Asymptotic results for topological data analysis of point processes

May 10, 2023

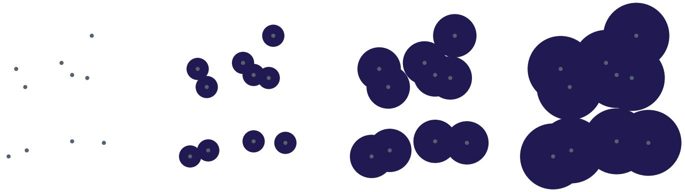
Anne Marie Svane
Joint with: Biscio, Chenavier, Hirsch, Otto

Department of Mathematical Sciences
Aalborg University
Denmark

AALBORG UNIVERSITY
DENMARK

Topological data analysis

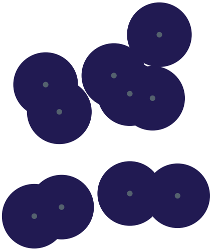
In 2D

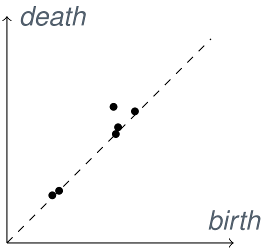


- $q = 0$ corresponds to connected components in the union of balls.
  - Interpretation: clusters.
  - A component dies when two components merge.
- $q = 1$ corresponds to connected components in the complement of the balls.
  - Interpretation: voids.
  - When a new component appears, we say that it is born.
  - When it gets covered again, we say it dies.

The persistence diagram

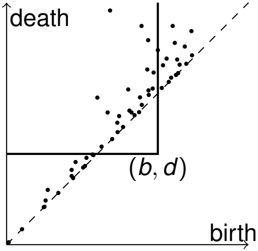
The information is summarized in the persistence diagram (q=1):





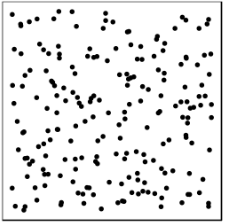
Persistent Betti numbers

- The persistent Betti numbers are given by the number of pairs born before *b* and dead after *d*.



Topological data analysis for point patterns

- TDA was invented for **deterministic** point patterns.
- Possibly measured with small amount of **noise**.
- In practice, many point patterns are realizations from a probability distribution.

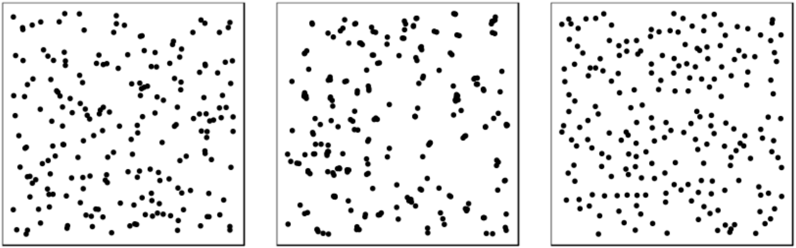


Point processes

A point process $\mathcal{P}$ is a random set of points in $\mathbb{R}^d$ which is a.s. locally finite:

- For any bounded $B \subseteq \mathbb{R}^d$, $\mathcal{P} \cap B$ is finite.

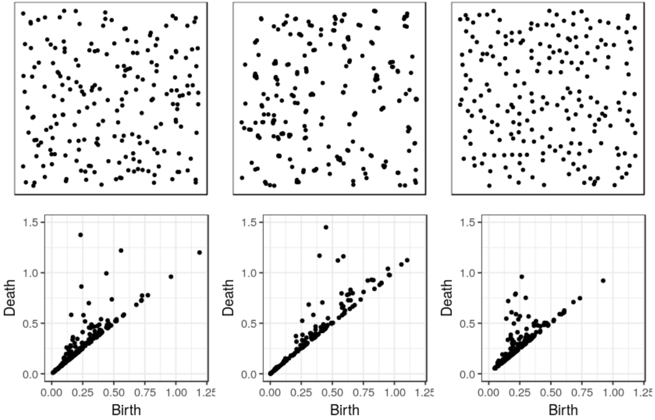
Examples:



Independence Clustering Repulsion

The persistence diagram

It seems that the persistence diagram can distinguish between different types of point patterns (q=1):



The Poisson process

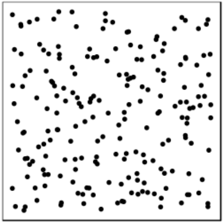
Complete spatial randomness

The (homogeneous) Poisson point process $\mathcal{P}$ is a point process on $\mathbb{R}^d$ given by:

1. For any bounded set $B \subseteq \mathbb{R}^d$:
$$\#(\mathcal{P} \cap B) \sim po(\rho|B|).$$
2. Given $\#(\mathcal{P} \cap B) = n$, the points in $\mathcal{P} \cap B$ are independent and uniformly distributed in B.

