Imperial College
London

11 July 2022

An introduction to Topological Data Analysis

Tutorial for Aalto-Imperial-TUM

Yueqi Cao and Anna Song

1. Overview of persistent homology
2. Data representation and filtration
3. Persistence diagrams and computation
4. Beyond persistent homology

Overview of persistent homology

Geometry v.s. Topology

Geometry is concerned with distances and rigid transformations. However biological data rarely enjoy natural metrics, are often noisy, high-dimensional and with only few useful coordinates.

Topology is concerned with qualitative geometric information. It ignores geometric quantitative measurements but deals with neighborhoods and connectivity of the objects. In particular, the k-**th homology group** of a topological space X describes the number of k-dim holes in X, such as connected components, loops, and cavities.

But often the data is a point cloud. How to build an interesting topological space on them to study its homology?

Simplicial complex

A k-**simplex** generalizes the notion of triangle to arbitrary dims: it is the k-dim convex hull of $k+1$ affinely-independent vertices, e.g. point, segment, triangle, tetrahedron in 3D.

A **simplicial complex** K is a set of simplices that satisfies:

- Every face σ of a simplex from K is also in K. E.g. any edge of a triangle in K is also in K.
- The non-empty intersection $\sigma_1 \cap \sigma_2 \neq \emptyset$ of any two simplices in K is a face of both simplices. E.g. two intersecting triangles necessarily share an edge or a vertex.

Examples

Imperial College London

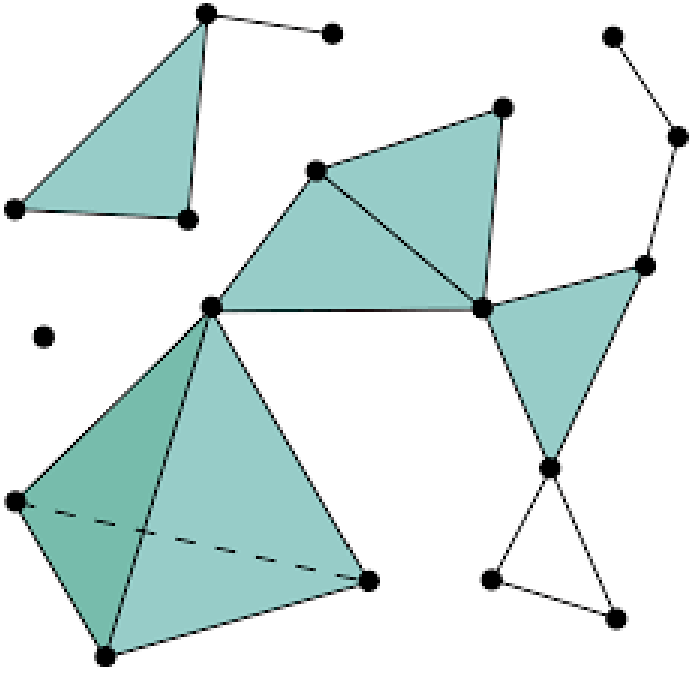


Figure 1: A simplicial complex.

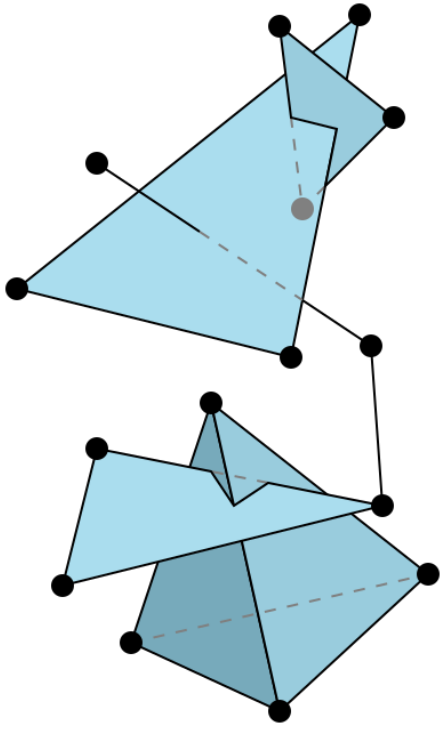


Figure 2: Not valid.

Homology of a simplicial complex

Imperial College London

Let K be a simplicial complex and the field $F = \mathbb{Z}/2\mathbb{Z}$.
$C_p(K)$ is the F-vector space generated by the p-simplices of K. An element of $C_p(K)$ is a p-**chain** $c = \sum_{i \in I} \sigma_i$. The **boundary** of c is $\partial(c) = \sum \partial(\sigma_i)$ where $\partial(\sigma)$ is the sum of the $(p-1)$-dim faces of a p-dim simplex (w.r.t. addition mod 2, $1 + 1 = 0$).

Z_p is the set of p-**cycles**, i.e. p-chains with zero boundary. B_p is the set of p-**boundaries**, i.e. the boundary of $(p+1)$-chains. Because $\partial \circ \partial = 0$, we have $B_p \subseteq Z_p$.

Homology of a simplicial complex

Then, the quotient group

$$H_p(K) = Z_p(K)/B_p(K)$$

is called the p-**th simplicial homology group** of K with $\mathbb{Z}/2\mathbb{Z}$ coefficients. The p-**th Betti number** of K is $\beta_p(K) = \dim(H_p(K))$.

Informally, a **homology** p-**cycle** (element of $H_p(K)$) can be represented as a p-cycle up to p-boundaries of $(p+1)$ chains. E.g., a 1-cycle up to the boundaries of a sum of triangles. Note: the sum of two loops can represent a homology cycle.

Examples

Imperial College London

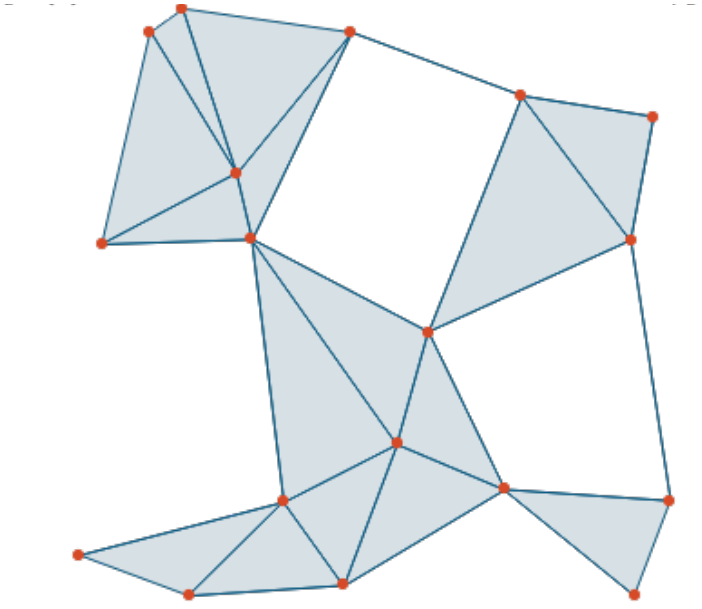


Figure 3: A simplicial (alpha) complex, with $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 0$.

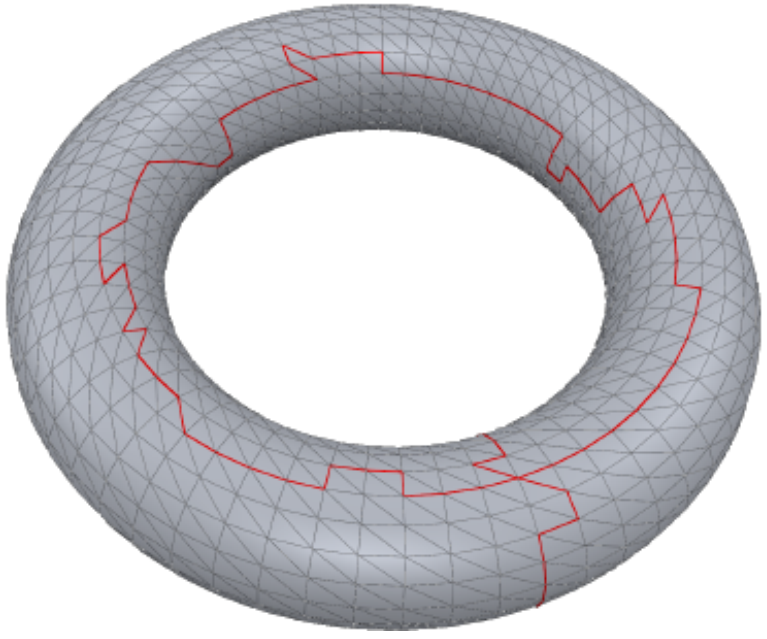


Figure 4: Triangulated torus with $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$. Showing two generators of $H_1(K)$.

