Topological Data Analysis II

Applications in Spatial statistics

Christophe Ange Napoléon Biscio

Statistics of Persistence Diagram

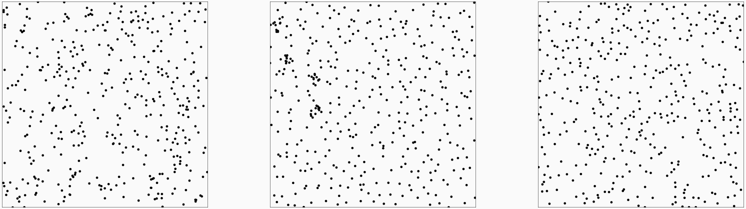
As any statisticians we should ask ourselves:

- How persistence diagrams behave under randomness/perturbation of the data?
- What is the behaviour of the mean persistence diagram? The average over iid realisations?
- Do I know the distribution of the PDs under suitable assumptions on my data/observations?
- How confident may I be in the ”random” distribution of a persistence diagram under random perturbation?

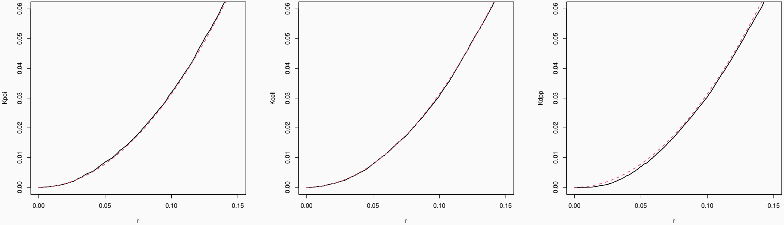
Analysing the persistence diagram

Visually

Let's look at 3 realisations of point processes.

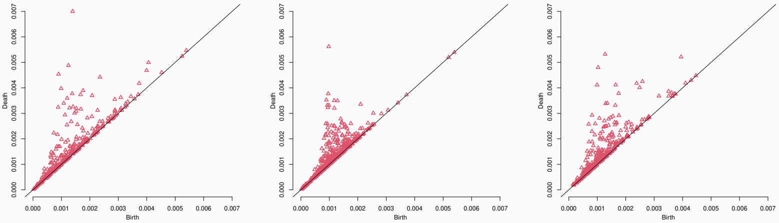


The estimate of the Ripley's K function against the one of a Poisson point process (in red):



What about the PD?

Here are the corresponding PDs of the loops for the three point pattern above.



- Visually we observe a difference.
- Are they different enough – ”far away/distant” from each others?
- Is this difference statistically significant?

Metric Aspects

- The space of persistence diagram is a metric space.
- The metric is defined by Bottleneck distance.
- Matching: Let A and B be two finite sets in $\mathbb{R}^2$. A perfect matching between A and B is a set of edges (a, b) with $a \in A$ and $b \in B$ such that each vertex is incident to exactly one edge.
- Edge Cost: of (a,b) in the matching is

$$d_\infty(a, b) = |a - b|_\infty = \max\{|a_x - b_x|, |a_y - b_y|\}.$$

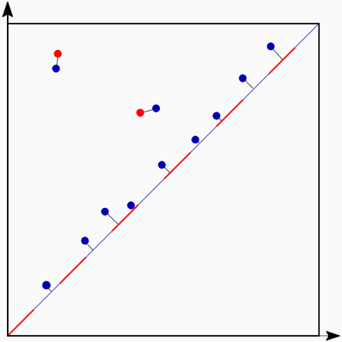
- Matching Cost is the sum of the cost of all the edges.
- Let D_1 and D_2 be two persistence diagrams. The bottleneck distance between them is defined by:

$$W_\infty(D_1, D_2) = \min_P \max_{(a,b)\in P} d_\infty(a, b).$$

for all possible matching P.

Bottleneck Distance – Optimal matching

The matching works because we add the diagonal to the PDs.



Issues with the Persistence Diagram

- As said, the persistence diagrams lay in a metric space. But we do not know much more.
- Consequently, it is not easy to define the mean of PDs. See Fasy et al. (2014).
- It is also unknown how to take the average of several PDs.

A solution:

- Define summary statistics, possibly functional, of the PD.
- Numbers and functions live in more convenient mathematical spaces, for example $L^2(\mathbb{R})$.
- Hence it becomes easier to take the average, define confidence intervals ...

