC手法：最長距離法）

6.2.1 AG-curveを用いた空間点パターンの分類方法

観測された点配置をAG-curveを用いて分類する手順を以下にまとめて示す。

AG-curveを用いた空間点パターンの分類手順

- Step 1：乱数を用いて領域内にCSR点過程に従う n 個の点を m 回発生させる。
- Step 2：Step 1で得られた m 個の点配置に対してAG-curveを計算する。
- Step 3：Step 2で得られた m 本のAG-curveのエンベロープ（envelope）を描く。
- Step 4：観測された（分類対象の）点配置に対するAG-curveを計算する。
- Step 5：Step 4で得られたAG-curveをStep 3のエンベロープの上に描く。
- Step 6：観測されたAG-curveとCSR点過程に従うエンベロープを比較する。
- Step 7：下記に示す規準に従い、観測された点配置を分類する。

またAG-curveを用いた空間点パターンの分類規準を以下にまとめて示す。

AG-curve を用いた空間点パターンの分類規準

Case 1 : 観測された AG-curve がエンベロープ内に完全に収まる場合。

⇒ 観測された点配置をランダム型に分類する。

Case 2 : 観測された AG-curve がエンベロープからはみ出して水平線に近づく場合。

⇒ 観測された点配置を規則型に分類する。

Case 3 : 観測された AG-curve がエンベロープからはみ出して L (エル) 字型に近づく場合。

⇒ 観測された点配置を集中型(凝集型)に分類する。

6.2.2 AG-curve を用いた空間点パターンの分類結果

CSR 点過程に対応するエンベロープの作成は全て $m = 100$ 回で行った。AGsi-curve の結果を図 6.4 に、AGav-curve の結果を図 6.5 に、AGco-curve の結果を図 6.6 に示す。これら 3 種類の AG-curve による分類結果は、全て先行研究 (Diggle, 1983, pp.1-2) と一致する。一方、分類感度を比較するため 3 種類の AG-curve が CSR 点過程のエンベロープから逸脱する程度を観察する。規則型の点パターン (Biological cells) の場合、AGsi-curve (図 6.4 の中右) はグラフの左端と右端の部分でエンベロープからの逸脱が観察される。一方、AGav-curve (図 6.5 の中右) と AGco-curve (図 6.6 の中右) の逸脱は右端の部分だけに観察される。また集中型 (凝集型) の点パターン (Redwood seedlings) の場合、AGsi-curve (図 6.4 の右下) はグラフの中央と左側の部分でエンベロープからの逸脱が観察される。一方、AGav-curve (図 6.5 の右下) と AGco-curve (図 6.6 の右下) の逸脱は中央部分だけに観察される。これらの結果は AGsi-curve の分類感度が一番高いことを示している。なお、3 種類の AG-curve のグラフは縦軸のスケールが異なるので、図中に縦方向の矢印 (↓) を用いて 0.10 の幅を表示した。

分類感度の高さを比較すると、

$$\text{AGsi-curve} > \text{AGav-curve} > \text{AGco-curve}$$

となる。

以降では AGsi-curve に的を絞って集中的に議論する。

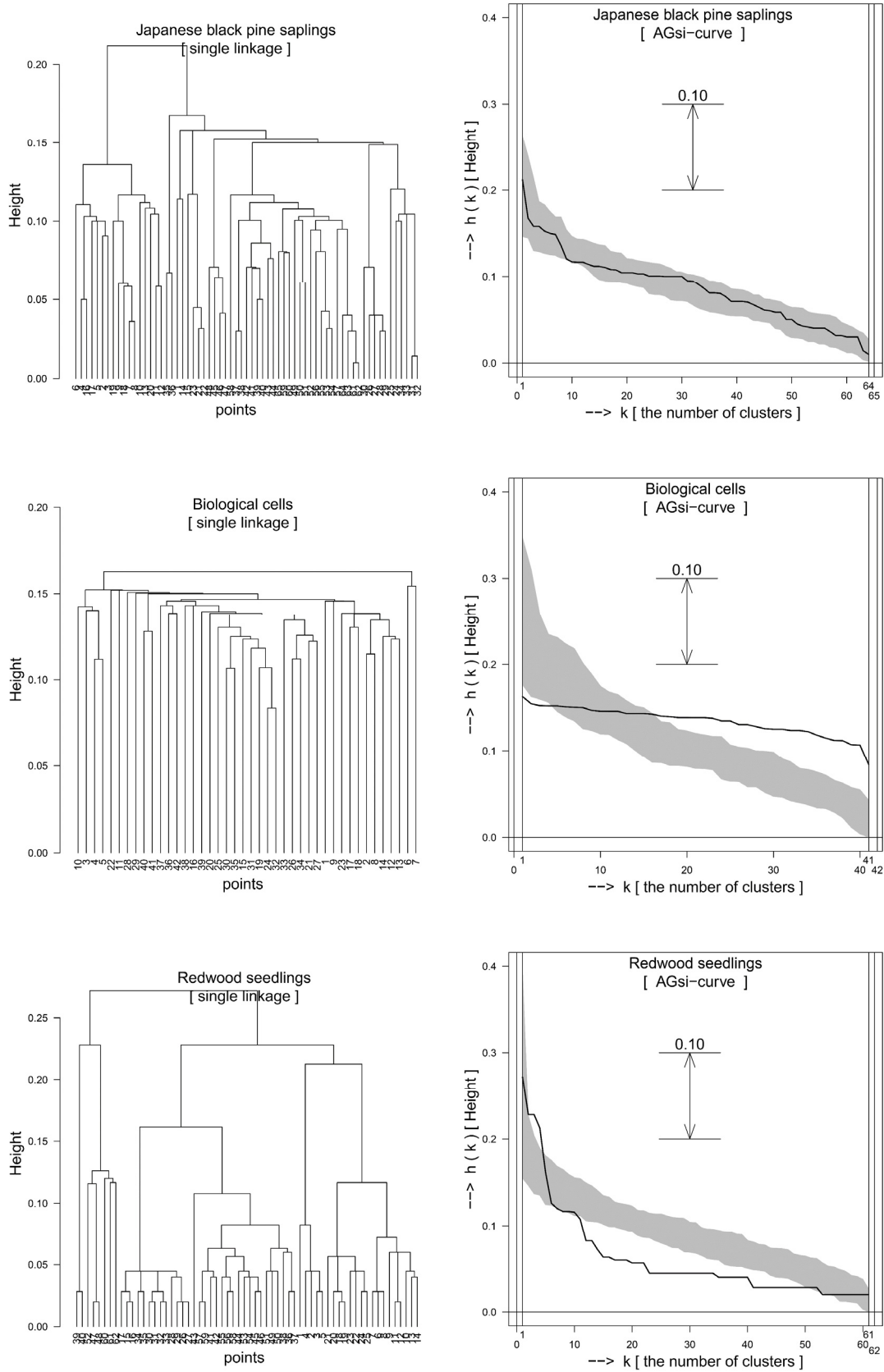


図6.4 AGsi-curveによる空間点パターンの分類

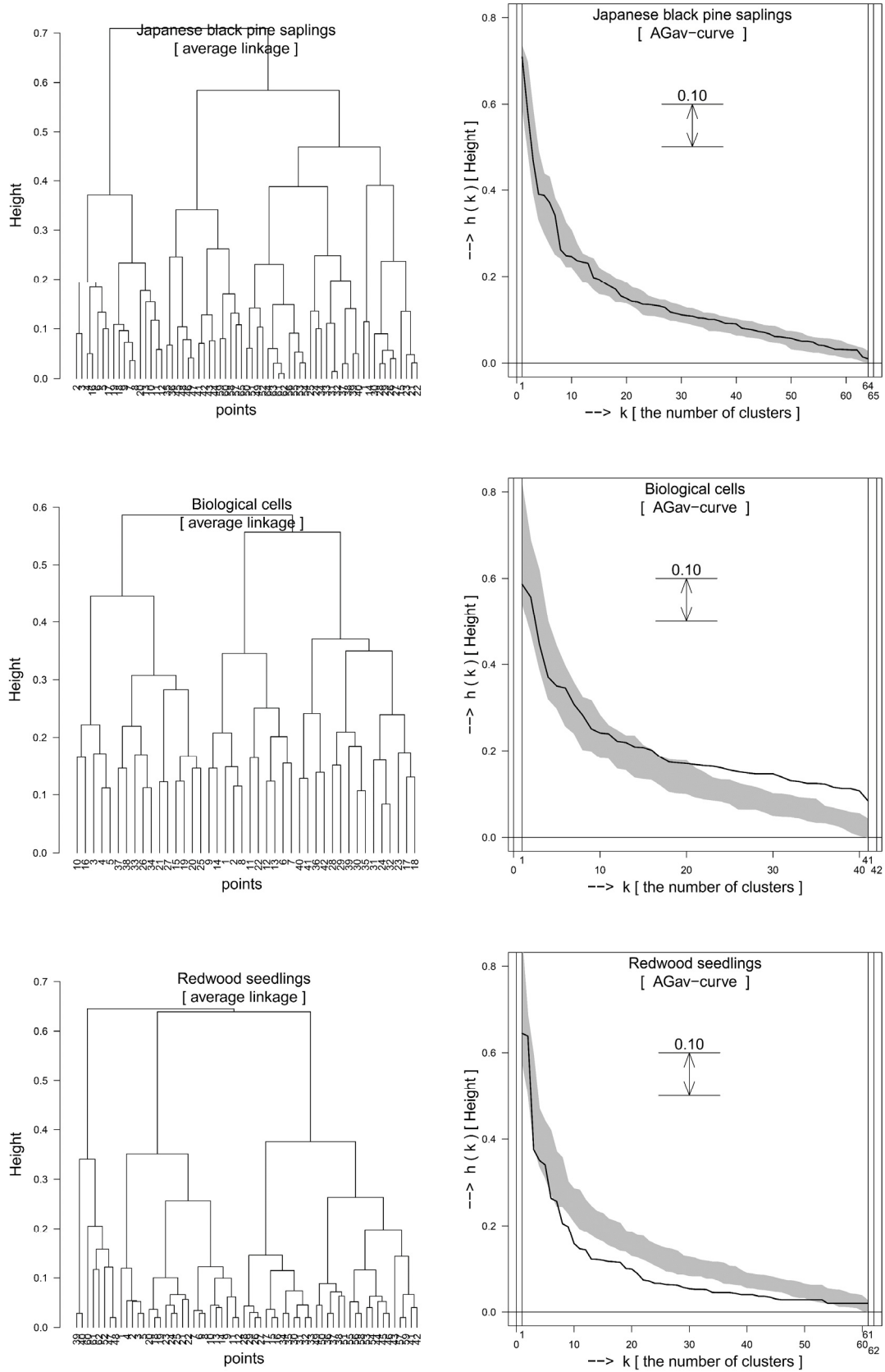


図6.5 AGav-curveによる空間点パターンの分類

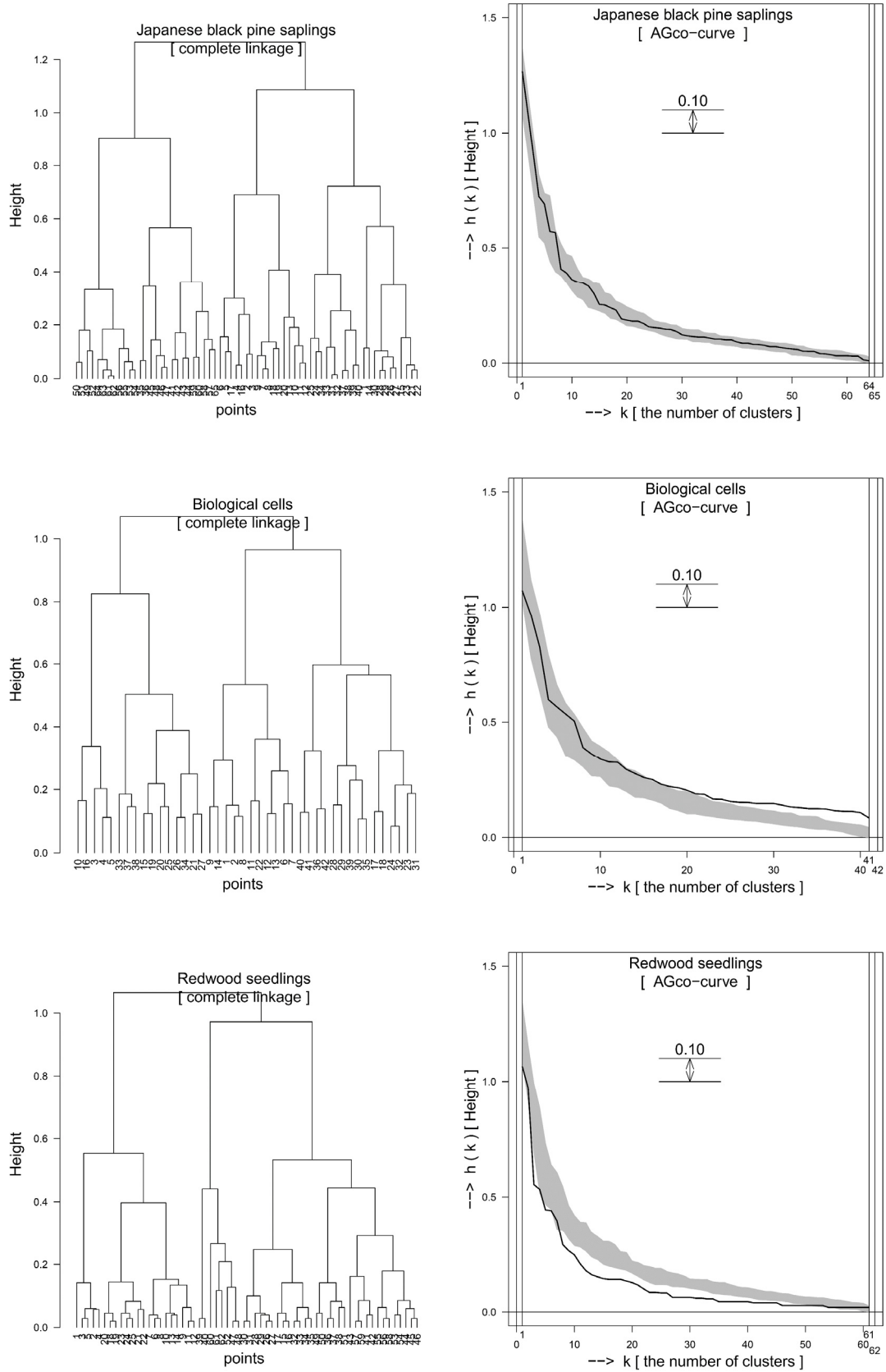


図 6.6 AGco-curve による空間点パターンの分類

6.3 AGsi-curve と既存のグラフィカルな手法との比較

4.3 節に於いて、G 関数と K 関数の問題点を示した。これらを AGsi-curve が、どのくらい改善しているのかを具体的な数値例（点パターン）を用いて示す。

6.3.1 G 関数との比較

AGsi-curve と G 関数を比較するため、以下のように人工的な 2 次元の点パターンを用意する。

人工的な点パターンの作成

- $l \times l$ 個の正方形の小ブロックを $B_1, B_2, \dots, B_{l \times l}$ と表す。
- これらの小ブロックに含まれる点の数を $N_1, N_2, \dots, N_{l \times l}$ とする。
- 各ブロックは 2 個以上の点を含むと仮定する。

$$N_1, N_2, \dots, N_{l \times l} \geq 2$$

ただし、これらは同じ個数である必要はない。

- これらの小ブロックを縦に l 個、横に l 個並べることによって、大きな正方形を構成する。
- このようにすると、小ブロックの配置を変えることにより、大きな正方形の領域内に異なる人工的な点パターンを数多く発生させることができる。

(以下では小ブロックを単にブロックと表記する)

また AGsi-curve と G 関数を比較するため、2 種類の最近隣点間距離 $Dbb(B)$ と $Dwb(B)$ を以下のように定義する。 $Dbb(B)$ は隣接するブロック間の最近隣点間距離 (point-to-point nearest-neighbor distance between blocks) を、 $Dwb(B)$ は同一ブロック内の最近隣点間距離 (point-to-point nearest-neighbor distance within blocks) を意味する。

$$Dbb(B) = \min_{\substack{p \in B \\ q \in B^{adj}}} \|p - q\|,$$

ここで点 p はブロック B に属し、点 q は B に隣接するブロック B^{adj} に属する。また、 $\| \cdot \|$ はユークリッド距離を表す。

$$Dwb(B) = \min_{\substack{p, q \in B \\ p \neq q}} \|p - q\|,$$

ここで点 p と点 q は同一ブロックに属する異なる点である。また、 $\| \cdot \|$ はユークリッド距離を表す。

ここからは、これまでの準備を利用して、 G 関数の問題点を明らかにするために次の状況を考える。

$$D_{bb}(B) > D_{wb}(B)$$

この状況下では、大きな正方形内に於ける最近隣点同士のリンクは、同一ブロック内の2点間でしか発生しない（隣接するブロック間でリンクは発生しない）。これはブロックの配置（回転も含める）に依存しないので、最近隣点間距離は不変である。

ここで、大きな正方形内の点パターンに対する G 関数と AG_{si} -curve を考える。 G 関数は第1近隣木の辺長の経験分布関数であった (4.1.1 節)。第1近隣木の辺長とは各点からみた最近隣点までの距離のことである。つまり、ブロックの配置を変えて様々な点パターンを発生させても、 $D_{bb}(B) > D_{wb}(B)$ である限り G 関数は不変である。一方、 AG_{si} -curve は AHC 手法の1つである最短距離法 (5.5.1 節) を利用している。最短距離法は MST の辺長を併合距離として用いるので、 AG_{si} -curve は領域内の全ての点をリンクすることができる。つまり、たとえ $D_{bb}(B) > D_{wb}(B)$ であったとしても、ブロックの配置を変えた様々な点パターンに対して異なる形状を示す。

次の事例に進む前に、これまでの説明を可視化する。 $D_{bb}(B) > D_{wb}(B)$ という条件下で、点間に発生する第1近隣木と MST のリンクを簡単な図形を用いて比較する。内部に3点を配置させた 2×2 個の正方形の小ブロック B_1, B_2, B_3, B_4 を用意する。これらを組み合わせて2種類の大きな正方形を作成し、点パターン $SPP1, SPP2$ とする。2種類の点パターンに対する第1近隣木を図 6.7 に、MST を図 6.8 に示す。

図 6.7 と図 6.8 を比較したときに、見かけ上の大きな違いは、第1近隣木が全点をリンクしないのに対して、MST が全点をリンクすることである。さらに辺長（点間距離）のばらつきを、第1近隣木と MST で比較すると以下ようになる。なお、第1近隣木で両側に矢印のある辺は2本が重なっている。また、第1近隣木の辺数は点数と同じだが、MST の辺数は点数より1少ない。

- SPP1 の第1近隣木 : 0.10×8 本、 0.15×4 本
- SPP2 の第1近隣木 : 0.10×8 本、 0.15×4 本
- SPP1 の MST : 0.10×4 本、 0.15×4 本、 0.35×1 本、 0.40×2 本
- SPP2 の MST : 0.10×4 本、 0.15×4 本、 0.35×3 本

第1近隣木では SPP1 と SPP2 の辺長のばらつきは同じであるが、MST では SPP1 と SPP2 の辺長のばらつきは異なっている。言い換えれば、第1近隣木は辺長のばらつきを用いて SPP1 と SPP2 を区別することはできないが、MST は辺長のばらつきを用いて SPP1 と SPP2 を区別することができることになる。

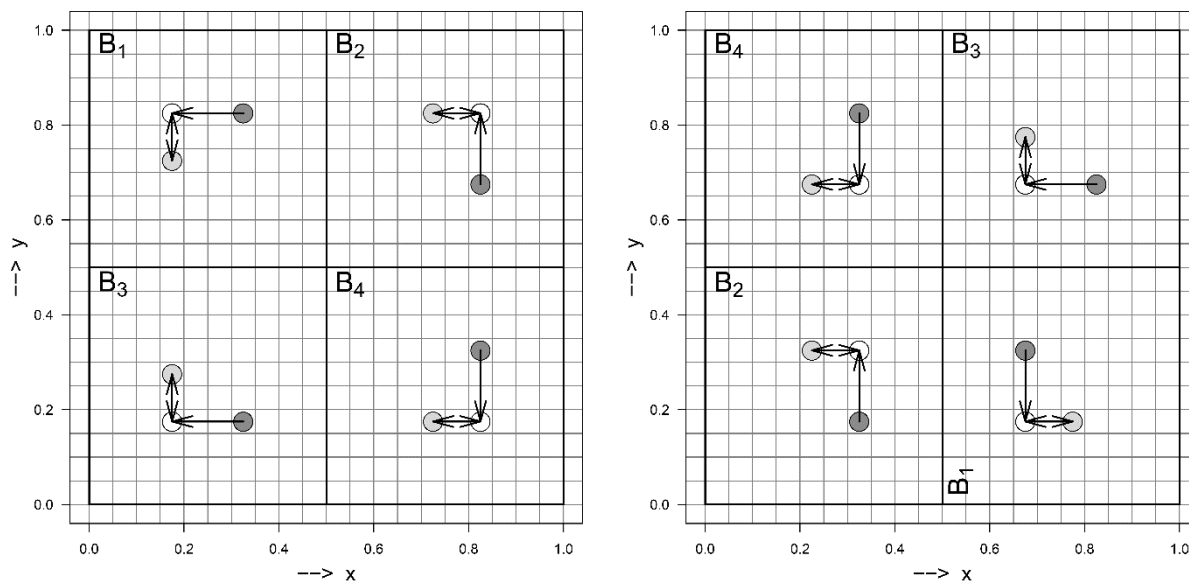


図 6.7 第1近隣木 (左 : SPP1、右 : SPP2)

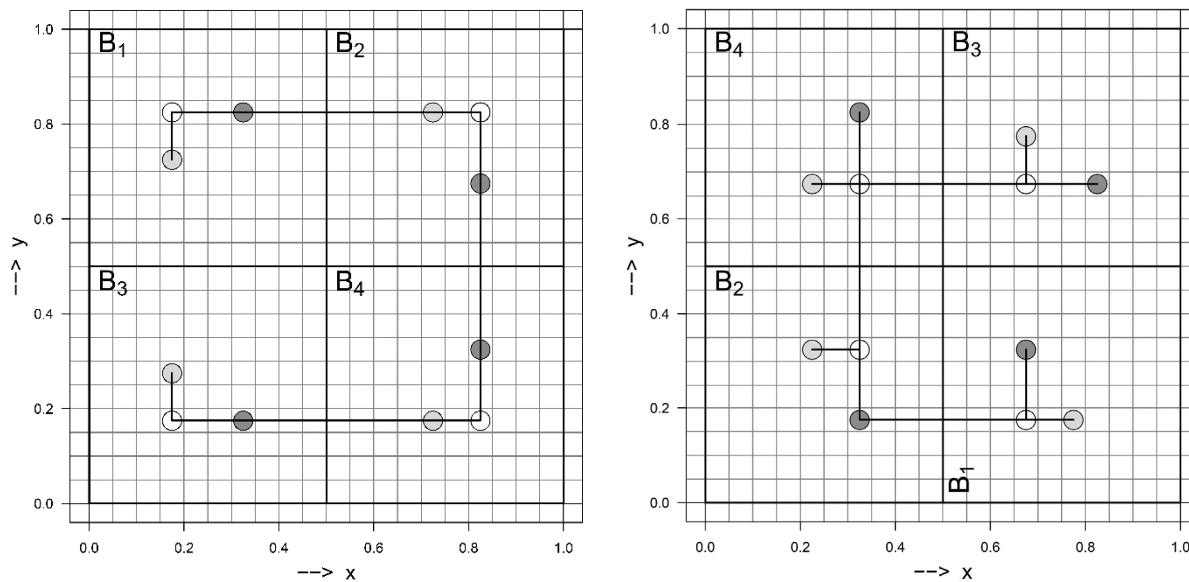


図 6.8 MST (左 : SPP1、右 : SPP2)

これまでに G 関数には区別できない点パターンが存在することが明らかになった。これを具体的な数値例で示す。 $l = 4$ として 16 個のブロックを考える。12 個のブロック (B_1, B_2, \dots, B_{12}) は 2 個の点を含み、4 個のブロック ($B_{13}, B_{14}, B_{15}, B_{16}$) は 19 個の点を含むとする。また、これらは条件 $\text{Dbb}(B) > \text{Dwb}(B)$ を満たしているものとする。これら 16 個のブロックの配置を変えて 2 つの点パターン SPP1 (図 6.9) と SPP2 (図 6.10) をつくる。ここでは分かり易さを優先し回転は考えない。

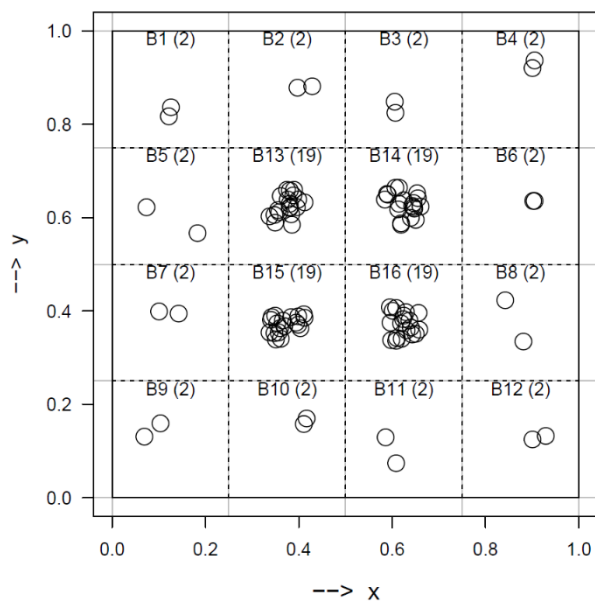


図 6.9 SPP1

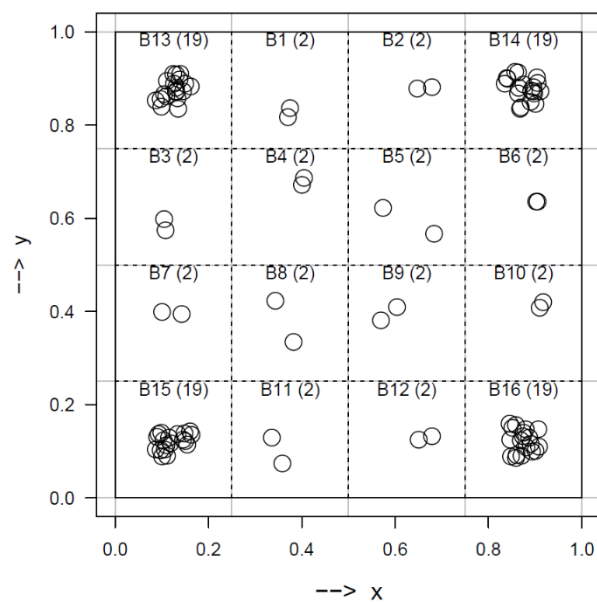


図 6.10 SPP2

SPP1 と SPP2 に対する G 関数と AGsi-curve を求める。図 6.12 は G 関数の結果だが、SPP1 と SPP2 は全く同じ形状になっている。一方、図 6.11 に示す AGsi-curve の結果は SPP1 と SPP2 が異なった形状を示している。G 関数では SPP1 と SPP2 を区別できないが、AGsi-curve を用いれば、SPP1 と SPP2 を区別できることが示された。

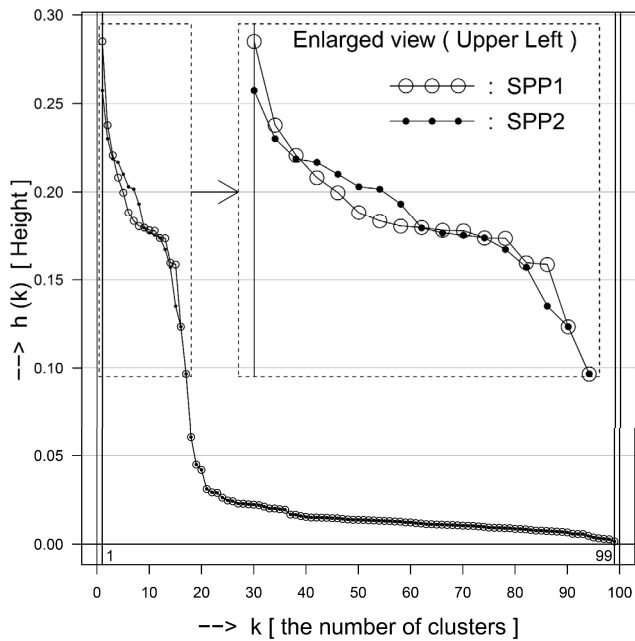


図 6.11 AGsi-curve

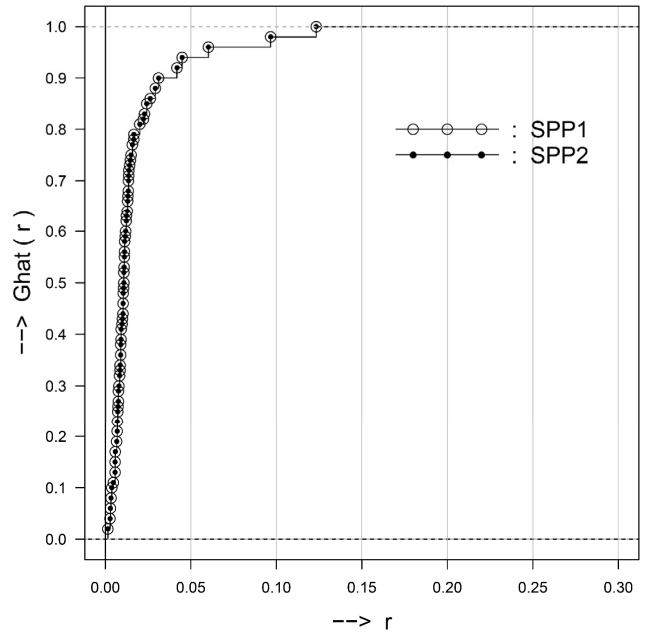


図 6.12 G 関数 Ghat(r)