Persistent homology

Imperial College London

The main idea of **persistent homology** in TDA is to study qualitative features that persist across multiple scales of analysis, e.g. in a nested family of simplicial complexes $K_1 \subseteq ... \subseteq K_n$. Instead of focusing on one particular scale to cluster points or connect regions in the space (and how to choose the scale?), one is interested in how the homology "evolves" when varying some parameter.

The resulting **persistence module** can be uniquely summarized with **persistence barcodes**, thanks to the Structure Theorem (explained later) from algebraic topology.

PH of data: what does it mean?

Imperial College London

Data can be **continuous** or **discrete** ; **images** or **point clouds**. Computing PH is a way to interpret the underlying phenomenon that produced such data. The choice of data **representation** and **filtration** is sometimes crucial to revealing different aspects of the phenomenon. Examples:

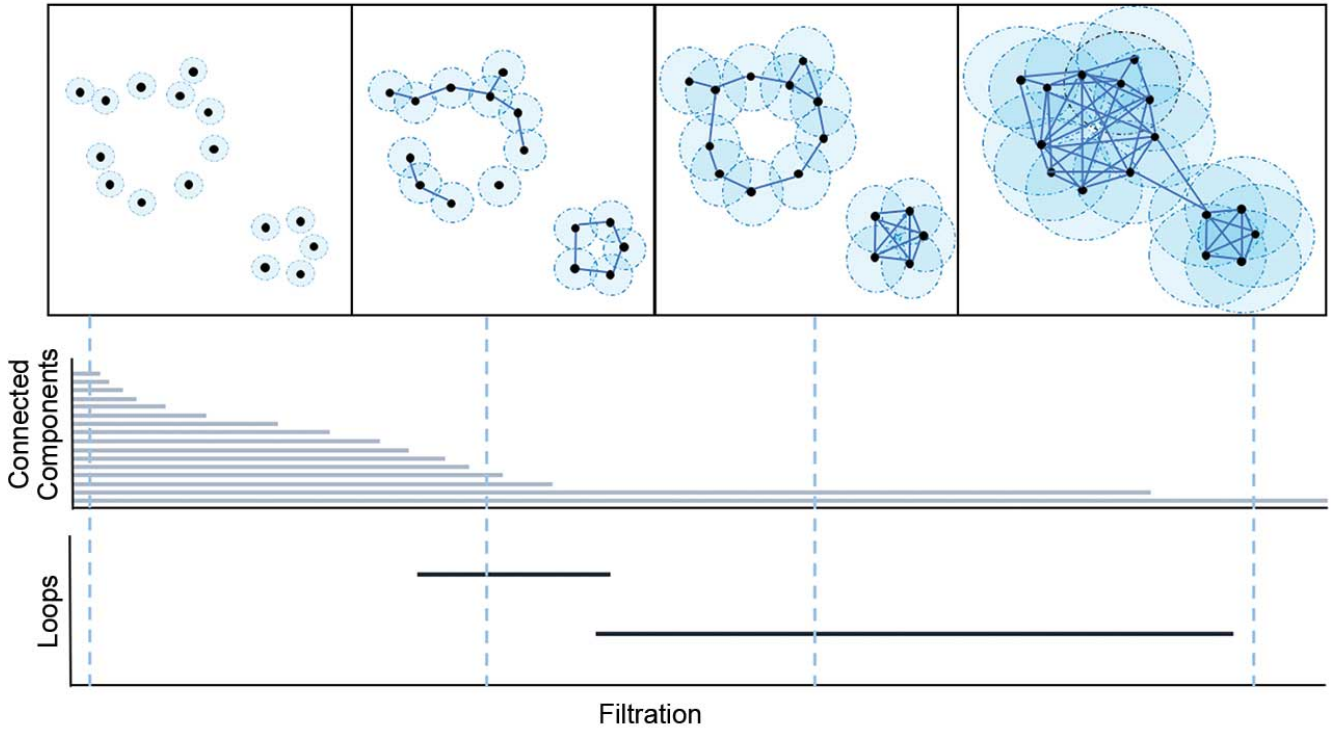
- A space can be filtered by the **sublevel sets** $f^{-1}(-\infty, t]$ of a function, e.g. curvature on a manifold, intensity in images, density of points. Approximations: sample values on a point cloud, graph vertices, or a grid, giving e.g. the **clique or cubical complexes**.
- Given a point cloud, one can build a simplicial complex filtered by a parameter of proximity $\epsilon > 0$. There are many definitions, e.g. **Rips, Čech, alpha complexes**.

Birth and death

Imperial College London

Intuitively, barcodes encode **births** and **deaths** of homology cycles.

- A new homology cycle is 'born' when it is formed;
- A homology cycle is 'dead' when it merges into an older one, in particular, if it becomes a boundary. Fig from [LR21].



Birth and death

Imperial College London

Formally, consider a sequence of subcomplexes $K_0 \hookrightarrow \cdots \hookrightarrow K_n$.
Let $[c] \in H(K_i)$ be a homology class (dimension omitted).

- $[c]$ is born at i if it is not in the image of $H(K_{i-1}) \to H(K_i)$
- $[c]$ is dead at j if
  - the image of $[c]$ via $H(K_i) \to H(K_{j-1})$ is not in $\mathrm{Im}(H(K_{i-1}) \to H(K_{j-1}))$
  - but the image of $[c]$ via $H(K_i) \to H(K_j)$ is in $\mathrm{Im}(H(K_{i-1}) \to H(K_j))$.

Meaning: what "remains" of $[c]$ at $j-1$ is still different from what remains of the classes present before $i-1$; however at j it can no more be distinguished from them.

Persistence modules

Imperial College London

Question

Why does the geometric filtration finally produce a set of barcodes?

Consider a finite filtration of simplicial complexes:

$$K_0 \hookrightarrow K_1 \hookrightarrow K_2 \hookrightarrow \cdots \hookrightarrow K_n$$

Apply the homology functor $H(\cdot, F)$ (F is a field!):

$$H(K_0, F) \xrightarrow{\phi_0} H(K_1, F) \xrightarrow{\phi_1} H(K_2, F) \xrightarrow{\phi_2} \cdots \xrightarrow{\phi_{n-1}} H(K_n, F)$$

The collection of homology vector spaces $H(K_i, F)$, together with vector space homomorphisms $\phi_i : H(K_i, F) \to H(K_{i+1}, F)$, is called a *Persistence Module*.

Persistence modules

Imperial College London

Define a $F[t]$-module as follows:

$$\mathcal{H}(K, F) = \bigoplus_{i=0}^{\infty} H(K_i, F)$$

where $H(K_i, F) = H(K_n, F)$ for $i \geq n$.
The action of t is given by

$$t \cdot (x^0, x^1, x^2, \cdots) = (0, \phi_0(x^0), \phi_1(x^1), \phi_2(x^2), \cdots)$$

$\mathcal{H}(K, F)$ is a graded module over a graded ring.

Structure theorem

Imperial College London

Recall that for a finitely generated module M over a principal ideal domain R, we have the following structure theorem (wiki page)

$$M \cong \left(\bigoplus_{i=1}^{\alpha} R \right) \oplus \left(\bigoplus_{j=1}^{\beta} R/(r_j) \right)$$

- F is a field so $F[t]$ is P.I.D.;
- Graded ideals of $F[t]$ are homogeneous of form (t^k);
- $\mathcal{H}$ is finitely generated;

Structure theorem for persistence modules

$$\mathcal{H}(K, F) \cong \left(\bigoplus_{i=1}^{\alpha} \mathscr{P}^{a_i} F[t] \right) \oplus \left(\bigoplus_{j=1}^{\beta} \mathscr{P}^{b_j} F[t]/(t^{c_j}) \right)$$

where $\mathscr{P}$ is shifting operator for gradings.

Structure theorem

Imperial College London

The invariants a_i, b_j, c_j are stored as sets of intervals, known as *barcodes*:

$$\{(a_i, \infty), \quad (b_j, b_j + c_j), \quad i = 1, \cdots, \alpha, j = 1, \cdots, \beta\}$$

Generalizations of the structure theorem:

- Persistence modules of general index sets [Cha+16];
- Structure of multidimensional persistence modules [CZ09];
- Categorification of persistent homology [BS14];
- and more...

