Statistics of Random Graphs on Clustering Point Sets

Joe Yukich

Lehigh University

Random Geometric Graphs and Their Applications to Complex Networks, Banff November 6-11, 2016

Based on joint work with B. Błaszczyszyn and D. Yogeshwaran

Introduction

Questions pertaining to geometric statistics on input (data) $\mathcal{X} \subset \mathbb{R}^d$ often involve analyzing sums

$$\sum_{x \in \mathcal{X}} \xi(x, \mathcal{X}),$$

where the $\mathbb{R}$-valued score function ξ, defined on pairs $(x, \mathcal{X})$, represents the interaction of x with respect to $\mathcal{X}$. The sums describe some global feature of the data in terms of local contributions $\xi(x, \mathcal{X}),\ x \in \mathcal{X}$.

Introduction

Questions pertaining to geometric statistics on input (data) $\mathcal{X} \subset \mathbb{R}^d$ often involve analyzing sums

$$\sum_{x \in \mathcal{X}} \xi(x, \mathcal{X}),$$

where the $\mathbb{R}$-valued score function ξ, defined on pairs $(x, \mathcal{X})$, represents the interaction of x with respect to $\mathcal{X}$. The sums describe some global feature of the data in terms of local contributions $\xi(x, \mathcal{X}),\ x \in \mathcal{X}$.

When $\mathcal{X}$ is random the scores $\xi(x, \mathcal{X})$ are spatially correlated.

Ex. 1: Statistics of geometric graphs

Clique counts. $\mathcal{X} \subset \mathbb{R}^d$ finite, $r \in (0, \infty)$.

· Join two points of $\mathcal{X}$ iff they are at distance at most r. Vietoris-Rips complex (with parameter r) is simplicial complex whose k-simplices correspond to unordered $(k+1)$-tuples of points in $\mathcal{X}$ all pairwise within r of each other. For $k \in \mathbb{N}$ and $x \in \mathcal{X}$ put

· $\xi_k(x, \mathcal{X}) := \frac{\text{number of } k\text{-simplices in V-R complex containing } x}{k+1}$

Ex. 1: Statistics of geometric graphs

Clique counts. $\mathcal{X} \subset \mathbb{R}^d$ finite, $r \in (0, \infty)$.

· Join two points of $\mathcal{X}$ iff they are at distance at most r. Vietoris-Rips complex (with parameter r) is simplicial complex whose k-simplices correspond to unordered $(k+1)$-tuples of points in $\mathcal{X}$ all pairwise within r of each other. For $k \in \mathbb{N}$ and $x \in \mathcal{X}$ put

· $\xi_k(x, \mathcal{X}) := \frac{\text{number of } k\text{-simplices in V-R complex containing } x}{k+1}$

· Total number of k-simplices in V-R complex: $\sum_{x \in \mathcal{X}} \xi_k(x, \mathcal{X})$.

· Chatterjee; Decreusefond et al.; Kahle + Meckes; Lachièze-Rey + Peccati; Penrose; Penrose + Y; Reitzner + Schulte; Thäle; Yogeshwaran + Adler.

Ex. 2: Statistics of nearest neighbor graphs

Total edge length. $\mathcal{X} \subset \mathbb{R}^d$ finite. Given $x \in \mathcal{X}$, let $x^{NN} \in \mathcal{X}$ be the nearest neighbor of x.

· Undirected nearest neighbor graph on $\mathcal{X}$: include an edge $\{x, y\}$ if $y = x^{NN}$ and/or $x = y^{NN}$.

· For $x \in \mathcal{X}$, put

$$\xi_{NN}(x, \mathcal{X}) := \begin{cases} \frac{1}{2}||x - x^{NN}|| & \text{if } \; x, x^{NN} \; \text{ are mutual n.n.} \\ ||x - x^{NN}|| & \text{otherwise.} \end{cases}$$

Ex. 2: Statistics of nearest neighbor graphs

Total edge length. $\mathcal{X} \subset \mathbb{R}^d$ finite. Given $x \in \mathcal{X}$, let $x^{NN} \in \mathcal{X}$ be the nearest neighbor of x.

· Undirected nearest neighbor graph on $\mathcal{X}$: include an edge $\{x, y\}$ if $y = x^{NN}$ and/or $x = y^{NN}$.

· For $x \in \mathcal{X}$, put

$$\xi_{NN}(x, \mathcal{X}) := \begin{cases} \frac{1}{2}||x - x^{NN}|| & \text{if} \quad x, x^{NN} \quad \text{are mutual n.n.} \\ ||x - x^{NN}|| & \text{otherwise.} \end{cases}$$

· Total edge length of NN graph on $\mathcal{X}$: $\sum_{x \in \mathcal{X}} \xi_{NN}(x, \mathcal{X})$.

· Bickel + Breiman; Barbour + Xia; Chatterjee; Last, Peccati + Schulte; Penrose + Y; Quiroz; Steele.

Ex. 3: Minimal spanning tree

$\mathcal{X} \subset \mathbb{R}^d$ finite. $\mathcal{E}(x) :=$ edges in $MST(\mathcal{X})$ containing x.

· For $x \in \mathcal{X}$, put

$$\xi_{MST}(x, \mathcal{X}) := \frac{1}{2} \sum_{e \in \mathcal{E}(x)} |e|.$$

Ex. 3: Minimal spanning tree

$\mathcal{X} \subset \mathbb{R}^d$ finite. $\mathcal{E}(x) :=$ edges in $MST(\mathcal{X})$ containing x.