6.3.2 K 関数との比較

K 関数の問題点の 1 つとして、強度 λ が大きくなるにつれて、規則型の点配置とランダム型の点配置の区別が曖昧になると述べた (4.3.2 節)。この観点で AGsi-curve と K 関数を比較するため、第 2 章において説明した人工的な 2 次元の点過程を 2 種類用いる。1 つは規則型の代表として SSI 点過程を、もう 1 つはランダム型の代表として CSR 点過程である。

単位正方形内に CSR 点過程に従う個数 200 ($\lambda = 200$) の点配置を 100 回発生 ($m = 100$) させ、これらに対応する AGsi-curve と $\hat{K}(r)$ の 100 本分のエンベロップを作図する。次に禁止半径 $r = 0.0375$ の SSI 点過程に従う個数 200 ($\lambda = 200$) の点配置を 1 回発生させ、これに対応する AGsi-curve と $\hat{K}(r)$ を求める。それぞれの CSR 点過程に対応するエンベロップの上に、SSI 点過程の AGsi-curve と $\hat{K}(r)$ を重ね描きする。禁止半径 $r = 0.0375$ の SSI 点過程に従う個数 200 ($\lambda = 200$) の点配置を図 6.13 に示す。

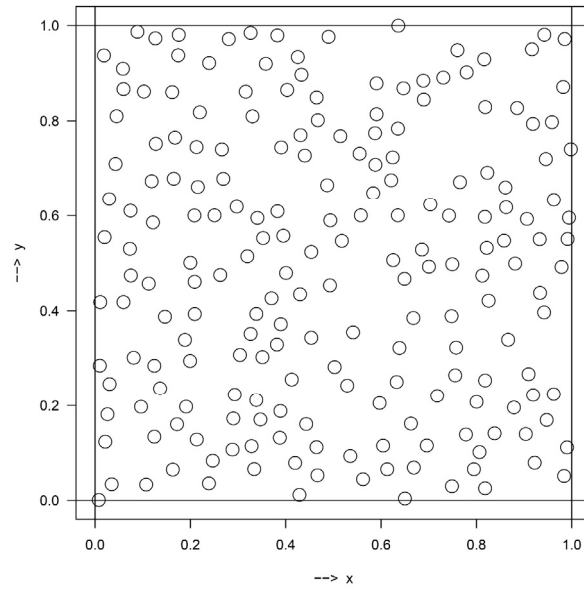


図 6.13 SSI 点過程に従う点配置 (強度 $\lambda = 200$ 、禁止半径 $r = 0.0375$)

$\hat{K}(r)$ の結果を図 6.14 に、AGsi-curve の結果を図 6.15 に示す。なお、両者の比較を正しく行うため、 $\hat{K}(r)$ の横軸の変域と AGsi-curve の縦軸の変域は $[0.00, 0.15]$ に合わせている。 $\hat{K}(r)$ の場合は $r = 0.0375$ の近傍で CSR のエンベロップからの逸脱が見られるが、すぐにエンベロップに吸収されてしまう。一方、AGsi-curve の場合は、左端と右端で CSR のエンベロップからの逸脱が見られる。エンベロップからの逸脱は AGsi-curve の方が分かり易い。

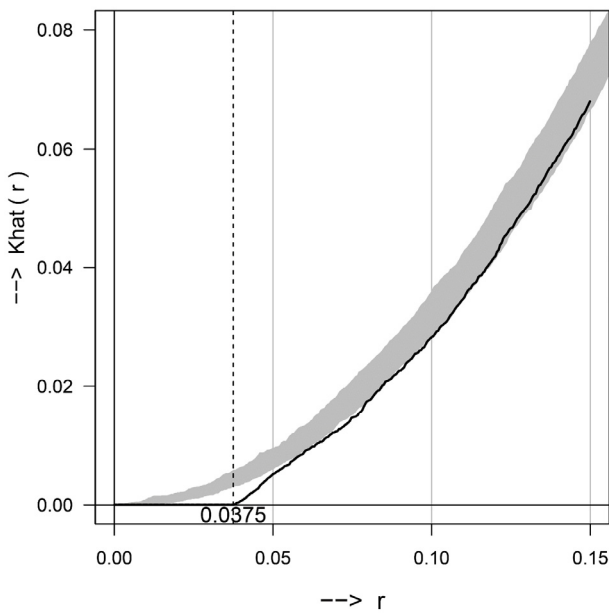


図 6.14 Khat(r)

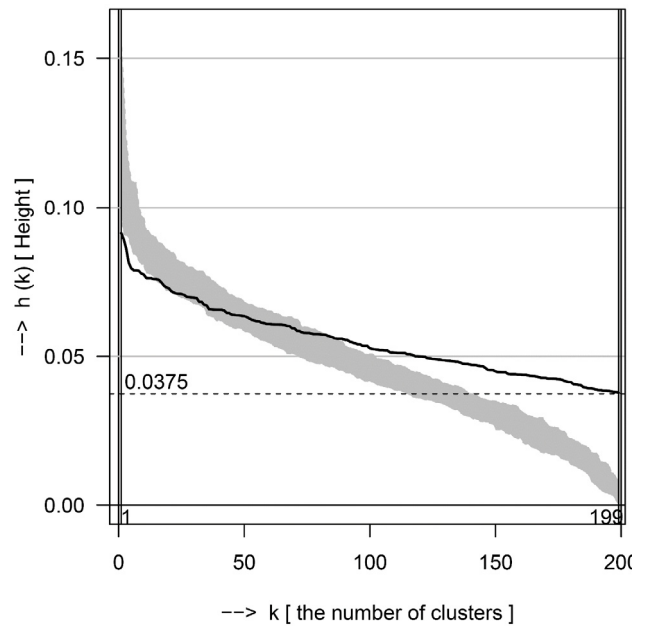


図 6.15 AGsi-curve

6.3.3 境界効果に対するロバスト性 (K 関数および G 関数との比較)

K 関数のもう 1 つの問題点として、(提案時から境界補正を含んでいることから、) 境界効果に対して極めて敏感であると述べた (4.3.2 節)。また、境界効果は領域内の点数が少ない場合と、領域の形状が細長い多角形の場合に顕著になると述べた (4.2.1.1 節)。ここでは領域の形状の違いを利用して、境界効果に対するロバスト性を調べることにする。参考のため、比較対象に G 関数も加える。Ripley (1976) の境界補正は、観測領域の外側も、内側と同じように点が分布していることが前提になっている。したがって、境界効果に対してはロバストであることが望ましい。

領域の形状として 2 種類の図形を用意する。1 つは面積が 1 で縦横比が 1 : 4 の長方形、もう 1 つは単位正方形である。これら 2 つの領域内に、CSR 点過程に従う個数 100 ($\lambda = 100$) の点パターンを 100 回発生 ($m = 100$) させ、2 つの形状に対応する境界補正無しのエンベロープを作成する。K 関数に関しては、比較のため境界補正有りのエンベロープを追加する。境界補正無しのエンベロープが長方形と正方形で一致すれば、境界効果に対してロバストであるという結論が得られる。正方形の領域を図 6.16 に、長方形の領域を図 6.17 に示す。

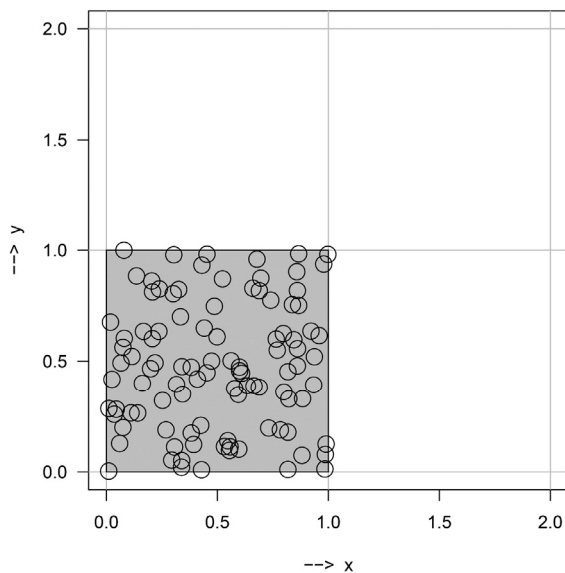


図 6.16 単位正方形

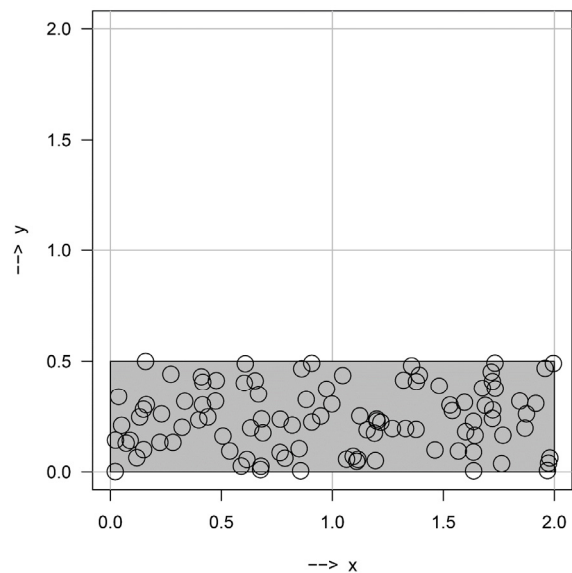


図 6.17 縦横比 1:4 の長方形 (面積は 1)

長方形と正方形のエンベロープを比較するため、両者を重ね描きする。K 関数の境界補正無しを図 6.18 に、追加した境界補正有りを図 6.19 に示す。また境界補正無しの G 関数を図 6.20 に、境界補正無しの

AGsi-curve を図 6.21 に示す。K 関数の境界補正無しは、長方形と正方形のエンベロップと期待値に明らかな差が見える。また長方形のエンベロップと期待値は、正方形に対して下方への偏りが発生している。これらの差は境界補正をすると、(当然のことだが) 殆んど解消される。一方、G 関数と AGsi-curve に於ける長方形と正方形の差は、エンベロップ・期待値ともに小さい。

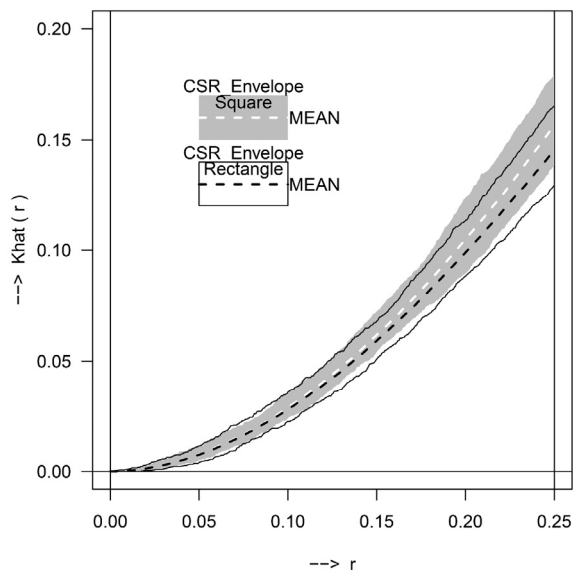


図 6.18 境界補正無し Khat(r)

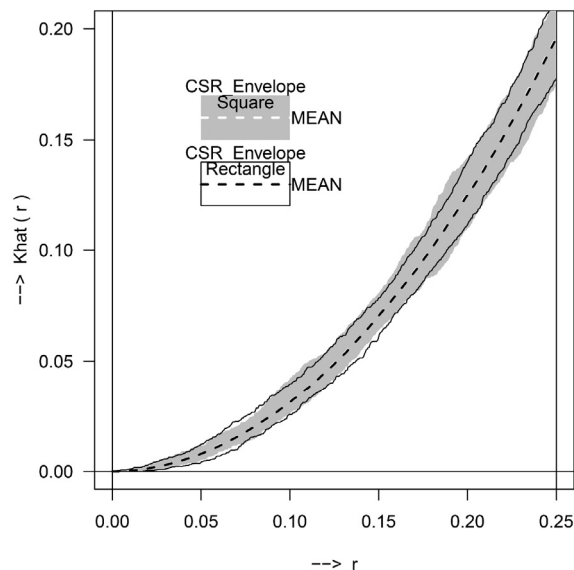


図 6.19 境界補正有り Khat(r)

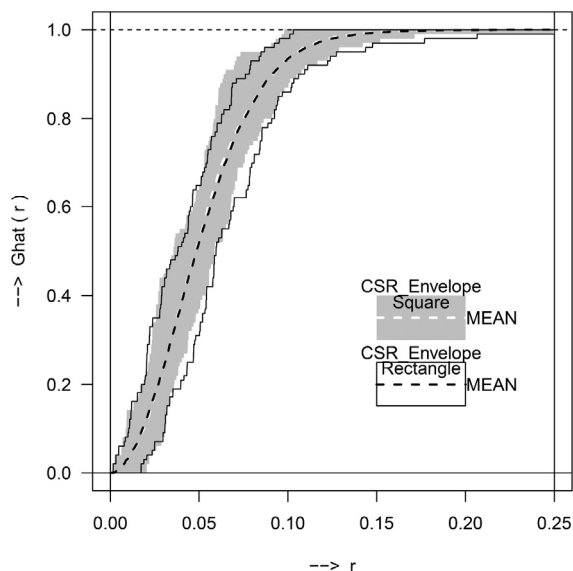


図 6.20 境界補正無し G 関数 Ghat(r)

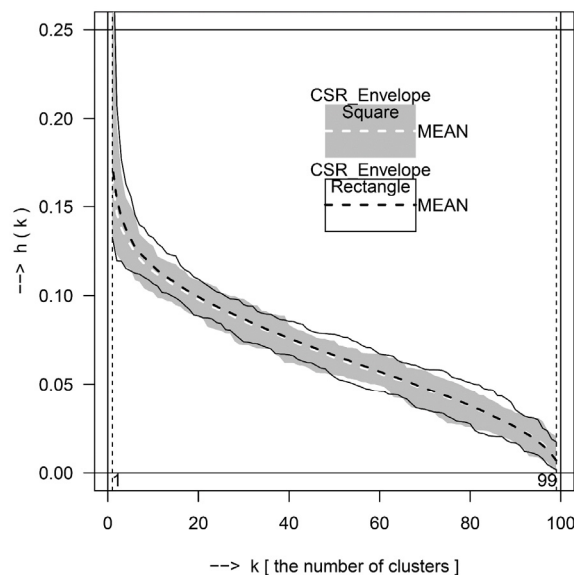


図 6.21 境界補正無し AGsi-curve

6.4 AGsi-curve と MST の関係

5.4.1 節に於いて、最短距離法の併合距離と MST の辺長が等価であることを述べた (Gower & Ross, 1969)。また、6.2 節に於いて、AGsi-curve は最短距離法のデンドログラムが持つ情報を利用していることを述べた。ここでは、AGsi-curve が MST の辺長の経験分布関数 $F_n(x)$ と等価であることを示す。併せて、両者の関係を確認するための簡単な数値例を示す。

6.4.1 AGsi-curve と MST の辺長の経験分布関数 $F_n(x)$ との関係

領域内に n 点が配置された事例を考える。これらの点を連結した MST の辺長 $D(k)$ を大きさの順に並べると以下のようなになる。

$$D(1) < D(2) < \dots < D(k) < \dots < D(n-2) < D(n-1)$$

一方、これらの点から求めた AGsi-curve の $h(k)$ を大きさの順に並べると以下のようなになる。

$$h(n-1) < h(n-2) < \dots < h(k) < \dots < h(2) < h(1)$$

Gower & Ross (1969) は両者の関係を以下のように示している。

$$h(k) = D(n-k) \quad k = 1, 2, \dots, n-1$$

したがって、 $h(k)$ と MST の辺長は単調に 1 対 1 の対応をしている。ただし逆順である。

直線で補完する前の AGsi-curve の点群の座標を以下のように表す。

$$\{(1, h(1)), (2, h(2)), \dots, (k, h(k)), \dots, (n-2, h(n-2)), (n-1, h(n-1))\}$$

MST の辺長の経験分布関数 $F_n(x)$ は、これらの座標を用いて以下のように構成される。

$$F_n(x) = \begin{cases} 0, & 0 \leq x < h(n-1) \\ \frac{1}{n-1}, & h(n-1) \leq x < h(n-2) \\ \frac{2}{n-1}, & h(n-2) \leq x < h(n-3) \\ \dots, & \dots \\ \frac{n-k}{n-1}, & h(n-k) \leq x < h(n-k-1) \\ \dots, & \dots \\ \frac{n-2}{n-1}, & h(2) \leq x < h(1) \\ \frac{n-1}{n-1} = 1, & h(1) \leq x \\ & 0 \leq x < \infty. \end{cases}$$

$F_n(x)$ が、 $\{h(1), h(2), \dots, h(k), \dots, h(n-2), h(n-1)\}$ の経験分布関数であることは明らかである。

逆に、 $F_n(x)$ が与えられているとき、その逆関数を以下のように定義する。

$$F_n^{-1}(u) = \inf\{x : F_n(x) \geq u\} \quad 0 \leq u \leq 1,$$

この逆関数を用いて、 $\{h(1), h(2), \dots, h(k), \dots, h(n-2), h(n-1)\}$ は以下のように計算することができる。

$$\begin{cases} h(1) = F_n^{-1}(1) \\ h(2) = F_n^{-1}\left(\frac{n-2}{n-1}\right) \\ \dots \\ h(k) = F_n^{-1}\left(\frac{n-k}{n-1}\right) \\ \dots \\ h(n-2) = F_n^{-1}\left(\frac{2}{n-1}\right) \\ h(n-1) = F_n^{-1}\left(\frac{1}{n-1}\right). \end{cases}$$

上記の結果から、点列 $\{(1, h(1)), (2, h(2)), \dots, (k, h(k)), \dots, (n-2, h(n-2)), (n-1, h(n-1))\}$ を求めることができ、これらの点列を順番に繋いでいけば、AGsi-curve が得られる。

6.4.2 AGsi-curve と MST の関係を確認するための簡単な数値例

単位正方形内に一様ランダムに 10 点を配置した点パターンを考える。この数値例に対して MST の辺長の経験分布関数と、AGsi-curve を求め比較する。点パターンに対する MST の辺長を図 6.22 に、MST の辺長の経験分布関数を図 6.23 に、AGsi-curve を図 6.24 に示す。MST の辺長の経験分布関数がジャンプす

る位置の横軸の値と、AGsi-curve が折れ曲がる位置の縦軸の値（最短距離法の各併合距離）が、それぞれ MST の各辺長と一致することが確認できる。MST の辺長の経験分布関数と AGsi-curve を比較するには、どちらかを 90 度回転すれば良いことが分かる。

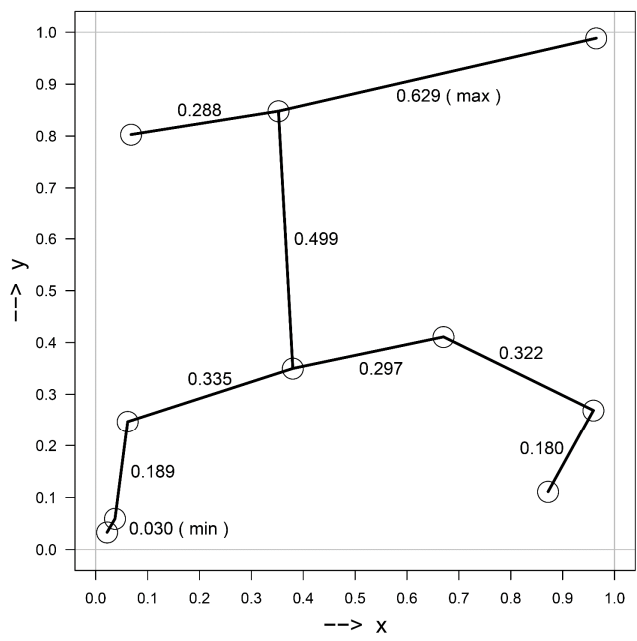


図 6.22 点パターンに対する MST の辺長

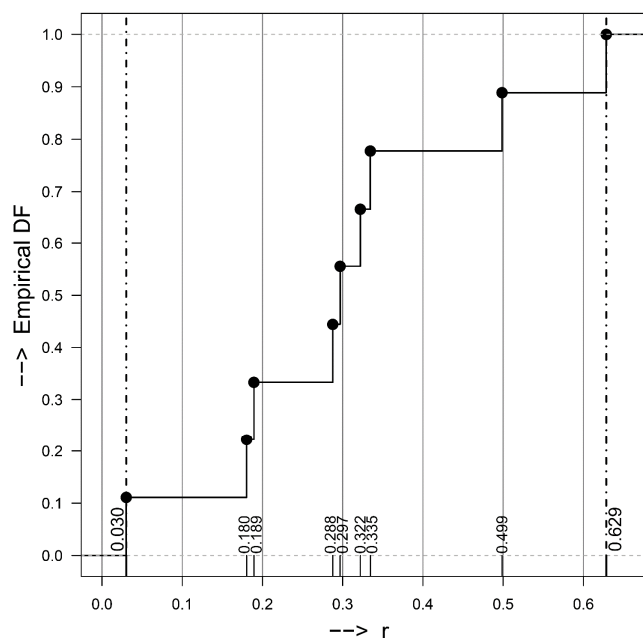


図 6.23 MST の辺長の経験分布関数

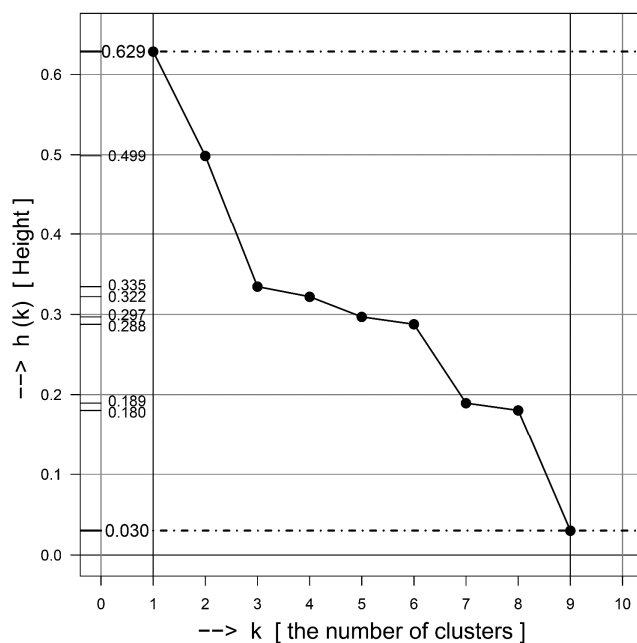


図 6.24 AGsi-curve

6.5 更に踏み込んだ比較

6.3 節で行った既存のグラフィカルな手法との比較は、Takai, Tamura & Motoyama (2017) に基づいている。ここでは、これまでとは異なる点パターンを用いて、AGsi-curve と既存のグラフィカルな手法との比較を行う。

6.5.1 集中型（凝集型）点配置の分類能力（親子点過程を用いた比較）

親子点過程の 1 つである Matérn Cluster 点過程を用いて単位正方形内に 3 種類の点パターンを発生させる。領域内の点数は確率変動するので、複数回の試行により点数を 100 に揃える。これらを用いて AGsi-curve と既存のグラフィカルな手法を比較する。

- Case 1 … クラスタ数 : 3 , $n = 100$
- Case 2 … クラスタ数 : 4 , $n = 100$
- Case 3 … クラスタ数 : 5 , $n = 100$

結果を図 6.25(A)と図 6.25(B)に示す。図 6.25(A)の第 1 列が用いた点パターン (Case 1, Case 2, Case 3) を、第 2 列が点パターンに対応する AGsi-curve を示す。また図 6.25(B)の第 1 列が点パターンに対応する G 関数を、第 2 列が K 関数を示す。どの手法もクラスタ数を増やすにしたがい、ランダム型に近づいていく。分類感度は K 関数が抜きん出て高い。一方、AGsi-curve と G 関数を用いても十分に分類できる。

分類感度の高さを比較すると、

$$K \text{ 関数} > \text{AGsi-curve} > G \text{ 関数}$$

となる。

参考として、AGsi-curve の基になった MST と Dendrogram を図 6.26 に示す。Dendrogram の適切な位置に水平線を引くことによりクラスタ数が分かる。

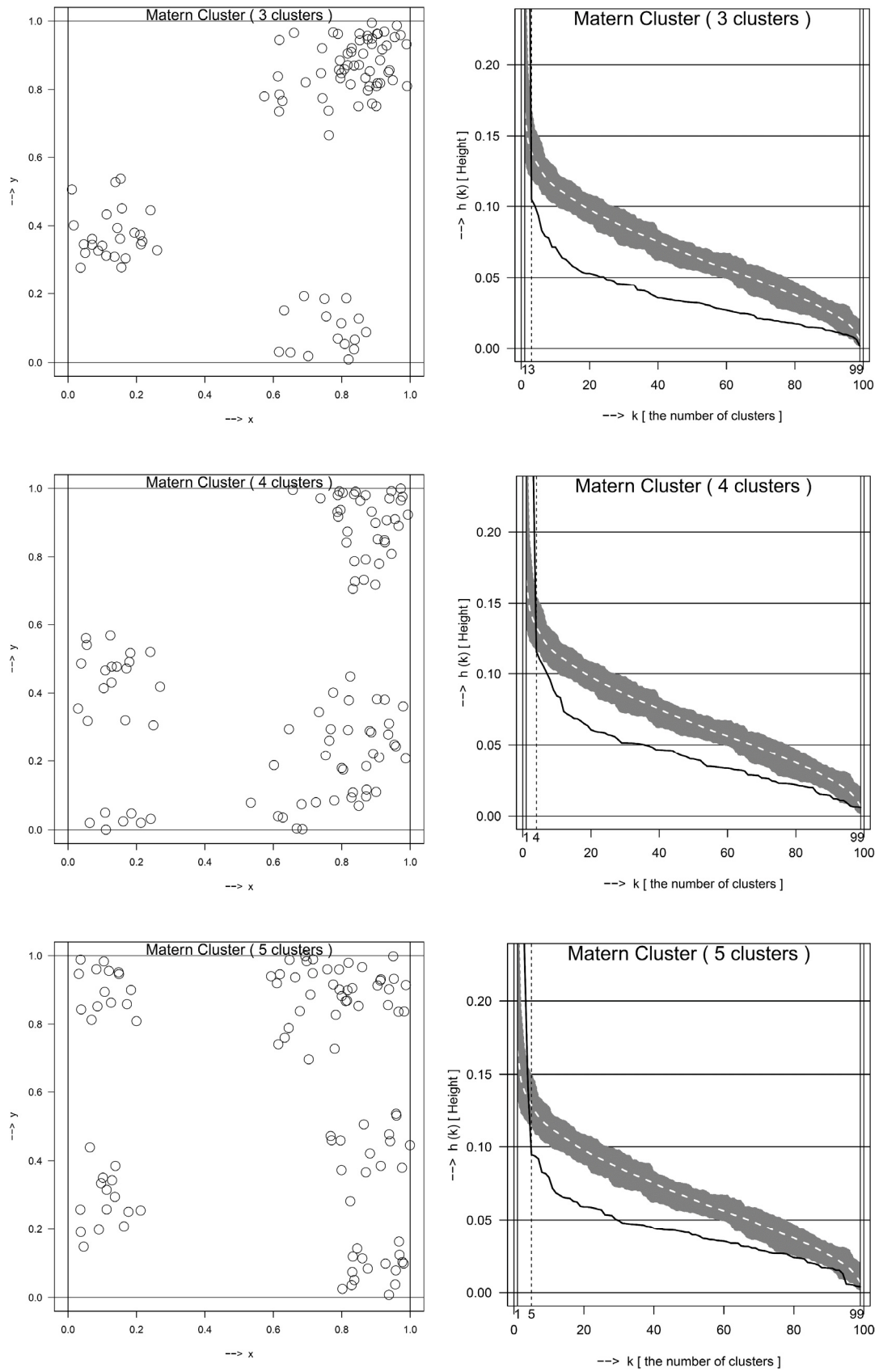


図 6.25(A) 集中型 (凝集型) 点配置の分類能力 (左: 点配置、右: AGsi-curve)