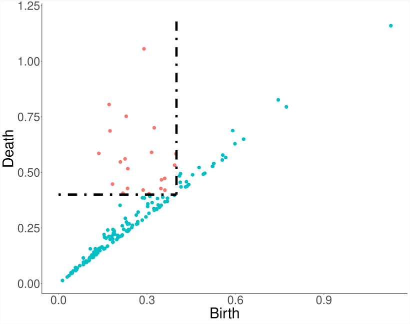
An other solution: kernel techniques, see Carriere et al. (2017).

Betti Numbers

Definition: Let $b, d > 0$ with $b < d$ and D be a PD. The persistent Betti number is

$$\beta_{b,d}^{D} = \#\{(x, y) \in D,\ x \leq b, y \geq d\}.$$

Example: The number of point in red is $\beta_{0.4,0.4}^{D}$.



Important: The knowledge of $\beta_{b,d}^{D}$ for all b, d defines completely D.

Landscapes

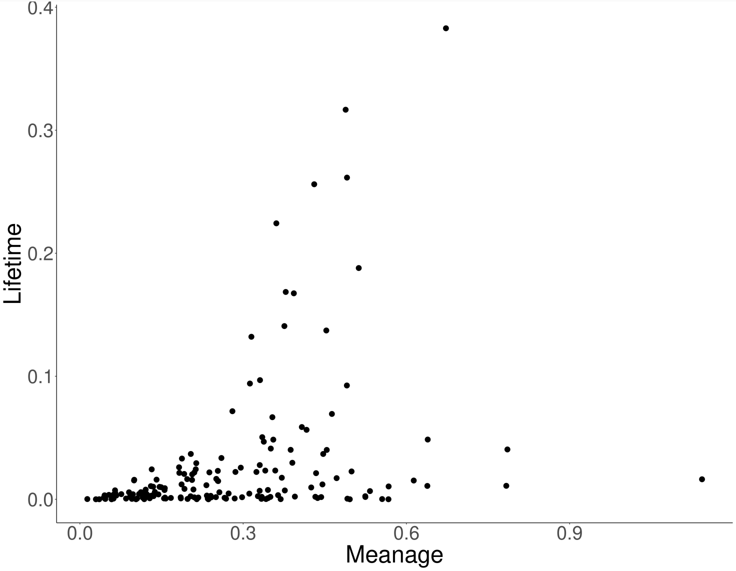
Persistence landscapes have been introduced by in Bubenik (2015). They are a collection of continuous piecewise linear function index by $p \in \mathbb{N}$.

To define them, introduce for each point $p = (x, y) = \left(\frac{b+d}{2}, \frac{d-b}{2}\right)$ representing a birth-death pair (b, d) in the persistence diagram.

Let

$$\Lambda_p(t) = \begin{cases} t - x + y, & t \in [x - y, x] \\ x + y - t, & t \in (x, x + y] \\ 0, & \text{otherwise} \end{cases} = \begin{cases} t - b, & t \in [b, \frac{b+d}{2}] \\ d - t, & t \in (\frac{b+d}{2}, d] \\ 0, & \text{otherwise.} \end{cases}$$

Rotated Persistence Diagram



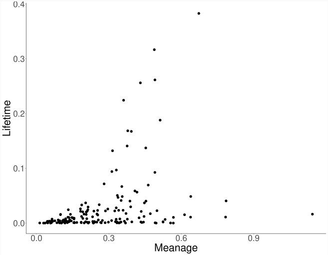
Definition Landscapes

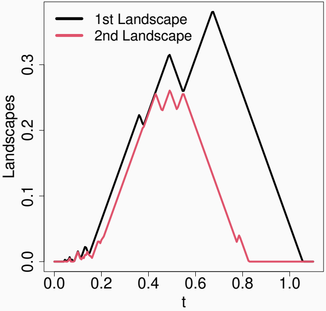
Definition: Let $k \in \mathbb{N}$ and $T > 0$. The k-th persistence landscape of a persistence diagram is

$$\lambda(k, t) = k\max_{p} \Lambda_p(t), \quad t \in [0, T].$$

- It characterizes completely the persistence diagram.
- It focus on the topological features with longer lifetimes.
- This is good for many applications like support estimation but not for spatial statistics.

