Inference in Topological Data Analysis

Wolfgang Polonik
Department of Statistics, UC Davis

joint with

Johannes Krebs and Benjamin Roycraft

LIKE23, June 25-29, 2023

Synopsis

Synopsis

(i) brief intro to TDA (persistence homology);

(ii) large sample distribution of persistent Betti numbers (and Euler characteristic process);

(iii) statistical inference via bootstrap for persistent Betti numbers (and Euler characteristics);

(iv) discuss statistical insights

TDA - persistent homology

TDA - persistent homology

- set of feature extraction methods;
- features are topological/geometric in nature;

Introductory texts with different foci:

Edelsbrunner and Harer (2010), Otter et al. (2017), Wasserman (2018), Boissonnat et al. (2018), Rabadan and Blumberg (2019), Chazal and Michel (2021), Virk (2022), and Dey and Wang (2022)

TDA - persistent homology

- set of feature extraction methods;
- features are topological/geometric in nature;

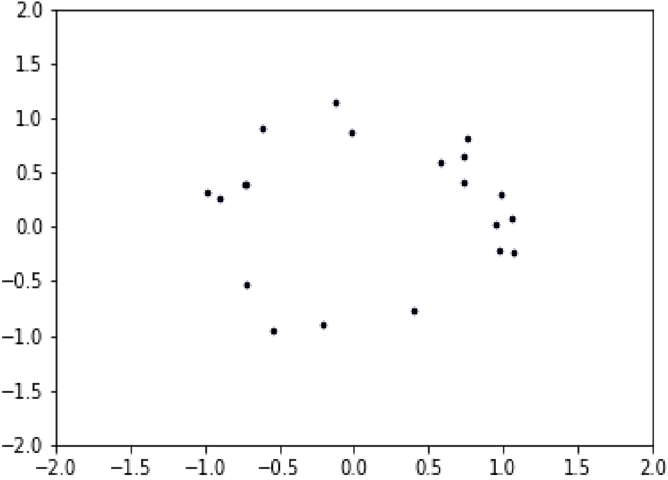
Introductory texts with different foci:

Edelsbrunner and Harer (2010), Otter et al. (2017), Wasserman (2018), Boissonnat et al. (2018), Rabadan and Blumberg (2019), Chazal and Michel (2021), Virk (2022), and Dey and Wang (2022)

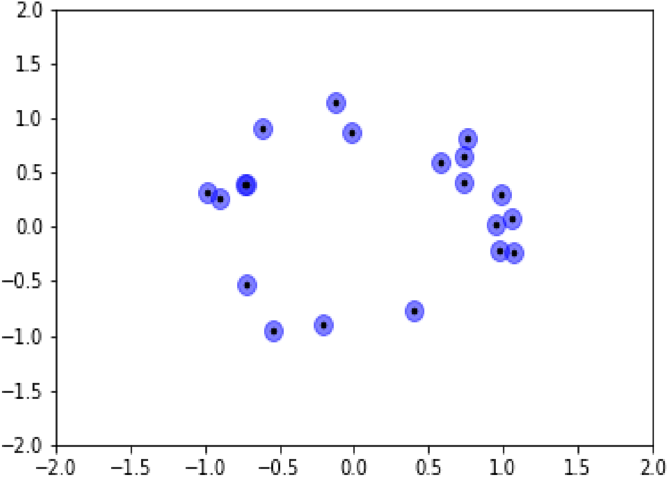
Recent surveys of applications in specific fields:

- Bukkuri et al. (2021) in oncology,
- Dłotko et al. (2019) in financial time series,
- Rabadan and Blumberg (2019) in genomics and evolution,
- Davies (2022) in cyber security,
- Amézquita et al. (2020) in biology,
- Smith et al. (2021) in chemical engineering,
- Joshi and Joshi (2019) in big data in health care, and
- Salch et al. (2021) discuss TDA methods in biomedical imaging.

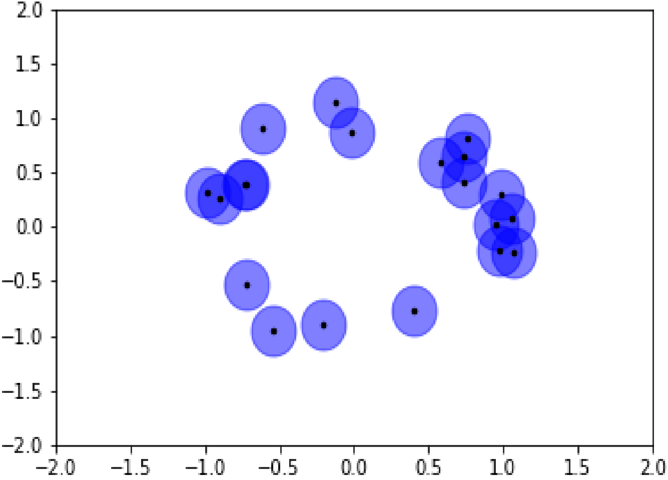
The persistence diagram based on the Čech filtration



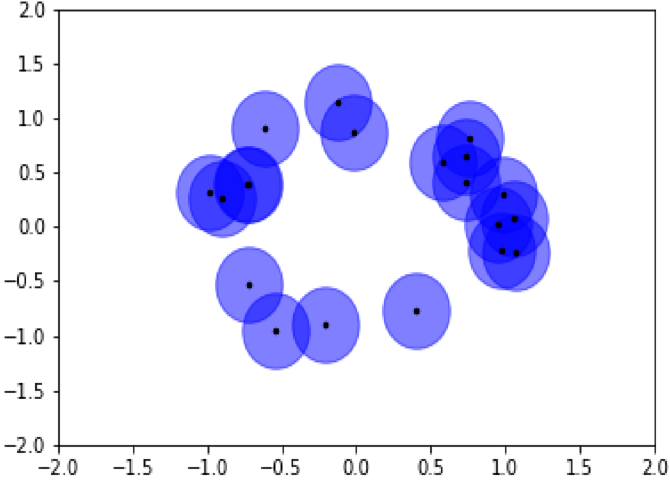
The persistence diagram based on the Čech filtration



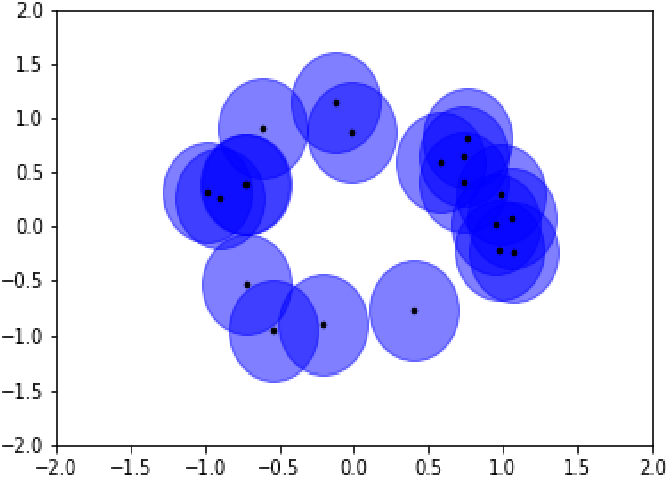
The persistence diagram based on the Čech filtration



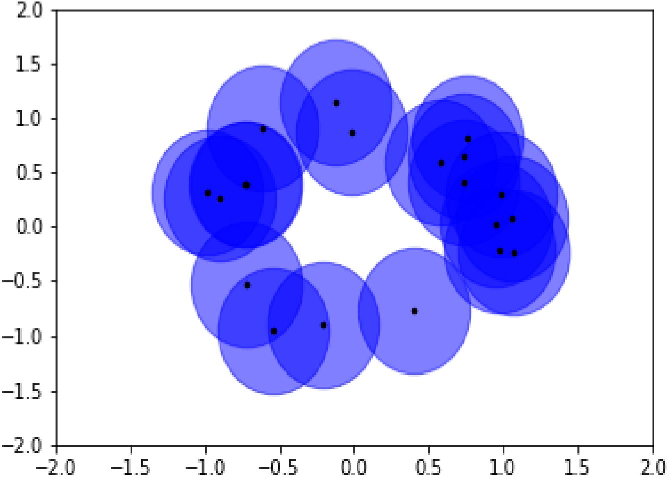
The persistence diagram based on the Čech filtration



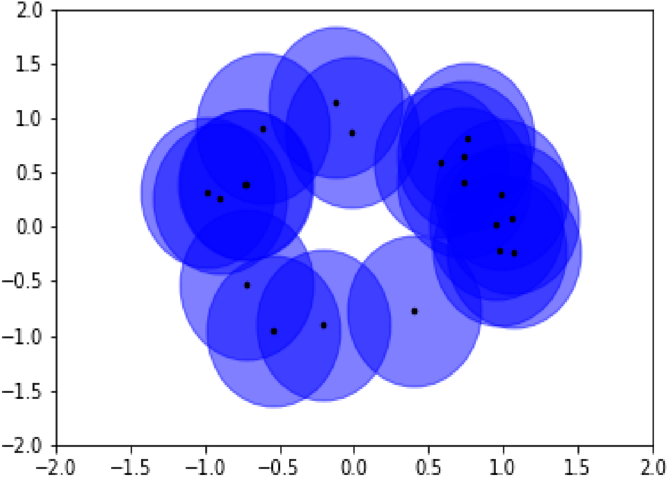
The persistence diagram based on the Čech filtration



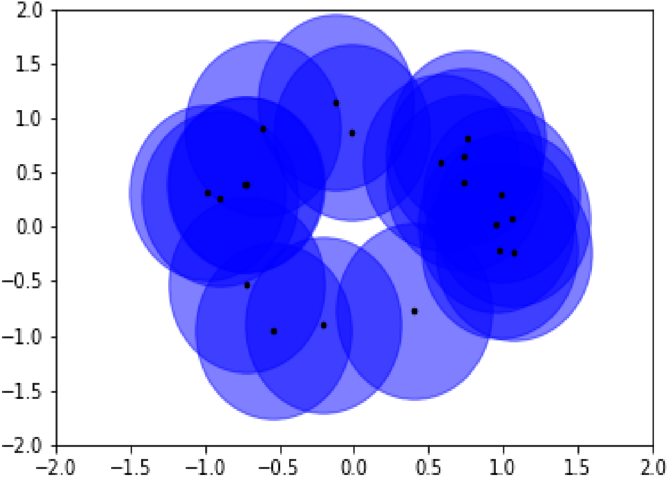
The persistence diagram based on the Čech filtration



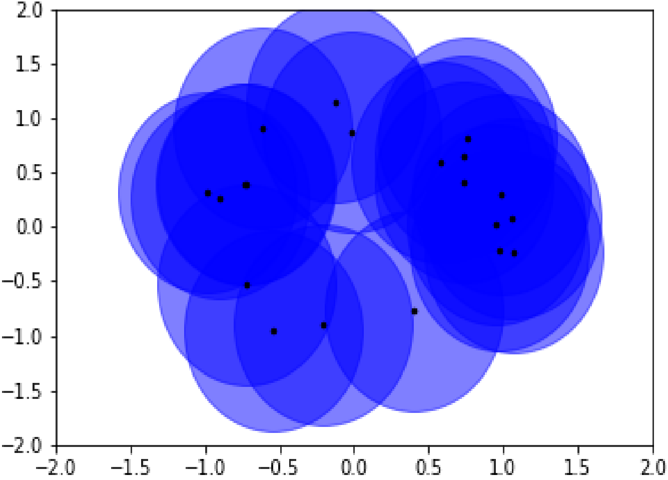
The persistence diagram based on the Čech filtration



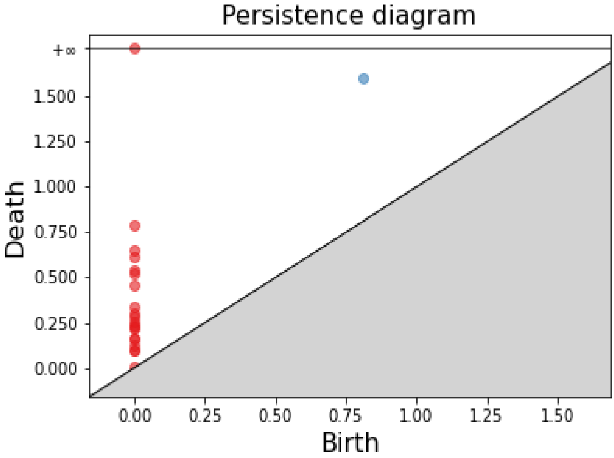
The persistence diagram based on the Čech filtration



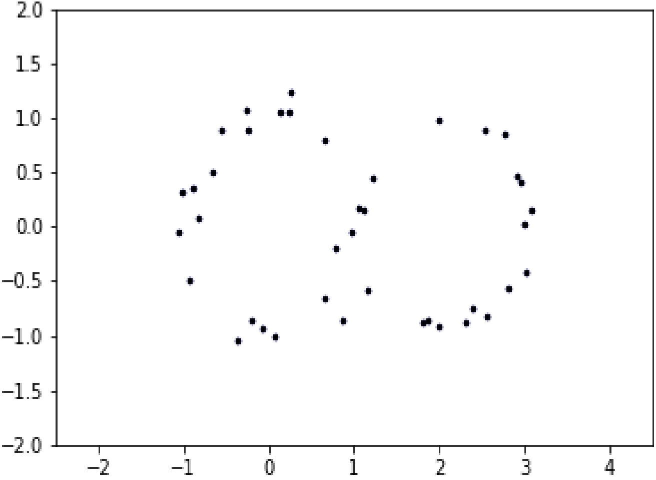
The persistence diagram based on the Čech filtration



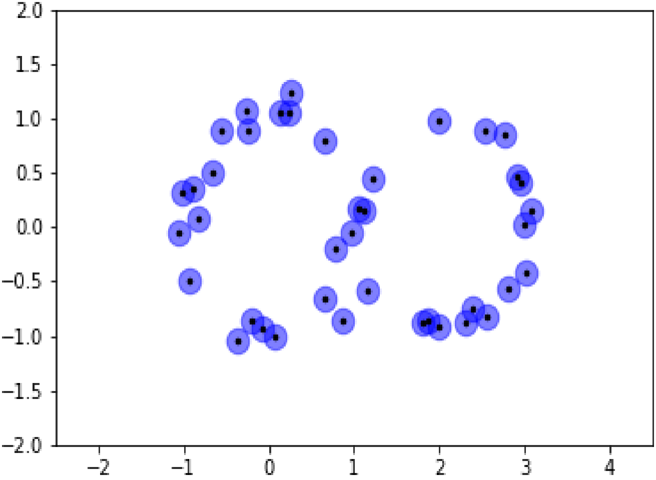
The persistence diagram based on the Čech filtration