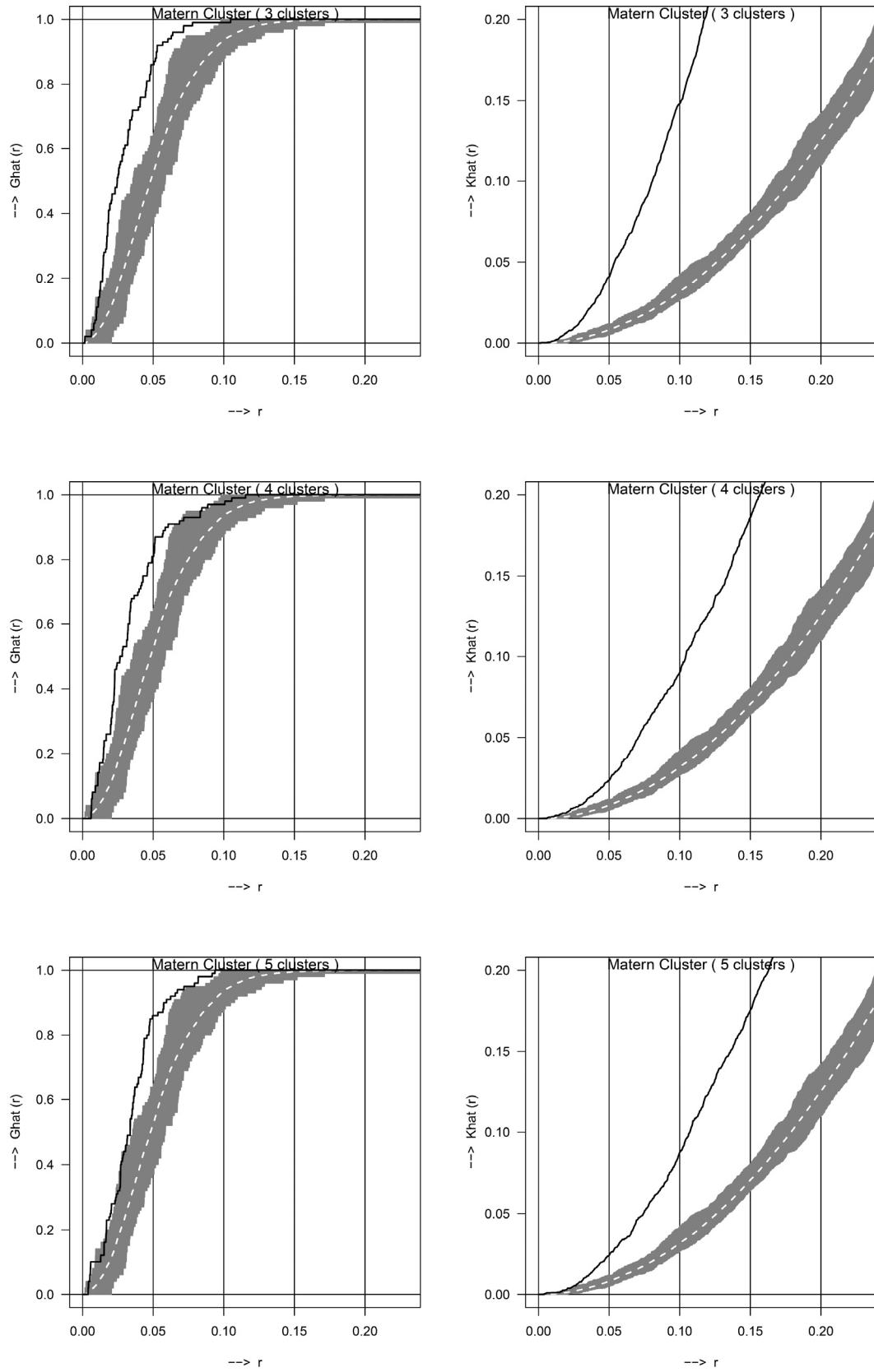


図 6.25(B) 集中型（凝集型）点配置の分類能力（左：G 関数、右：K 関数）

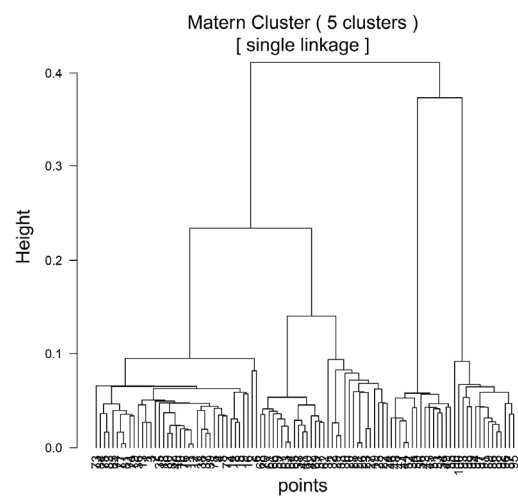
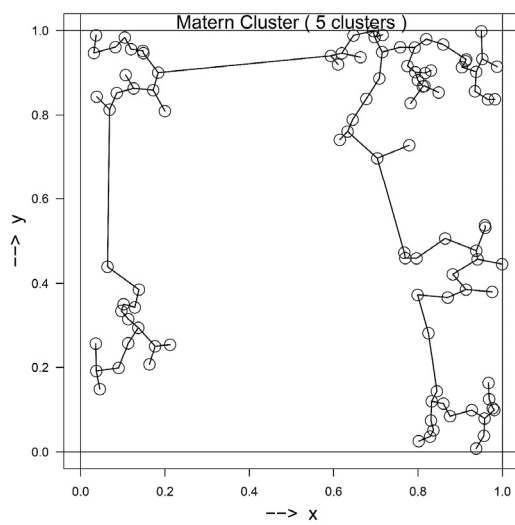
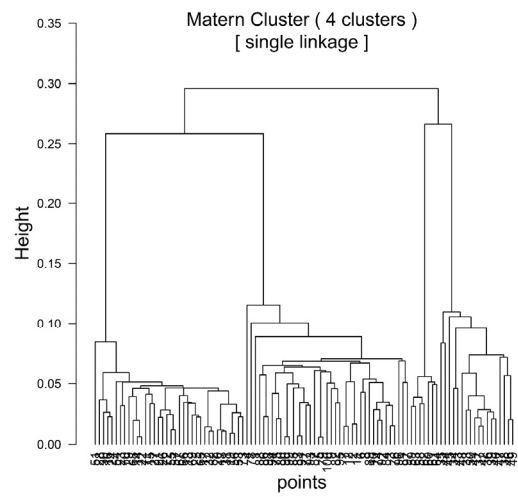
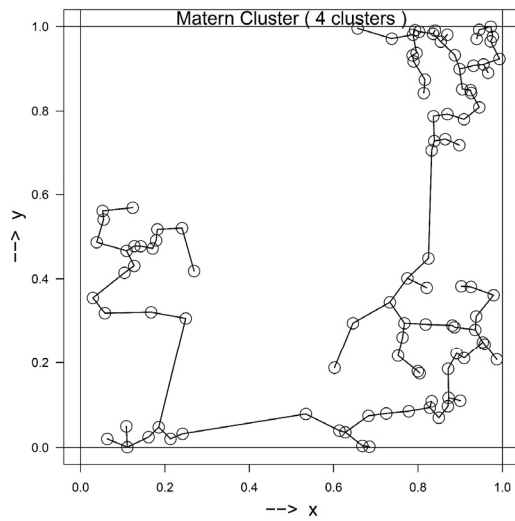
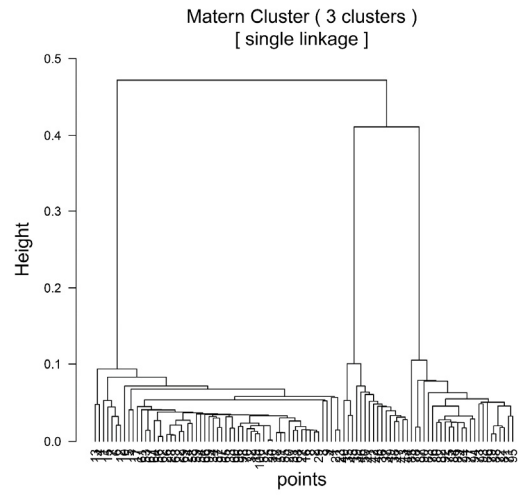
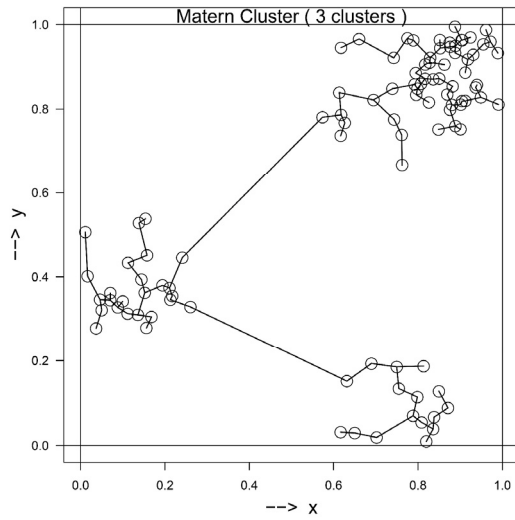


図 6.26 AGsi-curve の基になった MST (左) と Dendrogram (右)

6.5.2 規則型点配置の分類能力 (SSI 点過程を用いた比較)

SSI 点過程を用いて規則型の点パターンを発生させる。SSI 点過程の場合、強度 λ を固定し、禁止半径 (radius of inhibition)を小さくしていくと点パターンはランダム型に近づいていく。単位正方形内に発生させた以下の3種類の点パターンを用いて AGsi-curve と既存のグラフィカルな手法を比較する。

- Case 1 … 禁止半径 : 0.075, $\lambda = 100$
- Case 2 … 禁止半径 : 0.050, $\lambda = 100$
- Case 3 … 禁止半径 : 0.025, $\lambda = 100$

結果を図 6.27(A)と図 6.27(B)に示す。図 6.27(A)の第1列が用いた点パターン (Case 1, Case 2, Case 3) を、第2列が点パターンに対応する AGsi-curve を示す。また図 6.27(B)の第1列が点パターンに対応する G 関数を、第2列が K 関数を示す。どの手法も禁止半径を小さくするにしたがい、ランダム型に近づいていく。K 関数は Case 3 の場合をランダム型と分類してしまう。一方、AGsi-curve と G 関数は、かろうじて規則型の痕跡を残している。

分類感度の高さを比較すると、

$$\text{AGsi-curve} > \text{G 関数} > \text{K 関数}$$

となる。

参考として、AGsi-curve の基になった MST と Dendrogram を図 6.28 に示す。禁止半径が 0.075 の場合、MST の辺長は揃っているが、禁止半径が小さくなるにつれて MST の辺長はばらつくようになる。これと等価な現象が Dendrogram にも表れる。禁止半径が 0.075 の場合、Dendrogram の分岐位置は横並びだが、禁止半径が小さくなるにつれて分岐位置は上下にばらつくようになる。

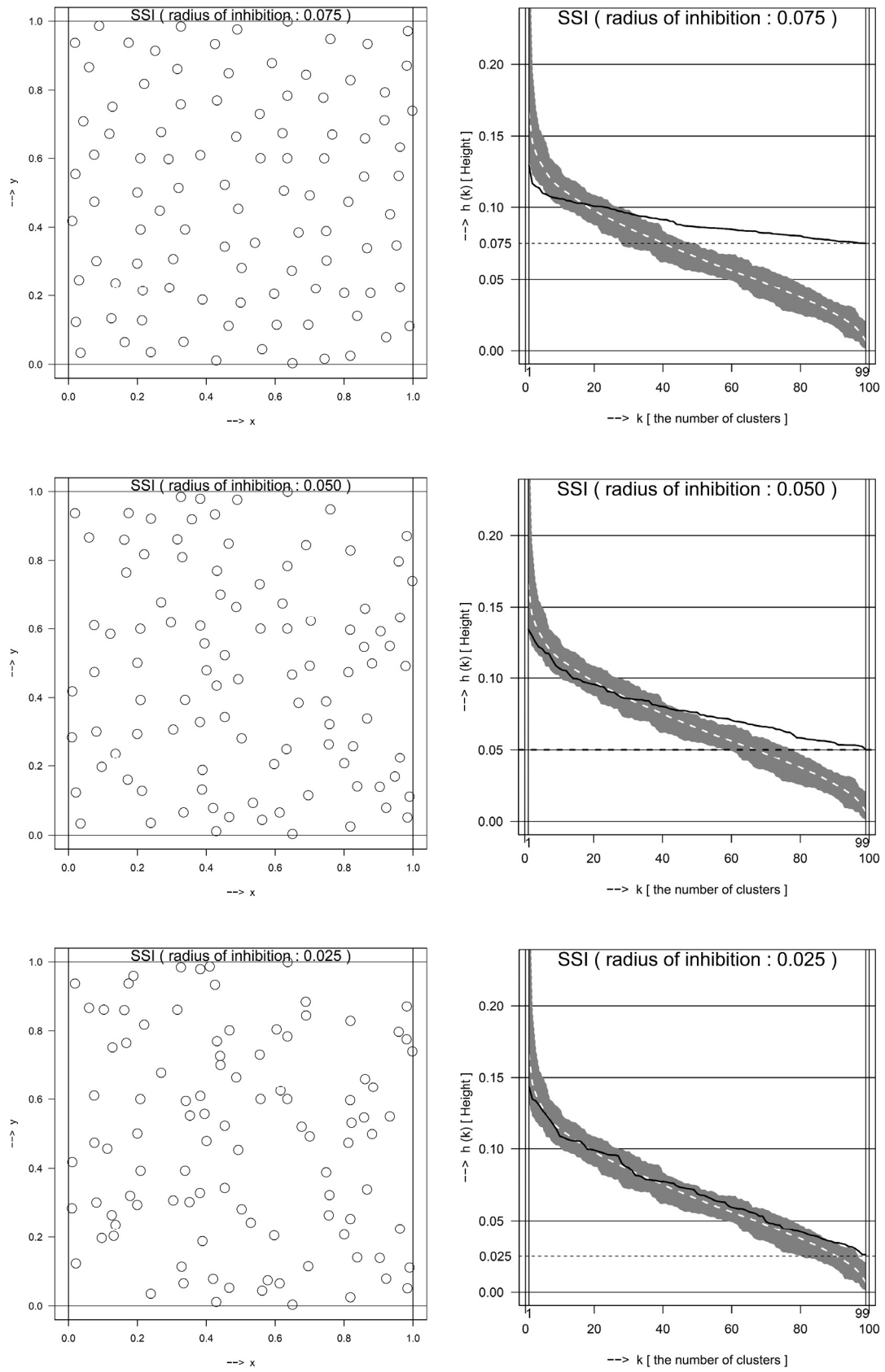


図 6.27(A) 規則型点配置の分類能力 (左: 点配置、右: AGsi-curve)

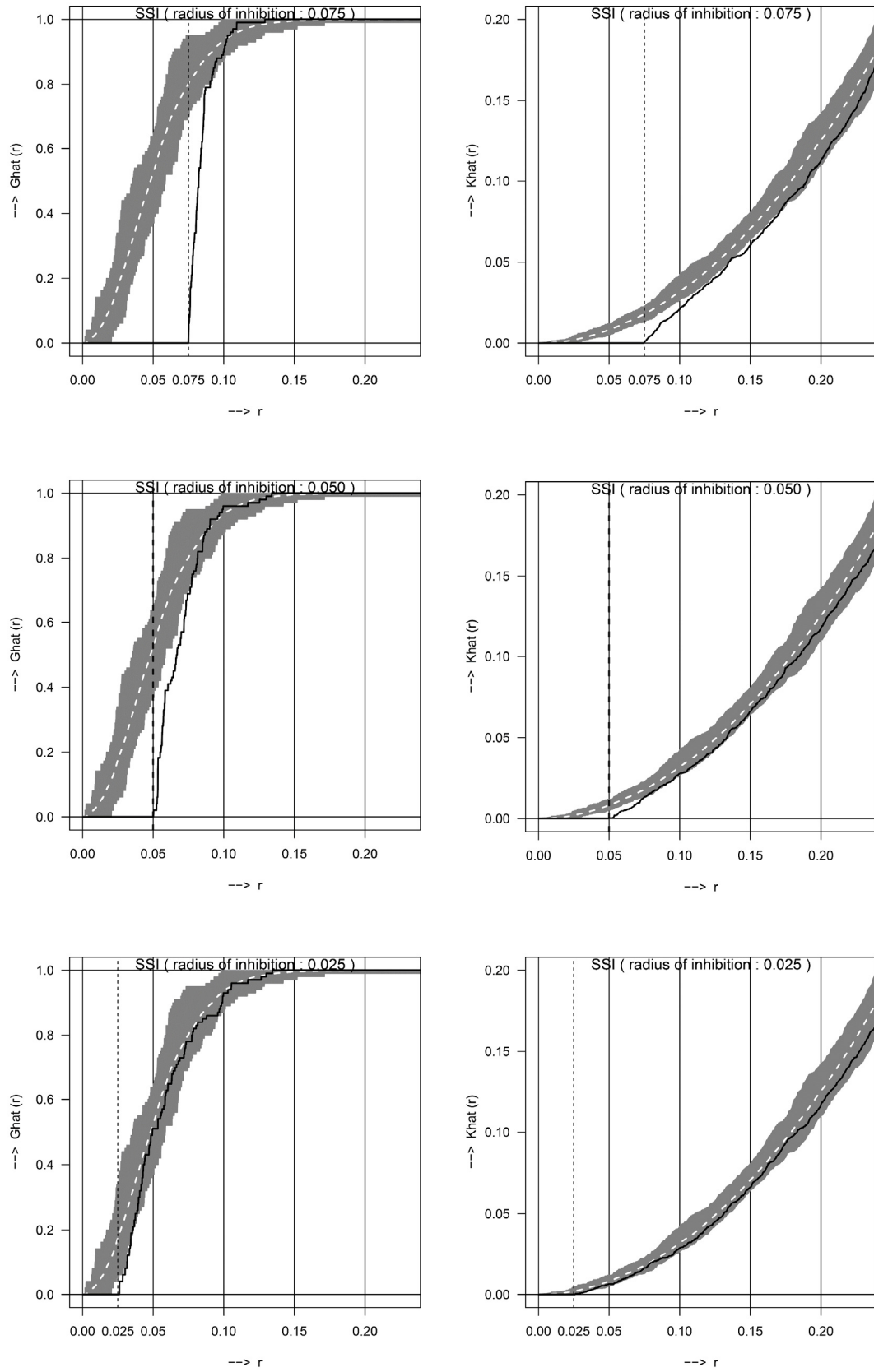


図 6.27(B) 規則型点配置の分類能力 (左: G 関数、右: K 関数)

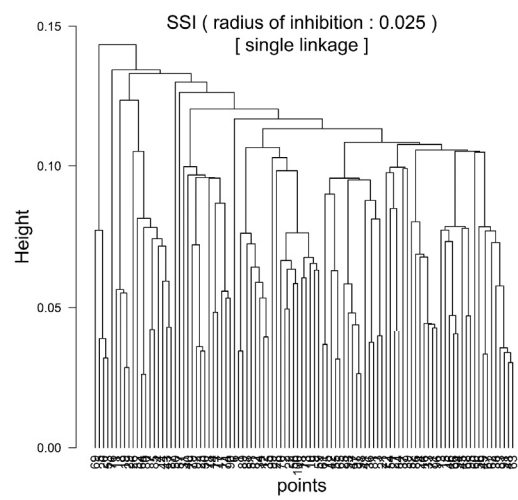
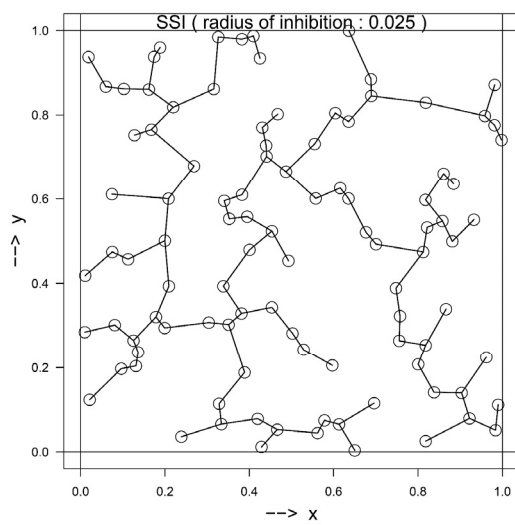
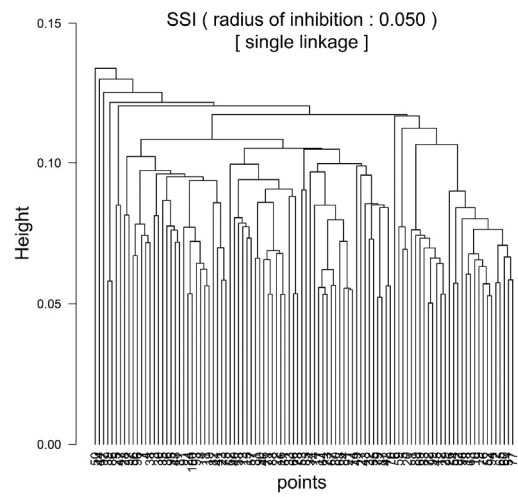
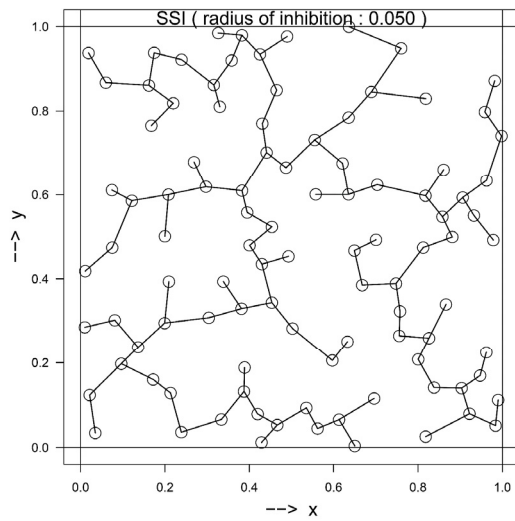
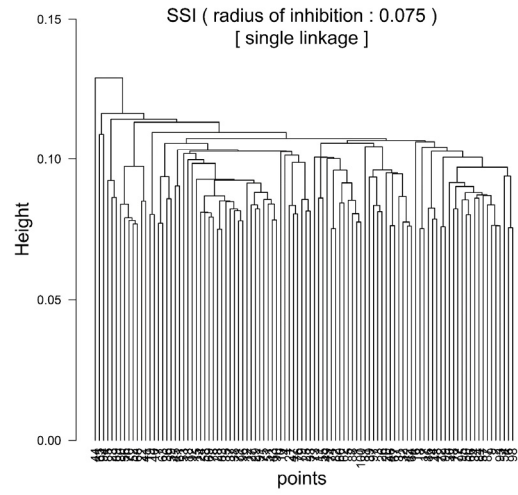
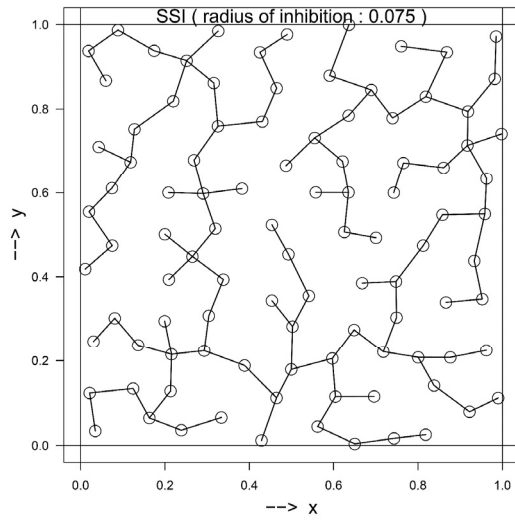


図 6.28 AGsi-curve の基になった MST (左) と Dendrogram (右)

6.6 【補足】 MST を空間点パターンの分類へ応用する

6.6.1 μ 関数 $\mu(r)$ の提案

AGsi-curve が MST の辺長の経験分布関数と等価であることを示した。したがって AGsi-curve を空間点パターンの分類に用いることができるのなら、MST の辺長の経験分布関数も空間点パターンの分類に用いることができる。これを μ 関数 $\mu(r)$ と呼ぶことにする。MST の辺長が従う確率分布を解析的に求めることはできないが、MST の探索アルゴリズムは非常に多く提案されている (Kruskal, 1956 など)。

MST の辺長の経験分布関数は以下のように定義される。これを用いて観測データから $\hat{\mu}(r)$ を推定することができる。等号に挟まれた表記は Diggle (1983, p.16) の G 関数の表記を踏襲したものである。

$$\hat{\mu}(r) = (n-1)^{-1} \#(\delta_i \leq r) = (n-1)^{-1} \sum_{\delta_i \leq r} 1$$

ここで、 n は領域内の点数、 δ_i は MST の辺長、 $\#$ は個数の意味である。ただし、領域内の点数が n であった場合の MST の辺数は $(n-1)$ であることに注意する。

一方、 $\hat{\mu}(r)$ は AGsi-curve の計算過程からも求めることができる。

$$\hat{\mu}(r) = (n-1)^{-1} \#(h_i \leq r) = (n-1)^{-1} \sum_{h_i \leq r} 1$$

ここで、 n は領域内の点数、 h_i は i 番目のクラスター併合時の距離、 $\#$ は個数の意味である。

観測された空間点パターンを、 $\hat{\mu}(r)$ を用いて分類する手順を以下にまとめて示す。

$\hat{\mu}(r)$ を用いた空間点パターンの分類手順

- Step 1 : 乱数を用いて領域内に CSR 点過程に従う n 個の点を m 回発生させる。
- Step 2 : Step 1 で得られた m 個の点配置に対して $\hat{\mu}(r)$ を計算する。
- Step 3 : Step 2 で得られた m 本の $\hat{\mu}(r)$ のエンベロープ (envelope) を描く。
- Step 4 : 観測された (分類対象の) 点配置に対する $\hat{\mu}(r)$ を計算する。
- Step 5 : Step 4 で得られた $\hat{\mu}(r)$ を Step 3 のエンベロープの上に描く。
- Step 6 : 観測された $\hat{\mu}(r)$ と CSR 点過程に従うエンベロープを比較する。
- Step 7 : 下記の規準に従い、観測された点配置を分類する。

また $\hat{\mu}(r)$ を用いた空間点パターンの分類規準を以下にまとめて示す。

$\hat{\mu}(r)$ を用いた空間点パターンの分類規準

Case 1 : 観測された $\hat{\mu}(r)$ がエンベロープ内に完全に収まる場合。

⇒ 観測された点配置をランダム型に分類する。

Case 2 : 観測された $\hat{\mu}(r)$ がエンベロープからはみ出して垂直線に近づく場合。

⇒ 観測された点配置を規則型に分類する。

Case 3 : 観測された $\hat{\mu}(r)$ がエンベロープからはみ出して Γ (ガンマ) 字型に近づく場合。

⇒ 観測された点配置を集中型(凝集型)に分類する。

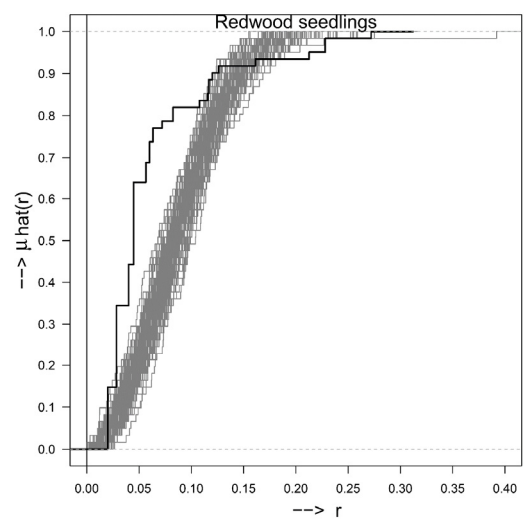
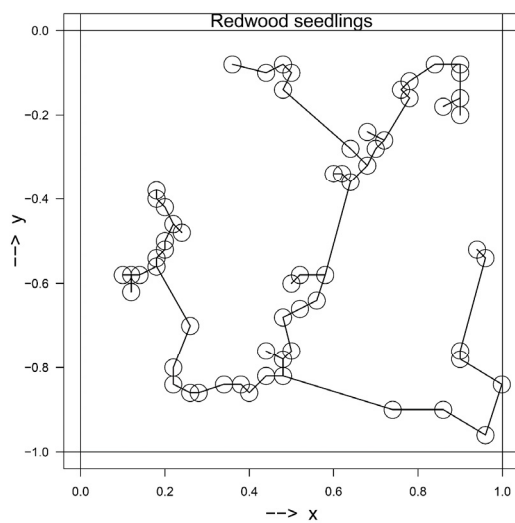
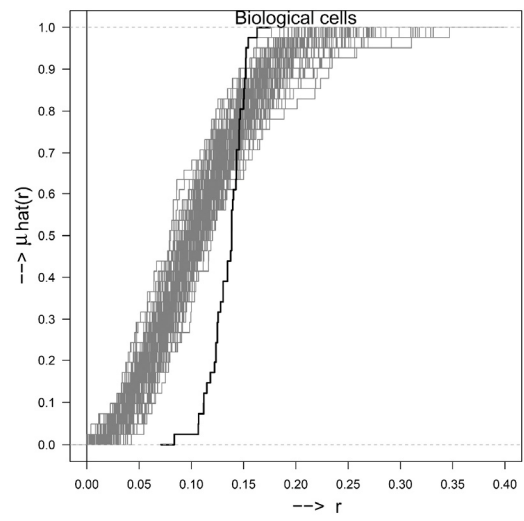
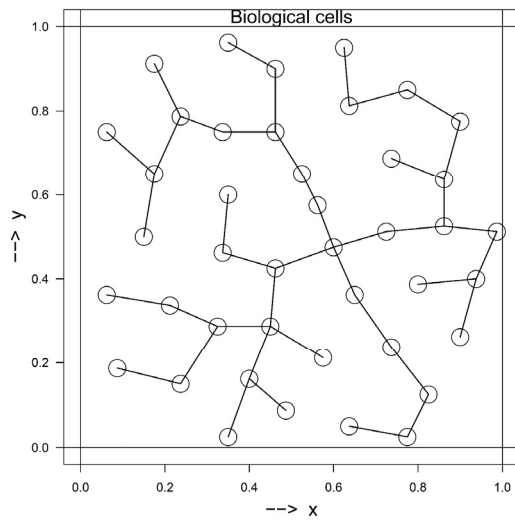
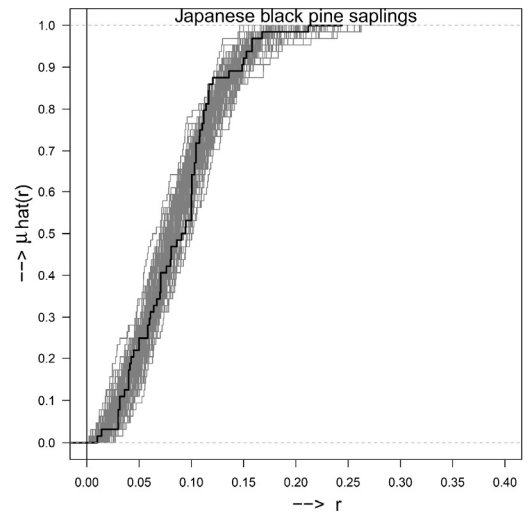
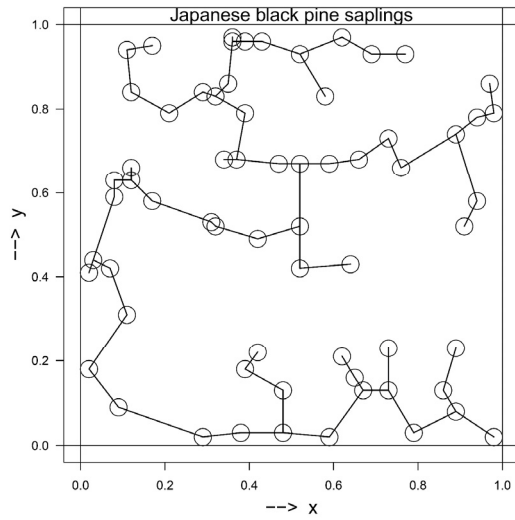
$\hat{\mu}(r)$ の分類規準は AGsi-curve の分類規準を 90 度回転したものである。