References

Overview

[Car09] Gunnar Carlsson. 'Topology and data'. In: *Bulletin of the American Mathematical Society* 46.2 (2009), pp. 255–308.

[EH08] Herbert Edelsbrunner and John Harer. 'Persistent homology-a survey'. In: *Contemporary mathematics* 453 (2008), pp. 257–282.

[LR21] Nicole Lazar and Hyunnam Ryu. 'The Shape of Things: Topological Data Analysis'. In: *CHANCE* 34.2 (2021), pp. 59–64.

[Mun17] Elizabeth Munch. 'A user's guide to topological data analysis'. In: *Journal of Learning Analytics* 4.2 (2017), pp. 47–61.

References

Imperial College London

Theory

[BS14] Peter Bubenik and Jonathan A Scott. 'Categorification of persistent homology'. In: *Discrete & Computational Geometry* 51.3 (2014), pp. 600–627.

[Cha+16] Frédéric Chazal et al. *The structure and stability of persistence modules*. Springer, 2016.

[CZ09] Gunnar Carlsson and Afra Zomorodian. 'The theory of multidimensional persistence'. In: *Discrete & Computational Geometry* 42.1 (2009), pp. 71–93.

[ZC05] Afra Zomorodian and Gunnar Carlsson. 'Computing persistent homology'. In: *Discrete & Computational Geometry* 33.2 (2005), pp. 249–274.

Data representation and filtration

Čech complexes

Imperial College London

Let $X = \{x_0, \cdots, x_n\}$ be a finite set in the Euclidean space. Let $\epsilon > 0$ be a given parameter. The Čech complex is defined by

- The vertex set (0-simplices) is X;
- A subset $\{x_0, \cdots, x_k\} \subseteq X$ spans a k-simplex iff the intersection of balls $\cap_{i=0}^{k} B(x_i, \epsilon)$ is not empty.

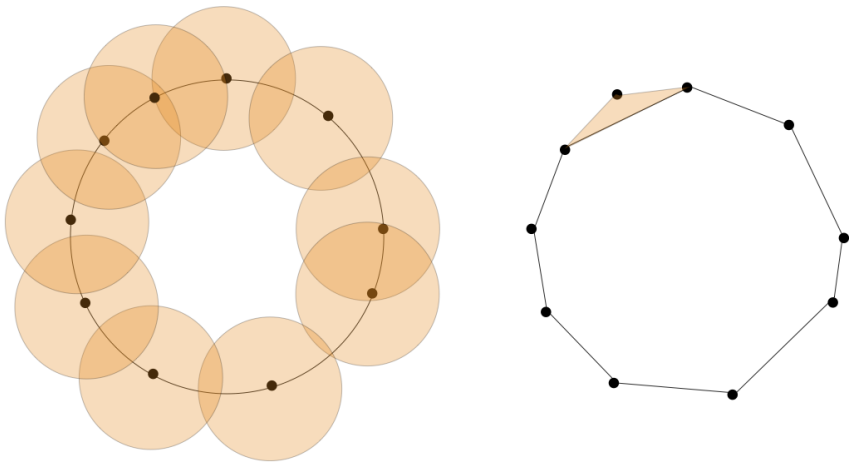


Figure 5: Image from Wikipedia

Rips complexes

Imperial College London

Let (X, d) be a finite metric space, and $\epsilon > 0$ be a given threshold. The Vietoris-Rips (VR) complex is an *abstract simplicial complex* defined according to the following rules:

- The vertex set (0-simplices) is X;
- A subset $\{x_0, \cdots, x_k\} \subseteq X$ spans a k-simplex iff $d(x_i, x_j) \leq \epsilon$, $\forall i, j \in \{0, \cdots, k\}$. i.e. The diameter $\text{diam}(\{x_0, \cdots, x_k\}) \leq \epsilon$.

We denote the VR complex by $\mathrm{VR}_\epsilon(X, d)$.

Remark:

- VR complex is a 'lazy' version of Čech: no need to check dimension by dimension;
- Some people prefer $\mathrm{VR}_\epsilon(X, d)$ to indicate that a k-simplex is spanned iff the diameter is no less than 2ϵ, so that the function sending every finite metric space to its persistence diagram has unit Lipschitz norm [Cha+09].

Rips complexes

Imperial College London

Nice things about VR complexes:

- $\mathrm{VR}_\epsilon(X, d) \subseteq \mathrm{VR}_{\epsilon'}(X, d)$ if $\epsilon \leq \epsilon'$;
- A general construction for many types of data;
- Clean stability results [Cha+09].

Not nice...

- Explicit computation (homotopy type, homology groups...) is difficult, even for a circle [AA17];
- The size of a VR complex explodes easily.

Parameters to control the size of a VR filtration:

- `max_edge_length`: set the maximal filtration threshold;
- `max_dimension`: set the dimension of a VR complex;
- `sparse`: set the sparsification ratio [She13].

Clique complexes

Imperial College London

Let $G = (V, E)$ be a graph. A clique is a complete subgraph of G. The collection of all cliques of G is called the clique complex of G. Equivalently, the clique complex $\mathrm{C}(G)$ of G can be constructed according to the following rules:

- The vertex set is V;
- A subset $\{v_0, \cdots, v_k\} \subseteq V$ spans a k-simplex iff v_i, v_j are connected by an element in E for all $i, j \in \{0, \cdots, k\}$.

Remark:

- The size of a clique complex can also explode if the graph is not sparse;
- A VR complex is the clique complex of its 1-skeleton;
- VR complexes and clique complexes are *flag complexes*, where the existence of a face is completely determined by its edges.

Clique complexes

Let $f : V \to \mathbb{R}$ be a function on the graph. Let $a_0 < a_1 < \cdots < a_n$. The sequence of sublevels

$$f(-\infty, a_0] \subseteq f(-\infty, a_1] \subseteq \cdots \subseteq f(-\infty, a_n]$$

induces sequence of subgraphs

$$G_0 \subseteq G_1 \subseteq \cdots \subseteq G_n$$

which yields the sequence of clique complexes

$$\mathrm{C}(G_0) \subseteq \mathrm{C}(G_1) \subseteq \cdots \subseteq \mathrm{C}(G_n)$$

Example: Take f to be the degree map.

Alpha complexes

Imperial College London

Let (X, d) be a finite metric space and $\epsilon > 0$ a parameter. The union of balls $\bigcup_{x \in X} B(x, \epsilon)$ can be decomposed using the intersections of each ball with the Voronoi cell containing x, namely

$$\bigsqcup_{x \in X} R(x, \epsilon)$$

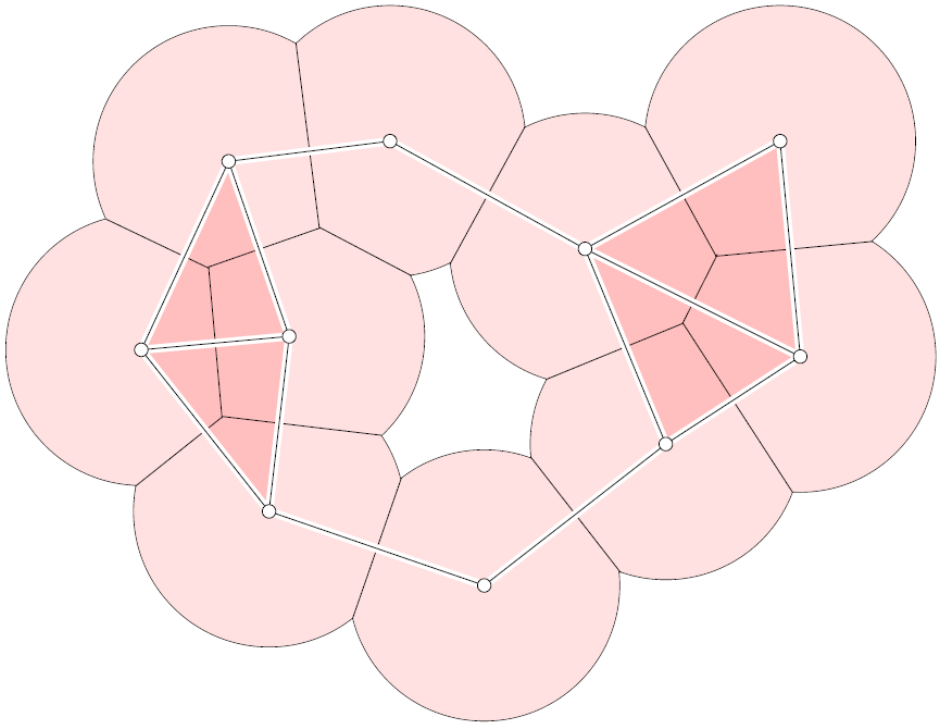
where $R(x, \epsilon) = B(x, \epsilon) \cap V(x)$. The **alpha complex** is then [Ede95]

$$\text{Alpha}_\epsilon(X, d) = \{\sigma \subseteq X \mid \bigcap_{x \in \sigma} R(x, \epsilon) \neq \emptyset\}$$

Alpha complexes

Imperial College London

It is a subcomplex of the Delaunay complex as $R(x, \epsilon) \subseteq V(x)$.
Furthermore $\text{Alpha}_\epsilon(X, d) \subseteq \check{\text{C}}\text{ech}_\epsilon(X, d)$.
Since the $R(x, \epsilon)$ are closed, convex and together cover the union, the Nerve Theorem implies that $\text{Alpha}_\epsilon(X, d)$ is homotopy equivalent to the union of the balls.