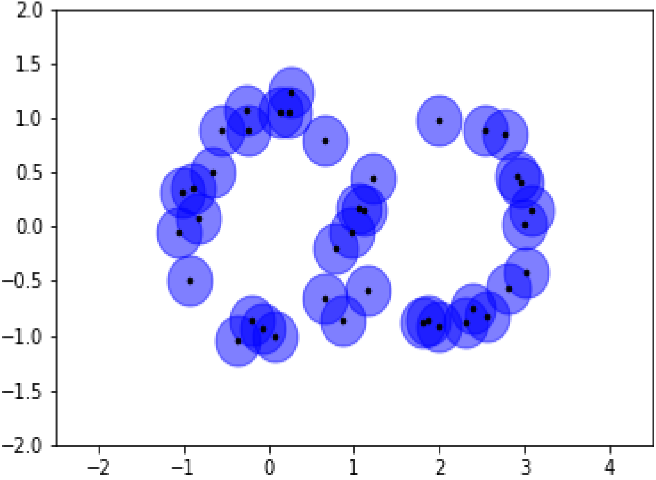
The persistence diagram based on the Čech filtration



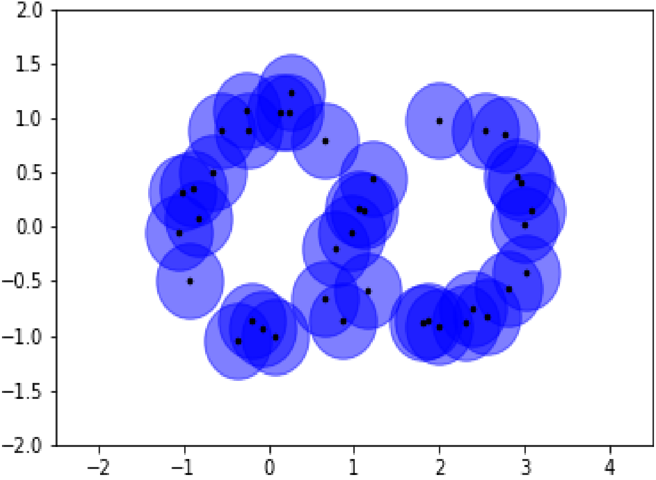
The persistence diagram based on the Čech filtration



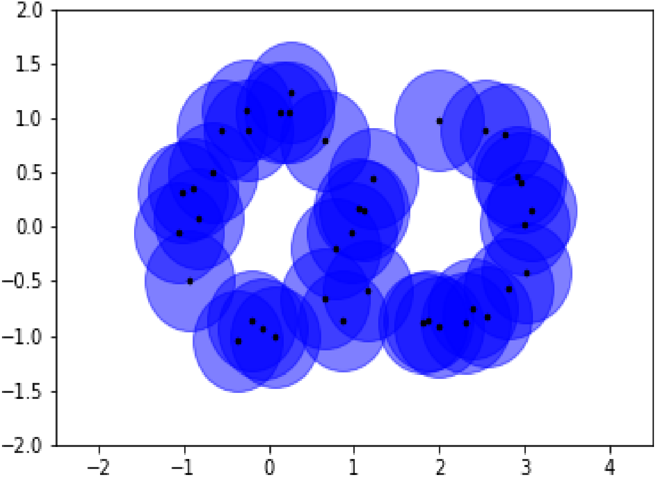
The persistence diagram based on the Čech filtration



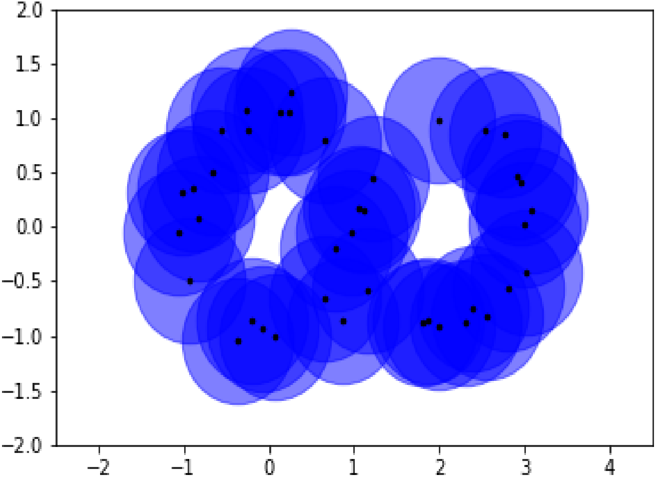
The persistence diagram based on the Čech filtration



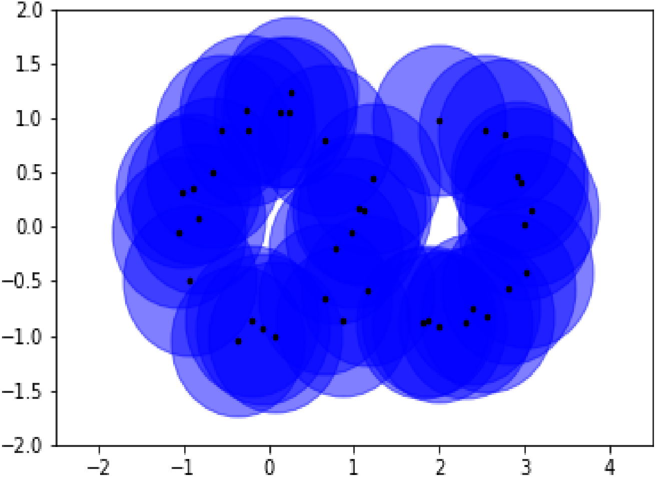
The persistence diagram based on the Čech filtration



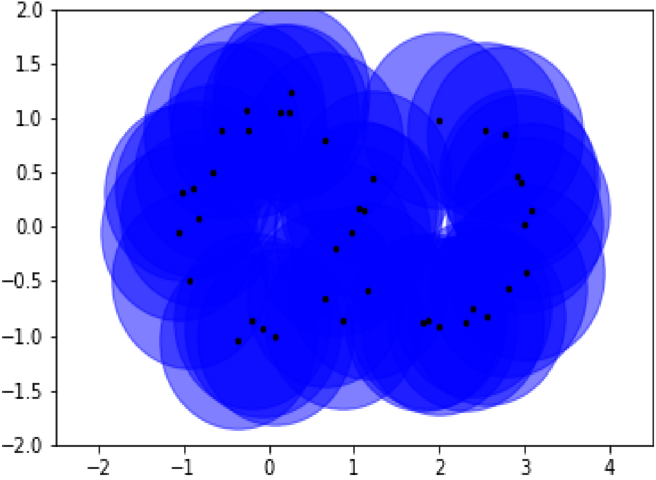
The persistence diagram based on the Čech filtration



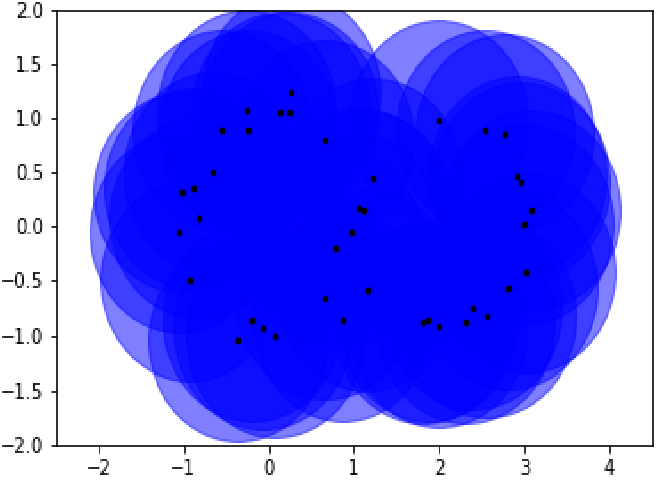
The persistence diagram based on the Čech filtration



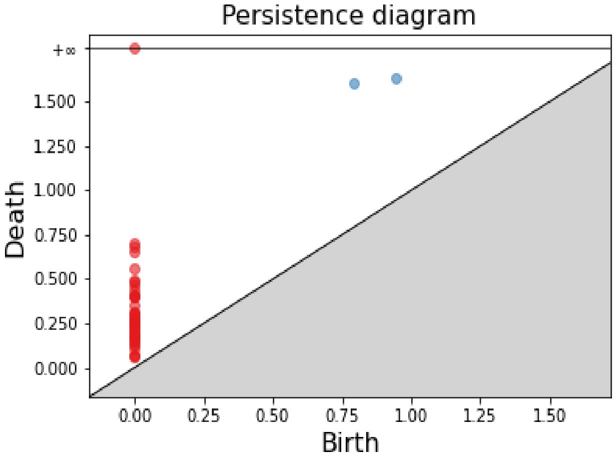
The persistence diagram based on the Čech filtration



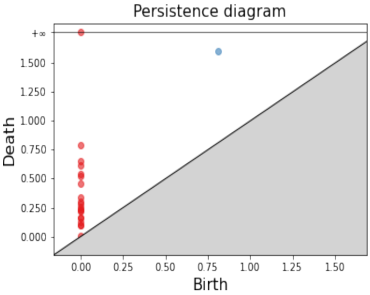
The persistence diagram based on the Čech filtration

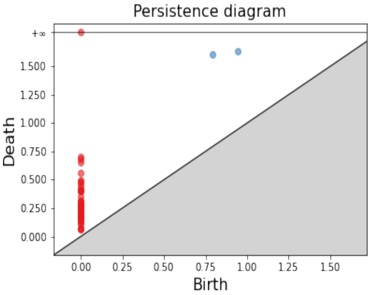


The persistence diagram based on the Čech filtration

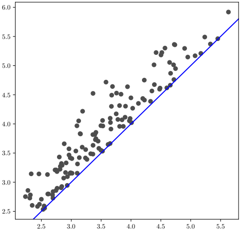


The persistence diagram based on the Čech filtration

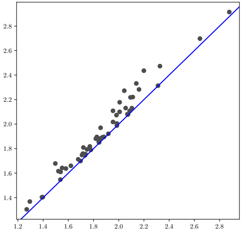




Persistence diagram for multivariate normal



1-dim. features



2-dim. features

- PD based on sample of size 200 from a 6-dimensional standard normal.
- Each point corresponds to a 1-dim. hole or a 2-dim. hole, respectively.
- There might be higher-dimensional features (holes)

Filtrations

- Let $\mathcal{B}_r = \bigcup_{i=1}^{n} B_d(x_i, r) \subset \mathbb{R}^d$; then

$$\mathcal{B} = \left\{\mathcal{B}_r : \ r \geq 0\right\}$$

defines a filtration of a topological space (in this case of $\mathbb{R}^d$): For any $r_1 \leq r_2 \cdots \leq r_N$,

$$\mathcal{B}_{r_1} \subseteq \mathcal{B}_{r_2} \subseteq \cdots \subseteq B_{r_N} \subseteq \cdots \subseteq \mathbb{R}^d$$

Filter functions

Filter functions $f : \mathbb{X} \to \mathbb{R}$ are often used to generate a filtration of a topological space $\mathbb{X}$ via sublevel sets (or superlevel sets).

Filter functions

Filter functions $f : \mathbb{X} \to \mathbb{R}$ are often used to generate a filtration of a topological space $\mathbb{X}$ via sublevel sets (or superlevel sets).

Filter function for union of balls (called Čech filtration), is

$$f(x) = D_{\mathbb{X}}(x) = \min_{i=1,\ldots,n} d(x, X_i) \qquad \text{distance function},$$

for

$$\{x \in \mathbb{X} : \; D_{\mathbb{X}}(x) \leq r\} = \bigcup_{i=1}^{n} B(X_i, r).$$

Filter functions

Filter functions $f : \mathbb{X} \to \mathbb{R}$ are often used to generate a filtration of a topological space $\mathbb{X}$ via sublevel sets (or superlevel sets).

Filter function for union of balls (called Čech filtration), is

$$f(x) = D_{\mathbb{X}}(x) = \min_{i=1,\dots,n} d(x, X_i) \qquad \text{distance function},$$

for

$$\left\{x \in \mathbb{X} : \ D_{\mathbb{X}}(x) \le r\right\} = \bigcup_{i=1}^{n} B(X_i, r).$$

Any function $f : \mathbb{X} \to \mathbb{R}$ defines a filtration $\mathcal{F}_{\mathbb{X}} = \left\{\mathbb{X}_t, t \in \mathbb{R}\right\}$ of $\mathbb{X}$ via

$$\mathbb{X}_t = \left\{x \in \mathbb{X} : f(x) \le t\right\}.$$

Filter functions

Filter functions $f : \mathbb{X} \to \mathbb{R}$ are often used to generate a filtration of a topological space $\mathbb{X}$ via sublevel sets (or superlevel sets).

Filter function for union of balls (called Čech filtration), is

$$f(x) = D_{\mathbb{X}}(x) = \min_{i=1,\dots,n} d(x, X_i) \qquad \text{distance function},$$

for

$$\{x \in \mathbb{X} : \; D_{\mathbb{X}}(x) \leq r\} = \bigcup_{i=1}^{n} B(X_i, r).$$

Any function $f : \mathbb{X} \to \mathbb{R}$ defines a filtration $\mathcal{F}_{\mathbb{X}} = \{\mathbb{X}_t, t \in \mathbb{R}\}$ of $\mathbb{X}$ via

$$\mathbb{X}_t = \{x \in \mathbb{X} : f(x) \leq t\}.$$

$$\mathbb{X}_{t_1} \subseteq \mathbb{X}_{t_2} \subseteq \cdots \subseteq \mathbb{X}_{t_N} \subseteq \cdots \subseteq \mathbb{X}, \qquad t_1 \leq t_2 \leq \ldots \leq t_N$$

Filter functions

Filter functions $f : \mathbb{X} \to \mathbb{R}$ are often used to generate a filtration of a topological space $\mathbb{X}$ via sublevel sets (or superlevel sets).

Filter function for union of balls (called Čech filtration), is

$$f(x) = D_{\mathbb{X}}(x) = \min_{i=1,\ldots,n} d(x, X_i) \qquad \text{distance function},$$

for

$$\{x \in \mathbb{X} : \; D_{\mathbb{X}}(x) \leq r\} = \bigcup_{i=1}^{n} B(X_i, r).$$

Any function $f : \mathbb{X} \to \mathbb{R}$ defines a filtration $\mathcal{F}_{\mathbb{X}} = \{\mathbb{X}_t, t \in \mathbb{R}\}$ of $\mathbb{X}$ via

$$\mathbb{X}_t = \{x \in \mathbb{X} : f(x) \leq t\}.$$

$$\mathbb{X}_{t_1} \subseteq \mathbb{X}_{t_2} \subseteq \cdots \subseteq \mathbb{X}_{t_N} \subseteq \cdots \subseteq \mathbb{X}, \qquad t_1 \leq t_2 \leq \ldots \leq t_N$$

Question is: Which filter function f is useful? (Depends on problem at hand.)