Rotated Persistence Diagram and Landscapes





Accumulated Persistence Function

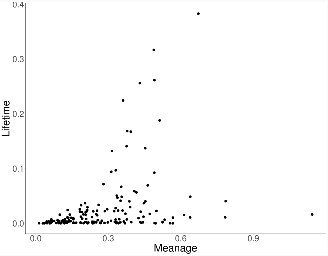
- Let's start from the rotated persistence diagram.
- For any point p in a persistence diagram D let m_p and l_p denotes the mean age and lifetime of the feature p.

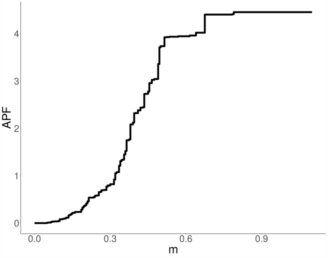
Definition: The APF of the persistence diagram D of the topological features of dimension k is defined for $m > 0$ by

$$\text{APF}_k(m) = \sum_{p \in D} l_p \mathbf{1}_{\{m_p \leq m\}}.$$

- There is one APF for the connected components, one for the loops, one for the voids ...
- It characterizes completely the persistence diagram, if no multiplicity on the points.
- It does not focus on feature with long or short lifetimes.
- Hence it is, to some extent, more suitable for spatial statistics.

Rotated persistence diagram and APF





Central Limit Theorem

CLT on Betti numbers

Preliminaries:

- As you may have seen, the asymptotics in spatial statistics is often different than in ”iid” setting.
- The asymptotic is not focused on the number of observations, often noted n, that goes to infinity.
- Instead, it considers asymptotic on ”increasing domains”, meaning the windows of observations grows.
- In practice, it means that you assume that you observe your points on a sufficiently large domain to capture the dependence structures of the observed phenomenon.

CLT on Betti numbers

Interest of a CLT:

- If you know the behaviour of the persistence diagram when the data follows a known distribution, then you can device a goodness of fit test for this distribution based on the persistence diagram.
- Thus we need a central limit theorem for persistence diagram.
- As the spaces of persistence diagram is too difficult to work in, we chose in place to work with Betti numbers.

Notation

This will be more mathematically involved. I apologize for the less mathematically involve among you.

- For any bounded function $f : \mathbb{R}^2 \to \mathbb{R}$ and persistence diagram D:

$$\langle f, D \rangle = \sum_{(b,d) \in D} f(b, d).$$

  You take the sum of the values of f evaluated at each (birth/death) point of the persistence diagram.
- $W_n = [-\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}]^2$, for $n \in \mathbb{N}$.
- $\mathbf{X}$ a stationary point process.
- $\mathbf{X}_n = \mathbf{X} \cap W_n$
- D_n the persistence diagram obtained from $\mathbf{X}_n$.

Assumptions I

- $\mathbf{X}$ is stationary.
- We implicitly considered only the topological feature with death time smaller than M. If not, we get problems related to unsolved problem in percolation theory.
- (Technical) We need to control the Palm expectation:

$$\forall p \in \mathbb{N}, \sup_{l \leq p, x \in \mathbb{R}^{2l}} \mathbb{E}_x^! [\mathbf{X}_1^p] < \infty.$$

- X exhibits exponential decay of correlations:

Assumptions II

- X exhibits exponential decay of correlations: For all $k \geq 1$, there exists $\rho^{(k)}$, $a < 1$, and $\phi : [0, \infty) \to [0, \infty)$ such that
  1. $\lim_{t\to\infty} t^n \phi(t) = 0$ for all $n \geq 1$,
  2. $\liminf_{t\to\infty} \log \phi(t)/t^b < 0$ for some $b > 0$,
  3. $$|\rho^{(p+q)}(\boldsymbol{x} \cup \boldsymbol{x}') - \rho^{(p)}(\boldsymbol{x})\rho^{(q)}(\boldsymbol{x}')| \leq (p+q)^{a(p+q)}\phi(\text{dist}(\boldsymbol{x}, \boldsymbol{x}'))$$

     for any $\boldsymbol{x} = \{x_1, \ldots, x_p\}, \boldsymbol{x}' = \{x_{p+1}, \ldots, x_{p+q}\} \subset \mathbb{R}^2$.
- For some $\nu > 0$:
  $$\liminf_{n\to\infty} \frac{\mathrm{Var}\langle f, D_n\rangle}{n^\nu} = \infty.$$

Central Limit Theorem

Under all the assumptions mentioned above,

$$\frac{\langle f, D_n \rangle - \mathbb{E}[\langle f, D_n \rangle]}{\sqrt{\mathrm{Var}(\langle f, D_n \rangle)}}$$

converges in distribution to a standard normal random variable $\mathcal{N}(0, 1)$ as $n \to \infty$.