Property:

- If B_1 and B_2 are disjoint, then $\mathcal{P} \cap B_1$ and $\mathcal{P} \cap B_2$ are independent.



Poisson cluster process

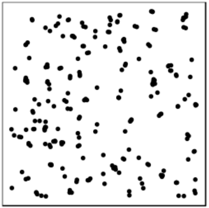
A clustered process

A Poisson cluster process is defined as follows:

- Start from a Poisson process $\mathcal{P}$.
- Replace each point in $\mathcal{P}$ by a random cluster (finite point set).

Matérn cluster process:

- Each cluster is Poisson distributed in $B_r(x)$.

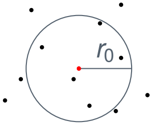


Gibbs point process

Often repulsive

For a finite point configuration $\eta \subseteq \mathbb{R}^d$, associate the energy $E(\eta)$.

- E is positive, increasing, translation + rotation invariant.
- Finite interaction range r_0 ("sufficiently small").

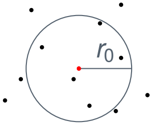


Gibbs point process

Often repulsive

For a finite point configuration $\eta \subseteq \mathbb{R}^d$, associate the energy $E(\eta)$.

- E is positive, increasing, translation + rotation invariant.
- Finite interaction range r_0 ("sufficiently small").



The finite volume Gibbs process on a bounded set $B \subseteq \mathbb{R}^d$ is absolutely continuous wrt. the Poisson process on B with intensity τ.

The density is

$$\eta \mapsto \exp(-\beta E(\eta))/Z(B).$$

Does not extend to all of $\mathbb{R}^d$!

Gibbs processes

Examples

Hardcore process:

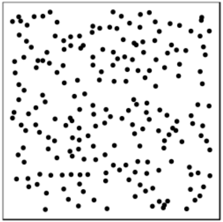
- If there exist $x, y \in \eta$ such that $|x - y| \leq r_0$ then $E(\eta) = \infty$.
- Otherwise $E(\eta) = 1$.

Strauss process:

$$E(\eta) = c_1 n_{r_0}(\eta),$$

where

- $n_{r_0}(\eta)$ is the number of pairs $x, y \in \eta$ with $0 < |x - y| \leq r_0$,
- $c_1 > 0$.



Statistics based on tda

Suppose we observe a point pattern P.

$$H_0 : P \sim \mathcal{P}_0.$$

Statistics based on tda

Suppose we observe a point pattern P.

$$H_0 : P \sim \mathcal{P}_0.$$

Possible summary statistics:

1. Compare $\beta_q^{b,d}(P)$ to $\mathbb{E}\beta_q^{b,d}(\mathcal{P}_0)$ for fixed b, d.

Statistics based on tda

Suppose we observe a point pattern P.

$$H_0 : P \sim \mathcal{P}_0.$$

Possible summary statistics:

1. Compare $\beta_q^{b,d}(P)$ to $\mathbb{E}\beta_q^{b,d}(\mathcal{P}_0)$ for fixed b, d.
2. Consider whole function $(b, d) \mapsto \beta_q^{b,d}(P)$ on $[0, T] \times [0, T]$.

Statistics based on tda

Suppose we observe a point pattern P.

$$H_0 : P \sim \mathcal{P}_0.$$

Possible summary statistics:

1. Compare $\beta_q^{b,d}(P)$ to $\mathbb{E}\beta_q^{b,d}(\mathcal{P}_0)$ for fixed b, d.
2. Consider whole function $(b, d) \mapsto \beta_q^{b,d}(P)$ on $[0, T] \times [0, T]$.
3. Consider a derived functional, e.g.
   - $\sum_{(b,d)} f(b, d)$, for f "nice".

Statistics based on tda

Suppose we observe a point pattern P.

$$H_0 : P \sim \mathcal{P}_0.$$

Possible summary statistics:

1. Compare $\beta_q^{b,d}(P)$ to $\mathbb{E}\beta_q^{b,d}(\mathcal{P}_0)$ for fixed b, d.
2. Consider whole function $(b, d) \mapsto \beta_q^{b,d}(P)$ on $[0, T] \times [0, T]$.
3. Consider a derived functional, e.g.
   - $\sum_{(b,d)} f(b, d)$, for f "nice".
   - $APF_q(r) = \sum_{(b,d)}(d - b)\mathbb{1}\{b \leq r\}$, (Biscio/Møller, 2019).