Tracking the ‘dynamics’ of the topological features (holes, homology) given by the filtration

Finding the persistence diagram

QUESTIONS:

Finding the persistence diagram

QUESTIONS:

- What exactly is understood by a 'hole" or a "topological feature"?

Finding the persistence diagram

QUESTIONS:

- What exactly is understood by a 'hole" or a "topological feature"?
- How to computationally find birth and death times of all the 'holes' (in any dimension) for high-dimensional data?

Finding the persistence diagram

QUESTIONS:

- What exactly is understood by a 'hole" or a "topological feature"?
- How to computationally find birth and death times of all the 'holes' (in any dimension) for high-dimensional data?

'ANSWERS':

- Bring in simplicial complexes (combinatorial objects; enables computation):

Finding the persistence diagram

QUESTIONS:

- What exactly is understood by a 'hole" or a "topological feature"?
- How to computationally find birth and death times of all the 'holes' (in any dimension) for high-dimensional data?

'ANSWERS':

- Bring in simplicial complexes (combinatorial objects; enables computation):

  - approximate each top. space in filtration by an abstract simplicial complex;
  - this results in a filtration of a simplicial complex

Finding the persistence diagram

QUESTIONS:

- What exactly is understood by a 'hole" or a "topological feature"?
- How to computationally find birth and death times of all the 'holes' (in any dimension) for high-dimensional data?

'ANSWERS':

- Bring in simplicial complexes (combinatorial objects; enables computation):
  - approximate each top. space in filtration by an abstract simplicial complex;
  - this results in a filtration of a simplicial complex
- What actually is computed are ranks of homology groups for each simplicial complex in the filtration.

Finding the persistence diagram

QUESTIONS:

- What exactly is understood by a 'hole" or a "topological feature"?
- How to computationally find birth and death times of all the 'holes' (in any dimension) for high-dimensional data?

'ANSWERS':

- Bring in simplicial complexes (combinatorial objects; enables computation):
  - approximate each top. space in filtration by an abstract simplicial complex;
  - this results in a filtration of a simplicial complex
- What actually is computed are ranks of homology groups for each simplicial complex in the filtration.
- Any change in (one of) the ranks corresponds to a birth or a death of a feature; then find pairings;

Finding the persistence diagram

QUESTIONS:

- What exactly is understood by a 'hole" or a "topological feature"?
- How to computationally find birth and death times of all the 'holes' (in any dimension) for high-dimensional data?

'ANSWERS':

- Bring in simplicial complexes (combinatorial objects; enables computation):
  - approximate each top. space in filtration by an abstract simplicial complex;
  - this results in a filtration of a simplicial complex
- What actually is computed are ranks of homology groups for each simplicial complex in the filtration.
- Any change in (one of) the ranks corresponds to a birth or a death of a feature; then find pairings;
- Persistence diagram 'summarizes' this dynamics

Finding the persistence diagram

QUESTIONS:

- What exactly is understood by a 'hole" or a "topological feature"?
- How to computationally find birth and death times of all the 'holes' (in any dimension) for high-dimensional data?

'ANSWERS':

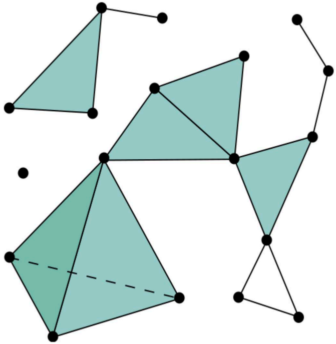
- Bring in simplicial complexes (combinatorial objects; enables computation):
  - approximate each top. space in filtration by an abstract simplicial complex;
  - this results in a filtration of a simplicial complex
- What actually is computed are ranks of homology groups for each simplicial complex in the filtration.
- Any change in (one of) the ranks corresponds to a birth or a death of a feature; then find pairings;
- Persistence diagram 'summarizes' this dynamics
- Each point in the persistence diagram corresponds to an equivalence classes of cycles, each of them 'encircling' the same 'hole'; (k-dimensional holes are encircled by cycles of k-dimensional simplices)

Simplicial complexes

Simplicial complexes are being used to facilitate computations.

Simplicial complexes

Simplicial complexes are being used to facilitate computations.



Definition (abstract simplical complex)

Given a finite set V, an abstract simplicial complex $\mathcal{C}$ with vertex set V is a collection of subsets of V such that

(i) each element of V lies in $\mathcal{C}$;

(ii) $D \in \mathcal{C}$ and $C \subset D \quad \Rightarrow \quad C \in \mathcal{C}$.

Each $D \in \mathcal{C}$ is called a *simplex*, and its dimension is $|D| - 1$. The *dimension of* $\mathcal{C}$ is the maximum dimension of the simplices in $\mathcal{C}$.

The Čech and the Vietoris-Rips filtration

Definition (Čech-complex)

Let $(\mathbb{X}, d)$ is a metric space, and let $\mathbb{X}_n = \{x_1, \ldots, x_n\} \subset \mathbb{X}$. For $r \geq 0$, the Čech-complex $\mathcal{C}_r(\mathbb{X}_n)$ at scale r over $\mathbb{X}_n$ is the abstract simplicial complex given by:

$$[x_0, \ldots, x_k] \in \mathcal{C}_r(\mathbb{X}_n) \quad \Leftrightarrow \quad \bigcap_{i=0}^{k} \overline{B}_r(x_i) \neq \emptyset.$$

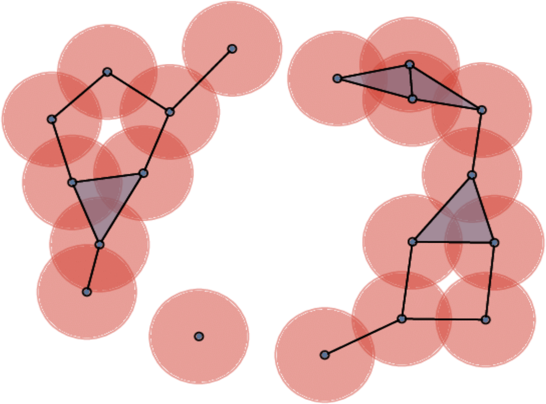
Definition (VR-complex)

Let $(\mathbb{X}, d)$ be a metric space, and let $\mathbb{X}_n = \{x_1, \ldots, x_n\} \subset \mathbb{X}$. For $r \geq 0$, the VR-complex $\mathrm{VR}_r(\mathbb{X}_n)$ at scale r over $\mathbb{X}_n$ is the abstract simplicial complex given by:

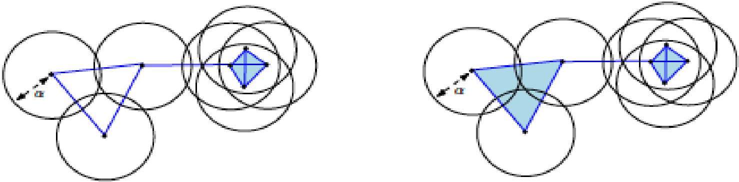
$$[x_0, \ldots, x_k] \in \mathrm{VR}_r(\mathbb{X}_n) \quad \Leftrightarrow \quad d(x_i, x_j) \leq r \quad \forall\, 0 \leq i \leq j \leq k.$$

The Čech complex

Note: The Čech complex $\mathcal{C}_r(\mathbb{X})$ is homotopy equivalent to the union of balls $\bigcup_{i=1}^{n} \overline{B}_r(x_i)$. (Nerve Theorem)

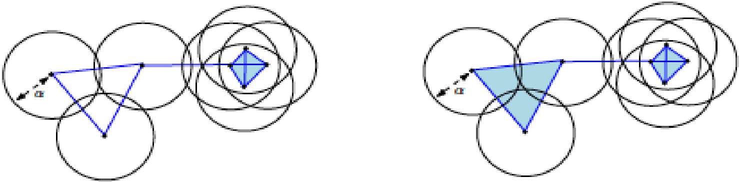


The Čech and the Vietoris-Rips filtration



Čech complex at scale r (left) and VR-complex at scale $2r$ (right).

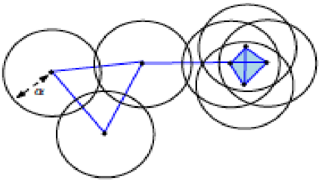
The Čech and the Vietoris-Rips filtration

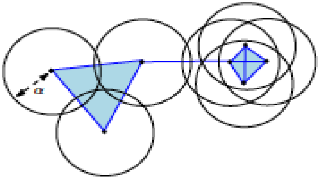


Čech complex at scale r (left) and VR-complex at scale $2r$ (right).

Important observation: VR-complex only relies on pairwise distances!

The Čech and the Vietoris-Rips filtration





Čech complex at scale r (left) and VR-complex at scale $2r$ (right).

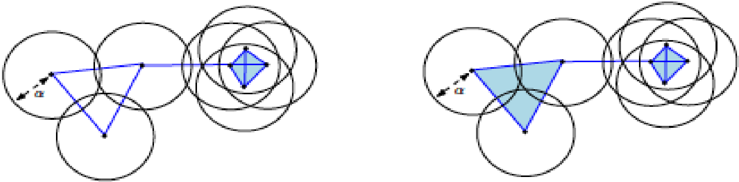
Important observation: VR-complex only relies on pairwise distances!

Proposition: *Let $\mathbb{X}_n$ be a finite set of points in $\mathbb{R}^d$. For any $\alpha \geq 0$,*

$$\mathrm{VR}_r(\mathbb{X}_n) \subset \mathcal{C}_r(\mathbb{X}_n) \subset \mathrm{VR}_{2r}(\mathbb{X}_n).$$

Filtrations: The collections $\{\mathcal{C}_r(\mathbb{X}_n) : r \geq 0\}$ and $\{\mathrm{VR}_r(\mathbb{X}_n) : r \geq 0\}$ are called Čech and VR-filtration, respectively.

The Čech and the Vietoris-Rips filtration



Čech complex at scale r (left) and VR-complex at scale $2r$ (right).

Important observation: VR-complex only relies on pairwise distances!

Proposition: *Let $\mathbb{X}_n$ be a finite set of points in $\mathbb{R}^d$. For any $\alpha \geq 0$,*

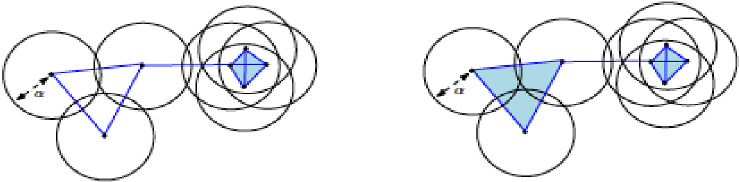
$$\mathrm{VR}_r(\mathbb{X}_n) \subset \mathcal{C}_r(\mathbb{X}_n) \subset \mathrm{VR}_{2r}(\mathbb{X}_n).$$

Filtrations: The collections $\{\mathcal{C}_r(\mathbb{X}_n) : r \geq 0\}$ and $\{\mathrm{VR}_r(\mathbb{X}_n) : r \geq 0\}$ are called Čech and VR-filtration, respectively.

General filtration of a simplicial complex C: Increasing sequence

$$C_1 \subseteq C_2 \subseteq \cdots \subseteq C_N = C.$$

The Čech and the Vietoris-Rips filtration



Čech complex at scale r (left) and VR-complex at scale $2r$ (right).

Important observation: VR-complex only relies on pairwise distances!

Proposition: *Let $\mathbb{X}_n$ be a finite set of points in $\mathbb{R}^d$. For any $\alpha \geq 0$,*

$$\mathrm{VR}_r(\mathbb{X}_n) \subset \mathcal{C}_r(\mathbb{X}_n) \subset \mathrm{VR}_{2r}(\mathbb{X}_n).$$

Filtrations: The collections $\{\mathcal{C}_r(\mathbb{X}_n) : r \geq 0\}$ and $\{\mathrm{VR}_r(\mathbb{X}_n) : r \geq 0\}$ are called Čech and VR-filtration, respectively.

General filtration of a simplicial complex C: Increasing sequence

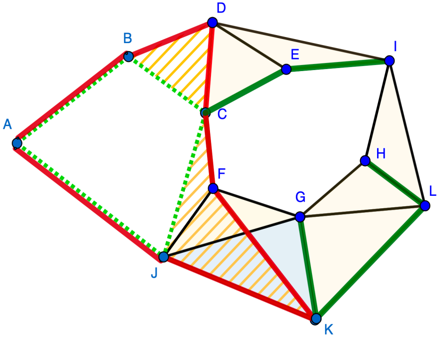
$$C_1 \subseteq C_2 \subseteq \cdots \subseteq C_N = C.$$

Add one simplex at a time $\rightsquigarrow$ algorithm.

Cycles and boundaries

Cycles and boundaries

- $Z_k(K) \subset C_k$ (kernel of $\partial_k : C_k \to C_{k-1}$); cycles.
- $B_k(K) \subset C_k$ (image of $\partial_{k+1} : C_{k+1} \to C_k$); boundaries of $(k+1)$-chains

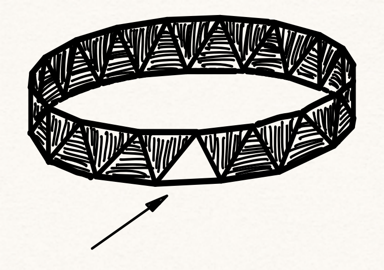


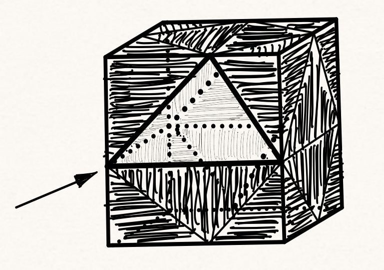
$H_k = Z_k / B_k$
(k-th homology group);

rank(H_k) = k-th Betti number

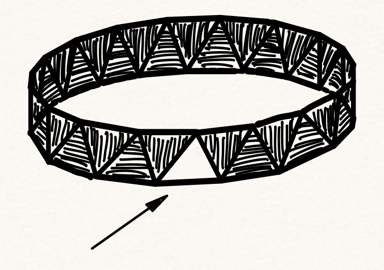
Both the solid red and the dashed green 1-chains are cycles. Their sum (or difference) is a boundary of a 2-chain (hatched). The 1-chain in bold green is not a cycle.

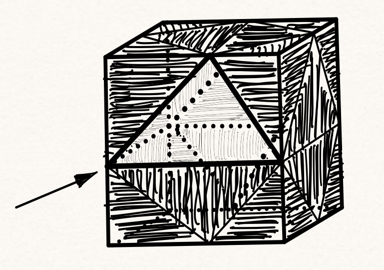
Example





Example





- How many 'holes' are here?
- What happens when add simplices closing the 'holes'?

Inference for persistent homology I

Biscio, C.A.N. and Møller, J. (2019): The accumulated persistence function, a new useful functional summary statistic for topological data analysis, with a view to brain artery trees and spatial point process applications, *J. Comput. Graphical Statist.*, **28**, 671 - 681.