Cubical complexes

A p-dim **cube** in $\mathbb{R}^d$ is a product of p non-degenerate intervals of the form $[k, k+1]$ and $d-p$ degenerate intervals $[k, k]$.

Examples: points, edges, pixels, voxels in 3D.

Similarly to simplicial complexes, a p-dim **cubical complex** K collects cubes of dim at most p and is closed by taking faces and intersections [WCV12].

Images

Rectangular cubical complexes are the natural representation of **images**. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function. Suppose it is sampled at the centers of maximal cubes (pixels or voxels for $d = 2$ or 3).

Then the filtration value assigned to any lower-dim cube is the minimal value of all maximal cubes containing it.

Example: an edge is assigned $\min f(v)$ among all voxels v containing it.

References

Imperial College London

Complex

[AA17] Michał Adamaszek and Henry Adams. 'The Vietoris–Rips complexes of a circle'. In: *Pacific Journal of Mathematics* 290.1 (2017), pp. 1–40.

[Ede95] H. Edelsbrunner. 'The union of balls and its dual shape'. In: *Discrete & Computational Geometry* 13.3 (1995), pp. 415–440.

[She13] Donald R Sheehy. 'Linear-size approximations to the Vietoris–Rips filtration'. In: *Discrete & Computational Geometry* 49.4 (2013), pp. 778–796.

[WCV12] Hubert Wagner, Chao Chen and Erald Vuçini. 'Efficient Computation of Persistent Homology for Cubical Data'. In: *Topological Methods in Data Analysis and Visualization II: Theory, Algorithms, and Applications*. 2012, pp. 91–106.

References

Imperial College London

Stability

[CEH07] David Cohen-Steiner, Herbert Edelsbrunner and John Harer. 'Stability of persistence diagrams'. In: *Discrete & computational geometry* 37.1 (2007), pp. 103–120.

[Cha+09] Frédéric Chazal et al. 'Gromov-Hausdorff stable signatures for shapes using persistence'. In: *Computer Graphics Forum*. Vol. 28. 5. Wiley Online Library. 2009, pp. 1393–1403.

[Coh+10] David Cohen-Steiner et al. 'Lipschitz functions have L_p-stable persistence'. In: *Foundations of computational mathematics* 10.2 (2010), pp. 127–139.

Persistence diagrams and computation

Boundary matrix

Imperial College London

Consider a filtered simplicial complex K. Order its simplices $\sigma_1, ..., \sigma_n$ by increasing filtration value.

Build the **boundary matrix** D as follows. Set $D[i,j] = 1$ if σ_i appears in the boundary $\partial(\sigma_j) = \sum_k \sigma_{i_k}$, $D[i,j] = 0$ otherwise.

The **lowest** index of column $D[:,j]$ is the largest non-zero row index and denoted by low(j). In other words, it indexes the younger (i.e. last appearing) simplex that intervenes in the boundary of σ_j.

Naive algorithm

Imperial College London

Computing the persistent homology of K consists in **reducing** D. We perform a Gaussian elimination $\mod 2$ of the columns from left to right.

$R \leftarrow D$
for $j = 1, ..., n$ **do**
  while $\exists j' < j$ such that $\text{low}(j') = \text{low}(j) \neq 0$ **do**
    $R[:, j] \leftarrow R[:, j] + R[:, j'] \mod 2$

After addition, the new column gives the boundary of $\sigma_j + \sigma_{j'_1} + ... + \sigma_{j'_l}$. Zero columns are boundaries of (created) cycles.

Reduction

Imperial College London

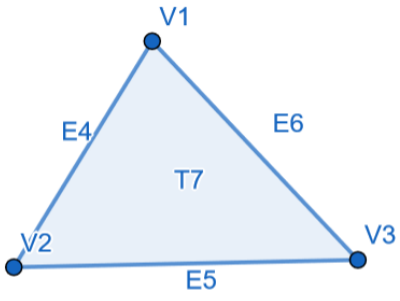


Figure 7

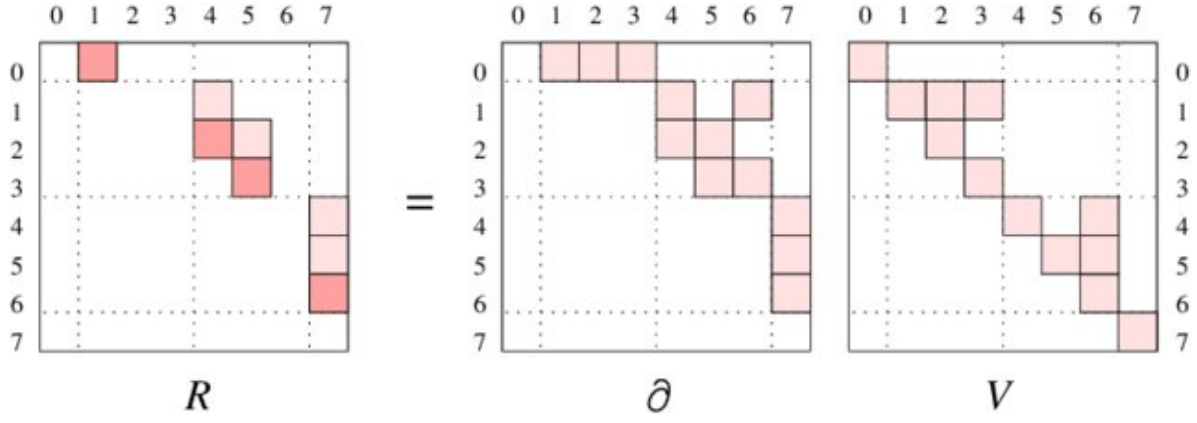


Figure 8: Reduced matrix for a filled triangle. Fig from [EH09].

Pairing simplices

Imperial College London

In the reduced matrix R:

- If $\text{low}(j) = i > 0$, then σ_j is a **negative simplex** that kills a $(p-1)$-cycle created by the positive simplex σ_i. Pairing them leads to the interval $[i, j)$.
- Otherwise $R[:, j] = 0$ and σ_j is a **positive simplex** that creates a p-cycle:
  - either σ_j is paired to some σ_k with $\text{low}(k) = j$, leading to $[j, k)$.
  - or σ_j creates an **essential** p-cycle involved in the homology group of K_n. We get $[j, +\infty)$.

One can read inside a non-zero column $R[:, j]$ one representative $(p-1)$-cycle of the homology cycle created at i and destroyed at j. It is the boundary of the sum of p-simplices induced by column operations.

Improved algorithms

Imperial College London

The naive algorithm is actually time- and memory-consuming. One avoids keeping the whole matrix, so with gudhi and giotto-tda one cannot retrieve representative p-cycles of persistence bars (not the case for Eirene.jl or Ripserer.jl).

Much time can be saved by not reducing columns i that become trivial afterwards and do not give rise to essential cycles of the form $[i, +\infty)$ but are paired with some j.

Instead, we rely on a much faster algorithm applying to cohomology, with fewer row operations, and combined to a "clearing" technique that appropriately ignores redundant reductions.

Main packages

Imperial College London

- Gudhi : more in the spirit of mathematics [The15]
- giotto-tda : machine learning spirit, compatible with scikit-learn [Tau+21]
- Ripser : ultra-fast computation of Rips persistence
- Eirene.jl and Ripserer.jl : possibility to retrieve geometric representatives from persistence bars

And many other packages.