Some limit theorems

- Suppose we observe $\mathcal{P}_n = \mathcal{P} \cap W_n$ in the window

$$W_n = \Big[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2} \Big]^d.$$

- Only one realization available.
- Consider the limit when $n \to \infty$.

Some limit theorems

- Suppose we observe $\mathcal{P}_n = \mathcal{P} \cap W_n$ in the window

$$W_n = \Big[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2} \Big]^d.$$

- Only one realization available.
- Consider the limit when $n \to \infty$.

Theorem (Hiraoka, Shirai, Trinh (2018))

For all point processes in this talk, there exists a constant $\beta_q^{b,d}$ such that

- $\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\beta_q^{b,d}(\mathcal{P}_n) \to \beta_q^{b,d}$
- $\lim_{n\to\infty} \frac{1}{n}\beta_q^{b,d}(\mathcal{P}_n) \to \beta_q^{b,d}$ *a.s.*

Central limit theorems for persistent Betti numbers

Poisson case:
Yogeshwaran, Subag, Adler (2017), Hiraoka, Shirai, Trinh (2018):

$$\frac{\beta_q^{b,d}(\mathcal{P}_n) - \mathbb{E}[\beta_q^{b,d}(\mathcal{P}_n)]}{\sqrt{n}} \implies N(0, \sigma^2).$$

Gibbs case:
Hirsch, Otto, S. (in progress):

$$\frac{\beta_q^{b,d}(\mathcal{P}_n) - \mathbb{E}[\beta_q^{b,d}(\mathcal{P}_n)]}{\sqrt{n}} \implies N(0, \sigma^2).$$

Note: $\sigma^2 > 0$ whenever the energy E is finite.

Central limit theorems for persistence diagrams

M-bounded persistence pairs: only allow components with spatial diameter less than M.

Let $f : [0, T]^2 \to \mathbb{R}$ be bounded and define

$$\xi_f(P) = \sum_{(b,d)} f(b, d).$$

Theorem (Biscio, Chenavier, Hirsch, S. (2020))

Let $\mathcal{P}$ be a Poisson, Matérn cluster, or log-Gaussian point process in $\mathbb{R}^2$. Fix $M, T > 0$. Then,

$$\frac{\xi_f(\mathcal{P}_n) - \mathbb{E}[\xi_f(\mathcal{P}_n)]}{\mathrm{Var}(\xi_f(\mathcal{P}_n))} \to N(0, 1).$$

Functional CLTs

M-bounded persistent Betti-numbers: $\beta_{1,M}^{b,d}(\mathcal{P}_n)$ only counts components with spatial diameter less than *M*.

Theorem (Biscio, Chenavier, Hirsch, S. (2020), Biscio, S. (2022))

Let $\mathcal{P}$ be a Poisson, Matérn cluster, log-Gaussian or Gibbs point process in $\mathbb{R}^2$. Fix $M, T > 0$. Then,

$$\left\{ \frac{\beta_{1,M}^{b,d}(\mathcal{P}_n) - \mathbb{E}[\beta_{1,M}^{b,d}(\mathcal{P}_n)]}{\sqrt{n}} \right\}_{b,d \in [0,T]^2}$$

converges in Skorokhod topology to a centered Gaussian process on $[0, T]^2$.

Botnan, Hirsch (2023): Generalizations in the Poisson case.

Derived results

New CLTs can be derived from the functional CLT via the continuous mapping theorem.

Derived results

New CLTs can be derived from the functional CLT via the continuous mapping theorem.

Example: The accumulated persistence function

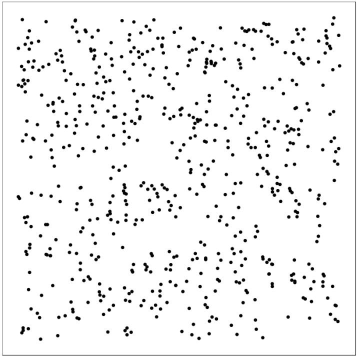
$$APF_q(r) = \sum_{(b,d)} (d - b)\mathbb{1}\{b \leq r\}.$$

Corollary

The process $\{\sqrt{n}(APF_q(r) - \mathbb{E}APF_q(r))\}_{r\in[0,R]}$ *converges to a centered Gaussian process.*

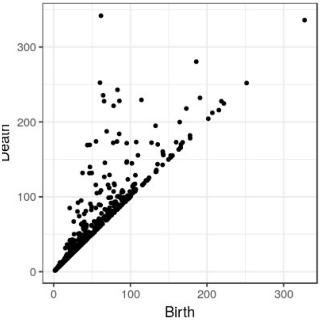
The minicolumn dataset (Nyengaard et al.)