Bubenik, P. (2015): Statistical topological data analysis using persistence landscapes, *J. Mach. Learn. Res*, **16**, 77 - 102.

Cericola, C., Johnson, I., Kiers, J., Krock, M., Purdy, J. and Torrence, J. (2018): Extending hypothesis testing with persistence homology to three or more groups, *Involve, a Journal of Mathematics*, **11**, 27 - 51.

Chazal, F., Fasy, B.T., Lecci, F., Rinaldo, A., Singh, A., and Wasserman, L. (2013): On the bootstrap for persistence diagrams and landscapes. *Modeling and Analysis of Information Systems.*, **20**, 96 - 105.

Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A. and Wasserman, L. (2015): Subsampling methods for persistent homology, *ICML*, 2143 - 2151.

Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A. and Wasserman, L. (2018): Robust topological inference: Distance to a measure and kernel distance. *J. Machine Learn. Res.*, **18**, 1 - 40.

DeWoskin, D, Climent, J., Cruz-White, I., Vazquez, M., Park, C., and Arsuaga, J. (2010): Applications of computational homology to the analysis of treatment response in breast cancer patients. *Topology and its Applications* **157**, 157-164.

Fasy, B. Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014): Confidence sets for persistence diagrams. *Ann. Statist.* **42**, 2301 - 2339.

Inference for persistent homology II

Garside, K., Gjoka,R., Henderson, H., Johson, H. and Makarenko, I. (2021): Event history and topological data analysis. *Biometrika*, **108**, 757 - 773.

Krebs, J. and WP (2019): On the asymptotic normality of persistent Betti numbers. `https://arxiv.org/abs/1903.03280`

Robinson, A. and Turner, K. (2017): Hypothesis testing for topological data analysis, *Journal of Applied and Computational Topology*, **1**, 241 - 261.

Krebs, J., Roycraft, B. and WP (2021): On approximation theorems for the Euler characteristic with applications to the bootstrap. *Electronic J. Stat.*, **15**, 4462-4509.

Roycraft, B., Krebs, J., and WP (2022): Bootstrapping Persistent Betti Numbers and Other Stabilizing Statistics. To appear in *Ann. Statist.* `https://arxiv.org/abs/2005.01417`

Vejdemo-Johansson, M. and Mukherjee, S. (2022): Multiple hypothesis testing with persistent homology. *Foundations of Data Science*, **4**, 667 - 705.

Analyses of TDA methodologies: Challenges

Analyses of TDA methodologies: Challenges

Conceptual:

1. What is the information contained in a persistence diagram for a given filtration? (Topology of support? Shape of filter function (e.g. density)? Dependence of observed data? Which one is important?...
2. Different filtrations result in different persistence diagrams with different behavior and different information; different tools are needed for their respective analysis;
3. How to compare persistence diagrams?

Analyses of TDA methodologies: Challenges

Conceptual:

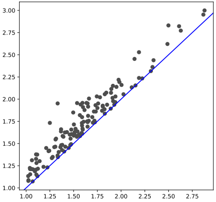
1. What is the information contained in a persistence diagram for a given filtration? (Topology of support? Shape of filter function (e.g. density)? Dependence of observed data? Which one is important?...
2. Different filtrations result in different persistence diagrams with different behavior and different information; different tools are needed for their respective analysis;
3. How to compare persistence diagrams?

Technical:

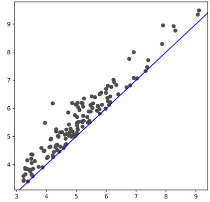
1. How to control *dependence* of points in persistence diagram?
2. What is the population counterpart of quantities of interest?
3. What is the 'right' asymptotic?
4. Can we conduct (asymptotically) valid statistical inference?

PDs for multivariate normals

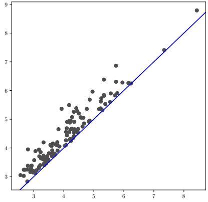
PDs (1-dim holes) for samples of size 200 from a 6-dimensional normal; diagonal covariance matrix; diagonal entries (variances) in captions



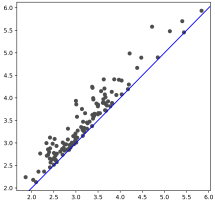
1, 1, 1, 1, 1, 1



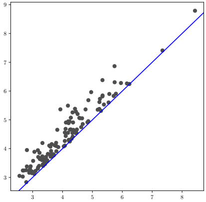
10, 10, 10, 10, 10, 10



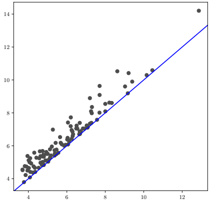
20, 1, 10, 10, 10, 2



1, 1, 10, 10, 10, 2



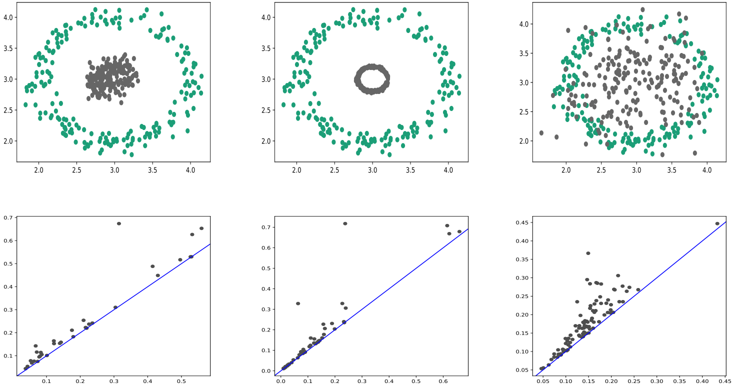
2, 1, 10, 10, 10, 20



1, 10, 10, 10, 10, 20

Nested circles

PDs (1-dim holes); samples of size 200



Dependent data; no signal

Dependent data; no signal

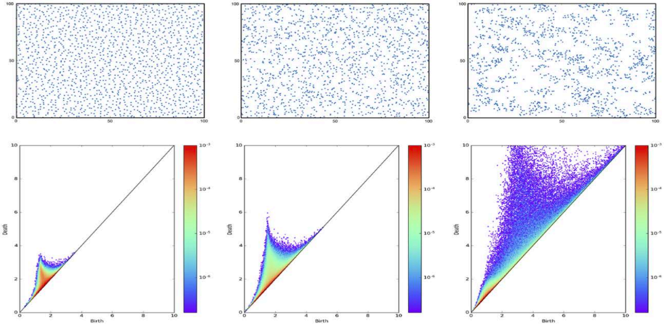


FIG. 1. *Top: Point processes with negative (Ginibre), zero (Poisson), and positive (Poisson cluster) correlations. In these three point processes, the number of points and the density are set to be* 1,000,000 *and* $1/2\pi$*, respectively. Bottom: The normalized persistence diagrams* $\xi_{1,L}/L^2$ *of the above.*

figure from Hiraoka et al. (2018)

One specific goal

- Gaining insight into how the sampling distribution influences the shape of the PD $\rightsquigarrow$ statistical inference.

One specific goal

- Gaining insight into how the sampling distribution influences the shape of the PD $\rightsquigarrow$ statistical inference.

See also Aaromi et al. (2021), who study the dependence of shape of PD (using persistence landscapes) on the trace of the covariance matrix of d-dimensional observations.

Our set-up

- persistence diagrams are based on either the Čech or the VR-filtration
- $X_1, X_2, \ldots, X_n, \ldots$ iid from F in $\mathbb{R}^d$.

Our set-up

- persistence diagrams are based on either the Čech or the VR-filtration
- $X_1, X_2, \ldots, X_n, \ldots$ iid from F in $\mathbb{R}^d$.

$\boxed{\text{We analyze topological noise based on iid data!}}$

think of a null-model

Signatures of PDs - or, extracting features from features

Signatures of PDs - or, extracting features from features

- Betti-curves (Bubenik and Kim, 2007)
- persistent Betti functions (Edelsbrunner et al., 2010)
- persistent homology transform (Curry et al., 2021)
- Euler characteristic curve
- Euler characteristic transform (Curry et al., 2021)
- persistence landscapes (Bubenik, 2015)
- persistence image (Adams et al., 2016)
- persistence surface/ persistence intensity (Chen et al., 2014)
- persistence terrace (Moon et al., 2018)
- methods based on kernel distance (Reininghaus et al., 2016)
- accumulated persistence function (Biscio and Møller, 2019)
- envelope embedding (Chevyrev et al. 2018)
- total persistence
-

We will consider (persistent) Betti functions (and Euler characteristics) in the one-sample set-up.

Persistent Betti curves

$\mathcal{K} = \{K_t, t \in \mathbb{R}\}$ filtration of a simplicial complex, i.e. $K_t \subset K_{t'} \subset K$ for $t \leq t'$.

Persistent Betti curves

$\mathcal{K} = \{K_t, t \in \mathbb{R}\}$ filtration of a simplicial complex, i.e. $K_t \subset K_{t'} \subset K$ for $t \leq t'$.

Definition (k-th Persistent Betti curve)

For $-\infty < s \leq t < \infty$

$$\boxed{\beta_k(K_s, K_t) = \beta_k(s,t) = \mathrm{rank}\, \frac{Z_k(K_s)}{Z_k(K_s) \cap B_k(K_t)}}.$$

Persistent Betti curves

$\mathcal{K} = \{K_t, t \in \mathbb{R}\}$ filtration of a simplicial complex, i.e. $K_t \subset K_{t'} \subset K$ for $t \leq t'$.

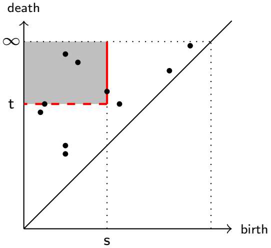
Definition (k-th Persistent Betti curve)

For $-\infty < s \leq t < \infty$

$$\boxed{\beta_k(K_s, K_t) = \beta_k(s,t) = \operatorname{rank} \frac{Z_k(K_s)}{Z_k(K_s) \cap B_k(K_t)}}.$$

We also have

$$\beta_k(s,t) = \sum_{(b,d) \in \mathrm{Dgm}_k} 1(b \leq s,\ d > t).$$



- $\beta_k(t) = \beta_k(t,t)$ is *Betti curve*.

Persistent Betti function

Remarks:

- persistent Betti function can be interpreted as multivariate survival function
  in particular, it characterizes the persistence diagram

- Betti curves are determined by the marginal distributions of the persistence diagram (distributions of birth and death times, respectively)

Asymptotics for (persistent) Betti curves for VR and Čech filtrations

One sample of size n and $n \to \infty$

One sample of size n and $n \to \infty$ technical challenges:

Asymptotics for (persistent) Betti curves for VR and Čech filtrations

One sample of size n and $n \to \infty$ technical challenges:

increasing n $\Rightarrow$ distances between observations decrease

$\Rightarrow$ all birth and death times shrink to zero (assuming noise)

$\Rightarrow$ persistence diagram degenerates asymptotically

Asymptotics for (persistent) Betti curves for VR and Čech filtrations

One sample of size n and $n \to \infty$ technical challenges:

$$
\begin{aligned}
\text{increasing } n \quad &\Rightarrow \text{ distances between observations decrease}\\
&\Rightarrow \text{ all birth and death times shrink to zero (assuming noise)}\\
&\Rightarrow \text{ persistence diagram degenerates asymptotically}
\end{aligned}
$$

- $\boxed{\text{to address this: rescale appropriately!}}$

Asymptotics for (persistent) Betti curves for VR and Čech filtrations

One sample of size n and $n \to \infty$ technical challenges:

increasing n $\Rightarrow$ distances between observations decrease
$\Rightarrow$ all birth and death times shrink to zero (assuming noise)
$\Rightarrow$ persistence diagram degenerates asymptotically

- $\boxed{\text{to address this: rescale appropriately!}}$

Here, critical (or thermodynamic) regime: Consider radii (filtration parameters) of the form $r = tn^{-1/d}$

Asymptotics for (persistent) Betti curves for VR and Čech filtrations

One sample of size n and $n \to \infty$ technical challenges:

increasing n $\Rightarrow$ distances between observations decrease
$\Rightarrow$ all birth and death times shrink to zero (assuming noise)
$\Rightarrow$ persistence diagram degenerates asymptotically

- $\boxed{\text{to address this: rescale appropriately!}}$

Here, critical (or thermodynamic) regime: Consider radii (filtration parameters) of the form $r = tn^{-1/d}$

Alternatively: Fix r and rescale data by $n^{1/d}$.

Asymptotics for (persistent) Betti curves for VR and Čech filtrations

One sample of size n and $n \to \infty$ technical challenges:

increasing n $\Rightarrow$ distances between observations decrease
$\Rightarrow$ all birth and death times shrink to zero (assuming noise)
$\Rightarrow$ persistence diagram degenerates asymptotically

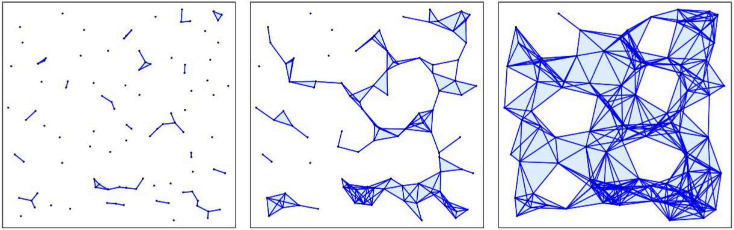
- $\boxed{\text{to address this: rescale appropriately!}}$

Here, critical (or thermodynamic) regime: Consider radii (filtration parameters) of the form $r = tn^{-1/d}$

Alternatively: Fix r and rescale data by $n^{1/d}$.

Motivation: behavior of nearest neighbor distance

Three different regimes



(a) Sparse regime — $r \ll tn^{-1/d}$

(b) Thermodynamic regime — $r = tn^{-1/d}$

(c) Dense regime — $r \gg tn^{-1/d}$

taken from Goel et al., (2019)

Asymptotic normality of Betti numbers for VR and Čech in critical regime

Theorem (Krebs and WP, 2019)

Let f be a bounded Lebesgue density on $[0,1]^d$.