· For $x \in \mathcal{X}$, put

$$\xi_{MST}(x, \mathcal{X}) := \frac{1}{2} \sum_{e \in \mathcal{E}(x)} |e|.$$

· Total edge length of MST: $L_{MST}(\mathcal{X}) := \sum_{x \in \mathcal{X}} \xi_{MST}(x, \mathcal{X})$.

· Aldous + Steele; Chatterjee + Sen; Kesten + Lee; Penrose + Y; Steele.

General questions

· When $\mathcal{P} \subset \mathbb{R}^d$ is a random pt configuration, the sums $\sum_{x \in \mathcal{P}} \xi(x, \mathcal{P})$ describe a global feature of the data.

General questions

· When $\mathcal{P} \subset \mathbb{R}^d$ is a random pt configuration, the sums $\sum_{x \in \mathcal{P}} \xi(x, \mathcal{P})$ describe a global feature of the data.

· **Question.** What is the distribution of these sums for large pt configurations $\mathcal{P}$? LLN? CLT? Second order asymptotics?

· We describe a methodology for answering these questions.

Goals

$\mathcal{P}$: a stationary point process on $\mathbb{R}^d$

Restrict to windows: $\mathcal{P}_n := \mathcal{P} \cap [-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$.

Goals

$\mathcal{P}$: a stationary point process on $\mathbb{R}^d$

Restrict to windows: $\mathcal{P}_n := \mathcal{P} \cap [-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$.

Goal. Given a score function $\xi(\cdot, \cdot)$ defined on pairs $(x, \mathcal{X})$, given a pt process $\mathcal{P}$, we seek the limit theory (LLN, CLT, variance asymptotics) for the total score

$$\sum_{x \in \mathcal{P}_n} \xi(x, \mathcal{P}_n)$$

and total measure

$$\mu_n^\xi := \sum_{x \in \mathcal{P}_n} \xi(x, \mathcal{P}_n) \delta_{n^{-1/d} x}.$$

Goals

$\mathcal{P}$: a stationary point process on $\mathbb{R}^d$

Restrict to windows: $\mathcal{P}_n := \mathcal{P} \cap [-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$.

Goal. Given a score function $\xi(\cdot,\cdot)$ defined on pairs $(x, \mathcal{X})$, given a pt process $\mathcal{P}$, we seek the limit theory (LLN, CLT, variance asymptotics) for the total score

$$\sum_{x \in \mathcal{P}_n} \xi(x, \mathcal{P}_n)$$

and total measure

$$\mu_n^\xi := \sum_{x \in \mathcal{P}_n} \xi(x, \mathcal{P}_n) \delta_{n^{-1/d}x}.$$

Tractable problems must be *local* in the sense that points far away from x should not play a role in the evaluation of the score $\xi(x, \mathcal{P}_n)$.

Stabilization

We assume translation invariant scores: $\xi(x, \mathcal{X}) = \xi(\mathbf{0}, \mathcal{X} - x)$.

Recall $\mathcal{P}_n := \mathcal{P} \cap [-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$

Stabilization

We assume translation invariant scores: $\xi(x, \mathcal{X}) = \xi(\mathbf{0}, \mathcal{X} - x)$.

Recall $\mathcal{P}_n := \mathcal{P} \cap [-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$

Key Definition. ξ is *stabilizing* wrt pt process $\mathcal{P}$ on $\mathbb{R}^d$ if for all $x \in \mathbb{R}^d$ there is $R := R^\xi(x, \mathcal{P}) < \infty$ a.s. (a 'radius of stabilization') such that

$$\xi(x, \mathcal{P} \cap B_R(x)) = \xi(x, (\mathcal{P} \cap B_R(x)) \cup \mathcal{A})$$

for any locally finite $\mathcal{A} \subset \mathbb{R}^d \setminus B_R(x)$.

Stabilization

We assume translation invariant scores: $\xi(x, \mathcal{X}) = \xi(\mathbf{0}, \mathcal{X} - x)$.

Recall $\mathcal{P}_n := \mathcal{P} \cap [-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$

Key Definition. ξ is *stabilizing* wrt pt process $\mathcal{P}$ on $\mathbb{R}^d$ if for all $x \in \mathbb{R}^d$ there is $R := R^{\xi}(x, \mathcal{P}) < \infty$ a.s. (a 'radius of stabilization') such that

$$\xi(x, \mathcal{P} \cap B_R(x)) = \xi(x, (\mathcal{P} \cap B_R(x)) \cup \mathcal{A})$$

for any locally finite $\mathcal{A} \subset \mathbb{R}^d \setminus B_R(x)$. ξ is *exponentially stabilizing* wrt $\mathcal{P}$ if there is a constant $c > 0$ such that

$$\sup_{x \in \mathbb{R}^d} \sup_{n \in \mathbb{N}} P[R^{\xi}(x, \mathcal{P}_n) \geq r] \leq c \exp(\frac{-r}{c}), \quad r \in [1, \infty).$$

Moment condition

$\mathcal{P}$: a pt process on $\mathbb{R}^d$; $\mathcal{P}_n := \mathcal{P} \cap [-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$.