6.6.2 μ 関数 $\hat{\mu}(r)$ を用いた空間点パターンの分類

6.2 節と全く同じ空間点パターンを分類対象として、 μ 関数 $\hat{\mu}(r)$ を用いて空間点パターンの分類を行った。その結果を図 6.29 に示す。AGsi-curve の結果である図 6.4 と比較するには、どちらかを 90 度回転すれば良い。当然のことながら、図 6.29 と図 6.4 は同じ分類結果を示している。

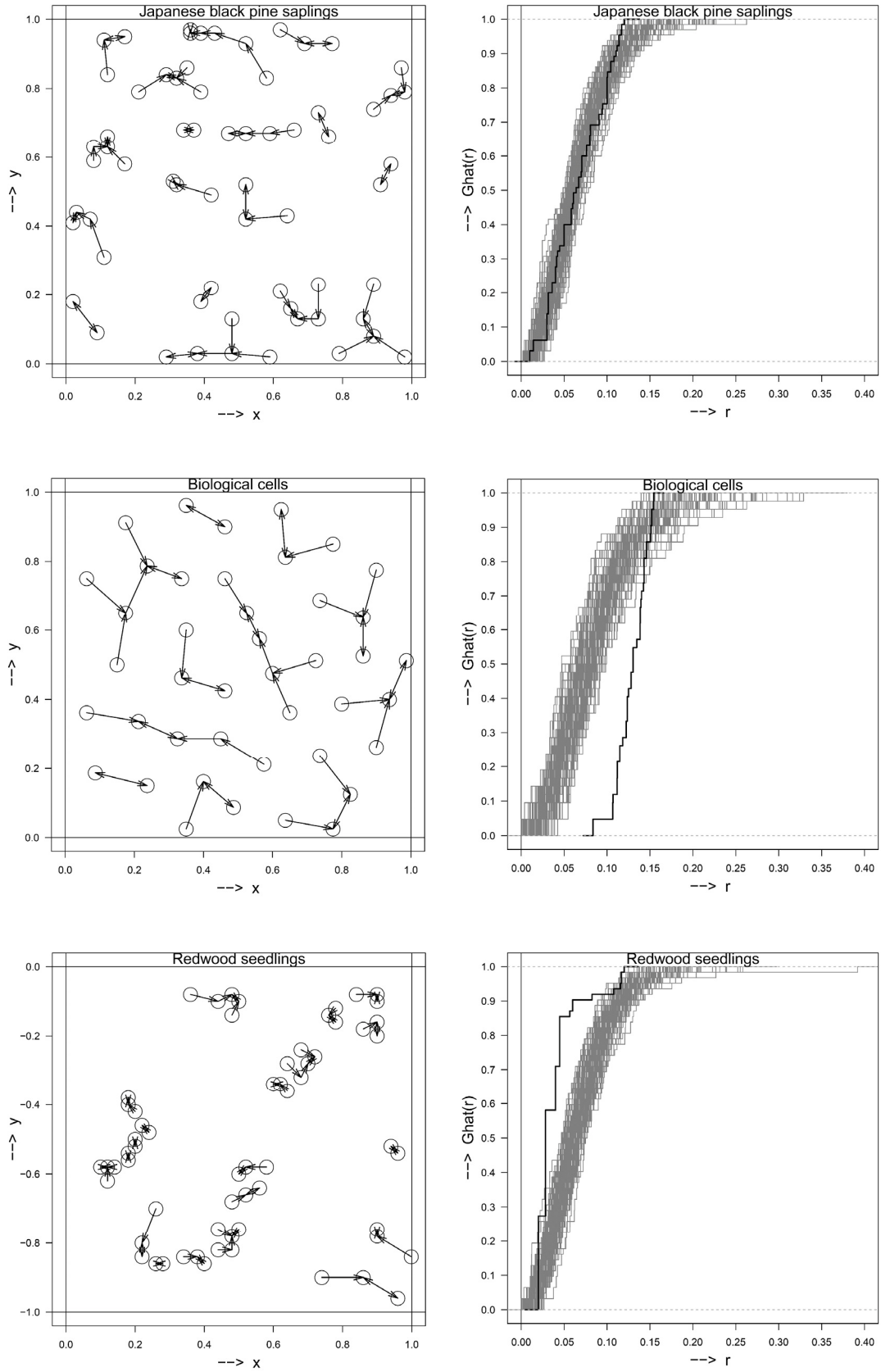
AGsi-curve の縦軸と μ 関数の横軸は完全に同じもので距離 r を表している。一方、AGsi-curve の横軸はクラスター数を表しているが、 μ 関数の縦軸は降順に並べ替えたクラスター数を区間 $[0, 1]$ に規格化したものと考えることが可能である。AGsi-curve の横軸は領域内の点数によって変化するが、 μ 関数の縦軸は領域内の点数によらず常に固定されている。AGsi-curve と μ 関数のどちらを選択するかは、応用分野に依存すると考えられる。

図 6.30 に示す G 関数による分類結果と比較する場合は、分布関数を基にした μ 関数を選択した方が適している。CSR 点過程 100 本分のエンベロープの幅を比較すると、 μ 関数の方が G 関数より狭くなっている。また分類対象に対する μ 関数の変化は G 関数より大きい。したがって総合的にみて、G 関数より μ 関数の方が分類精度が高いと言える。これは、G 関数と AGsi-curve を比較した結果と同じである。



注： μ 関数はエンベロープの内部が見えるように表示。

図 6.29 μ 関数を用いた空間点パターンの分類（左：MST、右： μ 関数）



注：G関数はエンベロープの内部が見えるように表示。

図 6.30 G関数を用いた空間点パターンの分類 (左：第1近隣木、右：G関数)

第7章 AGsi-curve の近似分布の導出

7.0 はじめに

第6章で AGsi-curve の振舞いは MST の辺長の順序統計量を用いて説明することができると述べた。したがって AGsi-curve の数理的性質を明らかにするためには、MST の辺長の順序統計量を求めれば良い。ところが MST の辺長が従う確率分布は解析的には求められていない。そのかわりに幾つかの近似分布が提案されている。本研究では Roberts (1968) の近似と Watanabe (2008) の近似を取り上げる。Watanabe (2008) によれば、Roberts (1968) の近似が MST の辺長の上限であり、Watanabe (2008) の近似が MST の辺長の下限である。これが正しければ AGsi-curve の近似分布を次のように導出することができる。

- AGsi-curve の近似分布 U :

MST の辺長に関する Roberts (1968) の近似分布の順序統計量を用いたもの。

- AGsi-curve の近似分布 L :

MST の辺長に関する Watanabe (2008) の近似分布の順序統計量を用いたもの。

この方針によって、7.1 の近似分布を導くことができる。ここでは多次元 (u 次元) 空間を対象とする。第7章の内容は、高井・田村・元山 (2018) に基づいている。

7.1 近似分布 U と近似分布 L

近似分布 U は命題 7.1、近似分布 L は次元の偶奇に応じて命題 7.2 ならびに命題 7.3 のようになる。全て MST の辺長の近似分布の第 j 順序統計量が基になっている。

【命題 7.1】 MST の辺長の分布として Roberts (1968) の近似を用いた近似分布 U の、

確率密度関数は (7.1)、原点まわりの s 次モーメントは (7.2) となる。

$$(7.1) \quad f_{(j)}^{R,u}(r) = D \sum_{l=0}^{j-1} u \lambda C r^{u-1} (-1)^l \binom{j-1}{l} \exp\{-(n-j+l)\lambda C r^u\}$$

$$(7.2) \quad E_{(j)}^{R,u}[r^s] = D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} (n-j+l)^{-\frac{s+u}{u}} (\lambda C)^{-\frac{s}{u}} \cdot \Gamma\left(\frac{s}{u} + 1\right)$$

$$\text{ここで、 } C = \frac{\pi^{\frac{u}{2}}}{2\Gamma\left(\frac{u}{2}+1\right)}$$

$$D = \frac{(n-1)!}{(j-1)!(n-1-j)!}$$

【命題 7.2】 MST の辺長の分布として Watanabe (2008) の近似を用いた偶数次元における、
近似分布 L の確率密度関数は (7.3)、原点まわりの s 次モーメントは (7.4) となる。

$$(7.3) \quad f_{(j)}^{W,2t}(r) = D \sum_{l=0}^{j-1} 2t \lambda C_e^{A^c \cap B} r^{2t-1} (-1)^l \binom{j-1}{l} \exp\{-(n-j+l) \lambda C_e^{A^c \cap B} r^{2t}\}$$

$$(7.4) \quad E_{(j)}^{W,2t}[r^s] = D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} (n-j+l)^{-\frac{s+2t}{2t}} (\lambda C_e^{A^c \cap B})^{-\frac{s}{2t}} \cdot \Gamma\left(\frac{s}{2t} + 1\right)$$

$$\text{ここで、 } C_e^{A^c \cap B} = \frac{V_{2t-1}(r)}{r^{2t-1}} \cdot \left\{ \frac{\sqrt{\pi} \Gamma\left(t + \frac{1}{2}\right)}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \right\}$$

$B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right)$ は不完全ベータ関数。

$$D = \frac{(n-1)!}{(j-1)!(n-1-j)!}$$

【命題 7.3】 MST の辺長の分布として Watanabe (2008) の近似を用いた奇数次元における、
近似分布 L の確率密度関数は (7.5)、原点まわりの s 次モーメントは (7.6) となる。

$$(7.5) \quad f_{(j)}^{W,2t+1}(r) = D \sum_{l=0}^{j-1} (2t+1) \lambda C_o^{A^c \cap B} r^{2t} (-1)^l \binom{j-1}{l} \exp\{-(n-j+l) \lambda C_o^{A^c \cap B} r^{2t+1}\}$$

$$(7.6) \quad E_{(j)}^{W,2t+1}[r^s] = D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} (n-j+l)^{-\frac{s+2t+1}{2t+1}} (\lambda C_o^{A^c \cap B})^{-\frac{s}{2t+1}} \cdot \Gamma\left(\frac{s}{2t+1} + 1\right)$$

$$\text{ここで、 } C_o^{A^c \cap B} = \frac{V_{2t}(r)}{r^{2t}} \cdot \left\{ \frac{\sqrt{\pi} \Gamma(t+1)}{\Gamma\left(t + \frac{3}{2}\right)} - B_{\frac{3}{4}}\left(t + 1, \frac{1}{2}\right) \right\}$$

$B_{\frac{3}{4}}\left(t + 1, \frac{1}{2}\right)$ は不完全ベータ関数。

$$D = \frac{(n-1)!}{(j-1)!(n-1-j)!}$$

7.2 命題の証明

ここでは、第 j 順序統計量の確率密度関数を導出したのち、第 j 順序統計量の原点まわりの s 次モーメントを導出する。第 j 順序統計量の確率密度関数 $f_{(j)}(r)$ は David & Nagaraja (2003, p.10) の式により求められるが、領域内の点数を n とすると、MST の辺長は $(n - 1)$ 本なので、本研究ではこれを (7.7) のように変形して用いる。

$$(7.7) \quad f_{(j)}(r) = \frac{(n-1)!}{(j-1)! \cdot (n-1-j)!} \cdot F(r)^{j-1} \{1 - F(r)\}^{n-1-j} f(r)$$

ここで、 $f(r)$ と $F(r)$ は、元の分布の確率密度関数と分布関数である。

また、第 j 順序統計量の原点まわりの s 次モーメント $E_{(j)}[r^s]$ は式 (7.8) を用いて求めることができる。

$$(7.8) \quad E_{(j)}[r^s] = \int_0^{\infty} r^s \cdot f_{(j)}(r) dr$$

7.2.1 命題 7.1 の証明

第3章の (3.8) 式において、Roberts (1968) の近似を用いた MST の辺長の確率密度関数を求めた。ここでは Watanabe (2008) と区別するため、右肩に添え字 R を追加して $f_u^R(r)$ と表記する。

$$(3.8) \quad f_u^R(r) = u\lambda \frac{\pi^{\frac{u}{2}}}{2\Gamma(\frac{u}{2} + 1)} r^{u-1} \cdot \exp\left(-\lambda \frac{\pi^{\frac{u}{2}}}{2\Gamma(\frac{u}{2} + 1)} r^u\right)$$

ここから分布関数 $F_u^R(r)$ は、

$$F_u^R(r) = 1 - \exp\left(-\lambda \frac{\pi^{\frac{u}{2}}}{2\Gamma(\frac{u}{2} + 1)} r^u\right)$$

と導出できる。

これらを式 (7.7) に代入することにより、第 j 順序統計量の確率密度関数 $f_{(j)}^{R,u}(r)$ を求めることができる。

$$f_{(j)}^{R,u}(r) = \frac{(n-1)!}{(j-1)! \cdot (n-1-j)!} \cdot F_u^R(r)^{j-1} \{1 - F_u^R(r)\}^{n-1-j} f_u^R(r)$$

ここでは以下のように、 C と D を導入し式の展開を簡略化する。

$$C = \frac{\pi^{\frac{u}{2}}}{2\Gamma(\frac{u}{2} + 1)}$$

$$D = \frac{(n-1)!}{(j-1)!(n-1-j)!}$$

したがって、 $f_{(j)}^{R,u}(r)$ は

$$\begin{aligned} f_{(j)}^{R,u}(r) &= D \cdot \{1 - \exp(-\lambda Cr^u)\}^{j-1} \cdot \{\exp(-\lambda Cr^u)\}^{n-1-j} \cdot u\lambda Cr^{u-1} \cdot \exp(-\lambda Cr^u) \\ &= u\lambda CD r^{u-1} \cdot \{1 - \exp(-\lambda Cr^u)\}^{j-1} \cdot \{\exp(-\lambda Cr^u)\}^{n-j} \\ &= u\lambda CD r^{u-1} \cdot \{1 - \exp(-\lambda Cr^u)\}^{j-1} \cdot \exp\{-(n-j)\lambda Cr^u\} \end{aligned}$$

となる。

ここで、先に $\{1 - \exp(-\lambda Cr^u)\}^{j-1}$ の2項展開を行うと、

$$\begin{aligned} \{1 - \exp(-\lambda Cr^u)\}^{j-1} &= 1 - \binom{j-1}{1} \cdot \exp(-\lambda Cr^u) + \binom{j-1}{2} \cdot \exp(-2\lambda Cr^u) \\ &\quad - \binom{j-1}{3} \cdot \exp(-3\lambda Cr^u) + \dots + (-1)^{j-1} \binom{j-1}{j-1} \cdot \exp\{-(j-1)\lambda Cr^u\} \end{aligned}$$

となる。

この結果を元の式に戻して整理すると、

$$\begin{aligned} f_{(j)}^{R,u}(r) &= u\lambda CD r^{u-1} [\exp\{-(n-j)\lambda Cr^u\} - \binom{j-1}{1} \cdot \exp\{-(n-j+1)\lambda Cr^u\} \\ &\quad + \binom{j-1}{2} \cdot \exp\{-(n-j+2)\lambda Cr^u\} - \binom{j-1}{3} \cdot \exp\{-(n-j+3)\lambda Cr^u\} \\ &\quad + \dots + (-1)^{j-1} \binom{j-1}{j-1} \cdot \exp\{-(n-1)\lambda Cr^u\}] \end{aligned}$$

となる。

さらに整理すると、第 j 順序統計量の確率密度関数 $f_{(j)}^{R,u}(r)$ は、

$$f_{(j)}^{R,u}(r) = D \sum_{l=0}^{j-1} u\lambda Cr^{u-1} (-1)^l \binom{j-1}{l} \exp\{-(n-j+l)\lambda Cr^u\}$$

ここで、 C と D は既出である。

となり、式(7.1)が導出できた。

次に $f_{(j)}^{R,u}(r)$ を式(7.8)に代入することにより、第 j 順序統計量の原点まわりの s 次モーメント $E_{(j)}^{R,u}[r^s]$ を求める。

$$E_{(j)}^{R,u}[r^s] = \int_0^{\infty} r^s \cdot f_{(j)}^{R,u}(r) dr$$

代入後、整理すると

$$\begin{aligned}
E_{(j)}^{R,u}[r^s] &= \int_0^\infty r^s \cdot D \sum_{l=0}^{j-1} u \lambda C r^{u-1} (-1)^l \binom{j-1}{l} \exp\{-(n-j+l)\lambda C r^u\} dr \\
&= D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} \int_0^\infty r^s u \lambda C r^{u-1} \exp\{-(n-j+l)\lambda C r^u\} dr
\end{aligned}$$

となる。

ここで $\tau = (n-j+l)\lambda C r^u$ として置換積分を行う。

$$\begin{aligned}
\frac{d\tau}{dr} &= (n-j+l)\lambda C u r^{u-1} \\
\frac{d\tau}{(n-j+l)} &= \lambda C u r^{u-1} dr \\
r &= \left\{ \frac{\tau}{(n-j+l)\lambda C} \right\}^{\frac{1}{u}}
\end{aligned}$$

積分範囲は変わらないので、

$$E_{(j)}^{R,u}[r^s] = D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} \int_0^\infty \left\{ \frac{\tau}{(n-j+l)\lambda C} \right\}^{\frac{s}{u}} \cdot \exp\{-\tau\} \cdot \frac{1}{(n-j+l)} d\tau$$

これを整理して

$$E_{(j)}^{R,u}[r^s] = D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} (n-j+l)^{-\frac{s+u}{u}} (\lambda C)^{-\frac{s}{u}} \int_0^\infty \tau^{\frac{s}{u}} \cdot \exp\{-\tau\} d\tau$$

となる。

この式の積分の項は、 Γ 関数の定義式となっているので、原点まわりの s 次モーメントは

$$E_{(j)}^{R,u}[r^s] = D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} (n-j+l)^{-\frac{s+u}{u}} (\lambda C)^{-\frac{s}{u}} \cdot \Gamma\left(\frac{s}{u} + 1\right)$$

ここで、 C と D は既出である。

となり、式 (7.2) が導出できた。

以上で、命題 7.1 が示された。

7.2.2 命題 7.2 の証明

Roberts (1968) とは異なり、Watanabe (2008) の場合は領域の形状が複雑なため、偶数次元と奇数次元を分ける必要がある。ここでは Watanabe (2008) の偶数次元に対応する命題 7.2 を証明する。

付録 1 の (付 1.1) 式に於いて、偶数次元 ($u = 2t$) に於ける Watanabe (2008) の近似を用いた MST の辺長の確率密度関数を求めた。ここでは Roberts (1968) と区別するため、右肩に添え字 W を追加して

$f_{2t}^W(r)$ と表記する。

$$(付 1.1) \quad f_{2t}^W(r) = 2t\lambda \cdot C_e^{Ac\cap B} \cdot r^{2t-1} \cdot \exp(-\lambda \cdot C_e^{Ac\cap B} \cdot r^{2t})$$

$$\text{ここで、} C_e^{Ac\cap B} = \frac{V_{2t-1}(r)}{r^{2t-1}} \cdot \left\{ \frac{\sqrt{\pi}\Gamma\left(t+\frac{1}{2}\right)}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t+\frac{1}{2}, \frac{1}{2}\right) \right\}$$

$B_{\frac{3}{4}}\left(t+\frac{1}{2}, \frac{1}{2}\right)$ は不完全ベータ関数。

ここから分布関数 $F_{2t}^W(r)$ は、

$$F_{2t}^W(r) = 1 - \exp(-\lambda \cdot C_e^{Ac\cap B} \cdot r^{2t})$$

と導出できる。

これらを式 (7.7) に代入することにより、第 j 順序統計量の確率密度関数 $f_{(j)}^{W,2t}(r)$ を求めることができる。

$$f_{(j)}^{W,2t}(r) = \frac{(n-1)!}{(j-1)! \cdot (n-1-j)!} \cdot F_{2t}^W(r)^{j-1} \cdot \{1 - F_{2t}^W(r)\}^{n-1-j} \cdot f_{2t}^W(r)$$

ここで以下のように、 D を導入し式の展開を簡略化する。

$$D = \frac{(n-1)!}{(j-1)! \cdot (n-1-j)!}$$

したがって、 $f_{(j)}^{W,2t}(r)$ は

$$\begin{aligned} f_{(j)}^{W,2t}(r) &= D \cdot \{1 - \exp(-\lambda C_e^{Ac\cap B} r^{2t})\}^{j-1} \cdot \{\exp(-\lambda C_e^{Ac\cap B} r^{2t})\}^{n-1-j} \cdot 2t\lambda C_e^{Ac\cap B} r^{2t-1} \cdot \exp(-\lambda C_e^{Ac\cap B} r^{2t}) \\ &= 2t\lambda C_e^{Ac\cap B} D r^{2t-1} \cdot \{1 - \exp(-\lambda C_e^{Ac\cap B} r^{2t})\}^{j-1} \cdot \{\exp(-\lambda C_e^{Ac\cap B} r^{2t})\}^{n-j} \\ &= 2t\lambda C_e^{Ac\cap B} D r^{2t-1} \cdot \{1 - \exp(-\lambda C_e^{Ac\cap B} r^{2t})\}^{j-1} \cdot \exp\{-(n-j)\lambda C_e^{Ac\cap B} r^{2t}\} \end{aligned}$$

となる。

ここで、先に $\{1 - \exp(-\lambda C_e^{Ac\cap B} r^{2t})\}^{j-1}$ の2項展開を行うと、

$$\begin{aligned} \{1 - \exp(-\lambda C_e^{Ac\cap B} r^{2t})\}^{j-1} &= 1 - \binom{j-1}{1} \cdot \exp(-\lambda C_e^{Ac\cap B} r^{2t}) + \binom{j-1}{2} \cdot \exp(-2\lambda C_e^{Ac\cap B} r^{2t}) \\ &\quad - \binom{j-1}{3} \cdot \exp(-3\lambda C_e^{Ac\cap B} r^{2t}) + \dots + (-1)^{j-1} \binom{j-1}{j-1} \cdot \exp\{-(j-1)\lambda C_e^{Ac\cap B} r^{2t}\} \end{aligned}$$

となる。

この結果を元の式に戻して整理すると、

$$\begin{aligned} f_{(j)}^{W,2t}(r) &= 2t\lambda C_e^{Ac\cap B} D r^{2t-1} \left[\exp\{-(n-j)\lambda C_e^{Ac\cap B} r^{2t}\} - \binom{j-1}{1} \cdot \exp\{-(n-j+1)\lambda C_e^{Ac\cap B} r^{2t}\} \right. \\ &\quad \left. + \binom{j-1}{2} \cdot \exp\{-(n-j+2)\lambda C_e^{Ac\cap B} r^{2t}\} - \binom{j-1}{3} \cdot \exp\{-(n-j+3)\lambda C_e^{Ac\cap B} r^{2t}\} \right] \end{aligned}$$

$$+ \dots + (-1)^{j-1} \binom{j-1}{j-1} \cdot \exp\{-(n-1)\lambda C_e^{A^c \cap B} r^{2t}\}]$$

となる。

さらに整理すると、第 j 順序統計量の確率密度関数 $f_{(j)}^{W,2t}(r)$ は、

$$f_{(j)}^{W,2t}(r) = D \sum_{l=0}^{j-1} 2t\lambda C_e^{A^c \cap B} r^{2t-1} (-1)^l \binom{j-1}{l} \exp\{-(n-j+l)\lambda C_e^{A^c \cap B} r^{2t}\}$$

ここで、 $C_e^{A^c \cap B}$ と D は既出である。

となり、式(7.3)が導出できた。

次に $f_{(j)}^{W,2t}(r)$ を式(7.8)に代入することにより、第 j 順序統計量の原点まわりの s 次モーメント $E_{(j)}^{W,2t}[r^s]$ を求める。

$$E_{(j)}^{W,2t}[r^s] = \int_0^{\infty} r^s \cdot f_{(j)}^{W,2t}(r) dr$$

代入後、整理すると

$$\begin{aligned} E_{(j)}^{W,2t}[r^s] &= \int_0^{\infty} r^s \cdot D \sum_{l=0}^{j-1} 2t\lambda C_e^{A^c \cap B} r^{2t-1} (-1)^l \binom{j-1}{l} \exp\{-(n-j+l)\lambda C_e^{A^c \cap B} r^{2t}\} dr \\ &= D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} \int_0^{\infty} r^s 2t\lambda C_e^{A^c \cap B} r^{2t-1} \exp\{-(n-j+l)\lambda C_e^{A^c \cap B} r^{2t}\} dr \end{aligned}$$

となる。

ここで $\tau = (n-j+l)\lambda C_e^{A^c \cap B} r^{2t}$ として置換積分を行う。

$$\frac{d\tau}{dr} = (n-j+l)\lambda C_e^{A^c \cap B} 2tr^{2t-1}$$

$$\frac{d\tau}{(n-j+l)} = \lambda C_e^{A^c \cap B} 2tr^{2t-1} dr$$

$$r = \left\{ \frac{\tau}{(n-j+l)\lambda C_e^{A^c \cap B}} \right\}^{\frac{1}{2t}}$$

積分範囲は変わらないので、

$$E_{(j)}^{W,2t}[r^s] = D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} \int_0^{\infty} \left\{ \frac{\tau}{(n-j+l)\lambda C_e^{A^c \cap B}} \right\}^{\frac{s}{2t}} \cdot \exp\{-\tau\} \cdot \frac{1}{(n-j+l)} d\tau$$

これを整理して

$$E_{(j)}^{W,2t}[r^s] = D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} (n-j+l)^{-\frac{s+2t}{2t}} (\lambda C_e^{A^c \cap B})^{-\frac{s}{2t}} \int_0^{\infty} \tau^{\frac{s}{2t}} \cdot \exp\{-\tau\} d\tau$$

となる。

この式の積分の項は、 Γ 関数の定義式となっているので、原点まわりの s 次モーメントは

$$E_{(j)}^{W,2t}[r^s] = D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} (n-j+l)^{-\frac{s+2t}{2t}} (\lambda C_e^{A^c \cap B})^{-\frac{s}{2t}} \cdot \Gamma\left(\frac{s}{2t} + 1\right)$$

ここで、 $C_e^{A^c \cap B}$ と D は既出である。

となり、式 (7.4) が導出できた。

以上で命題 7.2 が示された。

7.2.3 命題 7.3 の証明

Roberts (1968) とは異なり、Watanabe (2008) の場合は領域の形状が複雑なため、偶数次元と奇数次元を分ける必要がある。ここでは Watanabe (2008) の奇数次元に対応する命題 7.3 を証明する。

付録 1 の (付 1.2) 式に於いて、奇数次元 ($u = 2t + 1$) に於ける Watanabe (2008) の近似を用いた MST の辺長の確率密度関数を求めた。ここでは Roberts (1968) と区別するため、右肩に添え字 W を追加して $f_{2t+1}^W(r)$ と表記する。

$$(付 1.2) \quad f_{2t+1}^W(r) = (2t+1)\lambda \cdot C_o^{A^c \cap B} \cdot r^{2t} \cdot \exp(-\lambda \cdot C_o^{A^c \cap B} \cdot r^{2t+1})$$

$$\text{ここで、} C_o^{A^c \cap B} = \frac{V_{2t}(r)}{r^{2t}} \cdot \left\{ \frac{\sqrt{\pi}\Gamma(t+1)}{\Gamma(t+\frac{3}{2})} - B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right) \right\}$$

$$B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right) \text{ は不完全ベータ関数。}$$

ここから分布関数 $F_{2t+1}^W(r)$ は、

$$F_{2t+1}^W(r) = 1 - \exp(-\lambda \cdot C_o^{A^c \cap B} \cdot r^{2t+1})$$

と導出できる。

これらを式 (7.7) に代入することにより、第 j 順序統計量の確率密度関数 $f_{(j)}^{W,2t+1}(r)$ を求めることができる。

$$f_{(j)}^{W,2t+1}(r) = \frac{(n-1)!}{(j-1)! \cdot (n-1-j)!} \cdot F_{2t+1}^W(r)^{j-1} \cdot \{1 - F_{2t+1}^W(r)\}^{n-1-j} \cdot f_{2t+1}^W(r)$$