(i) *Let $\mathcal{P}_n$ be a Poisson process on $[0,1]^d$ with intensity nf. For $p = (s,t) \in \Delta$ let*

$$Z_{n,k}(p) = n^{-1/2}\left(\beta_k\left(\mathcal{C}_s(n^{1/d}\mathcal{P}_n), \mathcal{C}_t(n^{1/d}\mathcal{P}_n)\right) - \mathbb{E}\beta_k\left(\mathcal{C}_s(n^{1/d}\mathcal{P}_n), \mathcal{C}_t(n^{1/d}\mathcal{P}_n)\right)\right)$$

denote the centered and scaled persistent Betti numbers. For $k = 0, 1, \ldots, d-1$, there exist functions $\sigma_k : \Delta \times \Delta \to [0,\infty)$, such that for any choice $p_1 = (s_1, t_1), \ldots, p_m = (s_m, t_m) \in \Delta$, as $n \to \infty$

$$\left(Z_{n,k}(p_1), \ldots, Z_{n,k}(p_m)\right)' \to N\left(0, \Sigma(p_1, \ldots, p_m)\right) \qquad \textit{in distribution,}$$

where with $X \sim f$, the covariance matrix $\Sigma(p_1, \ldots, p_m) \in \mathbb{R}^{m\times m}$ is given by

$$\Sigma_{i,j}(p_1, \ldots, p_m) = \mathbb{E}\Big[\sigma_k\left[f^{\frac{1}{d}}(X)\,(s_i, t_i), f^{\frac{1}{d}}(X)\,(s_j, t_j)\right]\Big].$$

Asymptotic normality of Betti numbers for VR and Čech in critical regime

Theorem (Krebs and WP, 2019)

Let f be a bounded Lebesgue density on $[0,1]^d$.

(i) *Let $\mathcal{P}_n$ be a Poisson process on $[0,1]^d$ with intensity nf. For $p=(s,t)\in\Delta$ let*

$$Z_{n,k}(p)=n^{-1/2}\left(\beta_k(\mathcal{C}_s(n^{1/d}\mathcal{P}_n),\mathcal{C}_t(n^{1/d}\mathcal{P}_n))-\mathbb{E}\beta_k(\mathcal{C}_s(n^{1/d}\mathcal{P}_n),\mathcal{C}_t(n^{1/d}\mathcal{P}_n))\right)$$

denote the centered and scaled persistent Betti numbers. For $k=0,1,\dots,d-1$, there exist functions $\sigma_k:\Delta\times\Delta\to[0,\infty)$, such that for any choice $p_1=(s_1,t_1),\dots,p_m=(s_m,t_m)\in\Delta$, as $n\to\infty$

$$(Z_{n,k}(p_1),\dots,Z_{n,k}(p_m))'\to N(0,\Sigma(p_1,\dots,p_m)) \qquad \textit{in distribution,}$$

where with $X\sim f$, the covariance matrix $\Sigma(p_1,\dots,p_m)\in\mathbb{R}^{m\times m}$ is given by

$$\Sigma_{i,j}(p_1,\dots,p_m)=\mathbb{E}\Big[\sigma_k\big[f^{\frac{1}{d}}(X)\,(s_i,t_i),f^{\frac{1}{d}}(X)\,(s_j,t_j)\big]\Big].$$

(ii) *Replacing $\mathcal{P}_n$ in part (i) by a Binomial process $\mathbb{X}_n$ with density f gives a similar asymptotic normality result as in (i), but with a covariance matrix $\widetilde{\Sigma}\in\mathbb{R}^{m\times m}$ of the form*

$$\widetilde{\Sigma}_{i,j}(p_1,\dots,p_m)=\Sigma_{i,j}(p_1,\dots,p_m)-\mathbb{E}\Big[\alpha\big[f^{\frac{1}{d}}(X)\,(s_i,t_i)\big]\Big]\,\mathbb{E}\Big[\alpha\big[f^{\frac{1}{d}}(X)\,(s_j,t_j)\big]\Big],$$

for some function $\alpha:\Delta\to\mathbb{R}$, and with $\Sigma_{i,j}$ and X as in part (i).

Putting $s_i=t_i$ for all $i=1,\dots,m$ gives the joint asymptotic normality for Betti numbers.

Comments

Comments

1. Recall: we are analyzing topological noise

Comments

1. Recall: we are analyzing topological noise
2. Related results: Trinh (2017), Hiraoka et al. (2018), Trinh (2019), Goel et al. (2019), Owada and Thomas (2020)

Comments

1. Recall: we are analyzing topological noise
2. Related results: Trinh (2017), Hiraoka et al. (2018), Trinh (2019), Goel et al. (2019), Owada and Thomas (2020)
3. Novelties of our result:
   - multivariate result

Comments

1. Recall: we are analyzing topological noise
2. Related results: Trinh (2017), Hiraoka et al. (2018), Trinh (2019), Goel et al. (2019), Owada and Thomas (2020)
3. Novelties of our result:
   - multivariate result
   - no restriction on filtration parameter

Comments

1. Recall: we are analyzing topological noise
2. Related results: Trinh (2017), Hiraoka et al. (2018), Trinh (2019), Goel et al. (2019), Owada and Thomas (2020)
3. Novelties of our result:
    - multivariate result
    - no restriction on filtration parameter
    - allow for sampling from a Binomial process

Comments

1. Recall: we are analyzing topological noise

2. Related results: Trinh (2017), Hiraoka et al. (2018), Trinh (2019), Goel et al. (2019), Owada and Thomas (2020)

3. Novelties of our result:
    - multivariate result
    - no restriction on filtration parameter
    - allow for sampling from a Binomial process

4. Derivations rely on the notion of *stabilization* (see below)

Comments

1. Recall: we are analyzing topological noise
2. Related results: Trinh (2017), Hiraoka et al. (2018), Trinh (2019), Goel et al. (2019), Owada and Thomas (2020)
3. Novelties of our result:
   - multivariate result
   - no restriction on filtration parameter
   - allow for sampling from a Binomial process
4. Derivations rely on the notion of *stabilization* (see below)
5. Centering: expected Betti number

Convergence of expected Betti number in critical regime

Convergence of expected Betti number in critical regime

Theorem (convergence of expected values in the critical regime)

(i) *Let $\mathcal{P}_n$ be a homogeneous Poisson process on $[0,1]^d$ with intensity n. Then, for $k = 0, 1, 2, \ldots, d-1$ there exist functions $\gamma_k : \Delta \to [0, \infty)$, such that, as $n \to \infty$,*

$$\frac{1}{n}\,\mathbb{E}\beta_k\big(\mathcal{C}_s(n^{1/d}\mathcal{P}_n), \mathcal{C}_t(n^{1/d}\mathcal{P}_n)\big) \to \gamma_k(s,t).$$

(ii) *Let f be a bounded probability density on $[0,1]^d$ with compact support. Furthermore, let $\mathbb{X}_n$ be a Binomial process on $[0,1]^d$ with density f, and let $\mathcal{P}_n$ be an inhomogenous Poisson process on $[0,1]^d$ with intensity nf. Then, for $k = 0, 1, \ldots, d-1$ and with $\gamma_k(s,t)$ from part (i), as $n \to \infty$,*

$$\frac{1}{n}\,\mathbb{E}\beta_k\big(\mathcal{C}_s(n^{1/d}\mathbb{X}_n), \mathcal{C}_t(n^{1/d}\mathbb{X}_n)\big) \to \mathbb{E}\big[\gamma_k(s\,f^{1/d}(X),\ t\,f^{1/d}(X))\big]$$

and

$$\frac{1}{n}\,\mathbb{E}\beta_k\big(\mathcal{C}_s(n^{1/d}\mathcal{P}_n), \mathcal{C}_t(n^{1/d}\mathcal{P}_n)\big) \to \mathbb{E}\big[\gamma_k\big(s\,f^{1/d}(X),\ t\,f^{1/d}(X)\big)\big]$$

with $X \sim f$. Setting $s = t$ in either (i) or (ii) gives results for the Betti numbers.

See Trinh (2017), Hiraoka et al. (2018), Trinh (2019), Goel et al. (2019), Owada and Thomas (2020).

Discussion of asymptotic means and variances

- In general, the form of the functions γ_k, σ_k and α_k is unknown (only for $k = 0$ more is known).

- Note, however, that these functions are determined by the behavior under a homogeneous Poisson sampling $\rightsquigarrow$ they don't depend on the sampling density;

Consequences of this observation: The dependence of the limit on the sampling density f is through quantities of the form

$$\mathbb{E}_f\left[\Psi_{k,s,t}\left(f^{1/d}(X)\right)\right].$$

This will be discussed further below.

Statistical inference

For conducting statistical inference, one needs to know the (asymptotic) distribution. So we need to

estimate the limit variance of asymptotic normal.

We will do this using a bootstrap procedure.

The bootstrap

$\boxed{\text{A computational device to estimate sampling distributions.}}$

Efron (1979)

$\boxed{\text{A computational device to estimate sampling distributions.}}$

Efron (1979)

Basic idea:

Statistical Model

- $\mathbb{X}_n$ is drawn from F
- F unknown
- one sample of size n from F

- sampling distribution of $T_n(F)$ unknown

The bootstrap

A computational device to estimate sampling distributions.

Efron (1979)

Basic idea:

Statistical Model

- $\mathbb{X}_n$ is drawn from F
- F unknown
- one sample of size n from F
- sampling distribution of $T_n(F)$ unknown

Bootstrap World

- $\mathbb{X}_n^*$ is drawn from $\widehat{F}_n = F_n(\mathbb{X}_n)$
- $\widehat{F}_n$ is known
- draw as many bootstrap samples as desired
- estimate sampling distribution of $T_n(F)$ by $T_n(F_n)$
- $T_n(F_n)$ can be approximated arbitrarily well by Monte Carlo simulation

$\boxed{\text{A computational device to estimate sampling distributions.}}$

Efron (1979)

Basic idea:

Statistical Model

- $\mathbb{X}_n$ is drawn from F
- F unknown
- one sample of size n from F
- sampling distribution of $T_n(F)$ unknown

Bootstrap World

- $\mathbb{X}_n^*$ is drawn from $\widehat{F}_n = F_n(\mathbb{X}_n)$
- $\widehat{F}_n$ is known
- draw as many bootstrap samples as desired
- estimate sampling distribution of $T_n(F)$ by $T_n(F_n)$
- $T_n(F_n)$ can be approximated arbitrarily well by Monte Carlo simulation

Standard bootstrap: F_n = empirical distribution given by $\mathbb{X}_n$

Smooth bootstrap: Draw samples from a KDE $f_{n,h}(x)$ based on $\mathbb{X}_n$

Bootstrap confidence regions for Betti numbers for VR and Čech filtrations

Smooth bootstrap: Draw samples from a KDE $f_{n,h}(x)$ based on $\mathbb{X}_n$

- disadvantage: curse of dimensionality; bandwidth needs to be chosen

Bootstrap confidence regions for Betti numbers for VR and Čech filtrations

Smooth bootstrap: Draw samples from a KDE $f_{n,h}(x)$ based on $\mathbb{X}_n$

- disadvantage: curse of dimensionality; bandwidth needs to be chosen
- advantage in TDA context: no repeated observations (w.p. 1)
  - Repeated observations disregarded when building VR and Čech complexes
  - this means: effectively sample size is smaller and random
    (scaling issues; technical problems)

Bootstrap confidence regions for Betti numbers for VR and Čech filtrations

Smooth bootstrap: Draw samples from a KDE $f_{n,h}(x)$ based on $\mathbb{X}_n$

- disadvantage: curse of dimensionality; bandwidth needs to be chosen
- advantage in TDA context: no repeated observations (w.p. 1)
  - Repeated observations disregarded when building VR and Čech complexes
  - this means: effectively sample size is smaller and random (scaling issues; technical problems)

Recall, with $\mathcal{F}$ either VR or $\mathcal{C}$,

$$Z_{n,k}(p,\mathcal{F}) = n^{-1/2}\Big(\beta_k\big(\mathcal{F}_s(n^{\frac{1}{d}}\mathbb{X}_n),\mathcal{F}_t(n^{\frac{1}{d}}\mathbb{X}_n)\big) - \mathbb{E}\big[\beta_k\big(\mathcal{F}_s(n^{\frac{1}{d}}\mathbb{X}_n),\mathcal{F}_t(n^{\frac{1}{d}}\mathbb{X}_n)\big)\big]\Big).$$

The corresponding bootstrap version is

$$Z^*_{n,k}(p,\mathcal{F}) = n^{-1/2}\Big(\beta_k\big(\mathcal{F}_s(n^{\frac{1}{d}}\mathbb{X}^*_n),\mathcal{F}_t(n^{\frac{1}{d}}\mathbb{X}^*_n)\big) - \mathbb{E}\big[\beta_k\big(\mathcal{F}_s(n^{\frac{1}{d}}\mathbb{X}^*_n),\mathcal{F}_t(n^{\frac{1}{d}}\mathbb{X}^*_n)\big)\big|\mathbb{X}_n\big]\Big)$$

where $\mathbb{X}^*_n \sim \widehat{f}_{n,h}$.

Bootstrap confidence regions for Betti numbers for VR and Čech filtrations

Theorem (bootstrap for persistent Betti numbers)

Let $\mathbb{X}_n$ be a Binomial process in $\mathbb{R}^d$ with intensity f. Fix $k \geq 0$, and let $m \geq 1$ and $p_1, \ldots, p_m \in \Delta$. If

(i) $\|f\|_{2k+3} < \infty$,

(ii) $\|\widehat{f}_{n,h} - f\|_q \to 0$ *in probability (or a.s.) as* $n \to \infty$ *for some* $q > 2k+3$,

then, for $\mathcal{F} = \mathrm{VR}$ *and* $\mathcal{F} = \mathcal{C}$,

$$(Z_{n,k}(p_1, \mathcal{F}), \ldots, Z_{n,k}(p_m, \mathcal{F}))' \to_{\mathcal{D}} N(0, \Sigma_m) \quad \text{as } n \to \infty,$$

if and only if

$$(Z^*_{n,k}(p_1, \mathcal{F}), \ldots, Z^*_{n,k}(p_m, \mathcal{F}))' \to_{\mathcal{D}} N(0, \Sigma_m) \quad \text{in probab. (or a.s.) as } n \to \infty.$$

Sufficient conditions for convergence in probability (for all $q > 0$):

f and kernel K bounded, and $nh^{2d} \to \infty$.