- 634 brain neurons from human cerebral cortex
- Neurons tend to arrange in vertical columns
- 3D image projected to 2D
- Minicolumns would correspond to clusters in 2D

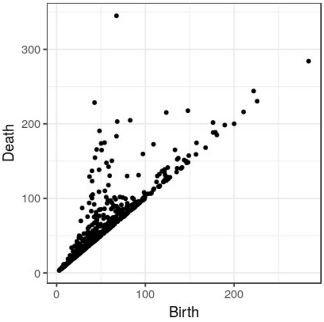


Persistence diagram for minicolumn data

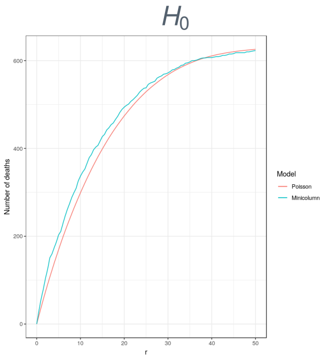
Dataset



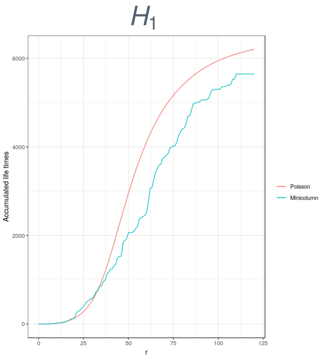
Poisson



Accumulated persistence function



$p = 0.0018$



$p = 0.00031$

About the proof

Rely on general CLTs for geometric functionals of the form:

- $H_n(\mathcal{P}_n) = \sum_{x \in \mathcal{P}_n} g(x, \mathcal{P}_n)$.

Gibbs processes:

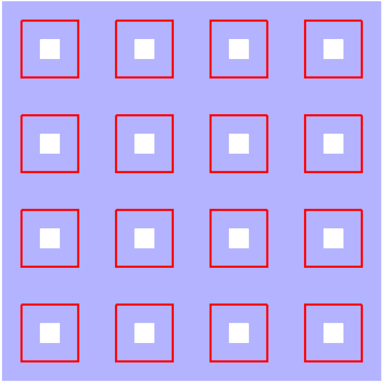
- CLT by Xia, Yukich (2015).

Other processes:

- CLT by Błaszczyszyn, Yogeshwaran, Yukich (2019)
- Requires variance is $\Omega(n)$

Variance lower bound

Idea of Xia and Yukich (2015), Biscio, Chenavier, Hirsch, S. (2020):



- Conditioned on the blue set and Λ, the red squares are independent.
- Lower bound on variance by sum of conditional variances of red squares.
- Each has conditional variance bounded from below.
- Number of red squares is $\Omega(n)$.

Summary

- TDA can be used to capture fine structure of a point pattern.
- CLTs are available, but still somewhat incomplete.
- Attempts to generalize to other random structures, e.g.
  - Random networks (Hirsch, Krebs 2022)
  - Random tesselations (Hirsch, Krebs, Redenbach 2023+)
- Disadvantages of TDA: results may be hard to interpret geometrically.

References

[1] Biscio, C. A. N., Chenavier, N., Hirsch, C., Svane, A. M. (2020) Testing goodness of fit for point processes via topological data analysis. *Electron. J. Statist.* **14**.

[2] Biscio, C. A. N., Svane, A. M. (2022): A functional central limit theorem for the empirical Ripley's K-function. *Electron. J. Statist.* **43**.

[3] Błaszczyszyn, B., Yogeshwaran, D., Yukich, J. E. (2019): Limit theory for geometric statistics of point processes having fast decay of correlations. *Ann. Probab.* **47**.

[4] Hiraoka, Y., Shirai, T., Trinh, K. D. (2018): Limit theorems for persistence diagrams. *Ann. Appl. Probab.* **5**.

[5] Schreiber, T. and Yukich, J. E. (2013): Limit theorems for geometric functionals of Gibbs point processes. *Ann. Henri Poincaré* **49**.

[6] Yogeshwaran, D., Subag, E., Adler, R. J. (2017): Random geometric complexes in the thermodynamic regime. *Probab. Theory Related Fields* **167**.

[7] Xia, A., Yukich, J. E. (2015): Normal approximation for statistics of Gibbsian input in geometric probability. *Adv. in Appl. Probab.* **47**.

Thank you for the attention!

AALBORG UNIVERSITY
DENMARK