ここで以下のように、 D を導入し式の展開を簡略化する。

$$D = \frac{(n-1)!}{(j-1)! \cdot (n-1-j)!}$$

したがって、 $f_{(j)}^{W,2t+1}(r)$ は

$$\begin{aligned}
f_{(j)}^{W,2t+1}(r) &= D \cdot \{1 - \exp(-\lambda C_o^{Ac \cap B} r^{2t+1})\}^{j-1} \cdot \{\exp(-\lambda C_o^{Ac \cap B} r^{2t+1})\}^{n-1-j} \\
&\quad \cdot (2t+1) \lambda C_o^{Ac \cap B} r^{2t} \exp(-\lambda C_o^{Ac \cap B} r^{2t+1}) \\
&= (2t+1) \lambda C_o^{Ac \cap B} D r^{2t} \cdot \{1 - \exp(-\lambda C_o^{Ac \cap B} r^{2t+1})\}^{j-1} \cdot \{\exp(-\lambda C_o^{Ac \cap B} r^{2t+1})\}^{n-j} \\
&= (2t+1) \lambda C_o^{Ac \cap B} D r^{2t} \cdot \{1 - \exp(-\lambda C_o^{Ac \cap B} r^{2t+1})\}^{j-1} \cdot \exp\{-(n-j) \lambda C_o^{Ac \cap B} r^{2t+1}\}
\end{aligned}$$

となる。

ここで、先に $\{1 - \exp(-\lambda C_o^{Ac \cap B} r^{2t+1})\}^{j-1}$ の 2 項展開を行うと、

$$\begin{aligned}
\{1 - \exp(-\lambda C_o^{Ac \cap B} r^{2t+1})\}^{j-1} &= 1 - \binom{j-1}{1} \cdot \exp(-\lambda C_o^{Ac \cap B} r^{2t+1}) + \binom{j-1}{2} \cdot \exp(-2\lambda C_o^{Ac \cap B} r^{2t+1}) \\
&\quad - \binom{j-1}{3} \cdot \exp(-3\lambda C_o^{Ac \cap B} r^{2t+1}) + \dots + (-1)^{j-1} \binom{j-1}{j-1} \cdot \exp\{-(j-1) \lambda C_o^{Ac \cap B} r^{2t+1}\}
\end{aligned}$$

となる。

この結果を元の式に戻して整理すると、

$$\begin{aligned}
f_{(j)}^{W,2t+1}(r) &= (2t+1) \lambda C_o^{Ac \cap B} D r^{2t} [\exp\{-(n-j) \lambda C_o^{Ac \cap B} r^{2t+1}\} - \binom{j-1}{1} \cdot \exp\{-(n-j+1) \lambda C_o^{Ac \cap B} r^{2t+1}\} \\
&\quad + \binom{j-1}{2} \cdot \exp\{-(n-j+2) \lambda C_o^{Ac \cap B} r^{2t+1}\} - \binom{j-1}{3} \cdot \exp\{-(n-j+3) \lambda C_o^{Ac \cap B} r^{2t+1}\} \\
&\quad + \dots + (-1)^{j-1} \binom{j-1}{j-1} \cdot \exp\{-(n-1) \lambda C_o^{Ac \cap B} r^{2t+1}\}]
\end{aligned}$$

となる。

さらに整理すると、第 j 順序統計量の確率密度関数 $f_{(j)}^{W,2t+1}(r)$ は、

$$f_{(j)}^{W,2t+1}(r) = D \sum_{l=0}^{j-1} (2t+1) \lambda C_o^{Ac \cap B} r^{2t} (-1)^l \binom{j-1}{l} \exp\{-(n-j+l) \lambda C_o^{Ac \cap B} r^{2t+1}\}$$

ここで、 $C_o^{Ac \cap B}$ と D は既出である。

となり、式 (7.5) が導出できた。

次に $f_{(j)}^{W,2t+1}(r)$ を式 (7.8) に代入することにより、第 j 順序統計量の原点まわりの s 次モーメント

$E_{(j)}^{W,2t+1}[r^s]$ を求める。

$$E_{(j)}^{W,2t+1}[r^s] = \int_0^{\infty} r^s \cdot f_{(j)}^{W,2t+1}(r) dr$$

代入後、整理すると

$$E_{(j)}^{W,2t+1}[r^s] = \int_0^{\infty} r^s \cdot D \sum_{l=0}^{j-1} (2t+1) \lambda C_o^{Ac \cap B} r^{2t} (-1)^l \binom{j-1}{l} \exp\{-(n-j+l) \lambda C_o^{Ac \cap B} r^{2t+1}\} dr$$

$$= D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} \int_0^{\infty} r^s (2t+1) \lambda C_0^{A^c \cap B} r^{2t} \exp\{-(n-j+l) \lambda C_0^{A^c \cap B} r^{2t+1}\} dr$$

となる。

ここで $\tau = (n-j+l) \lambda C_0^{A^c \cap B} r^{2t+1}$ として置換積分を行う。

$$\frac{d\tau}{dr} = (n-j+l) \lambda C_0^{A^c \cap B} (2t+1) r^{2t}$$

$$\frac{d\tau}{(n-j+l)} = \lambda C_0^{A^c \cap B} (2t+1) r^{2t} dr$$

$$r = \left\{ \frac{\tau}{(n-j+l) \lambda C_0^{A^c \cap B}} \right\}^{\frac{1}{2t+1}}$$

積分範囲は変わらないので、

$$E_{(j)}^{W, 2t+1}[r^s] = D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} \int_0^{\infty} \left\{ \frac{\tau}{(n-j+l) \lambda C_0^{A^c \cap B}} \right\}^{\frac{s}{2t+1}} \cdot \exp\{-\tau\} \cdot \frac{1}{(n-j+l)} d\tau$$

これを整理して

$$E_{(j)}^{W, 2t+1}[r^s] = D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} (n-j+l)^{-\frac{s+2t+1}{2t+1}} (\lambda C_0^{A^c \cap B})^{-\frac{s}{2t+1}} \int_0^{\infty} \tau^{\frac{s}{2t+1}} \cdot \exp\{-\tau\} d\tau$$

この式の積分の項は、 Γ 関数の定義式となっているので、原点まわりの s 次モーメントは、

$$E_{(j)}^{W, 2t+1}[r^s] = D \sum_{l=0}^{j-1} (-1)^l \binom{j-1}{l} (n-j+l)^{-\frac{s+2t+1}{2t+1}} (\lambda C_0^{A^c \cap B})^{-\frac{s}{2t+1}} \cdot \Gamma\left(\frac{s}{2t+1} + 1\right)$$

ここで、 $C_0^{A^c \cap B}$ と D は既出である。

となり、式 (7.6) が導出できた。

以上で、命題 7.3 が示された。

第 8 章 AGsi-curve の期待値の強度依存性

8.0 はじめに

第 8 章では第 7 章で導出した AGsi-curve の近似分布 U と近似分布 L の期待値、およびモンテカルロ法を用いて推定した厳密分布の期待値の特性を明らかにする。これら 3 種類の期待値の強度 (λ) への依存性には、以下の 2 つの特徴がある。

- 3 種類の期待値に共通する変化
- 3 種類の期待値の相対的な位置関係の変化

最初の特徴は、強度が大きくなると AGsi-curve の水準が低下することである。これは同一領域内で点数が増えると、MST の辺長の期待値が小さくなっていくことが原因である。一方、2 番目の特徴は、期待値の大小関係が強度に依存することである。これは MST の辺長の境界効果が原因であると考えられる。

8.1 節では、2 次元の AGsi-curve に関する 3 種類の期待値を求めたのち、その強度依存性の特徴を示す。次に、AGsi-curve の基になった MST の辺長の順序統計量との関連性を示す。8.2 節では、MST の辺長の期待値と中央値の強度依存性を詳しく調べる。8.3 節では、MST の辺長の境界効果を簡易的にモデル化し、簡単な計算機実験で確認する。8.4 節では、AGsi-curve の期待値に与える境界効果を検討する。

第 8 章の内容は、高井・田村・元山 (2018) に基づいている。

8.1 AGsi-curve の期待値の強度依存性 (2 次元)

先ず、単位正方形内に強度 $\lambda = 100$ で発生させた一様ランダムな点パターンに対する AGsi-curve を図 8.1 に、その確率密度関数と期待値とを図 8.2 に示す。ここで、近似分布 U の確率密度関数と期待値とは、第 7 章で導出した式 (7.1) と式 (7.2) に基づき、近似分布 L の確率密度関数と期待値とは、第 7 章で導出した式 (7.3) と式 (7.4) に基づき、厳密分布の確率密度関数と期待値とは、モンテカルロ法 ($m = 10^5$ 回) により推定できる。

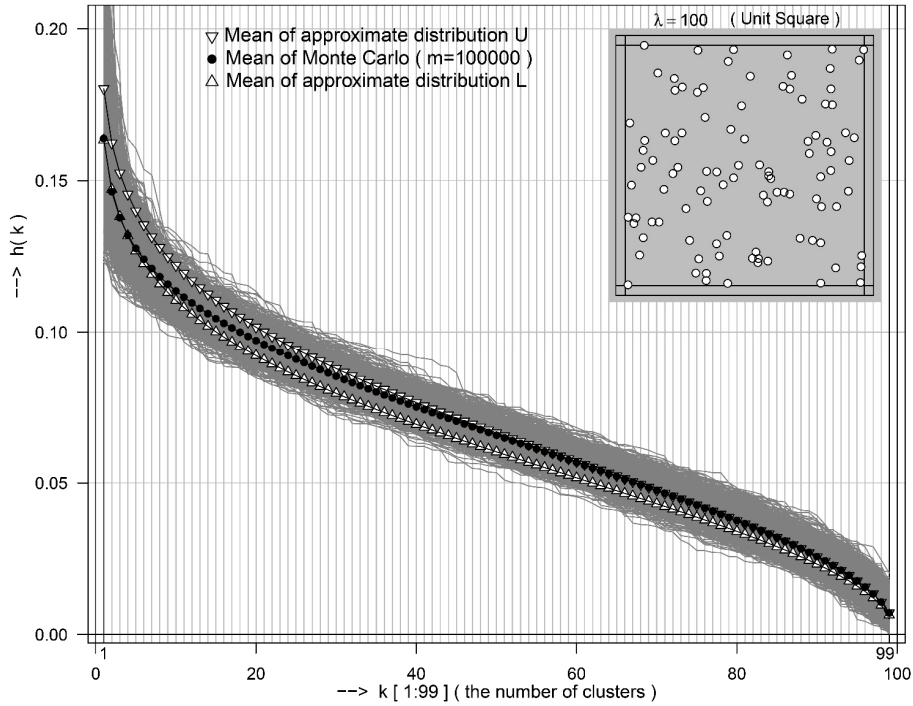


図 8.1 AGsi-curve の 3 種類の期待値 (∇ : 近似分布 U、 \bullet : 厳密分布、 ∇ : 近似分布 L) ($\lambda = 100$)
(モンテカルロ法による AGsi-curve は 10^5 本中 10^3 本分のみを表示)

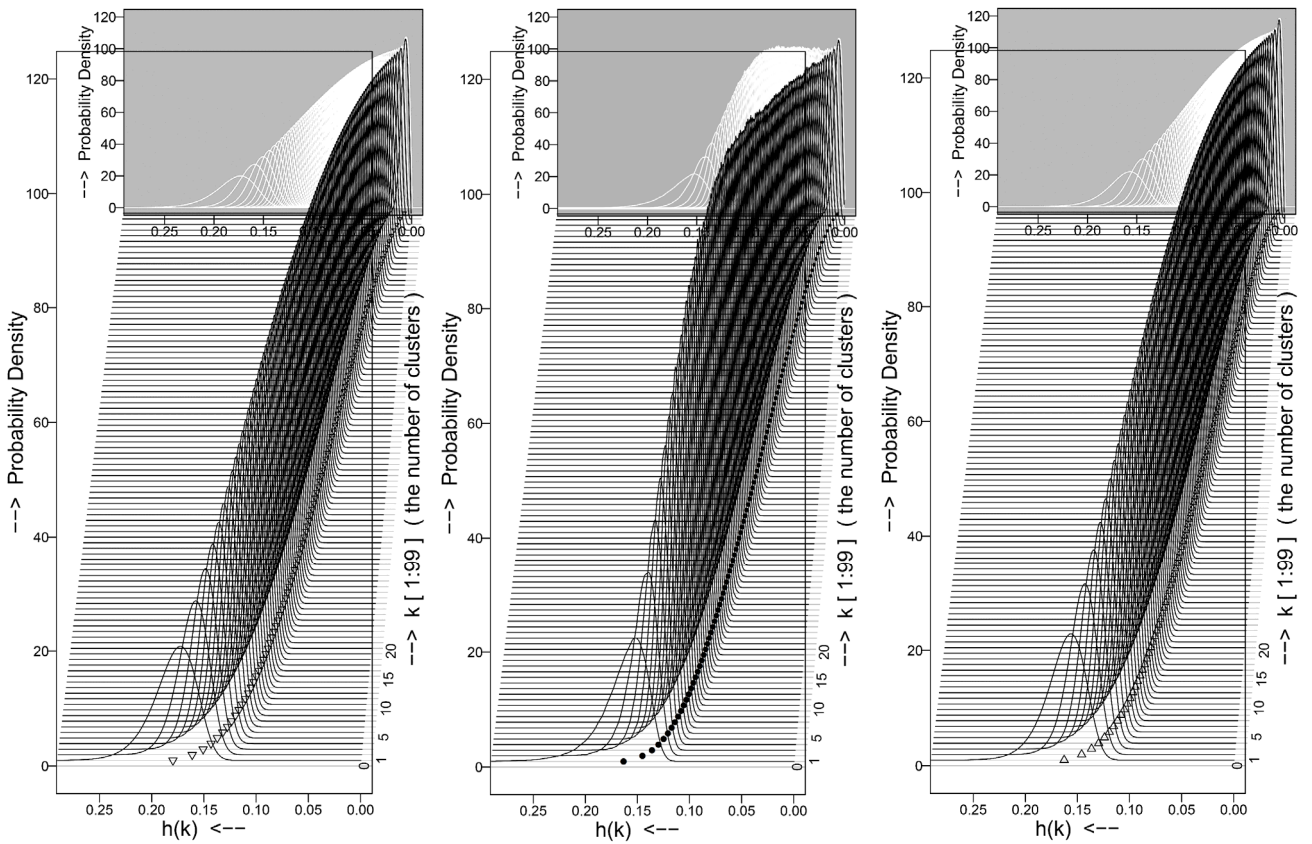


図 8.2 AGsi-curve の 3 種類の確率密度関数 (図 8.1 と同じ記号で期待値も表示) ($\lambda = 100$)
(左: 近似分布 U、中: カーネル法で推定した厳密分布、右: 近似分布 L)

8.1 節では、これらの期待値の強度依存性を調べる。このために、単位正方形内の一様ランダムな点パターンの強度 (λ) を 10 水準 ($\lambda = 10, 20, 30, \dots, 80, 90, 100$) に変化させ、AGsi-curve に関する 3 種類の分布の期待値を図 8.1 と同様な方法で求める。選択した 10 水準の強度は、4 章および 6 章において、分類手法の比較に用いた以下の現実の点パターンが、全て $\lambda < 100$ であることに基づいている。

- Japanese black pine sapling ($n = 65$)

- Biological cells ($n = 42$)

- Redwood seedlings ($n = 62$)

ここで n は単位正方形内の点数である。

これら 10 水準の強度 (λ) に対して求めた期待値を図 8.3 に示す。図 8.3 では、手前になる (強度 λ が大きくなる) ほど、AGsi-curve の 3 種類の期待値が (連動して) 下方に移動していく様子が分かる。これが強度依存性の第 1 の特徴である。また奥に行く (強度 λ が小さくなる) ほど、厳密分布の期待値が 2 つの近似分布の期待値より (相対的に) 大きくなる様子が分かる。これが強度依存性の第 2 の特徴である。

図 8.3 には全部で 30 本の曲線が描かれている。これら複数の曲線の変化を要約するため、図 8.3 の 3 断面 (最大値・中央値・最小値) を抽出して図 8.4 に示す。図 8.4 の横軸は強度 (λ)、縦軸は AGsi-curve の期待値になる。6.4 節などで述べたように、AGsi-curve は MST の辺長の順序統計量と等価であるので、図 8.4 は MST の辺長の順序統計量 (最大値・中央値・最小値) の強度依存性を示していることになる。

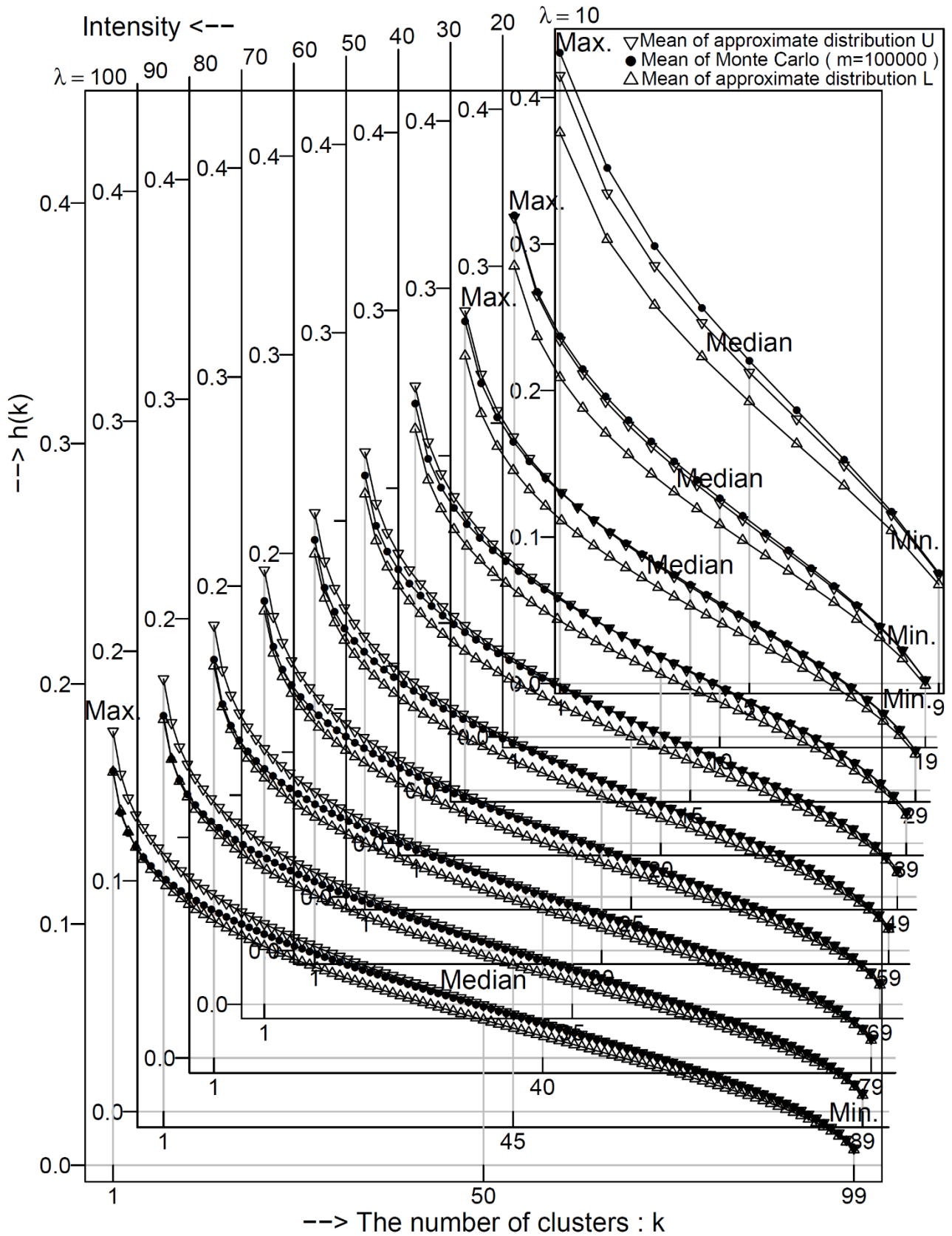


図 8.3 様々な強度に対する AGsi-curve の 3 種類の期待値 ($\lambda = 10, 20, 30, \dots, 80, 90, 100$)

(∇ : 近似分布 U、 \bullet : 厳密分布、 \triangle : 近似分布 L)

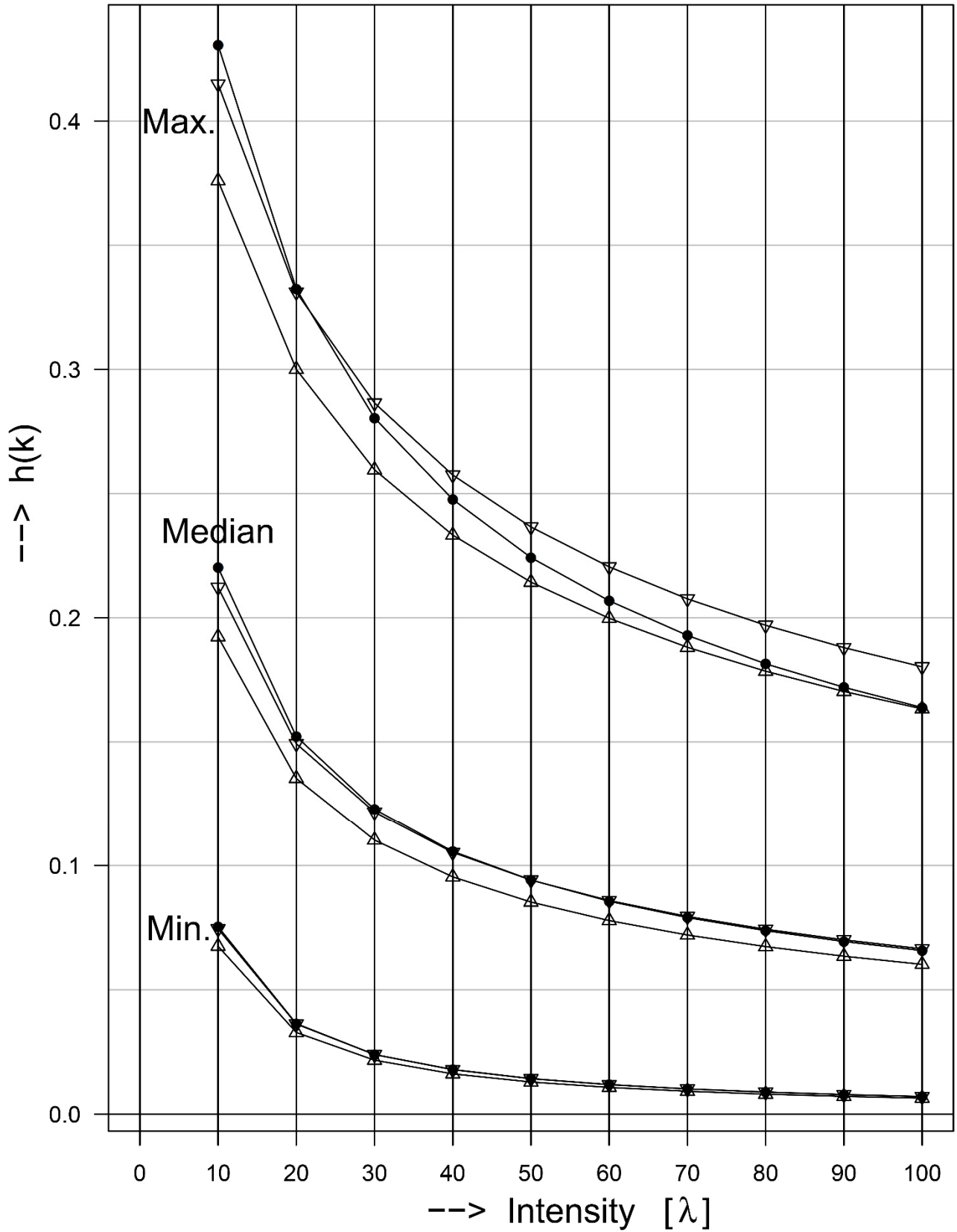


図 8.4 AGsi-curve の 3 種類の期待値 (図 8.3 の断面図: 最大値, 中央値, 最小値)

図 8.4 には、以下の 2 つの特徴が表れている。

- ・強度を大きくすると、最大値・中央値・最小値の値は小さくなる（全て右下がりの曲線）。
- ・強度が大きい水準では、厳密分布の期待値は近似分布 U と近似分布 L の期待値の間にあるが、強度が小さい水準では、厳密分布の期待値が近似分布 U の期待値を超える（長くなる）。

前者は同一領域内で点数が増えて行くと MST の辺長の期待値が小さくなっていくことから当然である。

一方、後者は MST の辺長の境界効果が原因であると考えられる。

8.2 MST の辺長の期待値と中央値（2 次元）

8.1 節では、AGsi-curve の強度依存性を示したが、8.2 節では MST の辺長分布の強度依存性を示す。このため、単位正方形内に強度 (λ) を変えて発生させる一様ランダムな点パターンに対して、MST の辺長に関する、Roberts (1968)、Watanabe (2008)、厳密分布の期待値と中央値を求める。MST の辺長の分布は右の裾が厚い非対称な形状なので期待値だけではなく中央値も求めることとした。

Roberts (1968) の近似分布の期待値と中央値は、第 3 章で導出した式 (3.2) に基づき、Watanabe (2008) の近似分布の期待値と中央値は付録 1 で導出した式 (付 1.3) に基づき、厳密分布の期待値と中央値は、モンテカルロ法 ($m = 10^4$ 回) により推定できる。ここで、モンテカルロ法は 20 水準 ($\lambda = 10, 20, 30, \dots, 180, 190, 200$) に対して行う。8.1 節より、水準を増やしたのは収束状況を観察するためである。

求めた 3 種類の期待値と中央値の結果を図 8.5 と図 8.6 に示す。また AGsi-curve は MST の辺長の順序統計量と等価なので、中央値を示すことは AGsi-curve の振舞いを調べる上でも有益である。(MST の辺長の期待値に対応する AGsi-curve の位置は指定できないが、MST の辺長の中央値に対応する AGsi-curve の位置は指定できる。)

図 8.5 の上図は期待値の計算結果である。図 8.5 の下図は Roberts (1968) と Watanabe (2008) の期待値を基準にした拡大図である。期待値の場合、 $\lambda > 30$ では厳密分布は Roberts (1968) と Watanabe (2008) の間にあるが、 $\lambda = 30$ で厳密分布は Roberts (1968) にほぼ一致し、 $\lambda < 30$ では厳密分布が Roberts (1968)

を超えてしまう。

図 8.6 の上図は中央値の計算結果である。図 8.6 の下図は Roberts (1968) と Watanabe (2008) の中央値を基準にした拡大図である。中央値の場合、 $\lambda > 50$ では厳密分布は Roberts (1968) と Watanabe (2008) の間にあるが、 $\lambda = 50$ で厳密分布は Roberts (1968) にほぼ一致し、 $\lambda < 50$ では厳密分布が Roberts (1968) を超えてしまう。図 8.6 の上図の左半分が図 8.4 の中央値と同じである。

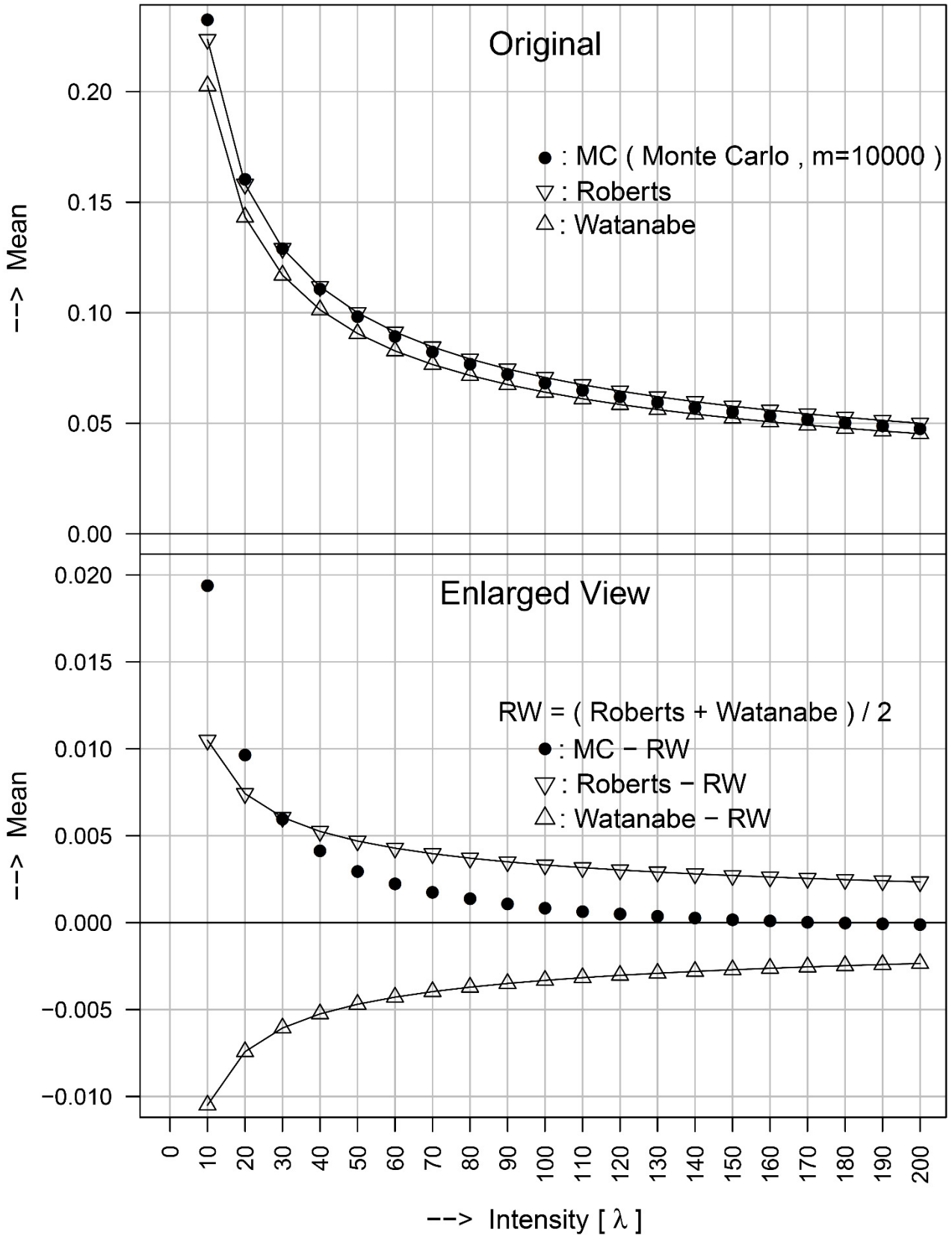


図 8.5 MST の辺長の期待値 (強度依存性)