Definition. Let $p \in [1, \infty)$. ξ satisfies the p moment condition wrt $\mathcal{P}$ if

$$\sup_{n \in \mathbb{N}} \sup_{x,y \in \mathbb{R}^d} \mathbb{E}\,|\xi(x, \mathcal{P}_n \cup \{y\})|^p < \infty.$$

Weak law of large numbers for Poisson input $\mathcal{H}$

Let $\mathcal{H}$ be a rate 1 Poisson pt process on $\mathbb{R}^d$; $\mathcal{H}_n := \mathcal{H} \cap [\frac{-n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$.

$$\mu_n^\xi := \sum_{x \in \mathcal{H}_n} \xi(x, \mathcal{H}_n) \delta_{n^{-1/d}x}.$$

Thm (WLLN): If ξ is stabilizing wrt $\mathcal{H}$, if ξ satisfies the p moment condition for some $p \in (1, \infty)$, then for all $f \in B([-\frac{1}{2}, \frac{1}{2}]^d)$ we have

$$|n^{-1}\mathbb{E}\,\langle \mu_n^\xi, f\rangle - \mathbb{E}\,\xi(\mathbf{0}, \mathcal{H} \cup \{\mathbf{0}\}) \int_{[-\frac{1}{2},\frac{1}{2}]^d} f(x)dx| \leq \epsilon_n.$$

Weak law of large numbers for Poisson input $\mathcal{H}$

Let $\mathcal{H}$ be a rate 1 Poisson pt process on $\mathbb{R}^d$; $\mathcal{H}_n := \mathcal{H} \cap [\frac{-n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$.

$$\mu_n^\xi := \sum_{x \in \mathcal{H}_n} \xi(x, \mathcal{H}_n) \delta_{n^{-1/d}x}.$$

Thm (WLLN): If ξ is stabilizing wrt $\mathcal{H}$, if ξ satisfies the p moment condition for some $p \in (1, \infty)$, then for all $f \in B([-\frac{1}{2}, \frac{1}{2}]^d)$ we have

$$|n^{-1} \mathbb{E} \langle \mu_n^\xi, f \rangle - \mathbb{E}\, \xi(\mathbf{0}, \mathcal{H} \cup \{\mathbf{0}\}) \int_{[-\frac{1}{2}, \frac{1}{2}]^d} f(x) dx| \leq \epsilon_n.$$

Penrose and Y (2003): $\epsilon_n = o(1)$.

Schulte + Y (2016): $\epsilon_n = O(n^{-1/d})$ if ξ is exponentially stabilizing wrt $\mathcal{H}$.

Gaussian fluctuations for Poisson input $\mathcal{H}$ on $\mathbb{R}^d$

Recall $\mu_n^\xi := \sum_{x \in \mathcal{H}_n} \xi(x, \mathcal{H}_n) \delta_{n^{-1/d} x}$.

Thm (CLT): Assume ξ is exponentially stabilizing wrt $\mathcal{H}$ and that ξ satisfies the p moment condition for some $p \in (4, \infty)$. If $f \in B([-\frac{1}{2}, \frac{1}{2}]^d)$ satisfies $\mathrm{Var}\langle \mu_n^\xi, f \rangle = \Omega(n)$, then

$$\sup_{t \in \mathbb{R}} \left| P\left[\frac{\langle \mu_n^\xi, f \rangle - \mathbb{E}\langle \mu_n^\xi, f \rangle}{\sqrt{\mathrm{Var}\langle \mu_n^\xi, f \rangle}} \leq t \right] - P[N \leq t] \right| \leq \epsilon_n.$$

Gaussian fluctuations for Poisson input $\mathcal{H}$ on $\mathbb{R}^d$

Recall $\mu_n^\xi := \sum_{x \in \mathcal{H}_n} \xi(x, \mathcal{H}_n)\delta_{n^{-1/d}x}$.

Thm (CLT): Assume ξ is exponentially stabilizing wrt $\mathcal{H}$ and that ξ satisfies the p moment condition for some $p \in (4, \infty)$. If $f \in B([-\frac{1}{2}, \frac{1}{2}]^d)$ satisfies $\mathrm{Var}\langle \mu_n^\xi, f\rangle = \Omega(n)$, then

$$\sup_{t \in \mathbb{R}} \left| P\left[\frac{\langle \mu_n^\xi, f\rangle - \mathbb{E}\langle \mu_n^\xi, f\rangle}{\sqrt{\mathrm{Var}\langle \mu_n^\xi, f\rangle}} \leq t \right] - P[N \leq t] \right| \leq \epsilon_n.$$

Penrose + Y (2005), Penrose (2007): $\epsilon_n = O((\log n)^{3d} n^{-1/2})$.

Last, Peccati + Schulte (2016): $\epsilon_n = \gamma_1 + ... + \gamma_5$.

Lachièze-Rey, Schulte + Y (2016): $\epsilon_n = O(n^{-1/2})$.

Variance asymptotics for Poisson input; volume order fluctuations

Given homogenous rate 1 Poisson pt process $\mathcal{H}$ on $\mathbb{R}^d$, and a score ξ, put

$$\sigma^2(\xi) = \mathbb{E}\,\xi^2(\mathbf{0},\mathcal{H}) + \int_{\mathbb{R}^d} \mathbb{E}\,\xi(\mathbf{0},\mathcal{H}\cup\{x\})\xi(x,\mathcal{H}\cup\{\mathbf{0}\}) - \mathbb{E}\,\xi(\mathbf{0},\mathcal{H})\mathbb{E}\,\xi(x,\mathcal{H})dx.$$

Variance asymptotics for Poisson input; volume order fluctuations

Given homogenous rate 1 Poisson pt process $\mathcal{H}$ on $\mathbb{R}^d$, and a score ξ, put

$$\sigma^2(\xi) = \mathbb{E}\,\xi^2(\mathbf{0},\mathcal{H}) + \int_{\mathbb{R}^d} \mathbb{E}\,\xi(\mathbf{0},\mathcal{H}\cup\{x\})\xi(x,\mathcal{H}\cup\{\mathbf{0}\}) - \mathbb{E}\,\xi(\mathbf{0},\mathcal{H})\mathbb{E}\,\xi(x,\mathcal{H})dx.$$