Can we do better? Yes we can have a functional CLT.

Functional Central Limit Theorem (FCLT)

Under an additional technical (but smooth) assumption $\rightarrow$ FCLT for the Betti numbers:

- Let D, D' be the persistence diagrams of the connected components and loops, respectively.
- The one-dimensional process

$$\left\{n^{-1/2}\left(\beta_{0,d}^{D_n} - \mathbb{E}[\beta_{0,d}^{D_n}]\right)\right\}_{d \leq r_f}$$

  converges weakly in Skorokhod topology to a centered Gaussian process.
- The two-dimensional process

$$\left\{n^{-1/2}\left(\beta_{b,d}^{D'_n} - \mathbb{E}[\beta_{b,d}^{D'_n}]\right)\right\}_{b,d \leq r_f}$$

  converges weakly in Skorokhod topology to a centered Gaussian process.

Application to (goodness of fit) deviation test.

- Thanks to the FCLT, we know the behaviour of ”any” functions of the betti numbers.
- This can be use to device deviation tests.

Setting:

- I observe the point pattern x.
- I want to test if x is a realisation of a point process $\mathbf{X}_0$.
- Let’s choose a summary statistics based on the persistence diagram (more details later): T.
- Thanks to the FCLT I know, up to some simulations, the behaviour of T and how likely it can deviate from its usual behaviour.
- If it deviates too much, I reject the assumption that x is a realisation of $\mathbf{X}_0$.

Statistics based on the persistence diagram

We may use almost any statistics based on the persistence diagram.

Using the connected components:

- For a given $r > 0$: $\int_0^r \sum_{i \leq d} \mathbf{1}_{(0,i) \in D} \mathrm{d}d$. I.e. the function of the number of deaths of connected components before time d for d varying from 0 to r.
- Intensity scaled version: $\frac{1}{\sqrt{\rho}|W|} \int_0^{r/\sqrt{\rho}} \sum_{i \leq d} \mathbf{1}_{(0,i) \in D} \mathrm{d}d$.
- $\sup_{m \in [0,R]} \mathrm{APF}_0(m)$ or $\int_{m \in [0,R]} \mathrm{APF}_0(m) \mathrm{d}m$.

Using the loops:

- $\int_{m \in [0,R]} \mathrm{APF}_1(m) \mathrm{d}m$.
- Intensity scaled version: $\frac{1}{|W|\sqrt{\rho}} \mathrm{APF}_1(\frac{r}{\sqrt{\rho}})$ where ρ is the (unknown) intensity of the process

Limitations

- Although we know a FCLT, we do not always know the mean and variance of the asymptotic Gaussian process.
- Hence it needs to be estimated by simulations which may be computationally difficult.
- Our results hold when the intensity of the point process is supposed to be known. It is never the case in practice.
- The best we could have done is to propose an ”intensity scaled” version which in simulation study provides better results.

Coding - R package TDA

References

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. Journal of Machine Learning Research, 16:77–102.

Carriere, M., Cuturi, M., and Oudot, S. (2017). Sliced wasserstein kernel for persistence diagrams. In International conference on machine learning, pages 664–673. PMLR.

Edelsbrunner, H. and Harer, J. L. (2010). Computational Topology. American Mathematical Society, Providence, RI.

Fasy, B., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014). Confidence sets for persistence diagrams. The Annals of Statistics, 42:2301–2339.

Additional References

Books on the theory:

- Munkres, J.R. (1984). Elements of Algebraic Topology (1st ed.). CRC Press.
- Hatcher, A. (2002). Algebraic Topology. Cambridge University Press. Freely available on the website of the author.

Online ressources:

- The master course *Foundations of Geometric Methods in Data Sciences* of Mathieu Carriere and Frederic Cazals: website
- The master course INF556 of Steve Oudot: website

Finally, the documentation of the various python libraries: gudhi, giotto-tda, dionysus may also provides you with many showcases of applications of TDA.

Thank you for your attention