上図 : 計算結果、下図 : Roberts と Watanabe の期待値を基準にした拡大図

(▽ : Roberts (1968) の期待値、● : 厳密分布の期待値、△ : Watanabe (2008) の期待値)

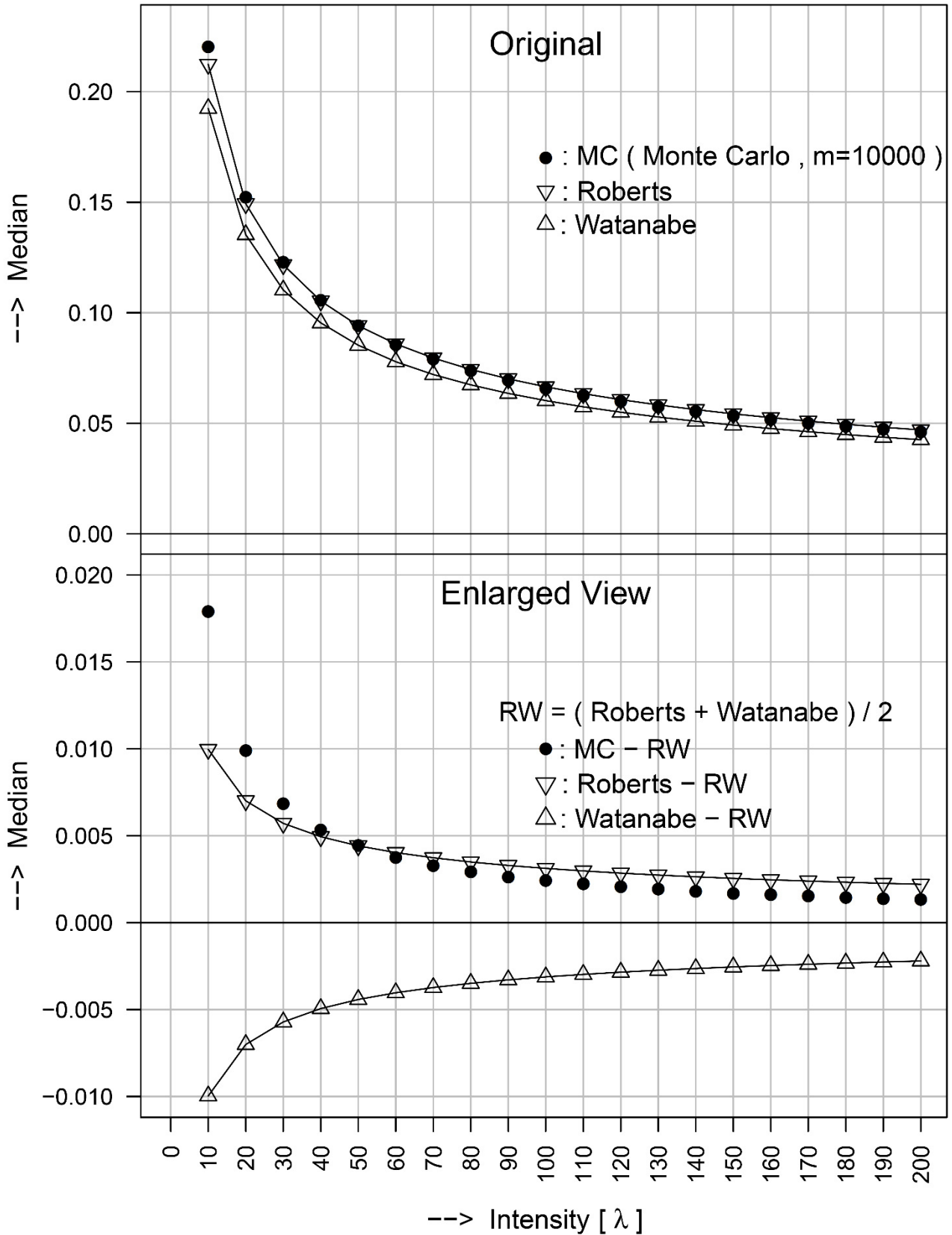


図 8.6 MST の辺長の中央値 (強度依存性)

上図 : 計算結果、下図 : Roberts と Watanabe の中央値を基準にした拡大図

(▽ : Roberts (1968) の中央値、● : 厳密分布の中央値、△ : Watanabe (2008) の中央値)

8.3 MST の辺長の境界効果 (2次元)

8.3 節では、8.1 節と 8.2 節で示された強度が小さい場合の境界効果について検討する。

笠原&古山 (1998) は、領域内の点数 (MST の辺数) が少ないほど、点間距離が離れる影響に加え、境界効果の影響も強く受けるようになり、MST の辺長の期待値が大きくなることを示している。本研究において強度が小さい水準では、AGsi-curve の厳密分布の期待値が近似分布 U の期待値を超えて大きくなる。この原因も MST の辺長の境界効果であると考えられる。

以下では、幾つかの仮定をおき、境界効果を受ける MST の辺長の期待値をモデル化する。厳密な議論は積分幾何学の助けを必要とするが、本研究では簡単な検討にとどめる。

まず、正方形の領域内に発生させた一様ランダムな点パターンに対する MST の辺長の期待値を求めるため、

L : 正方形 (領域) の 1 辺の長さ

λ : 強度

$\lambda L^2 - 1$: 領域内の MST の辺数

とおく。簡単のため境界効果を受ける辺数は、 L に比例すると仮定し、

γL : 境界効果を受ける MST の辺数

とする。これは境界がなければ、領域外の点と繋がる可能性があったのに、境界の存在により領域内の点と繋がった辺の数である。境界により最近隣点探索の範囲が制約を受け、最短経路ではなくなる可能性が高くなる。

MST の総延長距離 TL_{mst} は正方形 (領域) の面積 L^2 に比例すると仮定し、

δL^2 : MST の総延長距離

とする。

$\frac{\gamma L}{\lambda L^2 - 1}$: 境界効果を受ける辺の割合

とすると、MST の辺長の期待値は、

$$(8.1) \quad \frac{\delta L^2}{(\lambda L^2 - 1) \left(1 - \frac{\gamma L}{\lambda L^2 - 1}\right)} = \frac{\delta}{\lambda - \frac{1}{L^2} - \frac{\gamma}{L}}$$

と評価される。

ここで、仮に、 $\lambda = 10$, $\gamma = 1$, $\delta = 2$ と置き、正方形の1辺の長さ L を1~15まで変化させ、MSTの辺長の期待値(8.1)がどのように振舞うかを計算した結果を図8.7に示す。領域が拡大するに従い、MSTの辺長の期待値(8.1)が減少することが分かる。

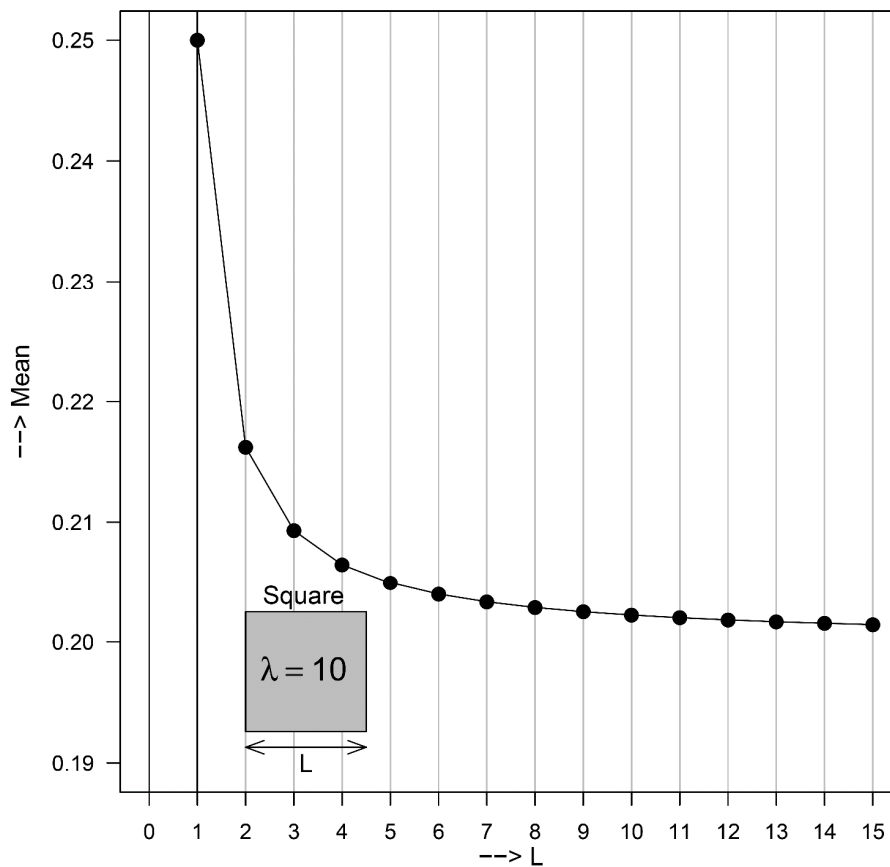


図8.7 MSTの辺長に対する境界効果の簡易的なモデル化の計算例

次に簡単な計算機実験によってモデルの検証を行う。強度固定 ($\lambda = 10$) で領域の面積を拡大し、MSTの辺長の変化を

- ・ 厳密分布の期待値と中央値
- ・ Roberts (1968) の期待値と中央値
- ・ Watanabe (2008) の期待値と中央値

を用いて比較する。領域は正方形の一辺 L を 15 水準 ($L = 1, 2, 3, \dots, 13, 14, 15$) 変化させる。MST の辺長の期待値を評価する上では、計算機実験においても期待値のみを算出すれば十分であるが、AGsi-curve と直接関係があるのは中央値であるので、参考として中央値も調べる。この背景には MST の辺長の分布は右の裾が厚い非対称な形状となっていることがあげられる。図 8.8 に強度固定 ($\lambda = 10$) で領域を広げた ($L = 1, 2, 3, 4, 5, 6$) 場合の MST を示す。図 8.8 の上図は、単位正方形内の MST を示しており、図 8.3 の一番上に配置された AGsi-curve に相当する。

Roberts (1968) と Watanabe (2008) の近似分布の期待値は、境界効果を無視しているので、強度 λ だけに依存し、領域の大きさや形状には依存しない。Roberts (1968) と Watanabe (2008) の期待値を以下に再掲する。一方、厳密分布の期待値と中央値は、モンテカルロ法の方法に依存して、領域の大きさや形状に大きく左右される。

$$E_2[r^1] = \frac{1}{\sqrt{2\lambda}} \approx \frac{0.707106\dots}{\lambda^{1/2}} \quad \text{[再掲：Roberts(1968)の期待値]}$$

$$E_2[r^1] = \frac{1}{2} \sqrt{\frac{\pi}{\lambda\left(\frac{\pi}{3} + \frac{\sqrt{3}}{2}\right)}} \approx \frac{0.640710\dots}{\lambda^{1/2}} \quad \text{[再掲：Watanabe(2008)の期待値]}$$

図 8.9 に計算機実験の結果を示す。図 8.9 の下図の $L = 1$ は単位正方形であり、その位置に於ける 3 つの中央値 (厳密分布・Roberts・Watanabe) は、

- ・ 図 8.4 の $\lambda = 10$ における 3 つの中央値
- ・ 図 8.6 の上図の $\lambda = 10$ における 3 つの中央値

と同じである。図 8.9 は以下の 2 項目を示している。

- ・ 領域を広げていくと厳密分布の期待値と中央値は Roberts (1968) より小さい値を示す。
- ・ 領域を広げていくと厳密分布の期待値と中央値は Roberts (1968) と Watanabe (2008) の間に在る或る値に収束していくように見える。

領域内の点数 (MST の辺数) が少ない [多い] ほど、境界効果を受ける点数 (MST の辺数) の割合が高く [低く] なり、その結果、MST の辺長の期待値が長く [短く] なることが確認できた。強度一定で領域が拡大すると、境界の長さは線形で増加するのに対して、MST の辺数は 2 乗で増加するため、MST の辺長の期待値や中央値に対する境界効果が緩和されると考えられる。

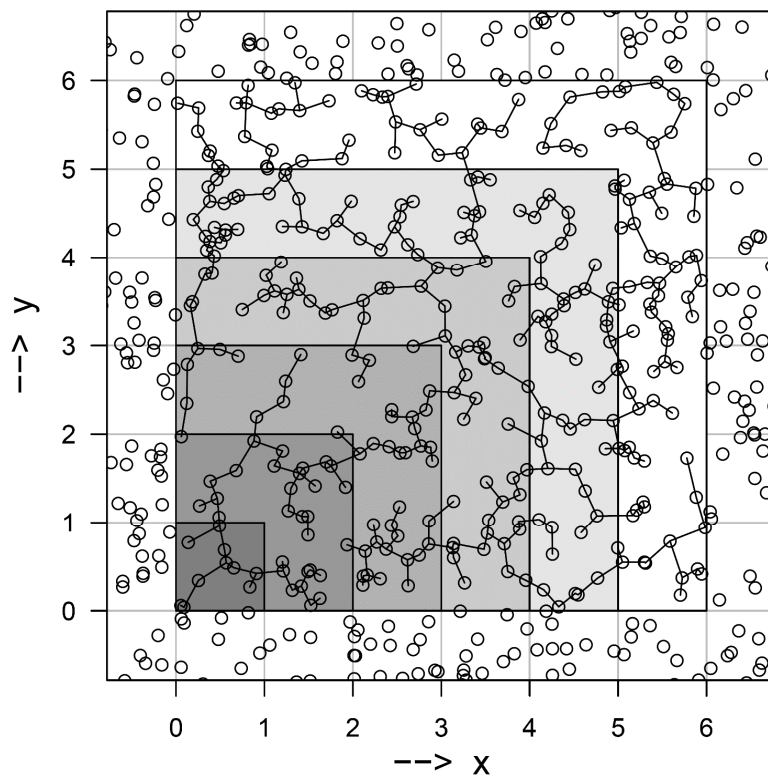
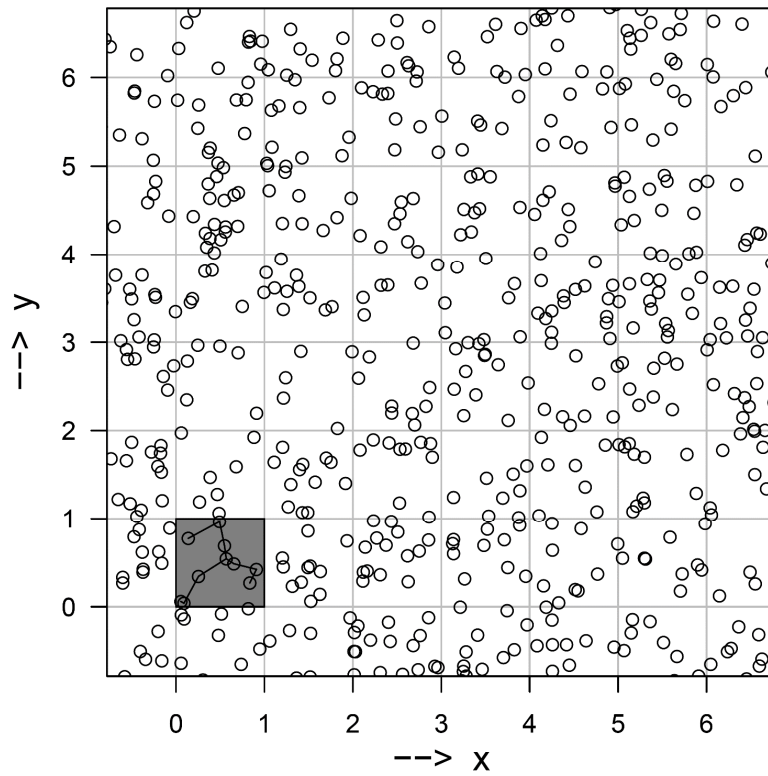


図 8.8 強度固定 ($\lambda = 10$) で領域を広げた場合の MST

(上図 : $L = 1$ (単位正方形)、下図 : $L = 1, 2, 3, 4, 5, 6$)

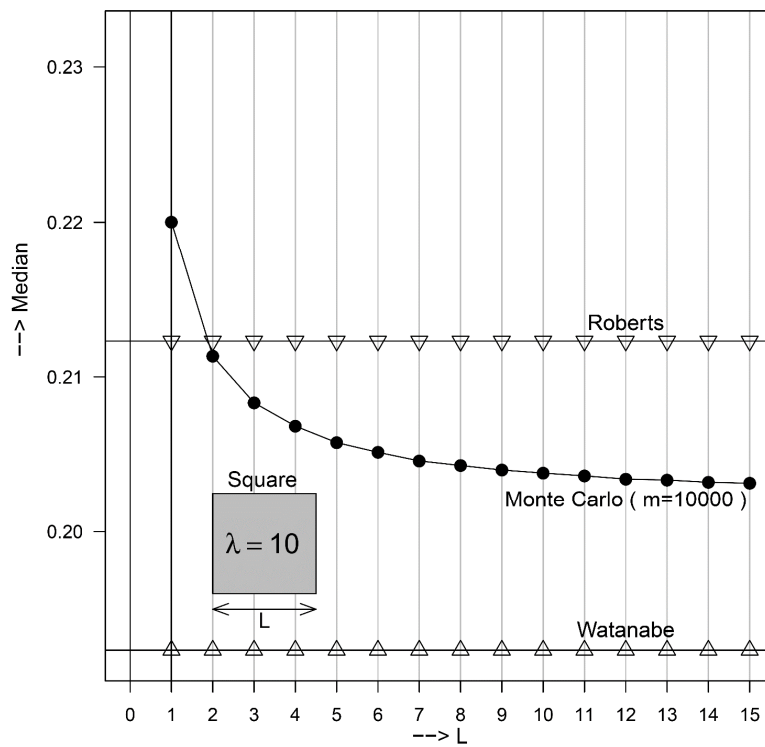
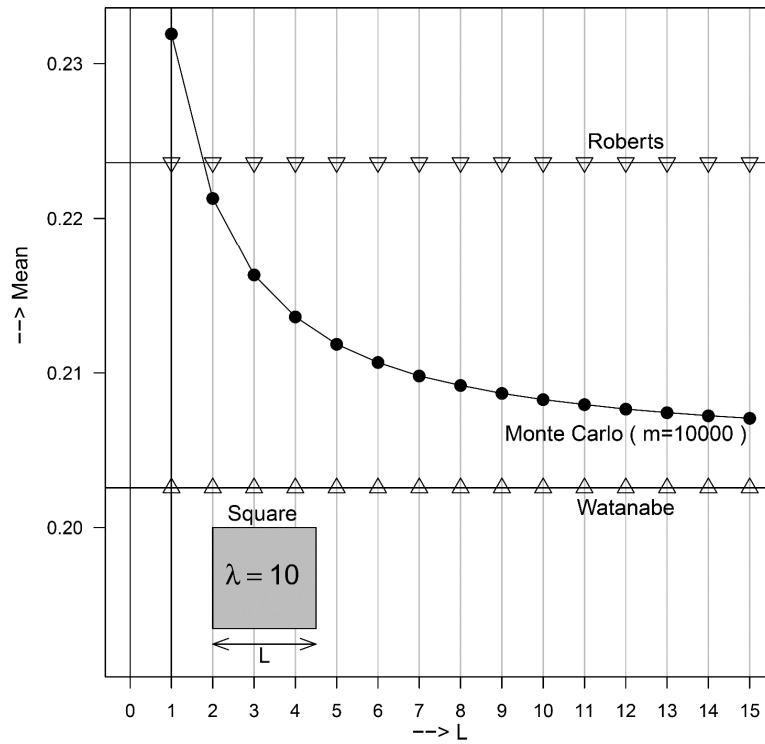


図 8.9 MST の辺長の境界効果 (計算機実験の結果)

(上図：期待値、下図：中央値)

境界効果が大きい領域では、Robert (1968) の近似が上限評価とはならないことは、この数値評価より明らかである。また、図 8.9 より図 8.7 に示した近似モデルによる期待値評価は厳密分布による境界効果を近似的に表現していることが分かる。

8.4 AGsi-curve の期待値に与える境界効果 (2次元)

これまで MST の辺長の振舞いに焦点を当ててきたが、本節では図 8.3 に示した AGsi-curve の期待値の強度依存性に戻り、その原因の 1 つである境界効果を検討する。AGsi-curve の厳密分布と 2 つの近似分布 U と L の期待値は、強度 (λ) が小さくなるにしたがい 3 つが連動して大きくなっていく。これは 8.1 節において述べたように、AGsi-curve の基礎を成す MST の辺長の振舞いを考えれば当然の現象である。さらに、 $50 \leq \lambda \leq 100$ の範囲では、近似分布 U と近似分布 L の期待値は、厳密分布の期待値を上下から挟み込んでおり、近似分布 U と近似分布 L が近似モデルとして有用であることが分かる。しかし、強度 (λ) が小さい範囲では、厳密分布の期待値は近似分布 U の期待値を超えてしまう。

8.3 節で述べたように、MST の辺長に関する Roberts (1968) の近似と Watanabe (2008) の近似は、境界効果を受けない。したがって、それらを基にしている AGsi-curve の近似分布 U と L の期待値も境界効果を受けない。一方、MST の辺長に関する厳密分布の期待値は境界効果を受ける。したがって、それらを基にしている AGsi-curve の厳密分布の期待値も境界効果を受ける。さらに 8.3 節では、MST の辺長の境界効果は領域内の点数 (MST の辺数) が少ないほど強くなることを、簡易的なモデル化と計算機実験で確認している。当然、AGsi-curve の厳密分布の期待値も、この影響を受けることになる。

強度 (λ) が小さくなるにしたがい、AGsi-curve の近似分布 U と L の期待値は、点間距離が離れる影響しか受けないのに対して、AGsi-curve の厳密分布の期待値は、点間距離が離れる影響に加えて、境界効果も、より強く受けるようになる。その結果、AGsi-curve の厳密分布の期待値は、近似分布 U の期待値を超えることになる。以上で、図 8.3 に示した AGsi-curve の期待値の強度依存性が説明できた。

8.5 【参考】3次元の AGsi-curve

簡易的に3次元の AGsi-curve を求めたので、参考として以下に示す。

体積 $1/8$ の立方体内に強度 $\lambda = 10^3$ で発生させた一様ランダムな点パターン ($n = 125$) に対する AGsi-curve を図 8.10 に、その確率密度関数と期待値とを図 8.11 に示す。ここで、近似分布 U の確率密度関数と期待値とは、第 7 章で導出した式 (7.1) と式 (7.2) に基づき、近似分布 L の確率密度関数と期待値とは、第 7 章で導出した式 (7.3) と式 (7.4) に基づき、厳密分布の確率密度関数と期待値とは、モンテカルロ法 ($m = 10^5$ 回) により推定できる。

図 8.10 から分かるように、AGsi-curve の厳密分布の期待値は、近似分布 U の期待値とほぼ一致している。これは 2 次元の AGsi-curve に関する検討結果から、境界効果が強く表れた結果であると考えられる。特に、8.3 節に示した、強度一定で領域の大きさを変えた計算機実験の結果と深い関係がある。高次元の場合の境界効果は非常に興味深い今後の検討課題とする。

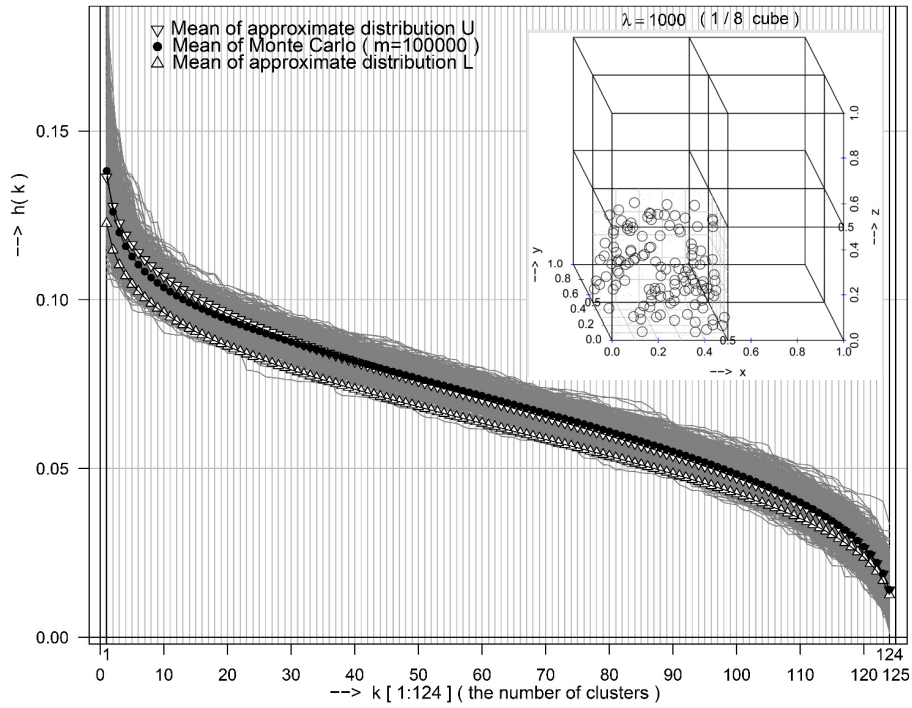


図 8.10 3次元の AGsi-curve ($\lambda = 1000$)

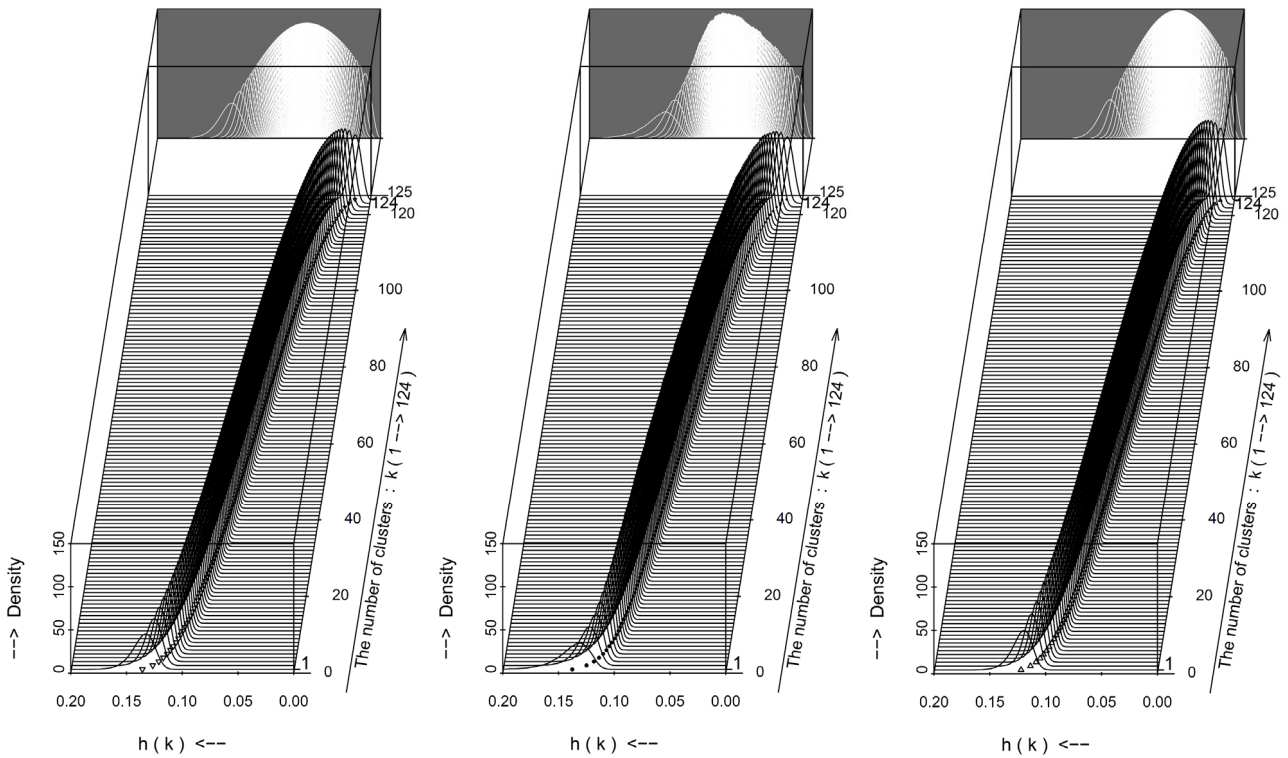


図 8.11 AGsi-curve の 3 種類の確率密度関数 (3 次元) ($\lambda = 1000$)