Thm (variance asymptotics): If ξ is exponentially stabilizing wrt $\mathcal{H}$, if ξ satisfies the p moment condition for some $p \in (2,\infty)$, then for all $f \in B([-\frac{1}{2},\frac{1}{2}]^d)$ we have

$$\lim_{n\to\infty} n^{-1}\mathrm{Var}\langle \mu_n^{\xi}, f\rangle = \sigma^2(\xi)\int_{[-\frac{1}{2},\frac{1}{2}]^d} f^2(x)dx \in [0,\infty).$$

Baryshnikov + Y (2005); Penrose (2007)

· **Question.** If the input pt process is neither a Poisson nor a binomial pt process, when do we get results which are qualitatively similar?

· Soshnikov (2002) and Shirai + Takahashi (2003): establish CLT for the *linear* statistics

$$\sum_{x \in \mathcal{P}_n} \delta_{n^{-1/d} x}$$

where $\mathcal{P}$ is determinantal pt process, $\mathcal{P}_n := \mathcal{P} \cap [\frac{-n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$.

· **Question.** If the input pt process is neither a Poisson nor a binomial pt process, when do we get results which are qualitatively similar?

· Soshnikov (2002) and Shirai + Takahashi (2003): establish CLT for the *linear* statistics

$$\sum_{x \in \mathcal{P}_n} \delta_{n^{-1/d}x}$$

where $\mathcal{P}$ is determinantal pt process, $\mathcal{P}_n := \mathcal{P} \cap [\frac{-n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$.

· Nazarov and Sodin (2012): establish CLT for the *linear* statistics

$$\sum_{x \in \mathcal{P}_n} \delta_{n^{-1/d}x}$$

where $\mathcal{P}$ is zero set of Gaussian analytic function, $\mathcal{P}_n := \mathcal{P} \cap [\frac{-n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$.

· We want to extend these results to non-linear statistics

$$\mu_n^{\xi} := \sum_{x \in \mathcal{P}_n} \xi(x, \mathcal{P}_n) \delta_{n^{-1/d}x}.$$

Clustering pt processes

Def. Given a pt process $\mathcal{P}$ on $\mathbb{R}^d$, the k pt correlation function $\rho^{(k)} : (\mathbb{R}^d)^k \to [0, \infty)$ is defined via

$$\mathbb{E}\left[\Pi_{i=1}^k \mathrm{card}(\mathcal{P} \cap B_i)\right] = \int_{B_1} ... \int_{B_k} \rho^{(k)}(x_1, ..., x_k) dx_1 ... dx_k,$$

where $B_1, ..., B_k$ are disjoint subsets of $\mathbb{R}^d$.

Clustering pt processes

Def. Given a pt process $\mathcal{P}$ on $\mathbb{R}^d$, the k pt correlation function $\rho^{(k)} : (\mathbb{R}^d)^k \to [0, \infty)$ is defined via

$$\mathbb{E}\left[\Pi_{i=1}^k \text{card}(\mathcal{P} \cap B_i)\right] = \int_{B_1} ... \int_{B_k} \rho^{(k)}(x_1, ..., x_k) dx_1 ... dx_k,$$

where $B_1, ..., B_k$ are disjoint subsets of $\mathbb{R}^d$.

Rk. $\rho^{(k)}(x_1, ..., x_k) = \Pi_{i=1}^k \rho^{(1)}(x_i)$ characterizes the Poisson pt process

Clustering pt processes

Key Definition. A pt process $\mathcal{P}$ on $\mathbb{R}^d$ *clusters* if there is a fast decreasing function $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ such that for all $k \in \mathbb{N}$ there are constants c_k and C_k such that for all $x_1, ..., x_{p+q} \in \mathbb{R}^d$,

$$|\rho^{(p+q)}(x_1, ..., x_{p+q}) - \rho^{(p)}(x_1, ..., x_p)\rho^{(q)}(x_{p+1}, ..., x_{p+q})| \leq C_{p+q}\phi(-c_{p+q}s),$$

where $s := \inf_{i \in \{1,...,p\},\ j \in \{p+1,...,p+q\}} ||x_i - x_j||$.

Clustering pt processes

Key Definition. A pt process $\mathcal{P}$ on $\mathbb{R}^d$ *clusters* if there is a fast decreasing function $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ such that for all $k \in \mathbb{N}$ there are constants c_k and C_k such that for all $x_1, ..., x_{p+q} \in \mathbb{R}^d$,

$$|\rho^{(p+q)}(x_1, ..., x_{p+q}) - \rho^{(p)}(x_1, ..., x_p)\rho^{(q)}(x_{p+1}, ..., x_{p+q})| \leq C_{p+q}\phi(-c_{p+q}s),$$

where $s := \inf_{i \in \{1,...,p\},\ j \in \{p+1,...,p+q\}} ||x_i - x_j||$.

Remarks.

· 'fast decreasing' means ϕ decays faster than any (negative) power,

· clustering does not imply 'clumping'. Better to replace 'clustering' with **'weakly correlated'**

Ex. 1: Determinantal pt process

A pt process is determinantal (DPP) if its correlation functions satisfy

$$\rho^{(k)}(x_1, ..., x_k) = \det(K(x_i, x_j))_{1 \leq i \leq j \leq k},$$

where $K(\cdot, \cdot)$ is Hermitian kernel of locally trace class integral operator from $L^2(\mathbb{R}^d)$ to itself.