Interpretation of the diagram

Imperial College London

Points close to the diagonal are generally attributed to **noise**, whereas those far from it are usually attributed to **persistent** features (we'll play with shapes and diagrams in the Jupyter notebook).

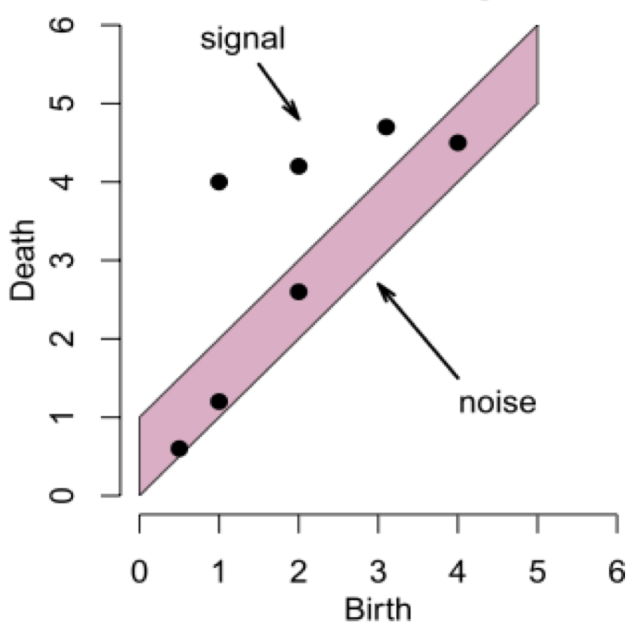


Figure 9: Image from GUDHI

Warning 1

Imperial College London

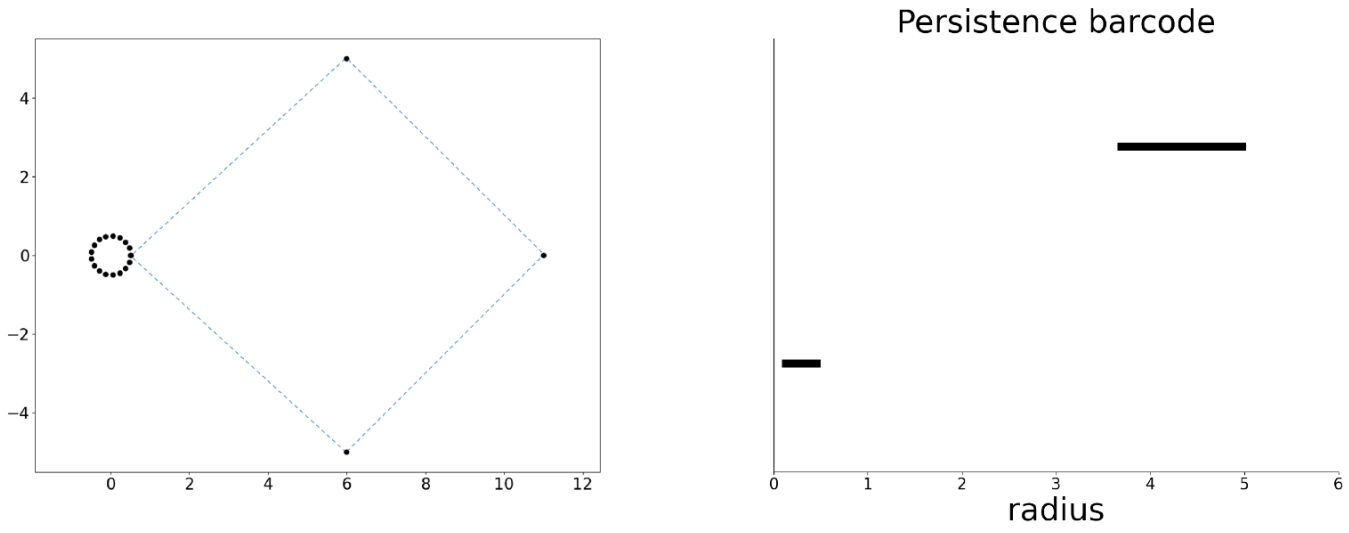
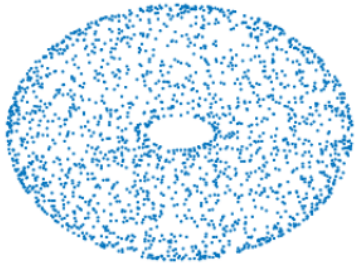


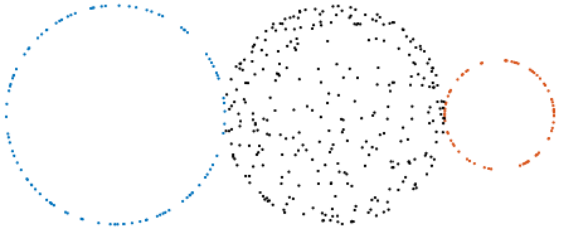
Figure 10: The smaller circle produces a less persistent lifetime than the 3 isolated points. Persistence measures a specific type of significance. Figure from [RB21]

Warning 2

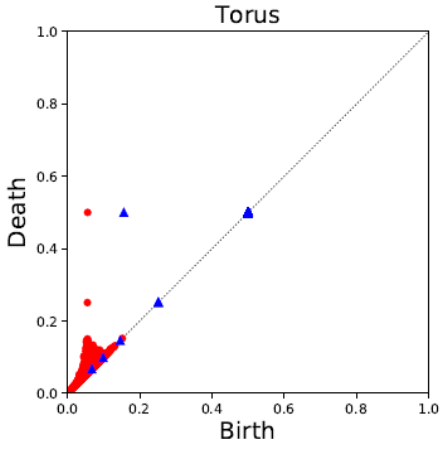
Imperial College London



(a)



(b)



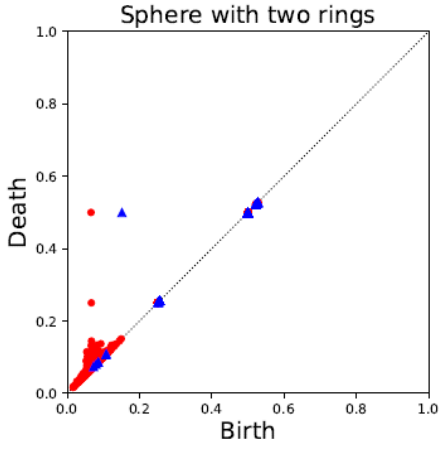


Figure 11: Holes and bubbles (PH1 and PH2). These two point clouds have quite the same diagram!! Figure from [RB21]

Wasserstein distance

Imperial College London

Let $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ be two persistence diagrams. How to measure the difference between X and Y?

Idea: Let x^* be the projection of x to the diagonal. Then consider $X' = X \cup \{y_1^*, \cdots, y_m^*\}$ and $Y' = Y \cup \{x_1^*, \cdots, x_n^*\}$. Range all matchings $x \longleftrightarrow y$, the difference is set to be the minimum of all the costs $\sum \|x - y\|$.

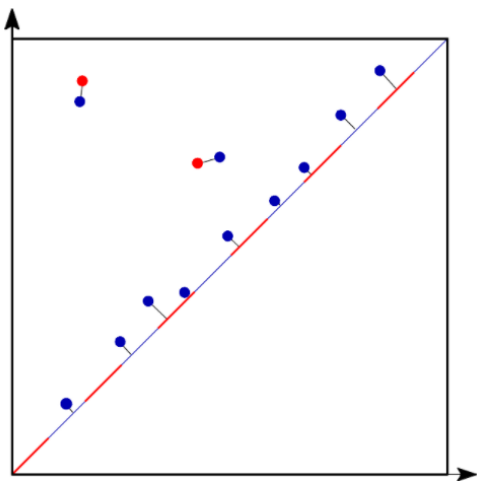


Figure 12: Image from GUDHI user manual

Wasserstein distance

Imperial College London

Formally, let d be any q-norm on $\mathbb{R}^2$ (often $q = \infty$). Let $1 \leq p < \infty$. The *p-Wasserstein distance* between X and Y is defined as

$$\mathrm{W}_p(X, Y) = \left(\inf_{\gamma} \sum_{x \in X'} d(x, \gamma(x))^p \right)^{\frac{1}{p}}$$

where γ ranges all bijections from X' and Y'.
For $p = \infty$, the *Bottleneck distance* is defined as

$$\mathrm{W}_\infty(X, Y) = \inf_{\gamma} \sup_{x \in X'} \{d(x, \gamma(x))\}$$

Stability

Bottleneck stability

Let $\mathcal{M}$ be a triangulable space and $f, g : \mathcal{M} \to \mathbb{R}$ be two tame functions. Let $D(\cdot)$ be the persistence diagram w.r.t levelset filtration. Then $\mathrm{W}_\infty(D(f), D(g)) \leq \|f - g\|_\infty$.