Real Data Example

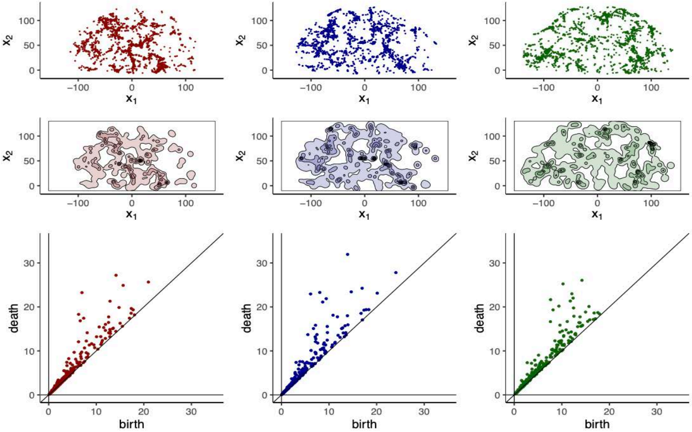
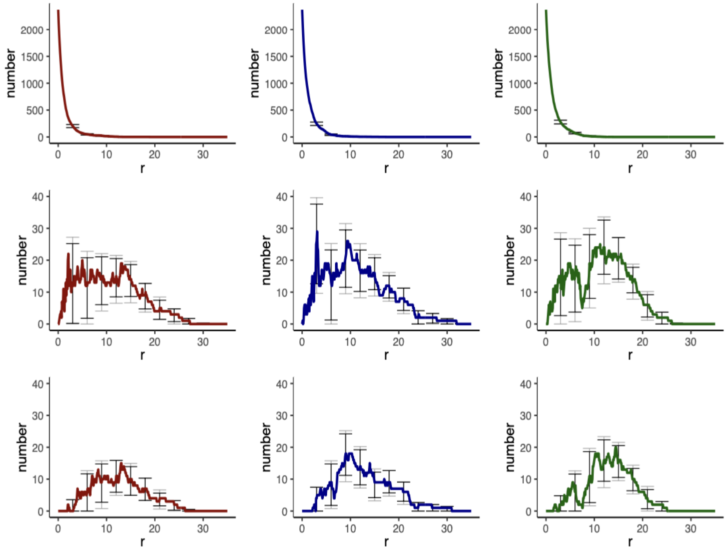


FIG 2. *Top row: Transformed point clouds. Middle row: Density estimates using adaptive bandwidth. Bottom row: Persistence diagrams in dimension $q = 1$ for the Vietoris-Rips complex. Columns from left to right: Galaxies with redshifts within* $(0.025, 0.026)$, $(0.027, 0.028)$, *and* $(0.029, 0.030)$, *respectively. Axis units are given in Megaparsecs (Mpc).*

Real Data Example



Galaxy data (three slices): top row: $\beta_0(r)$; middle row: $\beta_1(r)$; bottom row: $\beta_1(r, r+1)$

Theorem (Krebs & WP, 2021, Thomas and Owada, 2021)

Suppose that f can be approximated uniformly by a sequence of blocked density functions on $[0,1]^d$ of the form $f_m = \sum_{i=1}^{m^d} b_i \mathbb{1}_{B_i}$ with blocks B_i of the form $B_i = [a_1, b_1] \times \cdots \times [a_d, b_b]$ such that for all $i = 1, \ldots, d$ one has $c\frac{1}{m} \leq b_i - a_i \leq C\frac{1}{m}$ for some $c, C > 0$. Then, for a Poisson sampling scheme with intensity nf, as $n \to \infty$,

$$\boxed{\nu_n \Rightarrow G \qquad \text{weakly, in} \quad D[0,T],}$$

where $D[0,T]$ denotes Skorohod space, and G is a mean zero Gaussian process on $[0,T]$ with covariance of the form

$$\mathrm{Cov}(G(s), G(t)) = \mathbb{E}\big[\gamma(f(X)^{\frac{1}{d}}(s,t))\big].$$

where $X \sim f$. A similar result holds for a Binomial sampling scheme (with density f), where the covariance function changes to

$$\mathrm{Cov}(G(s), G(t)) = \mathbb{E}\big[\gamma(f(X)^{\frac{1}{d}}(s,t))\big] - \mathbb{E}\big[\alpha(f(X)^{\frac{1}{d}}s)\big]\,\mathbb{E}\big[\alpha(f(X)^{\frac{1}{d}}t)\big]$$

for some functions $\alpha : \Delta \times \Delta \to \mathbb{R}$, $\gamma : \Delta \times \Delta \to [0,\infty)$.

How does persistence diagram depend on sampling density?

How does persistence diagram depend on sampling density?

Recall:

- Limits of expected values and limit variances all depend on f through quantities of the form

$$\boxed{\int_{\mathbb{X}} \Psi(f^{1/d}(x))\, f(x) dx}$$

for some functions Ψ not depending on f.

Writing

$$\int_{\mathbb{X}} \Psi(f^{1/d}(x))\, f(x) dx = \mathbb{E}\Psi\big(f^{1/d}(X)\big),$$

we see that this value

depends on f only through the distribution of $f(X)$, where $X \sim f$.

See Vishwanath et al. (2020)

How does persistence diagram depend on sampling density?

What is this distribution? The survival function of $Y = f(X)$ with $X \sim f$ is

$$\boxed{S_f(t) = P_f\big(f(X) \geq t\big) = \int_{\Gamma(\lambda)} f(x)dx = F\big(\Gamma(t)\big),}$$

where $\Gamma(t) = \big\{x : f(x) \geq t\big\}$ (superlevel set of f); and F is distribution with density f.

$$S_f = S_g$$

$$\Rightarrow$$

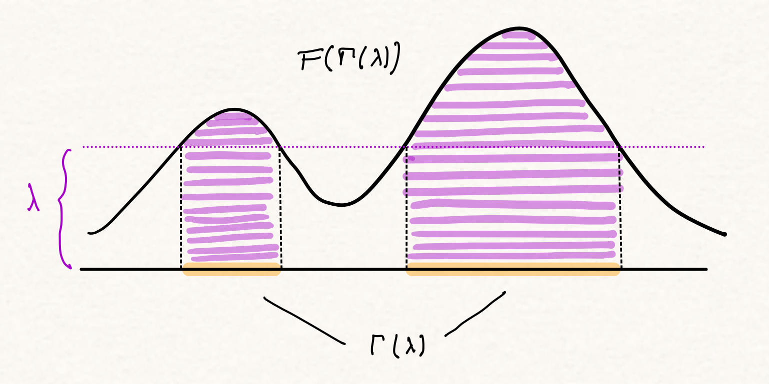
same asymptotic behavior of persistent Betti function (and Euler characteristic)

How does persistence diagram depend on sampling density?

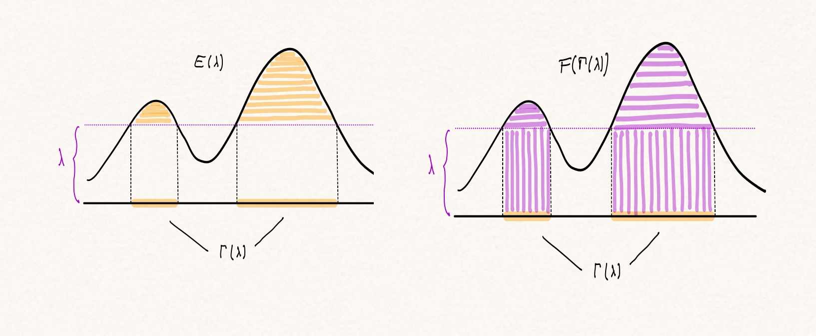
When is $S_f = S_g$?

How does persistence diagram depend on sampling density?

When is $S_f = S_g$?



How does persistence diagram depend on sampling density?



Excess mass function

We have: $S_f = S_g \quad \Leftrightarrow \quad E_f = E_g \quad \Leftrightarrow \quad \mathrm{Leb}(\Gamma_f(\lambda)) = \mathrm{Leb}(\Gamma_g(\lambda)) \ \forall \ \lambda$

E.g. f and g its monotonically decreasing rearrangement satisfy this!

Bandwidth selection for TDA

Bandwidth selection for TDA

Label	Description
F_1	Rotationally symmetric in $\mathbb{R}^2$, finite L_8 norm
F_2	Rotationally symmetric in $\mathbb{R}^2$, finite L_2 norm, infinite L_8 norm
F_3	$\mathbb{S}^1$ embedded in $\mathbb{R}^2$, additive Gaussian noise
F_4	Uniformly distributed over $B_0(1)$ in $\mathbb{R}^3$, additive Gaussian noise
F_5	5 clusters in $\mathbb{R}^3$, additive exponential noise
F_6	$\mathbb{S}^2$ embedded in $\mathbb{R}^5$, additive Cauchy noise
F_7	Flat figure-8 embedded in $\mathbb{R}^{10}$, additive Gaussian noise

Distr.	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_4	F_5	F_6	F_7
	$q=1$							$q=2$			
r	4.94	5.20	3.03	1.92	0.30	1.78	1.28	2.96	0.39	2.71	1.46
s	5.36	5.60	3.28	2.12	0.31	1.91	1.32	3.04	0.40	2.80	1.47
$n=100$	0.896	0.965	0.921	0.859	0.954	0.19		0.908	0.705	0.038	
	0.931	0.959	0.914	0.809	0.941	0.133		0.903	0.604	0.045	
	0.903	0.97	0.91	0.859	0.927	0.049		0.902	0.363	0.002	
	0.359	0.931	0.942	0.864	0	0	0.656	0.902	0	0	0.045
$n=200$	0.908	0.971	0.94	0.898	0.942	0.159		0.878	0.795	0.125	
	0.92	0.972	0.946	0.891	0.923	0.106		0.872	0.707	0.074	
	0.888	0.975	0.959	0.906	0.892	0.06		0.908	0.277	0.031	
	0.299	0.954	0.903	0.899	0	0	0.766	0.882	0	0	0.537
$n=300$	0.9	0.971	0.926	0.921	0.94	0.183		0.854	0.906	0.225	
	0.94	0.971	0.938	0.896	0.94	0.087		0.854	0.917	0.072	
	0.913	0.971	0.94	0.896	0.922	0.054		0.855	0.964	0.074	
	0.283	0.956	0.925	0.906	0	0	0.835	0.856	0	0	0.508
$n=400$	0.918	0.961	0.947	0.934	0.96	0.175		0.851	0.883	0.259	
	0.927	0.951	0.938	0.92	0.955	0.063		0.839	0.88	0.076	
	0.908	0.976	0.933	0.924	0.939	0.062		0.863	0.958	0.099	
	0.266	0.961	0.909	0.922	0.114	0	0.891	0.859	0	0	0.584

TABLE 2

Coverage proportions for 95% smoothed bootstrap confidence intervals on the mean persistent Betti numbers; coverage is estimated using $N = 1,000$ independent base samples with $B = 500$ bootstrap samples each. True mean persistent Betti numbers are estimated using a large ($N = 100,000$) number of independent samples from the true distribution. For each case, the values from top to bottom: Coverage proportions using "Hpi.diag", "Hlscv.diag", "Hscv.diag", and "bw.silv" bandwidth selectors, respectively (see Section 5).

Some comments on the proof of the above results

Some comments on the proof of the above results

Proofs heavily rely on the notion of stabilization of the add-one cost function.

Some comments on the proof of the above results

Proofs heavily rely on the notion of stabilization of the add-one cost function.

- $\mathcal{X}$ = set of all finite (multi) sets of $\mathbb{R}^d$
- $H : \mathcal{X} \to \mathbb{R}$ (such as Betti number or Euler characteristic)

Definition (add-one cost)

For $z \in \mathbb{R}^d$, the *add-one cost* function (for H) is

$$D_z(\mathbb{X}) = H(\mathbb{X} \cup \{z\}) - H(\mathbb{X}).$$

Examples: add-one cost

Example 1: Let

- $H(\mathbb{X}) = |S_k(\mathbb{X})|$ = number of k-simplices in $\mathrm{VR}_r(\mathbb{X})$ (for some fixed $r > 0$);

Examples: add-one cost

Example 1: Let

- $H(\mathbb{X}) = |S_k(\mathbb{X})| =$ number of k-simplices in $\mathrm{VR}_r(\mathbb{X})$ (for some fixed $r > 0$);
- $z = 0$;

Examples: add-one cost

Example 1: Let

- $H(\mathbb{X}) = |S_k(\mathbb{X})| =$ number of k-simplices in $\mathrm{VR}_r(\mathbb{X})$ (for some fixed $r > 0$);
- $z = 0$;
- $\psi(x_0, x_1, \ldots, x_k) =$ filtration time of $\sigma = [x_0, x_1, \ldots x_k]$.

Examples: add-one cost

Example 1: Let

- $H(\mathbb{X}) = |S_k(\mathbb{X})|$ = number of k-simplices in $\mathrm{VR}_r(\mathbb{X})$ (for some fixed $r > 0$);
- $z = 0$;
- $\psi(x_0, x_1, \ldots, x_k)$ = filtration time of $\sigma = [x_0, x_1, \ldots x_k]$.

Then

$$D_0(\mathbb{X}) = \sum_{\{x_{i_1}, \ldots, x_{i_k}\} \subset \mathbb{X}} \mathbf{1}\big(\psi(0, x_{i_1}, \ldots, x_{i_k}) \leq r\big).$$

Examples: add-one cost

Example 1: Let

- $H(\mathbb{X}) = |S_k(\mathbb{X})| =$ number of k-simplices in $\mathrm{VR}_r(\mathbb{X})$ (for some fixed $r > 0$);
- $z = 0$;
- $\psi(x_0, x_1, \ldots, x_k) =$ filtration time of $\sigma = [x_0, x_1, \ldots x_k]$.

Then

$$D_0(\mathbb{X}) = \sum_{\{x_{i_1}, \ldots, x_{i_k}\} \subset \mathbb{X}} \mathbf{1}\big(\psi(0, x_{i_1}, \ldots, x_{i_k}) \leq r\big).$$

Example 2: Let

- $H(k; \mathbb{X}) =$ length of k-NN graph over $\mathbb{X}$.

Example 1: Let

- $H(\mathbb{X}) = |S_k(\mathbb{X})|$ = number of k-simplices in $\mathrm{VR}_r(\mathbb{X})$ (for some fixed $r > 0$);
- $z = 0$;
- $\psi(x_0, x_1, \ldots, x_k)$ = filtration time of $\sigma = [x_0, x_1, \ldots x_k]$.

Then

$$D_0(\mathbb{X}) = \sum_{\{x_{i_1}, \ldots, x_{i_k}\} \subset \mathbb{X}} \mathbf{1}\big(\psi(0, x_{i_1}, \ldots, x_{i_k}) \leq r\big).$$

Example 2: Let

- $H(k; \mathbb{X})$ = length of k-NN graph over $\mathbb{X}$.
- $z = 0$;

Examples: add-one cost

Example 1: Let

- $H(\mathbb{X}) = |S_k(\mathbb{X})|$ = number of k-simplices in $\mathrm{VR}_r(\mathbb{X})$ (for some fixed $r > 0$);
- $z = 0$;
- $\psi(x_0, x_1, \dots, x_k)$ = filtration time of $\sigma = [x_0, x_1, \dots x_k]$.

Then

$$D_0(\mathbb{X}) = \sum_{\{x_{i_1}, \dots, x_{i_k}\} \subset \mathbb{X}} \mathbf{1}\big(\psi(0, x_{i_1}, \dots, x_{i_k}) \le r\big).$$

Example 2: Let

- $H(k; \mathbb{X})$ = length of k-NN graph over $\mathbb{X}$.
- $z = 0$;
- $\mathrm{kNN}(x, \mathbb{X})$ = set of k-nearest neighbors in $\mathbb{X}$ of x

Examples: add-one cost

Example 1: Let

- $H(\mathbb{X}) = |S_k(\mathbb{X})|$ = number of k-simplices in $\mathrm{VR}_r(\mathbb{X})$ (for some fixed $r > 0$);
- $z = 0$;
- $\psi(x_0, x_1, \ldots, x_k)$ = filtration time of $\sigma = [x_0, x_1, \ldots x_k]$.