Ex. 1: Determinantal pt process

A pt process is determinantal (DPP) if its correlation functions satisfy

$$\rho^{(k)}(x_1, ..., x_k) = \det(K(x_i, x_j))_{1 \le i \le j \le k},$$

where $K(\cdot, \cdot)$ is Hermitian kernel of locally trace class integral operator from $L^2(\mathbb{R}^d)$ to itself.

Fact (Błaszczyszyn, Yogeshwaran + Y (2016)). If

$$|K(x, y)| \le \phi(||x - y||), \quad x, y \in \mathbb{R}^d,$$

with ϕ fast decreasing, then the associated DPP clusters.

Infinite Ginibre ensemble

Infinite Ginibre ensemble on complex plane clusters with kernel

$$K(z_1, z_2) = \exp\left(i\mathrm{Im}(z_1\bar{z}_2) - \frac{1}{2}||z_1 - z_2||^2\right), \quad z_1, z_2 \in \mathbb{C}.$$

Infinite Ginibre ensemble

Infinite Ginibre ensemble on complex plane clusters with kernel

$$K(z_1, z_2) = \exp\left(i\mathrm{Im}(z_1\bar{z}_2) - \frac{1}{2}||z_1 - z_2||^2\right), \quad z_1, z_2 \in \mathbb{C}.$$

Ghosh, Khrishnapur, Peres (2016): Hole probabilities decay exponentially fast.

See also Błaszczyszyn, Yogeshwaran + Y (2016)

Ex. 2: Gaussian zero pt process

· Let $X_j, j \geq 1$, be i.i.d. standard complex Gaussians. Consider the Gaussian analytic function

$$F(z) := \sum_{j=1}^{\infty} \frac{X_j}{\sqrt{j!}} z^j, \quad z \in \mathbb{C}.$$

· Gaussian zero process $GAF := F^{-1}(\{0\})$ is trans. invariant

Ex. 2: Gaussian zero pt process

· Let $X_j, j \geq 1$, be i.i.d. standard complex Gaussians. Consider the Gaussian analytic function

$$F(z) := \sum_{j=1}^{\infty} \frac{X_j}{\sqrt{j!}} z^j, \quad z \in \mathbb{C}.$$

· Gaussian zero process $GAF := F^{-1}(\{0\})$ is trans. invariant

· GAF exhibits local repulsivity.

· GAF clusters (Nazarov and Sodin (2012)).

· Hole probabilities (Nishry (2010)):

$$\frac{-\log P(B(0,r) \cap GAF = \emptyset)}{r^4} \to c \in (0, \infty).$$

Other examples of clustering pt processes

· Permanental pt processes with fast decreasing kernel,

· Certain rarified Gibbs pt processes (Schreiber + Y, 2013),

· Convex geometry: Let $X_i, 1 \leq i \leq n$, be i.i.d. uniform on unit ball in $\mathbb{R}^d$. The angular coordinates of the extreme points, after re-scaling, converge to a clustering pt process on $\mathbb{R}^{d-1}$.

Weak law of large numbers for clustering input

Let $\mathcal{P}$ be clustering pt process on $\mathbb{R}^d$. Recall $\mathcal{P}_n := \mathcal{P} \cap [\frac{-n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$ and

$$\mu_n^\xi := \sum_{x \in \mathcal{P}_n} \xi(x, \mathcal{P}_n) \delta_{n^{-1/d}x}.$$

Thm (BYY '16): Assume

· ξ is stabilizing wrt $\mathcal{P}$

· ξ satisfies the p moment condition for some $p \in (1, \infty)$.

Weak law of large numbers for clustering input

Let $\mathcal{P}$ be clustering pt process on $\mathbb{R}^d$. Recall $\mathcal{P}_n := \mathcal{P} \cap [\frac{-n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$ and

$$\mu_n^\xi := \sum_{x \in \mathcal{P}_n} \xi(x, \mathcal{P}_n) \delta_{n^{-1/d}x}.$$

Thm (BYY '16): Assume

- ξ is stabilizing wrt $\mathcal{P}$
- ξ satisfies the p moment condition for some $p \in (1, \infty)$.

Then for all $f \in B([-\frac{1}{2}, \frac{1}{2}]^d)$ we have

$$\lim_{n \to \infty} n^{-1} \mathbb{E} \langle \mu_n^\xi, f \rangle = \mathbb{E}_{\mathbf{0}} \xi(\mathbf{0}, \mathcal{P} \cup \{\mathbf{0}\}) \int_{[-\frac{1}{2}, \frac{1}{2}]^d} f(x) dx \cdot \rho^{(1)}(\mathbf{0}).$$

Variance asymptotics for clustering input $\mathcal{P}$

- Given clustering input $\mathcal{P}$ on $\mathbb{R}^d$ and a score ξ, put

$$\sigma^2(\xi) := \mathbb{E}\,\xi^2(\mathbf{0}, \mathcal{P})\rho^{(1)}(\mathbf{0}) +$$

$$\int_{\mathbb{R}^d} \mathbb{E}\,\xi(\mathbf{0}, \mathcal{P} \cup x)\xi(x, \mathcal{P} \cup \mathbf{0})\rho^{(2)}(\mathbf{0}, x) - \mathbb{E}\,\xi(\mathbf{0}, \mathcal{P})\rho^{(1)}(\mathbf{0})\mathbb{E}\,\xi(x, \mathcal{P})\rho^{(1)}(x)dx.$$