Let F and G be two point clouds, and $f = d(\cdot, F)$ and $g = d(\cdot, G)$ be the distance functions, where $d = \|\cdot\|_\infty$.

- The levelset filtration is 'equal' to the VR filtration;
- By triangle inequality $\|f - g\|_\infty \leq \mathrm{H}(F, G)$ where H is the Hausdorff distance;

Vectorizations

Imperial College London

Persistence diagrams are nonlinear and difficult for statistical purposes.

Want to find feature maps sending PDs to vectors – vectorizations.

In `giotto-tda` we have

- Persistence landscapes;
- Betti curves;
- Persistence images;
- Persistence entropy;
- ...

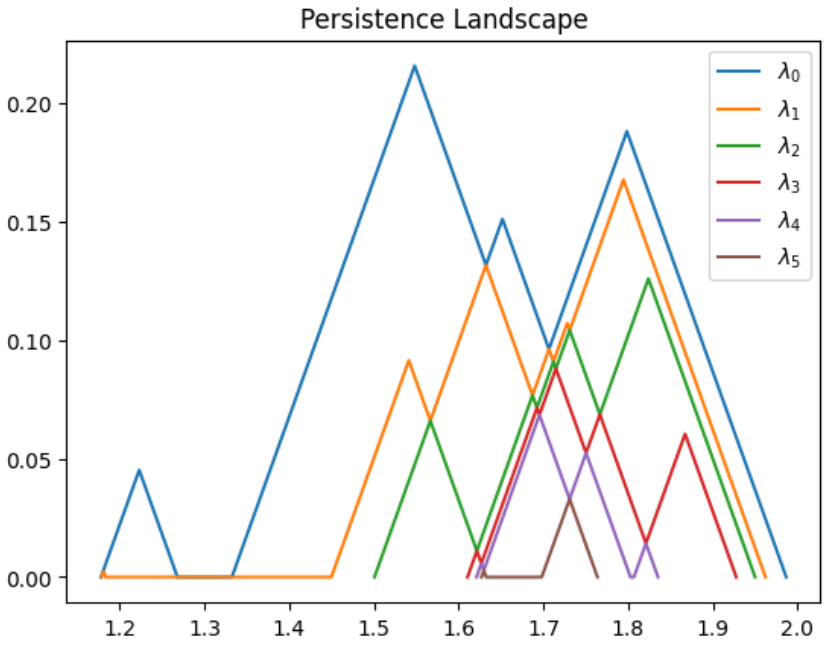
Persistence landscapes

Imperial College London

Let $\{b_i, d_i\}_{i \in I}$ be a persistence diagram. For each barcode define a function $\Lambda_i(t)$ as

$$\Lambda_i(t) = [\min\{t - b_i, d_i - t\}]_+$$

The k-th persistence landscape is a function $\lambda_k(t)$ at each t taking the k-th largest value of $\{\Lambda_i(t)\}_{i \in I}$.



References

Imperial College London

Algorithm

[EH09] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. Vol. 69. Miscellaneous Books. American Mathematical Society, Dec. 2009.

[Tau+21] Guillaume Tauzin et al. 'giotto-tda:: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration.'. In: *Journal of Machine Learning Research* 22 (2021), pp. 1–6.

[The15] The GUDHI Project. *GUDHI User and Reference Manual*. GUDHI Editorial Board, 2015.

Significance

[RB21] Yohai Reani and Omer Bobrowski. *Cycle Registration in Persistent Homology with Applications in Topological Bootstrap*. 2021. arXiv: 2101.00698 [cs.LG].

Beyond persistent homology

Reeb graphs

Consider X a topological space and $f : X \to \mathbb{R}$ a continuous function. Define an equivalence relation $\sim$ on X where $x \sim y$ iff x and y belong to the same connected component of a level set $f^{-1}(t)$ where $t \in \mathbb{R}$.

The **Reeb graph** is the quotient space $X/\sim$ endowed with the quotient topology. In other words, you "slice" the space like a bread and then record the connectivity of the slices into a graph.

Reeb graphs

Imperial College London

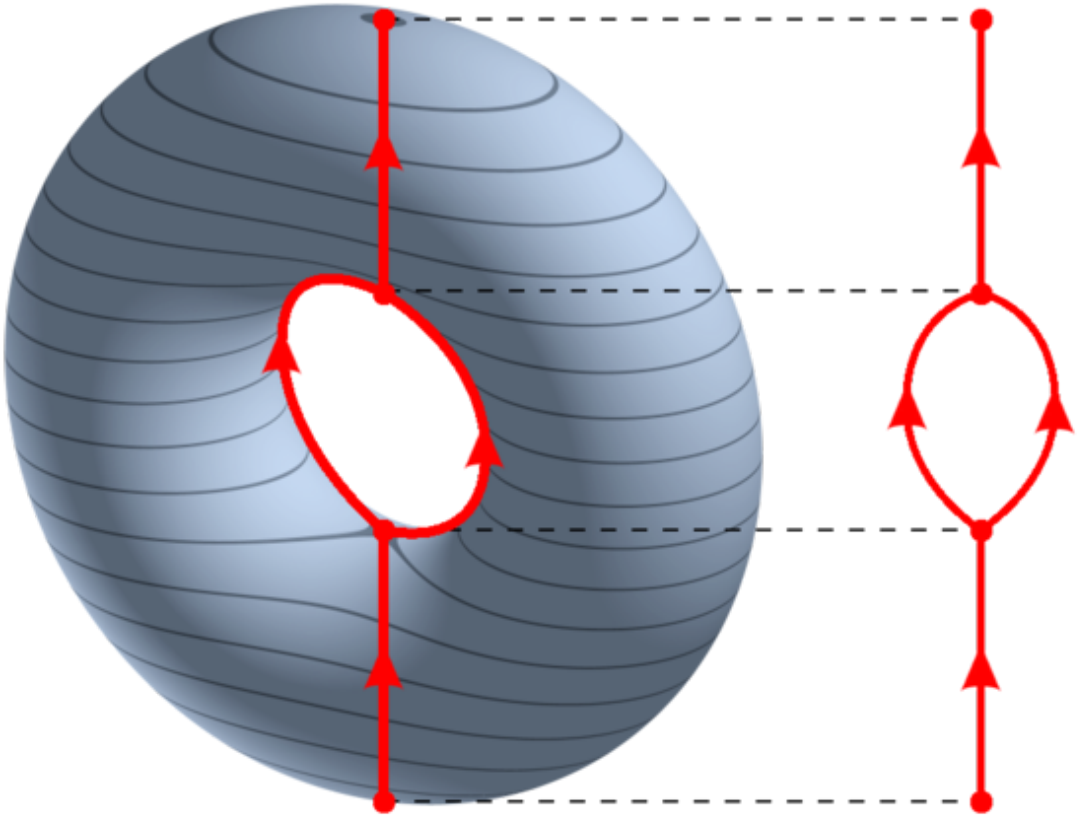


Figure 13: Reeb graph of the torus

Mapper

Imperial College London

Mapper is a very popular method that extracts the topology of data as a graph and constitutes a "pixelized" version of Reeb graphs [CO18]. Mapper is powerful to interpret point clouds in biological high-dim data [NLC11]. They are easy to compute and are implemented in giotto-tda.

Parameters:

- filter function f
- cover the range of f with overlapping intervals
- neighborhood size to determine "connected" points

Everything is about choosing them wisely, to reveal different properties of the data.