Then

$$D_0(\mathbb{X}) = \sum_{\{x_{i_1},\ldots,x_{i_k}\}\subset\mathbb{X}} \mathbf{1}\big(\psi(0, x_{i_1}, \ldots, x_{i_k}) \leq r\big).$$

Example 2: Let

- $H(k; \mathbb{X})$ = length of k-NN graph over $\mathbb{X}$.
- $z = 0$;
- $\mathrm{kNN}(x, \mathbb{X})$ = set of k-nearest neighbors in $\mathbb{X}$ of x

Then,

$$D_0(\mathbb{X}) = \sum_{x\in\mathbb{X}} d(0, x_j)\mathbf{1}\big(0 \in \mathrm{kNN}(x, \mathbb{X});\ x \in \mathrm{kNN}(0, \mathbb{X})\big).$$

Stabilization

Definition (weak stabilization)

A functional $H : \mathcal{X} \to \mathbb{R}$ is *weakly stabilizing* on a locally finite point process $\mathbb{Y}$ in $\mathbb{R}^d$, if there exists a random variable Δ_∞, such that, for any sequence $\{B_n\}_{n\geq 1}$ of (measurable) sets satisfying $B_n \to \mathbb{R}^d$ as $n \to \infty$, we have

$$D_z(\mathbb{Y} \cap B_n) \to \Delta_\infty \qquad \text{a.s. as } n \to \infty.$$

Definition (strong stabilization)

A functional $H : \mathcal{X} \to \mathbb{R}$ is *strongly stabilizing* on a locally finite point process $\mathbb{Y}$ in $\mathbb{R}^d$, if there exists random variables S (*radius of stabilization*) and Δ_∞ such that, for $z \in \mathbb{R}^d$, with probability 1,

$$D_z((\mathbb{Y} \cap B_S(z)) \cup A) = \Delta_\infty \qquad \text{for all } A \text{ finite with } A \subset \mathbb{R}^d \setminus B_S(z).$$

Stabilization

Definition (weak stabilization)

A functional $H : \mathcal{X} \to \mathbb{R}$ is *weakly stabilizing* on a locally finite point process $\mathbb{Y}$ in $\mathbb{R}^d$, if there exists a random variable Δ_∞, such that, for any sequence $\{B_n\}_{n\geq 1}$ of (measurable) sets satisfying $B_n \to \mathbb{R}^d$ as $n \to \infty$, we have

$$D_z(\mathbb{Y} \cap B_n) \to \Delta_\infty \qquad \text{a.s. as } n \to \infty.$$

Definition (strong stabilization)

A functional $H : \mathcal{X} \to \mathbb{R}$ is *strongly stabilizing* on a locally finite point process $\mathbb{Y}$ in $\mathbb{R}^d$, if there exists random variables S (*radius of stabilization*) and Δ_∞ such that, for $z \in \mathbb{R}^d$, with probability 1,

$$D_z((\mathbb{Y} \cap B_S(z)) \cup A) = \Delta_\infty \qquad \text{for all } A \text{ finite with } A \subset \mathbb{R}^d \setminus B_S(z).$$

- Clearly: strong stabilization $\Rightarrow$ weak stabilization;

Comments

- 'stabilization' formalized in Penrose and Yukich (2001); ideas go back to Kesten and Lee (1996)
- H stabilizes $\rightsquigarrow$ difference operator determined by 'local' information $\rightsquigarrow$ control of dependence
- adding moment conditions $\rightsquigarrow$ limit theorems

Examples

Example 1: Simplex counts:

$$D_0(\mathbb{Y} \cup B_n) = \sum_{\{X_{i_1},\dots,X_{i_k}\} \subset B_n} \mathbf{1}\big(\psi(0, X_{i_1}, \dots, X_{i_k}) \le r\big).$$

(Recall : $\Psi(\sigma)$ is filtration time, and we consider filtration VR_r.)

Examples

Example 1: Simplex counts:

$$D_0(\mathbb{Y} \cup B_n) = \sum_{\{X_{i_1},\ldots,X_{i_k}\} \subset B_n} \mathbf{1}\big(\psi(0, X_{i_1}, \ldots, X_{i_k}) \leq r\big).$$

(Recall : $\Psi(\sigma)$ is filtration time, and we consider filtration VR_r.)

Consider strong stabilization criterion.

Examples

Example 1: Simplex counts:

$$D_0(\mathbb{Y} \cup B_n) = \sum_{\{X_{i_1},\dots,X_{i_k}\} \subset B_n} \mathbf{1}\big(\psi(0, X_{i_1}, \dots, X_{i_k}) \le r\big).$$

(Recall : $\Psi(\sigma)$ is filtration time, and we consider filtration VR_r.)

Consider strong stabilization criterion.

Question to answer: Are there random variables S and $\Delta_\infty = \Delta_\infty(r)$, such that for all finite $A \subset \big(B_S(z)\big)^{\complement}$,

$$D_0\big((\mathbb{Y} \cap B_S(z)) \cup A\big) = \Delta_\infty(r) \quad \text{a.s.?}$$

Examples

Example 1: Simplex counts:

$$D_0(\mathbb{Y} \cup B_n) = \sum_{\{X_{i_1},\ldots,X_{i_k}\} \subset B_n} \mathbf{1}\big(\psi(0, X_{i_1}, \ldots, X_{i_k}) \leq r\big).$$

(Recall : $\Psi(\sigma)$ is filtration time, and we consider filtration VR_r.)

Consider strong stabilization criterion.

Question to answer: Are there random variables S and $\Delta_\infty = \Delta_\infty(r)$, such that for all finite $A \subset \big(B_S(z)\big)^{\complement}$,

$$D_0\big((\mathbb{Y} \cap B_S(z)) \cup A\big) = \Delta_\infty(r) \quad \text{a.s.?}$$

Observe:

$d(0, A) > 2r \quad \Rightarrow \quad$ points in A cannot be part of a simplex containing 0

$\Rightarrow \quad S = 2r$ is a radius of stabilization.

Note: S is not random; holds for both VR and Čech complex!

- this translates to Euler characteristic (as alternating sum of simplex counts)

Example 2: Length of k-NN graph; $d = 2$ (for simplicity)

Again, we show strong stabilization; here stabilization radius is random.

Better control of the tail behavior of the stabilization radius

$\Rightarrow$ stronger results

A key technical result

A key technical result

A key ingredient for the proofs of bootstrap validity:

Proposition

Let $\mathbb{X}_n$ and $\mathbb{Y}_n$ be Binomial processes with densities f and g, respectively, and suppose that $\|f\|_p < \infty$ for $p \geq 2$. Further let ψ be strongly stabilizing over $\mathbb{X}_n$. Then there exists a coupling between $\mathbb{X}_n$ and $\mathbb{Y}_n$ such that

$$\sup_{n\in\mathbb{N}} \mathrm{Var}\left[\frac{1}{\sqrt{n}}\left[\left(\psi\left(n^{1/d}\mathbb{Y}_n\right) - \psi\left(n^{1/d}\mathbb{X}_n\right)\right]\right] \leq \gamma\left(\|f - g\|_p\right),$$

where the rate function $\gamma\colon \mathbb{R}_+ \to \mathbb{R}$ is increasing with $\lim_{\delta\to 0}\gamma(\delta) = 0$ and depends only on f and p.

This applies to both persistent Betti numbers and Euler characteristic (pointwise)

Another key ingredient - the Geometric Lemma

Lemma (Corollary of Hiraoka et al. (2018), Lemma 2.11)

Let $A \subseteq B$ be two finite point sets of $\mathbb{R}^d$. Then, with $s \leq t$,

$$\left|\beta_k\left(\mathcal{F}_s(n^{1/d}B),\mathcal{F}_t(n^{1/d}B)\right) - \beta_k\left(\mathcal{F}_s(n^{1/d}A), \mathcal{F}_t(n^{1/d}A)\right)\right|$$

$$\leq \sum_{j=q}^{q+1} \left|S_j(\mathcal{F}_t(B) \setminus S_j(\mathcal{F}_t(A)\right|.$$

References I

Amézquita, E. J., Quigley, M. Y., Ophelders, T., Munch, E., and Chitwood, D. H. (2020). The shape of things to come: Topological data analysis and biology, from molecules to organisms. *Developmental Dynamics*, 249(7):816–833.

Bobrowski, O. and Mukherjee, S. (2015). The topology of probability distributions on manifolds. *Probab. Theory Relat. Fields*, 161(3):651–686.

Boissonnat, J.-D., Chazal, F., and Yvinec, M. (2018). *Geometric and Topological Inference*. Cambridge University Press, 1 edition.

Bukkuri, A., Andor, N., and Darcy, I. K. (2021). Applications of Topological Data Analysis in Oncology. *Front. Artif. Intell.*, 4:659037.

Chazal, F. and Michel, B. (2021). An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. arXiv:1710.04019 [cs, math, stat].

Davies, T. (2022). A Review of Topological Data Analysis for Cybersecurity. arXiv:2202.08037 [cs].

Decreusefond, L., Ferraz, E., Randriambololona, H., and Vergne, A. (2014). Simplicial homology of random configurations. *Advances in Applied Probability*, 46(2):325–347. Publisher: Applied Probability Trust.

Dey, T. K. and Wang, Y. (2022). *Computational Topology for Data Analysis*. Cambridge University Press, 1 edition.

Dłotko, P., Qiu, W., and Rudkin, S. (2019). Cyclicality, Periodicity and the Topology of Time Series. arXiv:1905.12118 [cs, math].

References II

Edelsbrunner, H. and Harer, J. (2010). *Computational Topology: An Introduction*. American Mathematical Society, Providence, R.I. OCLC: ocn427757156.

Goel, A., Trinh, K. D., and Tsunoda, K. (2019). Strong Law of Large Numbers for Betti Numbers in the Thermodynamic Regime. *J Stat Phys*, 174(4):865–892.

Hiraoka, Y., Shirai, T., and Trinh, K. D. (2018). Limit theorems for persistence diagrams. *Ann. Appl. Probab.*, 28(5).

Joshi, M. and Joshi, D. (2019). A survey of Topological Data Analysis Methods for Big Data in Healthcare Intelligence. *International Journal of Applied Engineering Research*, 14(2):5.

Kerscher, M. (2000). Statistical Analysis of Large-Scale Structure in the Universe. In Mecke, K. R. and Stoyan, D., editors, *Statistical Physics and Spatial Statistics*, Lecture Notes in Physics, pages 36–71, Berlin, Heidelberg. Springer.

Kesten, H. and Lee, S. (1996). The central limit theorem for weighted minimal spanning trees on random points. *The Annals of Applied Probability*, 6(2):495–527. Publisher: Institute of Mathematical Statistics.

Khanamiri, H. H., Berg, C. F., Slotte, P. A., Schlüter, S., and Torsæter, O. (2018). Description of Free Energy for Immiscible Two-Fluid Flow in Porous Media by Integral Geometry and Thermodynamics. *Water Resources Research*, 54(11):9045–9059.

Kilner, J. M., Kiebel, S. J., and Friston, K. J. (2005). Applications of random field theory to electrophysiology. *Neurosci Lett*, 374(3):174–178.

References III

Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., and Harrington, H. A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Sci.*, 6(1):1–38. Number: 1 Publisher: SpringerOpen.

Owada, T. and Thomas, A. M. (2020). Limit theorems for process-level Betti numbers for sparse and critical regimes. *Advances in Applied Probability*, 52(1):1–31. Publisher: Cambridge University Press.

Penrose, M. D. and Yukich, J. (2001). Central Limit Theorems for Some Graphs in Computational Geometry. *Ann. Appl. Probab.*, 11(4).

Rabadan, R. and Blumberg, A. J. (2019). *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press, Cambridge.

Salch, A., Regalski, A., Abdallah, H., Suryadevara, R., Catanzaro, M. J., and Diwadkar, V. A. (2021). From mathematics to medicine: A practical primer on topological data analysis (TDA) and the development of related analytic tools for the functional discovery of latent structure in fMRI data. *PLoS ONE*, 16(8):e0255859.

Scholz, C., Wirner, F., Götz, J., Rüde, U., Schröder-Turk, G. E., Mecke, K., and Bechinger, C. (2012). Permeability of Porous Materials Determined from the Euler Characteristic. *Phys. Rev. Lett.*, 109(26):264504. Publisher: American Physical Society.

Smith, A. D., Dłotko, P., and Zavala, V. M. (2021). Topological data analysis: Concepts, computation, and applications in chemical engineering. *Computers & Chemical Engineering*, 146:107202.

References IV

Thomas, A. M. and Owada, T. (2021). Functional limit theorems for the euler characteristic process in the critical regime. *Advances in Applied Probability*, 53(1):57–80. Publisher: Cambridge University Press.

Trinh, K. D. (2017). A remark on the convergence of Betti numbers in the thermodynamic regime. *Pacific Journal of Mathematics for Industry*, 9(1):4.

Trinh, K. D. (2019). On central limit theorems in stochastic geometry for add-one cost stabilizing functionals. *Electronic Communications in Probability*, 24(none):1–15. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Virk, Z. (2022). *Introduction to Persistent Homology*. Založba UL FRI, Ljubljana.

Vishwanath, S., Fukumizu, K., Kuriki, S., and Sriperumbudur, B. (2020). Statistical Invariance of Betti Numbers in the Thermodynamic Regime. arXiv:2001.00220 [math, stat].

Wasserman, L. (2018). Topological Data Analysis. *Annual Review of Statistics and Its Application*, 5(1):501–532.

Yen, P. T.-W. and Cheong, S. A. (2021). Using Topological Data Analysis (TDA) and Persistent Homology to Analyze the Stock Markets in Singapore and Taiwan. *Frontiers in Physics*, 9.

Back-up slides

The Euler characteristic

Recall: The Euler characteristic $\chi(K)$ of a simplicial complex K is defined as

$$\boxed{\chi(K) = \sum_{k=0}^{\infty} (-1)^k n_k(K),}$$

where $n_k(K)$ denotes the number of k-simplices in K.

The Euler characteristic

Recall: The Euler characteristic $\chi(K)$ of a simplicial complex K is defined as

$$\boxed{\chi(K) = \sum_{k=0}^{\infty} (-1)^k n_k(K),}$$

where $n_k(K)$ denotes the number of k-simplices in K. In particular, for a polyhedron in $\mathbb{R}^d$,

$$\chi(P) = \#\{\text{vertices}\} - \#\{\text{edges}\} + \#\{\text{faces}\}.$$

The Euler characteristic

Recall: The Euler characteristic $\chi(K)$ of a simplicial complex K is defined as

$$\boxed{\chi(K) = \sum_{k=0}^{\infty} (-1)^k n_k(K),}$$

where $n_k(K)$ denotes the number of k-simplices in K. In particular, for a polyhedron in $\mathbb{R}^d$,

$$\chi(P) = \#\{\text{vertices}\} - \#\{\text{edges}\} + \#\{\text{faces}\}.$$

In general: $\dim(K) = d \quad \Rightarrow \quad$ sum in the definition of the Euler characteristic is finite.