Variance asymptotics for clustering input $\mathcal{P}$

- Given clustering input $\mathcal{P}$ on $\mathbb{R}^d$ and a score ξ, put

$$\sigma^2(\xi) := \mathbb{E}\, \xi^2(\mathbf{0}, \mathcal{P}) \rho^{(1)}(\mathbf{0}) +$$

$$\int_{\mathbb{R}^d} \mathbb{E}\, \xi(\mathbf{0}, \mathcal{P} \cup x) \xi(x, \mathcal{P} \cup \mathbf{0}) \rho^{(2)}(\mathbf{0}, x) - \mathbb{E}\, \xi(\mathbf{0}, \mathcal{P}) \rho^{(1)}(\mathbf{0}) \mathbb{E}\, \xi(x, \mathcal{P}) \rho^{(1)}(x) dx.$$

- **Thm (BYY '16)**: If ξ is exponentially stabilizing wrt $\mathcal{P}$, if ξ satisfies the p moment condition for some $p \in (2, \infty)$, then for all $f \in B([-\frac{1}{2}, \frac{1}{2}]^d)$ we have

$$\lim_{n \to \infty} n^{-1} \mathrm{Var} \langle \mu_n^{\xi}, f \rangle = \sigma^2(\xi) \int_{[-\frac{1}{2}, \frac{1}{2}]^d} f^2(x) dx \in [0, \infty).$$

- **Rk.** When $\mathcal{P}$ is determinantal with fast decreasing kernel this extends Soshnikov (2002), who assumes $\xi \equiv 1$.

Gaussian fluctuations for clustering input $\mathcal{P}$

We say that ξ obeys a power growth condition if

$$|\xi(x, \mathcal{X} \cap B_r(x))| \leq c(r \vee 1)^{\mathrm{card}(\mathcal{X} \cap \mathrm{B_r(x)})}, \quad r > 0,\ x \in \mathcal{X}.$$

We formulate two central limit theorems according to the localization properties of ξ.

Gaussian fluctuations for clustering input $\mathcal{P}$

Thm (BYY '16) $\mu_n^\xi := \sum_{x \in \mathcal{P}_n} \xi(x, \mathcal{P}_n) \delta_{n^{-1/d}x}$. Assume

· ξ has deterministic radius of stabilization wrt $\mathcal{P}$,

· ξ satisfies the power growth condition, p moment condition for some $p \in (2, \infty)$, and

· given $f \in B([-\frac{1}{2}, \frac{1}{2}]^d)$, $\mathrm{Var}\langle \mu_n^\xi, f \rangle = \Omega(n^\alpha)$ for some $\alpha > 0$.

Then as $n \to \infty$, we have

$$\frac{\langle \mu_n^\xi, f \rangle - \mathbb{E} \langle \mu_n^\xi, f \rangle}{\sqrt{\mathrm{Var}\langle \mu_n^\xi, f \rangle}} \xrightarrow{\mathcal{D}} N.$$

Gaussian fluctuations for clustering input $\mathcal{P}$

Thm (BYY '16) $\mu_n^\xi := \sum_{x \in \mathcal{P}_n} \xi(x, \mathcal{P}_n) \delta_{n^{-1/d}x}$. Assume

$\cdot$ ξ has deterministic radius of stabilization wrt $\mathcal{P}$,

$\cdot$ ξ satisfies the power growth condition, p moment condition for some $p \in (2, \infty)$, and

$\cdot$ given $f \in B([-\frac{1}{2}, \frac{1}{2}]^d)$, $\mathrm{Var}\langle \mu_n^\xi, f \rangle = \Omega(n^\alpha)$ for some $\alpha > 0$.

Then as $n \to \infty$, we have

$$\frac{\langle \mu_n^\xi, f \rangle - \mathbb{E} \langle \mu_n^\xi, f \rangle}{\sqrt{\mathrm{Var}\langle \mu_n^\xi, f \rangle}} \xrightarrow{\mathcal{D}} N.$$

Remarks. (a) When $\mathcal{P}$ is determinantal with fast decreasing kernel, this extends Soshnikov (2002) and Shirai + Takahashi (2003) who restrict to linear statistics $\sum_{x \in \mathcal{P}_n} \delta_{n^{-1/d}x}$, i.e., they put $\xi \equiv 1$.

Gaussian fluctuations for clustering input $\mathcal{P}$

Thm (BYY '16) $\mu_n^\xi := \sum_{x \in \mathcal{P}_n} \xi(x, \mathcal{P}_n) \delta_{n^{-1/d}x}$. Assume

$\cdot$ ξ has deterministic radius of stabilization wrt $\mathcal{P}$,

$\cdot$ ξ satisfies the power growth condition, p moment condition for some $p \in (2, \infty)$, and

$\cdot$ given $f \in B([-\frac{1}{2}, \frac{1}{2}]^d)$, $\mathrm{Var}\langle \mu_n^\xi, f \rangle = \Omega(n^\alpha)$ for some $\alpha > 0$.

Then as $n \to \infty$, we have

$$\frac{\langle \mu_n^\xi, f \rangle - \mathbb{E} \langle \mu_n^\xi, f \rangle}{\sqrt{\mathrm{Var}\langle \mu_n^\xi, f \rangle}} \xrightarrow{\mathcal{D}} N.$$

Remarks. (a) When $\mathcal{P}$ is determinantal with fast decreasing kernel, this extends Soshnikov (2002) and Shirai + Takahashi (2003) who restrict to linear statistics $\sum_{x \in \mathcal{P}_n} \delta_{n^{-1/d}x}$, i.e., they put $\xi \equiv 1$.
(b) When $\mathcal{P}$ is the Gaussian zero process GAF, this extends Nazarov and Sodin (2012), who also restrict to linear statistics.