Mapper

Imperial College London

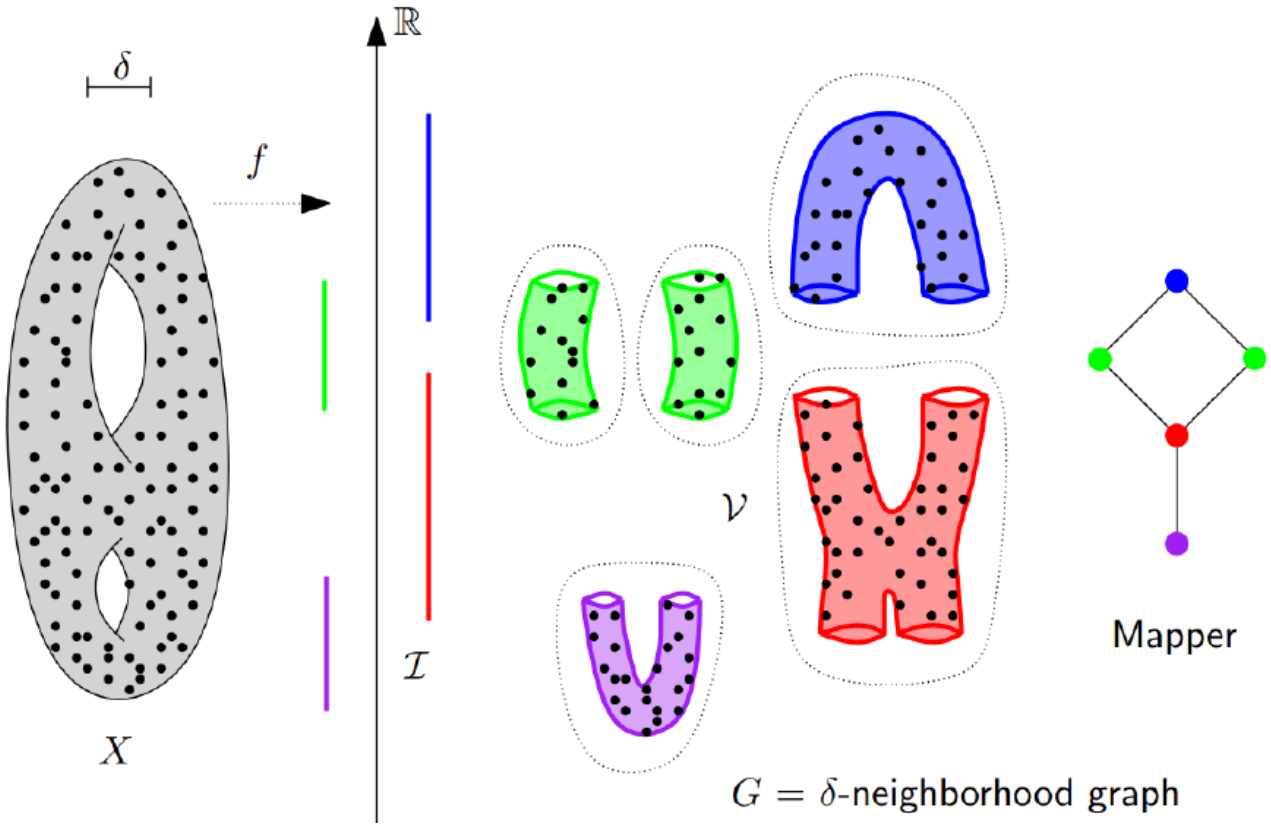


Figure 14: Building Mapper graph of a bitorus

Mapper

Imperial College London

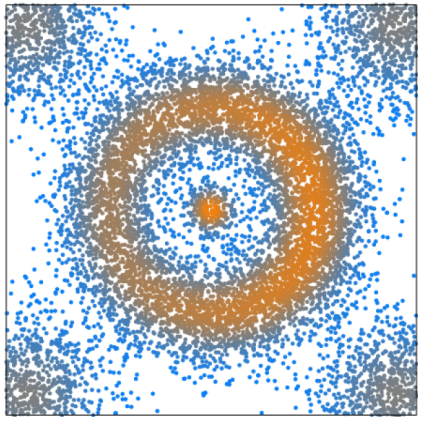


Figure 15: Sampling a known distribution.

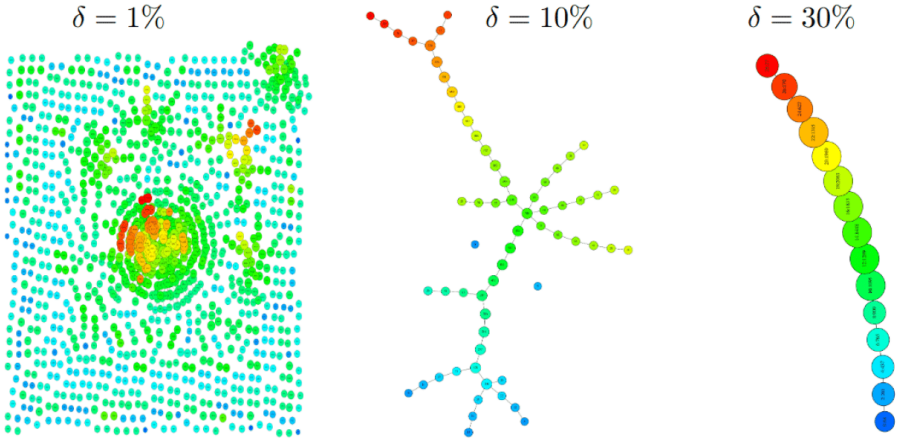


Figure 16: Mapper graphs for different parameters. f = density.

Mapper in biology

Imperial College London

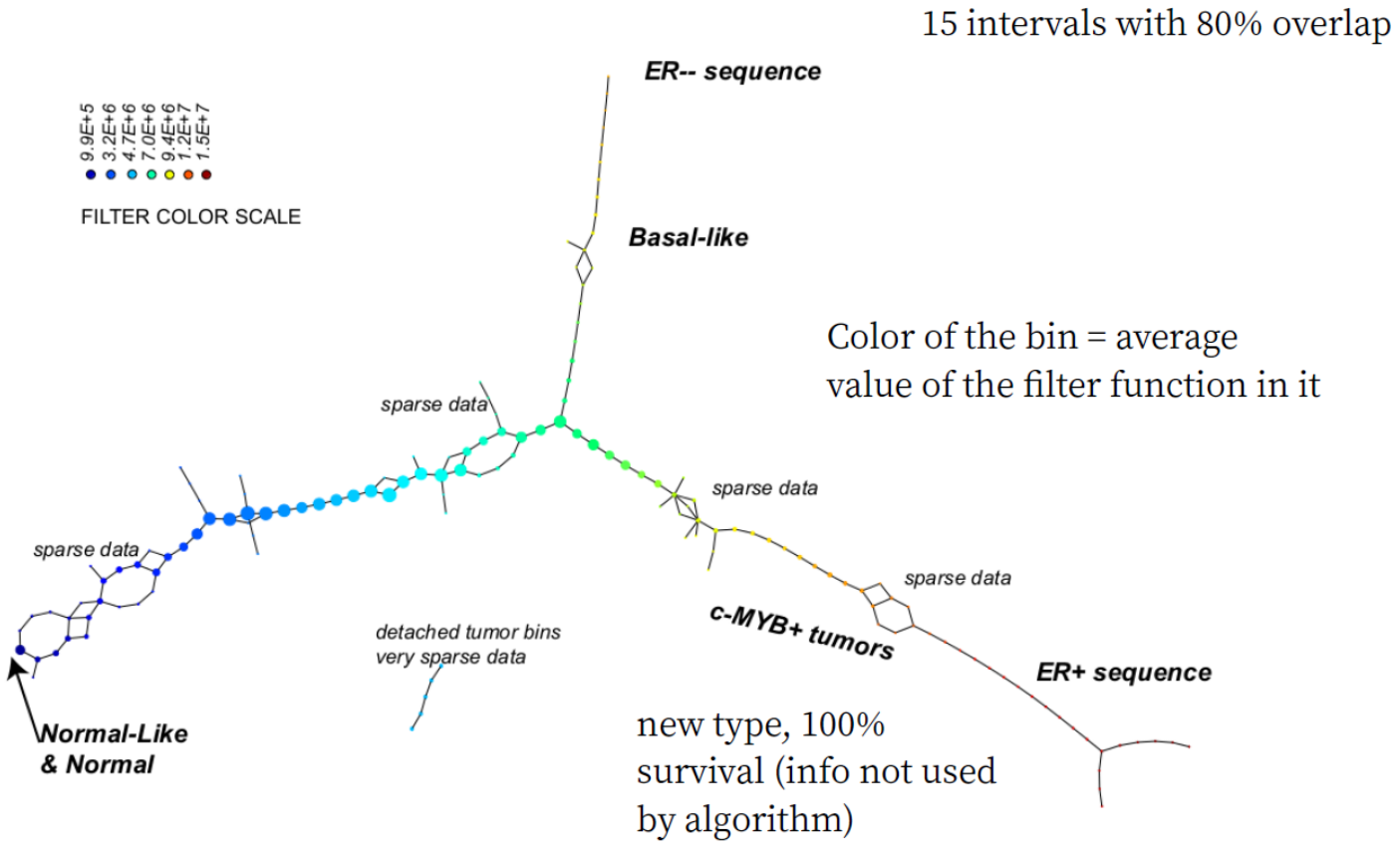


Figure 17: *Mathematical discovery* of a *biologically meaningful* subgroup (*ER*+) of breast cancers with Mapper! [NLC11]

UMAP

Imperial College London

UMAP, introduced in [MHM20], is an extremely popular tool among cell biologists for visualizing point cloud data spanned in a high-dim space, e.g. single-cell data.