(図 8.10 と同じ記号で期待値も表示)

(左 : 近似分布 U、中 : カーネル法で推定した厳密分布、右 : 近似分布 L)

第9章 第Ⅲ部のまとめと今後の課題

9.1 第Ⅲ部のまとめ

本研究の目的は、

- ①AG-curve を用いた空間点パターンの分類方法を提案する。
- ②空間点パターンの分類に適した AHC 手法を探索する。
- ③既存のグラフィカルな分類方法に対する AG-curve の優位性を検証する。
- ④AG-curve の数理的性質を明らかにする。

であった。

9.1.1 目的①と②に関して

本研究において提案する AG-curve は AHC 手法を応用して、空間点パターンの分類（ランダム型・規則型・集中型）を行うものである。AHC 手法の中から、単調性が保証されている最短距離法・最長距離法・（重み無し）群平均法の3手法を選び、それらを基にする AGsi-curve・AGco-curve・AGav-curve を分類感度の観点で比較した。その結果から、最短距離法を基にする AGsi-curve が最も良好であることを示した。したがって、これ以降の研究は対象を AGsi-curve に限定して行った。

9.1.2 目的③に関して

既存のグラフィカルな分類方法の中から、G 関数と K 関数を選び、それらと AGsi-curve を空間点パターンの分類性能の観点で比較した。その結果、以下の3項目示すことができた。

AGsi-curve は、

- ・ G 関数では区別できない空間点パターン进行分类することができる。
- ・ ランダム型と規則型の区別が K 関数より明瞭にできる（特に強度が大きい場合）。
- ・ K 関数より境界効果に関してロバストである。

9.1.3 目的④に関して

AGsi-curve の数理的性質は MST の辺長の順序統計量を用いて説明できることを示した。この結果から、MST の辺長の近似分布である Roberts (1968) と Watanabe (2008) を用いて、AGsi-curve の近似分布 U と近似分布 L を導出することができた。導出した 2 つの近似分布は、限定した強度の範囲内ではあるが、AGsi-curve の近似モデルとして有用であることを示した。

AGsi-curve の強度依存性を調べていく過程で、厳密分布と 2 つの近似分布の相対的な位置関係の変化から、厳密分布に現れる境界効果の存在に気が付いた。これは特に強度が小さい水準で顕著になる。AGsi-curve の境界効果に関しては、その基になった MST の辺長の厳密分布に遡って検討した。

MST の辺長に現れる境界効果の特性を、簡易的なモデル化と計算機実験を用いて明らかにすることができた。この結果を用いて、AGsi-curve の厳密分布と 2 つの近似分布の相対的な位置関係の変化を説明できることを示した。

9.2 今後の課題

9.2.1 AGsi-curve および（その基になった）MST の辺長の強度依存性

本研究における強度依存性の検討は、正方形の領域を対象として、強度は 100 以下に限定されている。この理由は分類の対象として用いた 3 種類の現実の空間点パターン (Japanese black pine sapling・Biological cells・Redwood seedlings) と関係があり、それらの領域の形状と領域内の点数に基づいている。研究結果の適用範囲を広げるためには、領域の形状を様々に変え (円・長方形・三角形など)、より広範囲な強度に関して特性を調べる必要がある。

9.2.2 AGsi-curve および（その基になった）MST の辺長の境界効果

本研究では、領域の大きさに関する境界効果を示したが、領域の形状に関する境界効果は、ほとんど検討されていない (領域の形状は正方形に限定されている)。研究結果の適用範囲を広げるためには、領域

の形状を様々に変え（円・長方形・三角形など）、境界効果を調べる必要がある。また、境界効果に関する本研究の数理的検討は、簡易的なレベルにとどまっている。積分幾何学を用いた、より厳密なアプローチが必要である。MST の辺長の境界効果が数理的に明らかになれば、MST の辺長の境界補正方法を検討することもできる。

9.2.3 AGsi-curve および（その基になった）MST の辺長の近似分布

これまで MST の効率的な探索方法は非常に多く提案されているが、MST の辺長が従う確率分布は解析的に求められていない。したがって本研究では AGsi-curve の近似分布を求めるために、MST の辺長の近似分布として Roberts (1968) と Watanabe (2008) を用いた。しかし、MST の辺長の厳密分布と比較した検討結果などから、Roberts (1968) と Watanabe (2008) の近似能力の限界も見えてきた。MST の辺長に関して、更に精度の高い近似分布が必要である。これを用いれば、AGsi-curve の近似分布の精度も向上する。

9.2.4 AGsi-curve の多次元への対応

AGsi-curve の近似分布 U と近似分布 L に関して、それらの期待値と原点まわりのモーメントを多次元で導出した。しかし、実際の計算例は主に 2 次元であり、3 次元は参考として取り上げた程度である。自然科学分野では高々 3 次元であるが、社会科学分野への応用を考えると更に高次元の特性を調べる必要がある。多次元の近似分布を具体的に計算する場合は、特に任意精度演算（桁落ち対策）に注意しなければならない。また高次元になるほど境界効果が強く表れることが予想される。

付録

付録 1 : Watanabe (2008) の近似 (多次元 MST の辺長)

1.0 はじめに

提案する手法の数理的性質を調べるためには、多次元 MST の辺長の近似分布が必要となる。このうち Roberts (1968) の近似に関しては Bertsimas and van Ryzin (1990) が存在する。一方、Watanabe (2008) の近似を多次元に拡張した文献は、少なくとも著者の知る限り存在しない。そこで、本研究中に Watanabe (2008) に関する多次元 MST の辺長の近似分布を導出した。その結果を付録 1 に記す。これは他文献で既出の可能性もあるため付録に記したが、著者が独立に導出したものである。

1.1 Watanabe (2008) の近似

Watanabe (2008) は都市計画の研究に於いて、2 次元 MST の辺長の近似分布を導出している。これを多次元に拡張する。 u 次元空間に距離 r 離れて存在する 2 点 a と b を考える。点 a を中心とする半径 r の u 次元球を A とし、点 b を中心とする半径 r の u 次元球を B とする。また、 u 次元空間の領域 Ω の体積を $V_u^\Omega(r)$ と表すと、 A と B の合併領域 (和集合 $A \cup B$) の体積は $V_u^{A \cup B}(r)$ 、 A と B の共通領域 (積集合 $A \cap B$) の体積は $V_u^{A \cap B}(r)$ 、 A と B の共通領域を B から除いた領域 ($A^c \cap B$) の体積($V_u^B(r) - V_u^{A \cap B}(r)$)は $V_u^{A^c \cap B}(r)$ と表せる。

多次元 (u 次元) に於ける Watanabe (2008) の近似分布は、領域 ($A^c \cap B$) に少なくとも 1 つ点があるという確率から導くことができる。

$$F_u(r) = 1 - \exp\{-\lambda(V_u^{A^c \cap B}(r))\}$$

したがって MST の辺の近似確率密度関数 $f_u(r)$ は、

$$f_u(r) = \frac{d}{dr}F_u(r) = \lambda \frac{d}{dr}(V_u^{A^c \cap B}(r)) \cdot \exp\{-\lambda(V_u^{A^c \cap B}(r))\}$$

となる。

ここで $V_u^{A^c \cap B}(r)$ に、付録 2 の (付 2.1) 式の $V_{2t}^{A^c \cap B}(r)$ を代入すると、偶数次元 ($u = 2t$) の近似確率密度関数 $f_{2t}(r)$ は、

(付 1.1)
$$f_{2t}(r) = 2t\lambda \cdot C_e^{A^c \cap B} \cdot r^{2t-1} \cdot \exp(-\lambda \cdot C_e^{A^c \cap B} \cdot r^{2t})$$

ここで、

$$C_e^{A^c \cap B} = \frac{V_{2t-1}(r)}{r^{2t-1}} \cdot \left\{ \frac{\sqrt{\pi}\Gamma\left(t + \frac{1}{2}\right)}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \right\}$$

$B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right)$ は不完全ベータ関数

となる。

また、 $V_u^{A^c \cap B}(r)$ に、付録 2 の (付 2.2) 式の $V_{2t+1}^{A^c \cap B}(r)$ を代入すると、奇数次元 ($u = 2t + 1$) の近似確率密度関数 $f_{2t+1}(r)$ は、

(付 1.2)
$$f_{2t+1}(r) = (2t + 1)\lambda \cdot C_o^{A^c \cap B} \cdot r^{2t} \cdot \exp(-\lambda \cdot C_o^{A^c \cap B} \cdot r^{2t+1})$$

ここで、

$$C_o^{A^c \cap B} = \frac{V_{2t}(r)}{r^{2t}} \cdot \left\{ \frac{\sqrt{\pi}\Gamma(t+1)}{\Gamma\left(t + \frac{3}{2}\right)} - B_{\frac{3}{4}}\left(t + 1, \frac{1}{2}\right) \right\}$$

$B_{\frac{3}{4}}\left(t + 1, \frac{1}{2}\right)$ は不完全ベータ関数。

となる。

また u 次元空間の原点まわりの s 次モーメントは、

$$E_u[r^s] = \int_0^\infty r^s \cdot f_u(r) dr$$

を用いて求められる。

ここで偶数次元 ($u = 2t$) の原点まわりの s 次モーメントは (付 1.1) 式を使って、

$$E_{2t}[r^s] = \int_0^\infty r^s \cdot f_{2t}(r) dr = (\lambda \cdot C_e^{A^c \cap B})^{\frac{-s}{2t}} \Gamma\left(1 + \frac{s}{2t}\right)$$

ここで、

$$C_e^{A^c \cap B} = \frac{V_{2t-1}(r)}{r^{2t-1}} \cdot \left\{ \frac{\sqrt{\pi}\Gamma\left(t + \frac{1}{2}\right)}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \right\}$$

$B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right)$ は不完全ベータ関数

となり、

奇数次元 ($u = 2t + 1$) の原点まわりの s 次モーメントは (付 1.2) 式を使って、

$$E_{2t+1}[r^s] = \int_0^\infty r^s \cdot f_{2t+1}(r) dr = (\lambda \cdot C_o^{Ac \cap B})^{\frac{-s}{2t+1}} \Gamma\left(1 + \frac{s}{2t+1}\right)$$

ここで、

$$C_o^{Ac \cap B} = \frac{V_{2t}(r)}{r^{2t}} \cdot \left\{ \frac{\sqrt{\pi} \Gamma(t+1)}{\Gamma\left(t + \frac{3}{2}\right)} - B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right) \right\}$$

$B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right)$ は不完全ベータ関数。

となる。

具体例として 1 次元・2 次元・3 次元の確率密度関数を求める。

1 次元の場合は奇数次元の式で $t=0$ と置き、2 次元の場合は偶数次元の式で $t=1$ と置き、3 次元の場合は奇数次元の式で $t=1$ と置くことによって、以下のように求められる。

$$f_1(r) = \lambda \cdot \exp(-\lambda r)$$

$$f_2(r) = 2\lambda \left(\frac{\pi}{3} + \frac{\sqrt{3}}{2}\right) r \cdot \exp\left\{-\lambda \left(\frac{\pi}{3} + \frac{\sqrt{3}}{2}\right) r^2\right\}$$

$$f_3(r) = 3\lambda \left(\frac{11\pi}{12}\right) r^2 \cdot \exp\left\{-\lambda \left(\frac{11\pi}{12}\right) r^3\right\}$$

次に 1 次元・2 次元・3 次元の期待値を求める。

1 次元の場合は奇数次元の式で $t=0, s=1$ と置き、2 次元の場合は偶数次元の式で $t=1, s=1$ と置き、3 次元の場合は奇数次元の式で $t=1, s=1$ と置くことによって、以下のように求められる。

$$E_1[r^1] = \frac{1}{\lambda}$$

(付 1.3)
$$E_2[r^1] = \frac{1}{2} \sqrt{\frac{\pi}{\lambda \left(\frac{\pi}{3} + \frac{\sqrt{3}}{2}\right)}} \approx \frac{0.640710 \dots}{\lambda^{1/2}}$$

(付 1.4)
$$E_3[r^1] = \frac{12^{1/3} \cdot \Gamma(1/3)}{3 \cdot 11^{1/3} \cdot \lambda^{1/3} \cdot \pi^{1/3}} \approx \frac{0.627655 \dots}{\lambda^{1/3}}$$

最後に 1 次元・2 次元・3 次元の標準偏差 (SD) を求める。標準偏差 (SD) は、

$$SD_u = \sqrt{E_u(r^2) - \{E_u(r^1)\}^2}$$

を用いて求められる。

また原点まわりの2次モーメントは、

$$E_1[r^2] = \frac{2}{\lambda^2}$$

$$E_2[r^2] = \frac{1}{\lambda \left(\frac{\pi}{3} + \frac{\sqrt{3}}{2} \right)}$$

$$E_3[r^2] = \frac{2 \cdot 12^{2/3} \cdot \Gamma(2/3)}{3 \cdot 11^{2/3} \cdot \lambda^{2/3} \cdot \pi^{2/3}}$$

である。

これらを用いて標準偏差 (SD) は、

$$SD_1 = \frac{1}{\lambda}$$

$$SD_2 = \frac{1}{2} \sqrt{\frac{4 - \pi}{\lambda \left(\frac{\pi}{3} + \frac{\sqrt{3}}{2} \right)}} \approx \frac{0.334914 \dots}{\lambda^{1/2}}$$

$$SD_3 = \frac{12^{1/3} \sqrt{2 \cdot 3 \cdot \Gamma(2/3) - \Gamma(1/3)^2}}{3 \cdot 11^{1/3} \cdot \lambda^{1/3} \cdot \pi^{1/3}} \approx \frac{0.228119 \dots}{\lambda^{1/3}}$$

と求められる。

付録 2 : 多次元空間に於ける部分球と結合球の体積

2.1 準備

2.1.1 領域の定義

u 次元空間に距離 r 隔てて存在する2点 a と b を考える。点 a を中心とする半径 r の u 次元球を A とし、点 b を中心とする半径 r の u 次元球を B とする。このとき、様々な領域の体積を以下のように表記する。

超球 A の体積： $V_u^A(r)$ 、超球 B の体積： $V_u^B(r)$

A と B の合併領域（和集合 $A \cup B$ ）の体積： $V_u^{A \cup B}(r)$

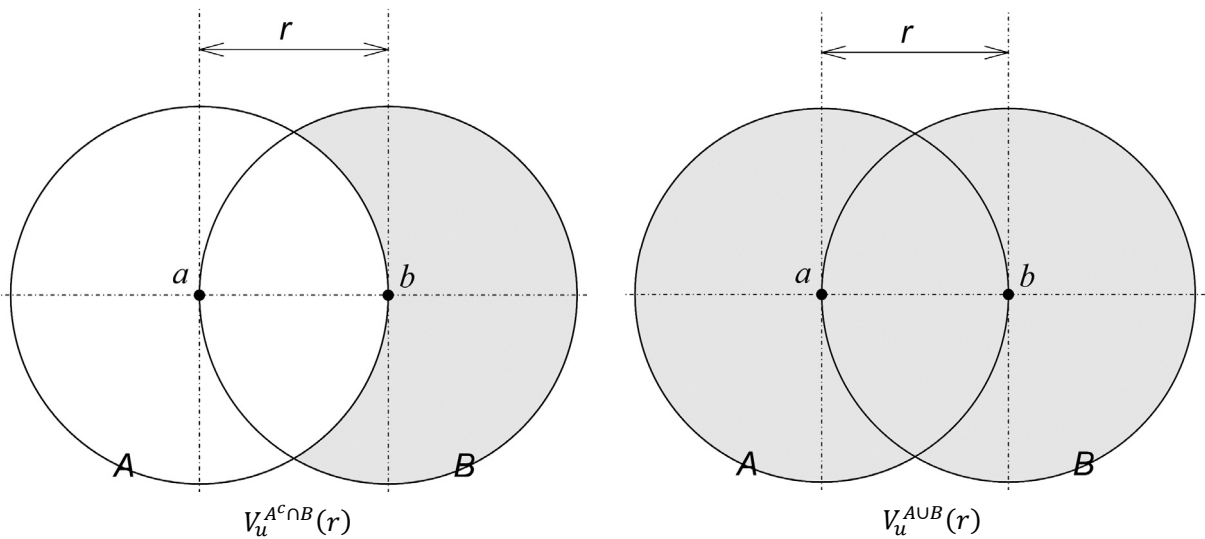
A と B の共通領域（積集合 $A \cap B$ ）の体積： $V_u^{A \cap B}(r)$

A と B の共通領域を B から除いた領域（ $V_u^B(r) - V_u^{A \cap B}(r)$ ）の体積： $V_u^{A^c \cap B}(r)$

$$V_u^A(r) = V_u^B(r) = V_u(r) = \frac{\pi^{\frac{u}{2}}}{\Gamma\left(\frac{u}{2} + 1\right)} r^u$$

以下では、部分球の体積 $V_u^{A^c \cap B}(r)$ と、結合球の体積 $V_u^{A \cup B}(r)$ を求める。

参考のため、2次元の場合の $V_u^{A^c \cap B}(r)$ と $V_u^{A \cup B}(r)$ を付図 2.1 に示す。



付図 2.1 $V_u^{A^c \cap B}(r)$ と $V_u^{A \cup B}(r)$ (2次元の場合)

2.1.2 定積分の値

$$\left\{ \begin{array}{l} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{2t} \theta d\theta = \frac{\sqrt{\pi} \Gamma\left(t + \frac{1}{2}\right)}{\Gamma(t + 1)} \\ \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^{2t} \theta d\theta = \frac{1}{2} B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \end{array} \right.$$

$$\left\{ \begin{array}{l} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{2t+1} \theta d\theta = \frac{\sqrt{\pi} \Gamma(t + 1)}{\Gamma\left(t + \frac{3}{2}\right)} \\ \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^{2t+1} \theta d\theta = \frac{1}{2} B_{\frac{3}{4}}\left(t + 1, \frac{1}{2}\right) \end{array} \right.$$

ここで $B_z(a, b)$ は不完全ベータ関数を表す。

$$B_z(a, b) = \int_0^z \tau^{a-1} (1 - \tau)^{b-1} d\tau$$

2.2 部分球の体積 : $V_u^{A^c \cap B}(r)$

$V_u^{A^c \cap B}(r)$ は以下のようにして求められる。

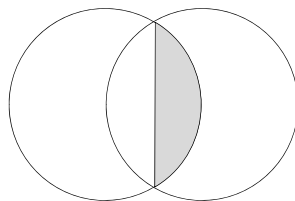
$$V_u^{A^c \cap B}(r) = V_u^B(r) - V_u^{A \cap B}(r)$$

また、 $V_u^B(r)$ と $\frac{1}{2} V_u^{A \cap B}(r)$ は以下の漸化式で表せる。

$$\begin{aligned} V_u^B(r) &= r V_{u-1}^B(r) \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^u \theta d\theta = r V_{u-1}(r) \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^u \theta d\theta \\ \frac{1}{2} V_u^{A \cap B}(r) &= r V_{u-1}^B(r) \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^u \theta d\theta = r V_{u-1}(r) \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^u \theta d\theta \end{aligned}$$

ここで、 $V_u^A(r) = V_u^B(r) = V_u(r)$ である。

参考のため 2 次元の場合の $\frac{1}{2} V_u^{A \cap B}(r)$ を付図 2.2 に示す。



付図 2.2 $\frac{1}{2} V_u^{A \cap B}(r)$ (2 次元の場合)

ここでは偶数次元と奇数次元に場合分けして部分球の体積 $V_u^{A^c \cap B}(r)$ を計算する。

【偶数次元の場合： $u = 2t$ 】

$$\begin{aligned}
 \text{(付 2.1)} \quad V_{2t}^{A^c \cap B}(r) &= V_{2t}^B(r) - V_{2t}^{A \cap B}(r) \\
 &= rV_{2t-1}(r) \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{2t} \theta d\theta - 2rV_{2t-1}(r) \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^{2t} \theta d\theta \\
 &= rV_{2t-1}(r) \left\{ \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{2t} \theta d\theta - 2 \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^{2t} \theta d\theta \right\} \\
 &= r \cdot V_{2t-1}(r) \cdot \left\{ \frac{\sqrt{\pi} \Gamma\left(t + \frac{1}{2}\right)}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \right\} \\
 &= r^{2t} \cdot \frac{V_{2t-1}(r)}{r^{2t-1}} \cdot \left\{ \frac{\sqrt{\pi} \Gamma\left(t + \frac{1}{2}\right)}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \right\} \\
 &= r^{2t} \cdot C_e^{A^c \cap B}
 \end{aligned}$$

ここで、

$$C_e^{A^c \cap B} = \frac{V_{2t-1}(r)}{r^{2t-1}} \cdot \left\{ \frac{\sqrt{\pi} \Gamma\left(t + \frac{1}{2}\right)}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \right\}$$

$B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right)$ は不完全ベータ関数

なお、 $V_{2t-1}(r)$ は $(2t-1)$ 次元超球（全球）の体積なので、 $\frac{V_{2t-1}(r)}{r^{2t-1}}$ は次元に依存する定数となる。例えば1次元なら 2 、3次元なら $\frac{4\pi}{3}$ である。

【奇数次元の場合： $u = 2t + 1$ 】

$$\begin{aligned}
 \text{(付 2.2)} \quad V_{2t+1}^{A^c \cap B}(r) &= V_{2t+1}^B(r) - V_{2t+1}^{A \cap B}(r) \\
 &= rV_{2t}(r) \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{2t+1} \theta d\theta - 2rV_{2t}(r) \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^{2t+1} \theta d\theta \\
 &= rV_{2t}(r) \left\{ \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{2t+1} \theta d\theta - 2 \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^{2t+1} \theta d\theta \right\} \\
 &= r \cdot V_{2t}(r) \cdot \left\{ \frac{\sqrt{\pi} \Gamma(t+1)}{\Gamma\left(t + \frac{3}{2}\right)} - B_{\frac{3}{4}}\left(t + 1, \frac{1}{2}\right) \right\}
 \end{aligned}$$

$$\begin{aligned}
&= r^{2t+1} \cdot \frac{V_{2t}(r)}{r^{2t}} \cdot \left\{ \frac{\sqrt{\pi}\Gamma(t+1)}{\Gamma\left(t+\frac{3}{2}\right)} - B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right) \right\} \\
&= r^{2t+1} \cdot C_0^{A^c \cap B}
\end{aligned}$$

ここで、

$$C_0^{A^c \cap B} = \frac{V_{2t}(r)}{r^{2t}} \cdot \left\{ \frac{\sqrt{\pi}\Gamma(t+1)}{\Gamma\left(t+\frac{3}{2}\right)} - B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right) \right\}$$

$B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right)$ は不完全ベータ関数。

なお、 $V_{2t}(r)$ は $(2t)$ 次元超球 (全球) の体積なので、 $\frac{V_{2t}(r)}{r^{2t}}$ は次元に依存する定数となる。例えば 2 次元なら π 、4 次元なら $\frac{\pi^2}{2}$ になる。

2.3 結合球の体積 : $V_u^{A \cup B}(r)$

$V_u^{A \cup B}(r)$ は以下のようにして求められる。

$$\begin{aligned}
V_u^{A \cup B}(r) &= V_u^A(r) + V_u^B(r) - V_u^{A \cap B}(r) \\
&= 2V_u(r) - V_u^{A \cap B}(r)
\end{aligned}$$

ここで、 $V_u^A(r) = V_u^B(r) = V_u(r)$ である。

また、 $V_u(r)$ と $\frac{1}{2}V_u^{A \cap B}(r)$ は以下の漸化式で表せる。

$$\begin{aligned}
V_u(r) &= rV_{u-1}(r) \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^u \theta d\theta \\
\frac{1}{2}V_u^{A \cap B}(r) &= rV_{u-1}(r) \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^u \theta d\theta
\end{aligned}$$

ここでは偶数次元と奇数次元に場合分けして結合球の体積 $V_u^{A \cup B}(r)$ を計算する。

【偶数次元の場合 : $u = 2t$ 】

$$\begin{aligned}
\text{(付 2.3)} \quad V_{2t}^{A \cup B}(r) &= 2V_{2t}(r) - V_{2t}^{A \cap B}(r) \\
&= 2rV_{2t-1}(r) \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{2t} \theta d\theta - 2rV_{2t-1}(r) \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^{2t} \theta d\theta
\end{aligned}$$

$$\begin{aligned}
&= rV_{2t-1}(r) \left\{ 2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{2t} \theta d\theta - 2 \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^{2t} \theta d\theta \right\} \\
&= r \cdot V_{2t-1}(r) \cdot \left\{ \frac{2\sqrt{\pi}\Gamma\left(t + \frac{1}{2}\right)}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right) \right\} \\
&= r^{2t} \cdot \frac{V_{2t-1}(r)}{r^{2t-1}} \cdot \left\{ \frac{2\sqrt{\pi}\Gamma\left(t + \frac{1}{2}\right)}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \right\} \\
&= r^{2t} \cdot C_e^{AUB}
\end{aligned}$$

ここで、

$$C_e^{AUB} = \frac{V_{2t-1}(r)}{r^{2t-1}} \cdot \left\{ \frac{2\sqrt{\pi}\Gamma\left(t + \frac{1}{2}\right)}{\Gamma(t+1)} - B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right) \right\}$$

$B_{\frac{3}{4}}\left(t + \frac{1}{2}, \frac{1}{2}\right)$ は不完全ベータ関数。

なお、 $V_{2t-1}(r)$ は $(2t-1)$ 次元超球 (全球) の体積なので、 $\frac{V_{2t-1}(r)}{r^{2t-1}}$ は次元に依存する定数となる。例えば

1次元なら 2 、3次元なら $\frac{4\pi}{3}$ である。

【奇数次元の場合 : $u = 2t + 1$ 】

$$\begin{aligned}
\text{(付 2.4)} \quad V_{2t+1}^{AUB}(r) &= 2V_{2t+1}(r) - V_{2t+1}^{ANB}(r) \\
&= 2rV_{2t}(r) \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{2t+1} \theta d\theta - 2rV_{2t}(r) \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^{2t+1} \theta d\theta \\
&= rV_{2t}(r) \left\{ 2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{2t+1} \theta d\theta - 2 \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \cos^{2t+1} \theta d\theta \right\} \\
&= r \cdot V_{2t}(r) \cdot \left\{ \frac{2\sqrt{\pi}\Gamma(t+1)}{\Gamma\left(t + \frac{3}{2}\right)} - B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right) \right\} \\
&= r^{2t+1} \cdot \frac{V_{2t}(r)}{r^{2t}} \cdot \left\{ \frac{2\sqrt{\pi}\Gamma(t+1)}{\Gamma\left(t + \frac{3}{2}\right)} - B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right) \right\} \\
&= r^{2t+1} \cdot C_o^{AUB}
\end{aligned}$$

ここで、

$$C_o^{AUB} = \frac{V_{2t}(r)}{r^{2t}} \cdot \left\{ \frac{2\sqrt{\pi}\Gamma(t+1)}{\Gamma\left(t+\frac{3}{2}\right)} - B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right) \right\}$$

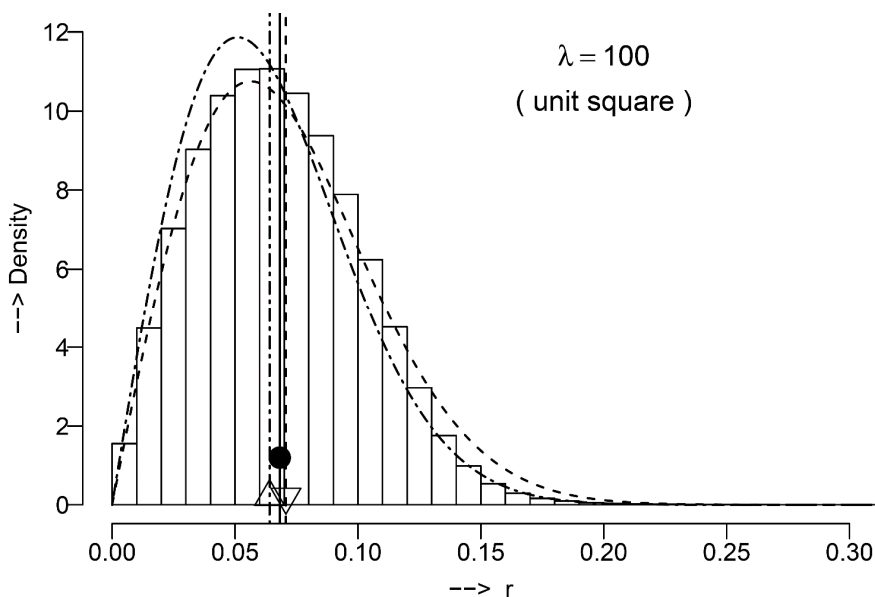
$B_{\frac{3}{4}}\left(t+1, \frac{1}{2}\right)$ は不完全ベータ関数。

なお、 $V_{2t}(r)$ は $(2t)$ 次元超球 (全球) の体積なので、 $\frac{V_{2t}(r)}{r^{2t}}$ は次元に依存する定数となる。例えば2次元なら π 、4次元なら $\frac{\pi^2}{2}$ になる。