Gaussian fluctuations for clustering input $\mathcal{P}$

Thm (BYY '16) $\mu_n^\xi := \sum_{x \in \mathcal{P}_n} \xi(x, \mathcal{P}_n) \delta_{n^{-1/d}x}$. Assume

· $\mathcal{P}$ clusters and clustering coeff. satisfy mild growth condition

· ξ is exponentially stabilizing wrt $\mathcal{P}$,

· ξ satisfies the power growth condition, p moment condition for some $p \in (2, \infty)$, and

· given $f \in B([-\frac{1}{2}, \frac{1}{2}]^d)$, $\mathrm{Var}\langle \mu_n^\xi, f \rangle = \Omega(n^\alpha)$ for some $\alpha > 0$. Then

$$\frac{\langle \mu_n^\xi, f \rangle - \mathbb{E} \langle \mu_n^\xi, f \rangle}{\sqrt{\mathrm{Var}\langle \mu_n^\xi, f \rangle}} \xrightarrow{\mathcal{D}} N.$$

Gaussian fluctuations for clustering input $\mathcal{P}$

Thm (BYY '16) $\mu_n^\xi := \sum_{x \in \mathcal{P}_n} \xi(x, \mathcal{P}_n) \delta_{n^{-1/d}x}$. Assume

· $\mathcal{P}$ clusters and clustering coeff. satisfy mild growth condition

· ξ is exponentially stabilizing wrt $\mathcal{P}$,

· ξ satisfies the power growth condition, p moment condition for some $p \in (2, \infty)$, and

· given $f \in B([-\frac{1}{2}, \frac{1}{2}]^d)$, $\mathrm{Var}\langle \mu_n^\xi, f \rangle = \Omega(n^\alpha)$ for some $\alpha > 0$. Then

$$\frac{\langle \mu_n^\xi, f \rangle - \mathbb{E}\langle \mu_n^\xi, f \rangle}{\sqrt{\mathrm{Var}\langle \mu_n^\xi, f \rangle}} \xrightarrow{\mathcal{D}} N.$$

Rk. If $\mathcal{P}$ is determinantal with fast decreasing kernel (e.g. Ginibre) then $\mathcal{P}$ satisfies stated condition.

Proof idea for CLT - cf Malyshev (1975)

· For k large, show that kth order cumulant for $\langle \mu_n^\xi, f \rangle / \sqrt{\mathrm{Var} \langle \mu_n^\xi, f \rangle}$ vanishes as $n \to \infty$.

· Given ξ, consider k mixed moment functions $m_{(k)} : (\mathbb{R}^d)^k \to \mathbb{R}$ given by

$$m_{(k)}(x_1, ..., x_k; \mathcal{P}_n) := \mathbb{E}\, \Pi_{i=1}^k \xi(x_i, \mathcal{P}_n) \rho^{(k)}(x_1, ..., x_k).$$

Proof idea for CLT - cf Malyshev (1975)

· For k large, show that kth order cumulant for $\langle \mu_n^\xi, f \rangle / \sqrt{\mathrm{Var}\langle \mu_n^\xi, f \rangle}$ vanishes as $n \to \infty$.

· Given ξ, consider k mixed moment functions $m_{(k)} : (\mathbb{R}^d)^k \to \mathbb{R}$ given by

$$m_{(k)}(x_1, ..., x_k; \mathcal{P}_n) := \mathbb{E}\, \Pi_{i=1}^k \xi(x_i, \mathcal{P}_n) \rho^{(k)}(x_1, ..., x_k).$$

· Need to show that the mixed moments 'cluster', that is for all $k \in \mathbb{N}$ there are constants c_k and C_k s.t. for all $x_1, ..., x_{p+q} \in \mathbb{R}^d$,

$$|m_{(p+q)}(x_1, ..., x_{p+q}) - m_{(p)}(x_1, ..., x_p) m_{(q)}(x_{p+1}, ..., x_{p+q})| \leq C_{p+q} \varphi(-c_{p+q} s$$

where φ is fast decreasing and

$$s := \inf_{i \in \{1,...,p\},\ j \in \{p+1,...,p+q\}} ||x_i - x_j||.$$

· $\mathcal{P}$ clusters and ξ exp. stabilizing $\Rightarrow$ mixed moments cluster.

Corollaries

1. Clique counts in geometric graph $G(\mathcal{X}, r)$.

$\cdot\ \xi_k(x, \mathcal{X}) := \frac{\text{number of } k\text{-simplices in V-R complex containing } x}{k+1}$

Corollaries

1. Clique counts in geometric graph $G(\mathcal{X}, r)$.

· $\xi_k(x, \mathcal{X}) := \frac{\text{number of } k\text{-simplices in V-R complex containing } x}{k+1}$

· k-simplex count: $N_k(\mathcal{X}) := \sum_{x \in \mathcal{X}} \xi_k(x, \mathcal{X})$.

Corollaries

1. Clique counts in geometric graph $G(\mathcal{X}, r)$.

· $\xi_k(x, \mathcal{X}) := \frac{\text{number of } k\text{-simplices in V-R complex containing } x}{k+1}$

· k-simplex count: $N_k(\mathcal{X}) := \sum_{x \in \mathcal{X}} \xi_k(x, \mathcal{X})$.