The Euler characteristic

Recall: The Euler characteristic $\chi(K)$ of a simplicial complex K is defined as

$$\boxed{\chi(K) = \sum_{k=0}^{\infty} (-1)^k n_k(K),}$$

where $n_k(K)$ denotes the number of k-simplices in K. In particular, for a polyhedron in $\mathbb{R}^d$,

$$\chi(P) = \#\{\text{vertices}\} - \#\{\text{edges}\} + \#\{\text{faces}\}.$$

In general: $\dim(K) = d \quad \Rightarrow \quad$ sum in the definition of the Euler characteristic is finite.

Euler characteristic is a topological invariant.

The Euler characteristic

Recall: The Euler characteristic $\chi(K)$ of a simplicial complex K is defined as

$$\boxed{\chi(K) = \sum_{k=0}^{\infty} (-1)^k n_k(K),}$$

where $n_k(K)$ denotes the number of k-simplices in K. In particular, for a polyhedron in $\mathbb{R}^d$,

$$\chi(P) = \#\{\text{vertices}\} - \#\{\text{edges}\} + \#\{\text{faces}\}.$$

In general: $\dim(K) = d \quad \Rightarrow \quad$ sum in the definition of the Euler characteristic is finite.

Euler characteristic is a topological invariant.

Relation to Betti numbers:

$$\boxed{\chi(K) = \sum_{k=0}^{\infty} (-1)^k \beta_k(K).}$$

(This follows from observing that $\beta_k(K) = |S_k^+(K)| - |S_{k+1}^-(K)|$, with $S_k^{\pm}(K)$ positive/negative simplices in filtration.)

The Euler characteristic

Recall: The Euler characteristic $\chi(K)$ of a simplicial complex K is defined as

$$\boxed{\chi(K) = \sum_{k=0}^{\infty} (-1)^k n_k(K),}$$

where $n_k(K)$ denotes the number of k-simplices in K. In particular, for a polyhedron in $\mathbb{R}^d$,

$$\chi(P) = \#\{\text{vertices}\} - \#\{\text{edges}\} + \#\{\text{faces}\}.$$

In general: $\dim(K) = d \quad \Rightarrow \quad$ sum in the definition of the Euler characteristic is finite.

Euler characteristic is a topological invariant.

Relation to Betti numbers:

$$\boxed{\chi(K) = \sum_{k=0}^{\infty} (-1)^k \beta_k(K).}$$

(This follows from observing that $\beta_k(K) = |S_k^+(K)| - |S_{k+1}^-(K)|$, with $S_k^\pm(K)$ positive/negative simplices in filtration.)

Advantages: Easier to compute - no need to find persistence diagram, just count number of simplices.

Euler characteristic and the Euler characteristic curve

Applications:

Kerscher (2000); Kilner et al. (2005), Scholz et al. (2012); Khanamiri et al. (2018); Amézquita et al. (2020), Yen and Cheong (2021)

Various contexts: porous matter; astronomy; expected Euler characteristic heuristic; VR-complex and biological shapes;...

Theory and methodology related to our work:

Decreusefond et al. (2014), Bobrowski and Mukherjee (2015), Thomas and Owada (2021)

Weak convergence of the Euler characteristic process

Again, we consider

- VR and Čech filtrations
- $\mathbb{X}_n$ either Binomial or Poisson point process
- critical regime

Weak convergence of the Euler characteristic process

Again, we consider

- VR and Čech filtrations
- $\mathbb{X}_n$ either Binomial or Poisson point process
- critical regime

The Euler characteristic process: With $\mathcal{F}$ either VR or $\mathcal{C}$:

$$\boxed{\nu_n(t) = \frac{1}{\sqrt{n}} \left(\chi(\mathcal{F}_t(n^{\frac{1}{d}}\mathbb{X}_n)) - \mathbb{E}\big[\chi(\mathcal{F}_t(n^{\frac{1}{d}}\mathbb{X}_n))\big]\right), \quad t \in [0, T].}$$

Weak convergence of the Euler characteristic process

Theorem (Krebs & WP, 2021, Thomas and Owada, 2021)

Suppose that f can be approximated uniformly by a sequence of blocked density functions on $[0,1]^d$ of the form $f_m = \sum_{i=1}^{m^d} b_i \mathbb{1}_{B_i}$ with blocks B_i of the form $B_i = [a_1, b_1] \times \cdots \times [a_d, b_b]$ such that for all $i = 1, \ldots, d$ one has $c\frac{1}{m} \le b_i - a_i \le C\frac{1}{m}$ for some $c, C > 0$. Then, for a Poisson sampling scheme with intensity nf, as $n \to \infty$,

$$\boxed{\nu_n \Rightarrow G \qquad \text{weakly, in } D[0,T],}$$

where $D[0,T]$ denotes Skorohod space, and G is a mean zero Gaussian process on $[0,T]$ with covariance of the form

$$\mathrm{Cov}(G(s), G(t)) = \mathbb{E}\big[\gamma(f(X)^{\frac{1}{d}}(s,t))\big].$$

where $X \sim f$. A similar result holds for a Binomial sampling scheme (with density f), where the covariance function changes to

$$\mathrm{Cov}(G(s), G(t)) = \mathbb{E}\big[\gamma(f(X)^{\frac{1}{d}}(s,t))\big] - \mathbb{E}\big[\alpha(f(X)^{\frac{1}{d}}s)\big]\,\mathbb{E}\big[\alpha(f(X)^{\frac{1}{d}}t)\big]$$

for some functions $\alpha : \Delta \times \Delta \to \mathbb{R}$, $\gamma : \Delta \times \Delta \to [0,\infty)$.

Discussion of weak convergence of Euler characteristic process

Discussion of weak convergence of Euler characteristic process

- The form of the functions γ_k, σ_k and α_k is not well understood.

- These functions are determined by the behavior under a homogeneous Poisson sampling $\rightsquigarrow$ they don't depend on the sampling density;

- The dependence of the limit on the sampling density f is through quantities of the form

$$\mathbb{E}_f\left[\Psi_{k,s,t} f^{1/d}(X)\right].$$

This will be discussed further below.

- We do not have a process-level result for the persistence Betti numbers!

- As for persistent Betti numbers, process is centered by its expected value. Does the expected value converge to a limit?

Expected value of the Euler characteristic

Proposition (Bobrowski and Mukherjee, 2015; Thomas and Owada, 2021)

Let f be a density w.r.t. uniform measure on an m-dimensional manifold $\mathbb{X}$ embedded in $\mathbb{R}^d$, and assume f to be bounded. Furthermore, let $\mathcal{P}_n$ be a Poisson process with intensity nf. Then

(i)
$$\frac{1}{n}\mathbb{E}\chi(\mathcal{C}_t(n^{1/m}\mathcal{P}_n)) \to \sum_{k=1}^{m}(-1)^k c_k(t) \qquad \text{as } n \to \infty,$$

where the non-negative $c_k(t)$ are of the form $c_k(t) = \int_{\mathbb{X}} \Psi_{k,t}(f(x))\, f(x)\, dx$ for some functions $\Psi_{k,t}$, $k = 0, 1, \ldots, m$ not depending on f. A similar result holds for $\mathcal{P}_n$ replaced by a Binomial process $\mathbb{X}_n$ with density f.

(ii) *If f is a density on $\mathbb{R}^d$, then*

$$\frac{1}{n}\mathbb{E}\chi(\mathrm{VR}_t(n^{1/d}\mathcal{P}_n)) \to \sum_{k=1}^{\infty}(-1)^k d_k(t) \qquad \text{as } n \to \infty,$$

where $d_k(t)$ are of the form $d_k(t) = \int_{\mathbb{X}} \widetilde{\Psi}_{k,t}(f(x))\, f(x)\, dx$ for some functions $\widetilde{\Psi}_{k,t}$, $k = 0, 1, \ldots$ not depending on f.

Again, smooth bootstrap:

- $N \sim \mathrm{Poiss}(n)$
- $\mathbb{X}_n^* = (X_1^*, \ldots, X_n^*), X_i^* \sim_{iid} \widehat{f}_n$ (density estimator - not necessarily KDE)
- $\mathcal{P}^* = (X_1^*, \ldots, X_N^*)$

With $\mathcal{F}$ being either VR or $\mathcal{C}$, let

$$\mathcal{F}_{n,t}^* \text{ being either } \mathcal{F}_t(n^{1/d}(\mathbb{X}_n^*)) \text{ or } \mathcal{F}_t(n^{1/d}\mathcal{P}_n^*).$$

Define bootstrap version of Euler characteristic process:

$$\boxed{\nu_n^*(t) = \frac{1}{\sqrt{n}} \left(\chi(\mathcal{K}_{n,t}^*) - \mathbb{E}^*\left[\chi(\mathcal{K}_{n,t}^*)\right]\right), \quad t \in [0, T].}$$

Pointwise bootstrap for the Euler characteristic curve in the critical regime

First we discuss point-wise results:

First we discuss point-wise results:

Theorem (Krebs & WP, 2021)

Let f be a density on $[0,1]^d$ and let $(\hat{f}_n : n \in \mathbb{N})$ be a sequence of density estimators with the property that $\lim_{n\to\infty} \|\hat{f}_n - f\|_\infty = 0$ a.s. (in probability). Then, for ν_n^* based on bootstrap samples drawn from $\hat{f}_n$,

$$\|\hat{f}_n - f\|_\infty^{-1/2} \cdot \sup_{t\in[0,T]} W_1(\nu_n^*(t), \nu_n(t)) = O(1) \quad a.s.\ (in\ probability),$$

where W_1 is the 1-Wasserstein distance.

Furthermore, for each $t \in [0,T]$

$$\{\|\hat{f}_n - f\|_\infty^{1/2} + n^{-1/2}\}^{-1} \cdot d_K(\nu_n^*(t), \nu_n(t)) = O(1) \quad a.s.\ (in\ probability),$$

where d_K denotes Kolmogorov-distance.

E.g.: For $\widehat{f}_n = \widehat{f}_{n,h}$ a KDE with p-th order kernel:

$$\|\widehat{f}_{n,h} - f\|_\infty = O\Big(\Big(\frac{\log n}{n}\Big)^{p/(d+2p)}\Big) \quad \text{a.s.}$$

Bootstrap for the Euler characteristic process in the critical regime

Now we consider a process-level result. For simplicity, smooth bootstrap based on the KDE:

Bootstrap for the Euler characteristic process in the critical regime

Now we consider a process-level result. For simplicity, smooth bootstrap based on the KDE:

Theorem (Krebs & WP, 2021)

For either the VR or the Čech filtration constructed over either a (rescaled) Binomial process with density f on $[0,1]^d$, or a (rescaled) Poisson process with intensity nf. Suppose that for some $p > d$,

- *f is p times continuously differentiable;*
- *bootstrap samples are based on a KDE $\widehat{f}_{n,h}$, based on a p-th order kernel;*

Then, for any fixed $T > 0$,

$$\boxed{W_1^{D[0,T]}(\nu_n^*, \nu_n) = O\big((\log n)^{\alpha}\, n^{-\beta}\big),}$$

where $\frac{1}{3} < \alpha = \frac{2p}{4d+2p} < 1$ and $0 < \beta = \frac{3p}{4d+8p} - \frac{1}{4} < \frac{1}{8}$, and where $W_1^{D[0,T]}$ denotes the 1-Wasserstein distance on the Skorohod space $D[0,T]$.

Discussion of the results

Discussion of the results

- Above results imply convergences in law $\rightsquigarrow$ bootstrap works;

Discussion of the results

- Above results imply convergences in law $\rightsquigarrow$ bootstrap works;
- Above results come with rates of approximation;

Discussion of the results

- Above results imply convergences in law $\rightsquigarrow$ bootstrap works;
- Above results come with rates of approximation;
- Bootstrap results are based on bounds for

$$W_1(\nu_{n,f}, \nu_{n,g}) \qquad \text{and} \qquad d_K(\nu_{n,f}, \nu_{n,g})$$

in terms of distances between f and g, where $\nu_{n,f}$ and $\nu_{n,g}$ denote the Euler characteristic process for samples from f and g, respectively. (See below.)

A key technical result

A key technical result

In case of the Euler characteristic (with finite stabilization radius(!)), we have a stronger result:

Theorem (Approximating property in the Wasserstein-Kantorovich distance)

Let f be an essentially bounded density on $[0,1]^d$, $r \in \mathbb{R}_+$, and $g \in B_\infty(f,r)$. There are coupled Poisson processes $(\mathcal{P}_n, \mathcal{Q}_n)$ with intensities (nf, ng) and coupled binomial processes $(\mathbb{X}_n, \mathbb{Y}_n)$ with densities f, g, respectively, and a constant $C_{0,f} \in \mathbb{R}_+$ depending on f and r but not on g (as long as $g \in B_\infty(f,r)$), such that

$$\sup_n \sup_{t \in [0,T]} \mathrm{Var}(\nu_{f,n}(t) - \nu_{g,n}(t)) \leq C_{0,\kappa} \|f - g\|_\infty .$$

In particular,

$$\sup_n \sup_{t \in [0,T]} W_1(\nu_{f,n}(t), \nu_{g,n}(t)) \leq C_{0,\kappa} \|f - g\|_\infty^{1/2} .$$

Filter functions

Union of balls with midpoints at the data points $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ are sublevel set of the distance function $d_{\mathbb{X}_n}(x)$:

$$\bigcup_{i=1}^{n} \overline{B}(r; X_i) = \{d_{\mathbb{X}_n}(x) \leq r\}$$

where $\overline{B}(r; X_i)$ = closed ball of radius r and midpoint X_i, and

$$d_{\mathbb{X}_n}(x) = \inf\{\|x - X_i\|, \, i = 1, \ldots, n\}.$$

Filter functions

Union of balls with midpoints at the data points $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ are sublevel set of the distance function $d_{\mathbb{X}_n}(x)$:

$$\bigcup_{i=1}^{n} \overline{B}(r; X_i) = \left\{ d_{\mathbb{X}_n}(x) \leq r \right\}$$

where $\overline{B}(r; X_i)$ = closed ball of radius r and midpoint X_i, and

$$d_{\mathbb{X}_n}(x) = \inf \left\{ \|x - X_i\|, \, i = 1, \ldots, n \right\}.$$

More generally: If a persistence diagram is based on the sublevel set filtration of the form $\mathcal{F}_f = \{f^{-1}((-\infty, t]), t \in \mathbb{R}\}$, then we call f a filter function.