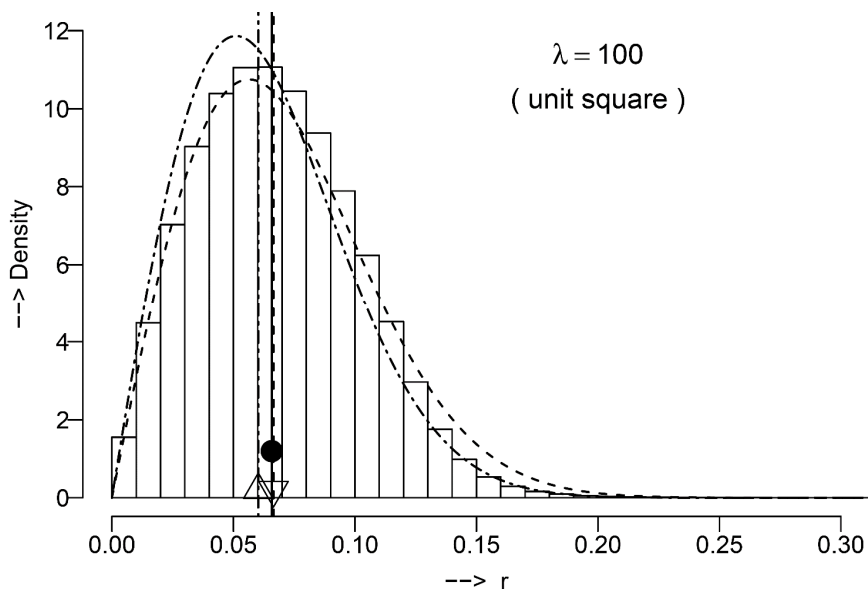
付録 3 : MST の辺長の期待値と中央値 (2 次元)

単位正方形内に強度 (λ) を変えて一様ランダムに発生させた点パターンに対する MST の辺長の分布を求めた。 ($\lambda = 10, 20, 30, \dots, 180, 190, 200$)

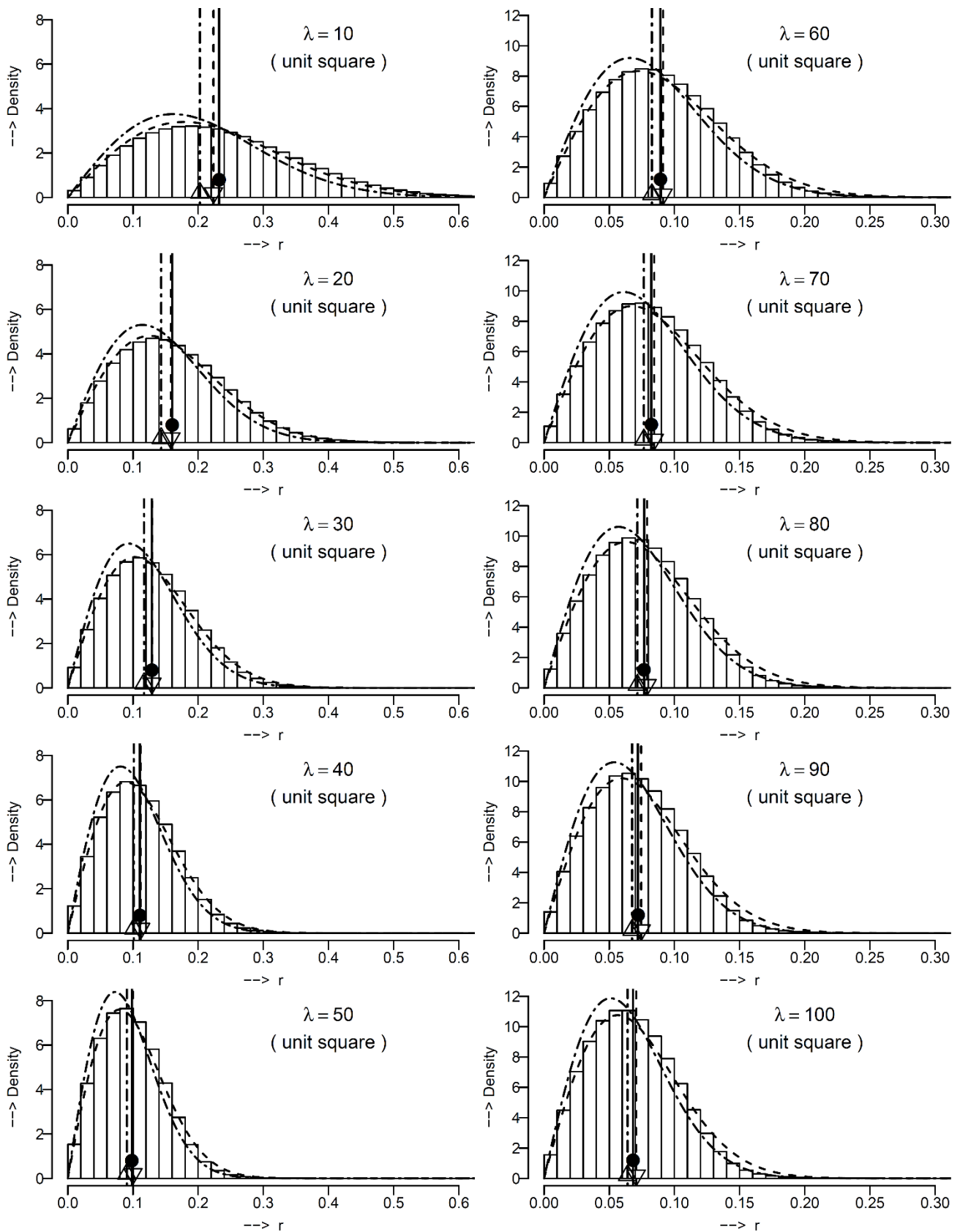
- 厳密分布 : ヒストグラム (モンテカルロ法, $m=10000$ 回)、期待値&中央値 (●)
- Roberts (1968) の近似分布 : 確率密度関数 (破線)、期待値&中央値 (▽)
- Watanabe (2008) の近似分布 : 確率密度関数 (一点鎖線)、期待値&中央値 (△)



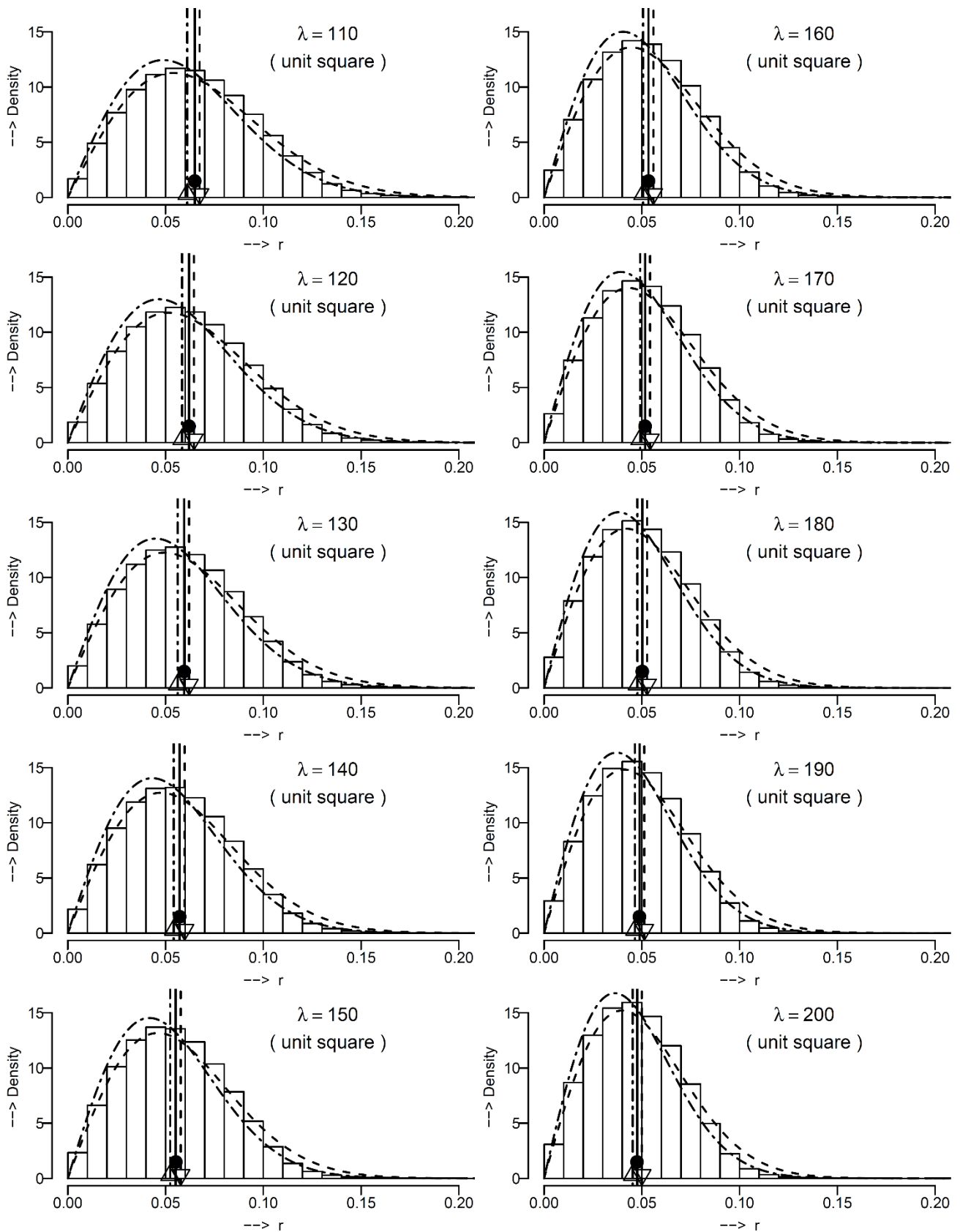
付図 3.1 MST の辺長の期待値 ($\lambda = 100$)



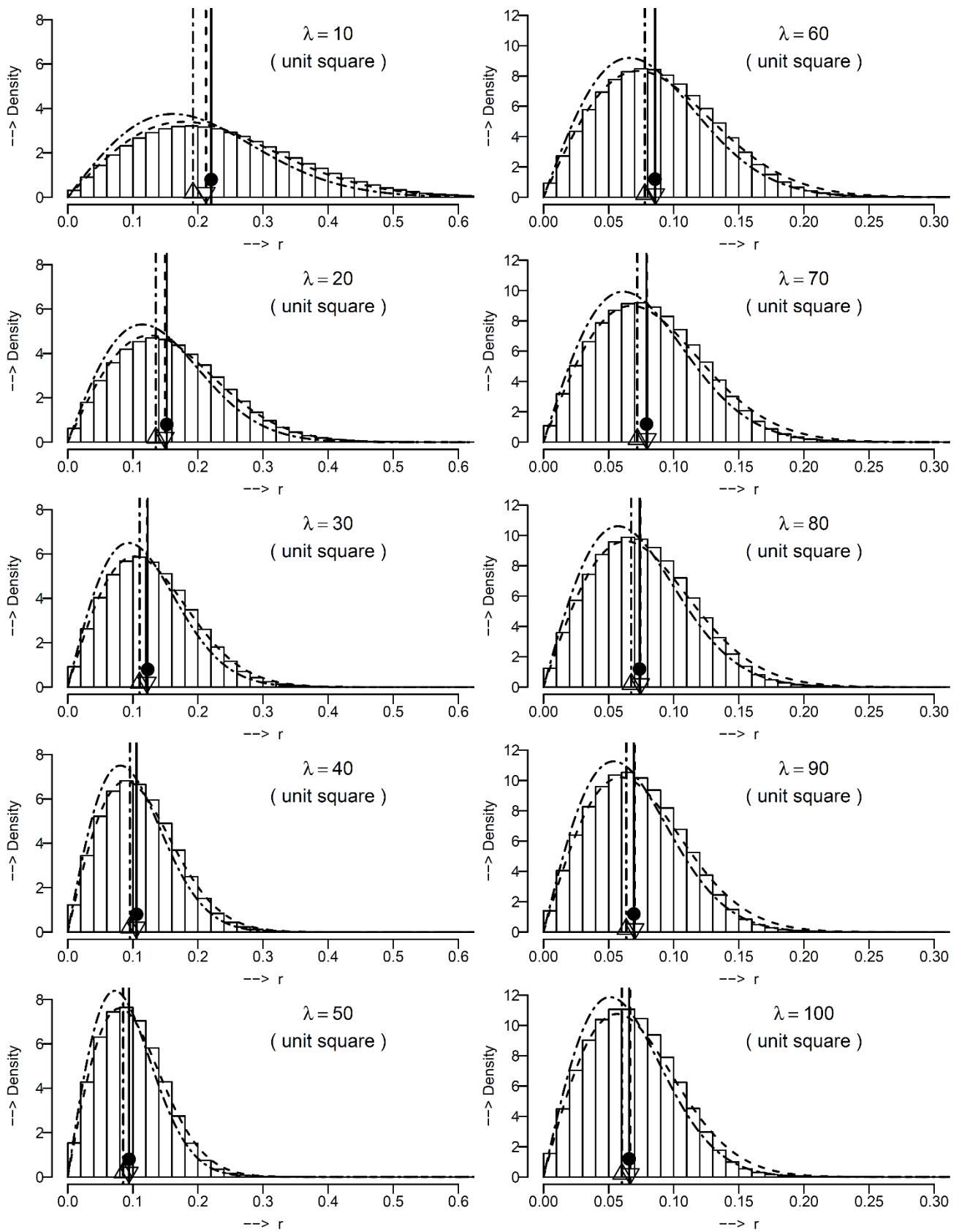
付図 3.2 MST の辺長の中央値 ($\lambda = 100$)



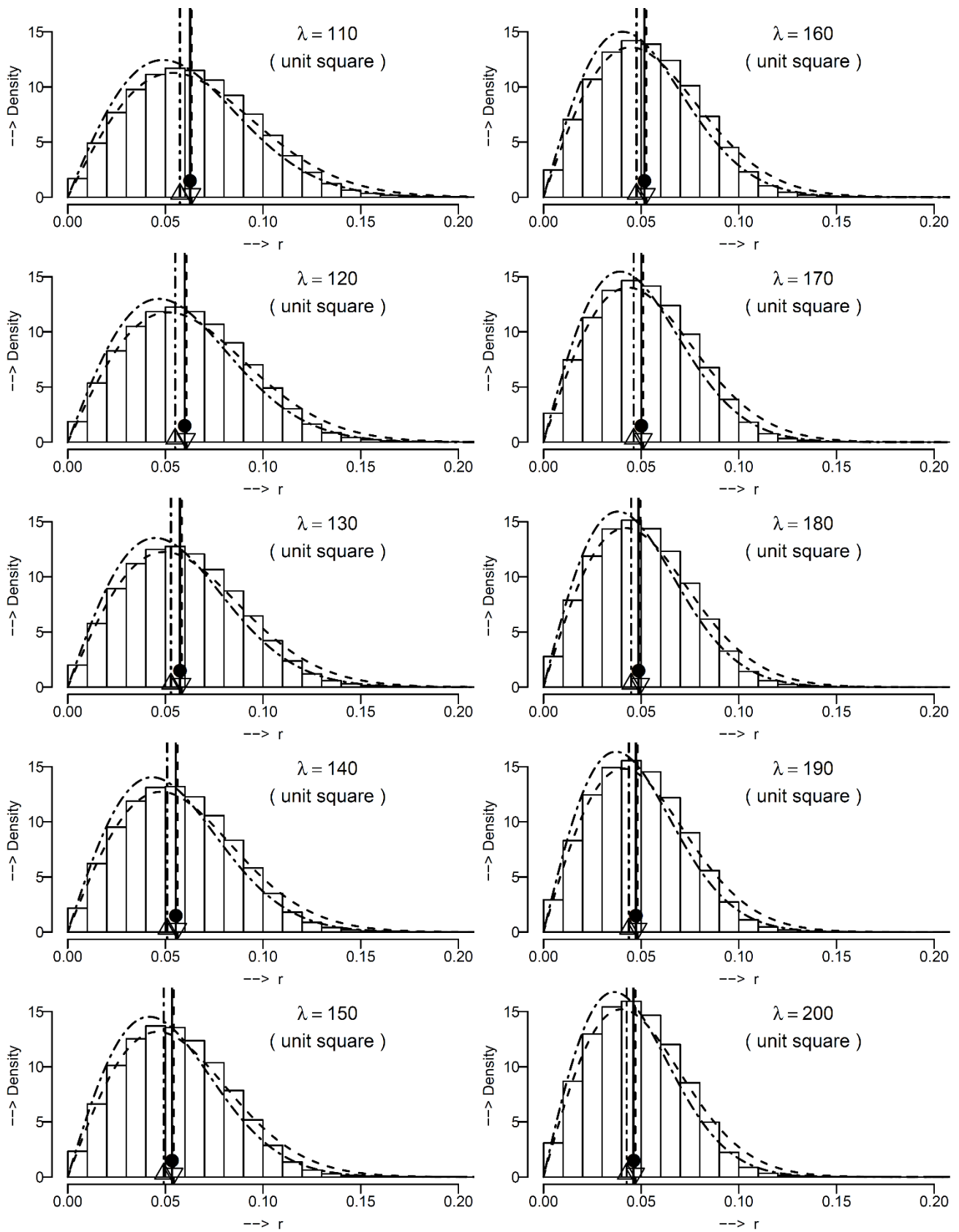
付図 3.3 (A) MST の辺長の期待値 ($\lambda = 10, 20, 30, \dots, 80, 90, 100$)



付図 3.3 (B) MST の辺長の期待値 ($\lambda = 110, 120, 130, \dots, 180, 190, 200$)



付図 3.4 (A) MST の辺長の中央値 ($\lambda = 10, 20, 30, \dots, 80, 90, 100$)



付図 3.4 (B) MST の辺長の中央値 ($\lambda = 110, 120, 130, \dots, 180, 190, 200$)

【謝辞】

はじめに、博士論文の審査員である 5 名の先生方にはご多忙のところ貴重な時間をさいて本論文の審査を引き受けて下さり、多くの貴重なご助言をいただきました。厚く御礼申し上げます。

【審査員の先生方】

総合研究大学院大学複合科学研究科統計科学専攻

(情報・システム研究機構 統計数理研究所)

教授 中野純司 (主査)

教授 田村義保

准教授 加藤昇吾

独立行政法人統計センター

理事長 椿 広計

青山学院大学経済学部

准教授 元山 斉

本研究の遂行および学位論文の執筆にあたり、主任指導教員の田村義保先生と副指導教員の加藤昇吾先生には、終始懇切丁寧な御指導を賜りました。謹んで感謝の意を表します。

大学院 1 年目の主任指導教員であった椿広計先生には、研究者としての心構えを教えていただきました。大学院 2 年目以降の主任指導教員であった田村義保先生には、計算機実験にあたり乱数の使い方に関して指導していただきました。主査になっていただいた中野純司先生からは、提案する手法と既存の手法の比較において、更に踏み込んだ比較を行うようにとの御指摘があり、新たな結論を導くことができました。加藤昇吾先生には学位論文中の全ての式を確認し、誤りを指摘していただきました。元山斉先生には提案する手法の他分野への応用に関して貴重な示唆をいただきました。これに関しては、ぜひ研究を進めていきたいと思っています。

統計科学に関心を持ち、大学院への進学を決めたのは、1990 年代に産学交流でお世話になった 3 名の先生方との出会いがきっかけでした。本研究で提案する方法の基になったアイデアも、この時に思いつ

いたものです。この場をお借りして御礼申し上げます。

【1990年代に産学交流でお世話になった先生方】

独立行政法人統計センター理事長 椿 広計（当時、慶応義塾大学理工学部講師）

電気通信大学副学長 椿 美智子（当時、電気通信大学電気通信学部助手）

電気通信大学講師 山本 渉（当時、慶応義塾大学理工学部大学院生）

総合研究大学院大学在学中は、統計数理研究所の教官の皆さま、企画グループ研究支援担当の皆さま、総合研究大学院大学複合科学研究科統計科学専攻の学生の皆さま、には大変お世話になりました。この場をお借りして御礼申し上げます。

最後に、ここまで見守ってくれた妻（朋子）に心から感謝します。

参考文献

- Anderberg, M.R. (1973). *CLUSTER ANALYSIS FOR APPLICATIONS*. Academic Press.
- Bertsimas, D. J. and van Ryzin. G. (1990). An asymptotic determination of the minimum spanning tree and minimum matching constants in geometrical probability. *Operations Research Letters*, Vol.9, Issue 4, pp.223-231.
- Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction*, Second Edition, John Wiley & Sons.
- Baker, F. B., Hubert, L. J. (1976). A Graph-Theoretic Approach to Goodness-of-Fit in Complete-Link Hierarchical Clustering. *Journal of the American Statistical Association*, Vol.71, No.356 (Dec., 1976), pp.870-878.
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. CRC, Boca Raton, FL.
- Clark, P. J. and Evans, F. C. (1954). Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations, *Ecology*, Vol.35, pp.445–453.
- Clark, P. J. and Evans, F. C. (1955). On Some Aspects of Spatial Pattern in Biological Populations, *Science*, New Series, Vol.121, No.3142, pp.397–398.
- Clark, P. J. (1956). Grouping in Spatial Distributions, *Science*, New Series, Vol.123, No.3192, pp.373–374.
- Clark, P. J. and Evans, F. C. (1979). Generalization of a Nearest Neighbor Measure of Dispersion for Use in K Dimensions, *Ecology*, Vol.60, No.2, pp.316–317.
- Cox, T. F. (1981). Reflexive Nearest Neighbours, *Biometrics*, Vol.37, No.2, pp.367-369.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data*, 2nd edition. John Wiley & Sons, Inc., New York.
- David, H. A. and Nagaraja, H. N. (2003). *Order Statistics* (3rd ed.), John Wiley & Sons.
- Diggle, P. J., Besag, J. & Gleaves, J.T. (1976). Statistical Analysis of Spatial Point Patterns by Means of Distance Methods. *Biometrics*, Vol.32, No.3, pp.659-667.
- Diggle, P. J. (1979). On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics*, Vol.35, pp.87-101.
- Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, London.

- Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*, 2nd edition. Academic Press, New York.
- Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, 3rd edition. Chapman and Hall/CRC.
- Dixon, P. M. (2002). Ripley's K function, in. *Encyclopedia of Environmetrics*, Vol.3, Ed. by. A.H. El-Shaarawi and W. W. Piegorsch. John Wiley & Sons, Ltd, Chichester, pp.1796–1803.
- Duranton, G., & Overman, H. G. (2005). Testing for localization using micro-geographic data. *The Review of Economic Studies*, Vol.72, No.4, pp.1077-1106.
- Everitt, B. S., Landau, S., Leese, M., Stahl, D. (2011). *Cluster Analysis*, 5th edition. John Wiley & Sons.
- 古山正雄. (2003). 地域間ネットワークにおける最短結合と近隣結合に関する理論的考察. 日本都市計画学会都市計画論文集. No.38-3. pp.379-384.
- Gatrell, A. C., Bailey, T. C., Diggle, P. J., & Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology, *Trans Inst Br Geogr NS*, Vol.21, pp.256-274.
- Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*, 2nd edition. John Wiley & Sons.
- Gilbert, E. N. (1965). RANDOM MINIMAL TREES, *J. Soc. INDUST. APPL. MATH.*, Vol.13, No.2, pp.376-387.
- Gower, J. C., Ross, J. S. (1969). Minimum Spanning Trees and Single Linkage Cluster Analysis. *Journal of the Royal Statistical Society*, Series C. Vol.18, No.1, pp.54-64
- Illian, J., Penttinen, A., Stoyan, H., Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*, John Wiley & Sons.
- Kaluzny, S. P., Vega, S. C., Cardoso, T. P. and Shelly, A. A. (1996). *S+SPATIALSTATS User's Manual for Windows and UNIX*. Springer-Verlag.
- 笠原一人・古山正雄. (1998). 最短木および階層を有する木の長さに関する考察. 日本建築学会計画系論文集. 第504号. pp.155-161.
- Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and a traveling salesman problem. *Proc. Amer. Math. Soc.*, Vol.7, pp.48-50.

- Legendre, P. and Legendre, L. (1998). *Numerical Ecology*, 2nd edition. Elsevier Science.
- Matérn, B. (1986). *Spatial Variation*, Lecture Notes in Statistics 36, 2nd edition. Springer-Verlag, Berlin
- Murtagh, F. (1983). A Survey of Recent Advances in Hierarchical Clustering Algorithms. *THE COMPUTER JOURNAL*, Vol.26, No.4. pp.354-359.
- 間瀬茂・武田純. (2001). 空間データモデリング. 共立出版.
- 間瀬茂. (2010). 地球統計学とクリギング法-RとgeoRによるデータ解析-. オーム社.
- Nakamura, N and Ohsumi, N. (1990). Space-distorting properties in agglomerative hierarchical clustering algorithms, Research memorandum #387, *Technical Report of The Institute of Statistical Mathematics*.
- Neyman, J. and Scott, E. L. (1958). Statistical Approach to Problems of Cosmology. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol.20, No.1, pp.1-43.
- Ogata, Y. (1999). Seismicity analysis through point-process modeling: A review, *Pure Appl. Geophys*, Vol.155, pp. 471-507.
- 大隅昇・中村隆. (1981). 分類系統樹の性質とその比較評価方式. 統計数理研究所彙報. 第28巻. 第1号. pp.117-133.
- 大隅昇・L.ルバール・A.モリノウ・K.M.ワーウィック・馬場康維. (1994). 記述的多変量解析法. 日科技連出版社.
- Okabe, A., and Yamada, I. (2001). The *K*-function method on a network and its computational implementation. *Geographical Analysis*, Vol.33, No.3, pp.271-290.
- 奥野忠一・久米均・芳賀敏郎・吉澤正. (1986). ≪改訂版≫ 多変量解析法. 日科技連出版社.
- 奥野忠一・片山善三郎・上郡長昭・伊藤哲二・入倉則夫・藤原信夫. (1986). 工業における多変量データの解析. 日科技連出版社.
- Page, R.D.M. and Holmes, E.C. (1998). *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, Oxford.
- Penrose, M.D. (1997). THE LONGEST EDGE OF THE RANDOM MINIMAL SPANNING TREE, *The Annals of*

Applied Probability, Vol.7, No.2, pp.340-361.

Perry, G. L. W., Miller, B. P., & Enright, N. J. (2006). A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant Ecology*, Vol.187, pp.59-82.

Piccarreta, R. and Billari, F. C. (2007). Clustering work and family trajectories by using a divisive algorithm. *Journal of the Royal Statistical Society A*, Vol.170, pp.1061–1078.

Pickard, D. K. (1982). Isolated nearest neighbors, *Journal of Applied Probability*, Vol.19, No.2, pp.444-449.

Pielou, E.C. (1969). *An Introduction to Mathematical Ecology*, John Wiley & Sons, Inc.

Prim, R. C. (1957). Shortest Connection Networks And Some Generalizations. *Bell System Technical Journal*, Vol.36, Issue 6, pp.1389-1401.

Pollard, J. H. (1971). On Distance Estimator of Density in Randomly Distributed Forests. *Biometrics*, Vol.27, No.4, pp.991-1002.

Ripley, B.D. (1976). The Second-Order Analysis of Stationary Point processes. *Journal of Applied Probability*, Vol.13, pp.255-266.

Ripley, B.D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society. Series B*, Vol.39, No.2, pp.172–212.

Roberts, F. D. K. (1968). Random minimal trees, *Biometrika*, Vol.55, pp.255–258.

Schneider, J.W. and Borlund, P. (2007a). Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology*, Vol.58, pp.1586–1595.

Schneider, J. W. and Borlund, P. (2007b). Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by the use of the Mantel and Procrustes statistics. *Journal of the American Society for Information Science and Technology*, Vol.58, pp.1596–1609.

Sneath, P.H.A. (1966). A comparison of different clustering methods as applied to randomly spaced points. *Classification Society Bulletin*, 1966, 1, pp.2-18.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy*. W. H. Freeman, San Francisco.

- Sokal, R., & Rohlf, F. (1962). The comparison of dendograms by objective methods. *Taxon*, Vol.11, pp.33–40.
- Stoyan, D., and Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields*. John Wiley & Sons, Ltd, Chichester.
- Strauss, D. J. (1975). A model for clustering. *Biometrika*, Vol.62, pp.467-475.
- Takai, T., Tamura, Y. and Motoyama, H. (2017). A NEW GRAPHICAL APPROACH TO CLASSIFY SPATIAL POINT PATTERNS BASED ON HIERARCHICAL CLUSTER ANALYSIS. *J. Jpn. Soc. Comp. Statist*, Vol.30, pp.1-14.
- 高井勉・田村義保・元山齊. (2018). AGsi-curve の数理的性質 (投稿中)
- Thompson, H. R. (1956). Distribution of Distance to Nth Neighbour in a Population of Randomly Distributed Individuals, *Ecology*, Vol.37, No.2, pp.391-394.
- Ward, J. H. Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, Vol. 58, No. 301, pp.236-244
- Watanabe, D. (2008). Evaluating the Configuration and the Travel Efficiency on Proximity Graphs as Transportation Networks, *Forma*, Vol.23, pp.81–87.
- Wiegand, T. & Moloney, K.A. (2004). Rings, circles, and null-models for point pattern analysis in ecology. *Oikos*, Vol.104, pp.209–229.
- Wishart, D. (1969). 256. Note: An Algorithm for Hierarchical Classifications. *Biometrics*, Vol.25, No.1, pp.165-170.
- Wilk, M. B., and Gnanadesikan, R. (1968). Probability Plotting Methods for the Analysis of Data. *Biometrika*, Vol.55, No.1, pp.1-17.
- 柳井春夫・高木廣文・市川雅教・服部芳明・佐藤俊哉・丸井英二. (1986). 多変量解析ハンドブック. 現代数学社.
- Yao, Y., and Simons, G. (1996). A large-Dimensional Independent and Identically Distributed Property for Nearest Neighbor Counts in Poisson Processes. *The Annals of Applied Probability*, Vol.6, No.2, pp.561-571.
- Zahn, C. T. (1971). Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers*, Vol.C-20, Issue.1, pp.68-86.