Theorem. Let $\mathcal{P}$ be any clustering point process (e.g., Ginibre ensemble, Gaussian zero process, permamental point process with fast decreasing kernel,...). Let $\mathcal{P}_n := \mathcal{P} \cap [-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$. Then

$$\lim_{n\to\infty} n^{-1} \mathbb{E}\, N_k(\mathcal{P}_n) = \mathbb{E}_{\mathbf{0}}[\xi_k(\mathbf{0}, \mathcal{P})] \rho^{(1)}(\mathbf{0}),$$

$$\lim_{n\to\infty} n^{-1} \mathrm{Var} N_k(\mathcal{P}_n) = \sigma^2(\xi_k) \in [0, \infty),$$

and, provided $\mathrm{Var} N_k(\mathcal{P}_n) = \Omega(n^\alpha)$, $\alpha > 0$, we have

$$\frac{N_k(\mathcal{P}_n) - \mathbb{E}\, N_k(\mathcal{P}_n)}{\sqrt{\mathrm{Var} N_k(\mathcal{P}_n)}} \xrightarrow{\mathcal{D}} N.$$

Corollaries

2. Total edge length in geometric graph $G(\mathcal{X}, r)$.

$\mathcal{P} \subset \mathbb{R}^d$ clustering pt process. $\mathcal{P}_n := \mathcal{P} \cap [-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$.
$\mathcal{E}(x) :=$ edges in $RGG(\mathcal{P}_n)$ containing $x \in \mathcal{P}_n$.

· For $x \in \mathcal{P}_n$, put

$$\xi_{RGG}(x, \mathcal{P}_n) := \frac{1}{2} \sum_{e \in \mathcal{E}(x)} |e|.$$

· Total edge length: $L_{RGG}(\mathcal{P}_n) := \sum_{x \in \mathcal{P}_n} \xi_{RGG}(x, \mathcal{P}_n)$.

2. Total edge length in geometric graph $G(\mathcal{X}, r)$.
$\mathcal{P} \subset \mathbb{R}^d$ clustering pt process. $\mathcal{P}_n := \mathcal{P} \cap [-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$.
$\mathcal{E}(x) :=$ edges in $RGG(\mathcal{P}_n)$ containing $x \in \mathcal{P}_n$.

· For $x \in \mathcal{P}_n$, put

$$\xi_{RGG}(x, \mathcal{P}_n) := \frac{1}{2} \sum_{e \in \mathcal{E}(x)} |e|.$$

· Total edge length: $L_{RGG}(\mathcal{P}_n) := \sum_{x \in \mathcal{P}_n} \xi_{RGG}(x, \mathcal{P}_n)$.

Theorem. Let $\mathcal{P}$ be any clustering pt process. Then

$$\lim_{n \to \infty} n^{-1} \mathbb{E}\, L_{RGG}(\mathcal{P}_n) = \mathbb{E}_{\mathbf{0}}[\xi_k(\mathbf{0}, \mathcal{P})] \rho^{(1)}(\mathbf{0}),$$

$$\lim_{n \to \infty} n^{-1} \mathrm{Var} L_{RGG}(\mathcal{P}_n) = \sigma^2(\xi_k) \in [0, \infty),$$

and, provided $\mathrm{Var} L_{RGG}(\mathcal{P}_n) = \Omega(n^\alpha)$, $\alpha > 0$, we have

$$\frac{L_{RGG}(\mathcal{P}_n) - \mathbb{E}\, L_{RGG}(\mathcal{P}_n)}{\sqrt{\mathrm{Var} L_{RGG}(\mathcal{P}_n)}} \xrightarrow{\mathcal{D}} N.$$

Corollaries

3. Total edge length in nearest neighbor graph.

· For $x \in \mathcal{X}$, put

$$\xi_{NN}(x, \mathcal{X}) := \begin{cases} \frac{1}{2}||x - x^{NN}|| & \text{if} \quad x, x^{NN} \quad \text{are mutual n.n.} \\ ||x - x^{NN}|| & \text{otherwise.} \end{cases}$$

· $L_{NN}(\mathcal{X}) := \sum_{x \in \mathcal{X}} \xi_{NN}(x, \mathcal{X})$.

Corollaries

3. Total edge length in nearest neighbor graph.

· For $x \in \mathcal{X}$, put

$$\xi_{NN}(x, \mathcal{X}) := \begin{cases} \frac{1}{2}||x - x^{NN}|| & \text{if } \ x, x^{NN} \ \text{ are mutual n.n.} \\ ||x - x^{NN}|| & \text{otherwise.} \end{cases}$$

· $L_{NN}(\mathcal{X}) := \sum_{x \in \mathcal{X}} \xi_{NN}(x, \mathcal{X})$.

Theorem. Let $\mathcal{P}$ be Ginibre ensemble $\mathcal{P}_n := \mathcal{P} \cap [-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d$. Then

$$\lim_{n \to \infty} n^{-1} \mathbb{E}\, L_{NN}(\mathcal{P}_n) = \mathbb{E}_{\mathbf{0}}[\xi_{NN}(\mathbf{0}, \mathcal{P})] \rho^{(1)}(\mathbf{0}),$$

$$\lim_{n \to \infty} n^{-1} \mathrm{Var} L_{NN}(\mathcal{P}_n) = \sigma^2(\xi_{NN}) \in [0, \infty),$$

and, provided $\mathrm{Var} L_{NN}(\mathcal{P}_n) = \Omega(n^{\alpha}),\ \alpha > 0$, we have

$$\frac{L_{NN}(\mathcal{P}_n) - \mathbb{E}\, L_{NN}(\mathcal{P}_n)}{\sqrt{\mathrm{Var} L_{NN}(\mathcal{P}_n)}} \xrightarrow{\mathcal{D}} N.$$

THANK YOU