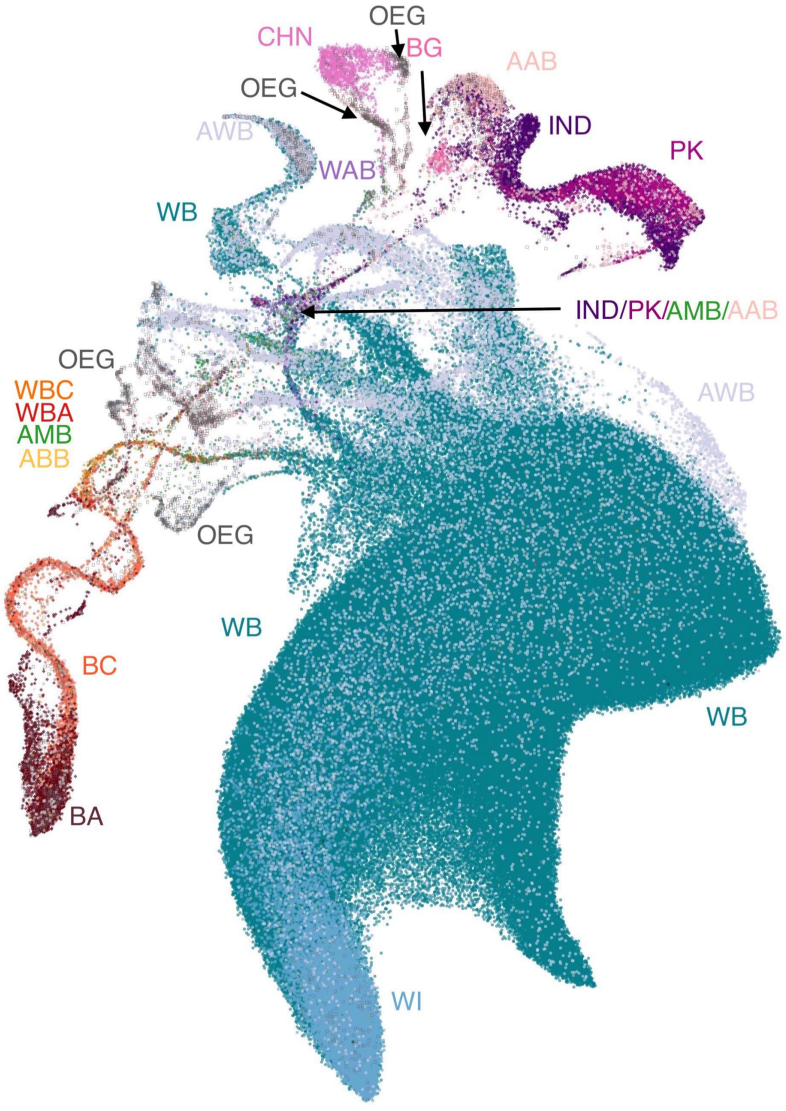
UMAP was theoretically inspired from TDA ("fuzzy simplicial sets"), but in practice the algorithm is part of the class of k-neighbour based graph learning algorithms, e.g. Laplacian Eigenmaps, Isomap and t-SNE (also used in biology).

The idea is to find a 2D representation of the point cloud by optimizing some objective function that preserves the topological structure of the k-neighbor graph.

UMAP

Imperial College London

From [Dia+19].



Multidimensional persistence

Imperial College London

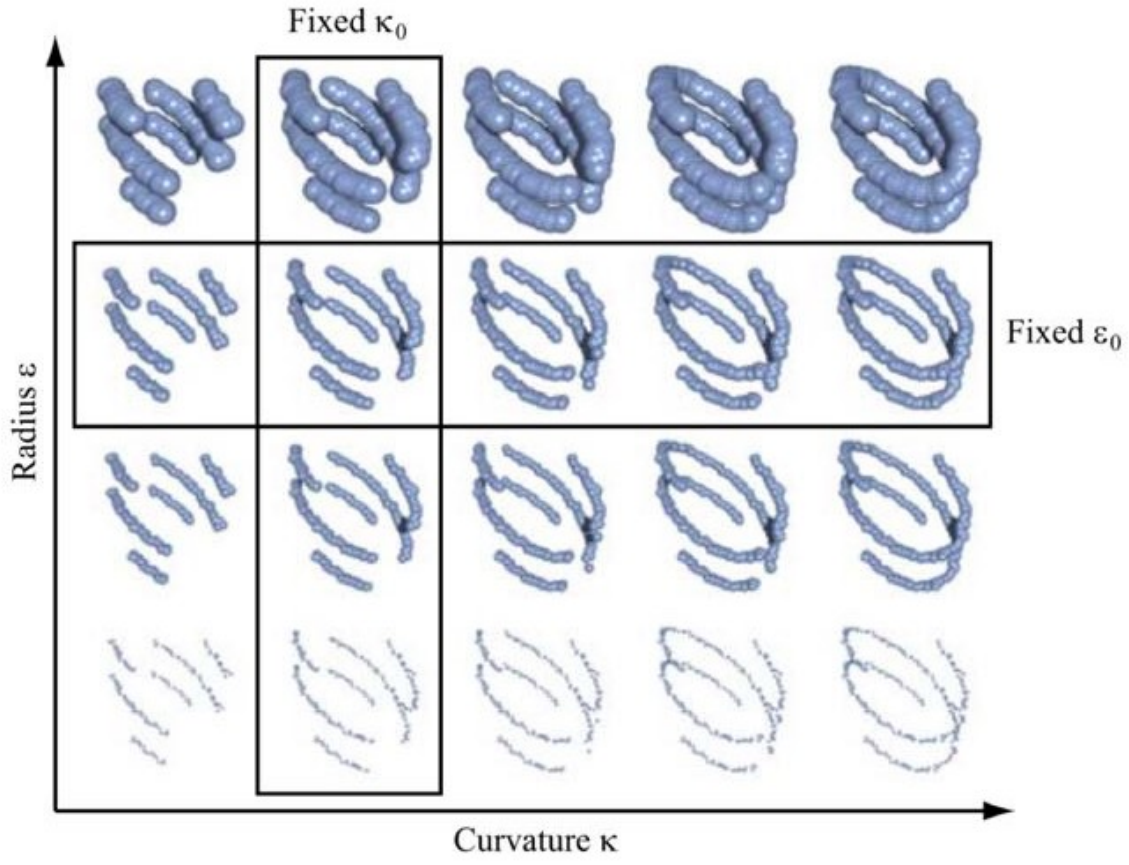


Figure 18: Example of a bifiltration. Image taken from [CZ09].

Multidimensional persistence

Imperial College London

- For $u, v \in \mathbb{N}^n$, say $u \prec v$ if $u_i \leq v_i$ for $i = 1, \cdots, n$;
- A multifiltration of simplicial complexes is a family $\{K_u\}_{u \in \mathbb{N}^n}$ such that $K_u \subseteq K_v$ if $u \prec v$;
- Similar to the one-dimensional case, we can define the persistence module as the collection of homology vector spaces $H(K_u, F)$ and homomorphisms $\phi_{u \to v} : H(K_u, F) \to H(K_v, F)$, and assign a true $F[t_1, \cdots, t_n]$-module structure to

$$\mathcal{H}(K, F) = \bigoplus_{u \in \mathbb{N}^n} H(K_u, F)$$

Multidimensional persistence

Imperial College London

- Unlike the one-dimensional case, the structure of multidimensional persistence module is complicated;
- *No* complete discrete invariant exists for multidimensional persistence module;
- The rank invariant is a function $\rho_{\mathcal{H}} : \mathbb{N}^n \times \mathbb{N}^n \to \mathbb{N}$ sending (u, v) to the rank of the map $\phi_{u \to v}$;
- Software for visualizing the rank invariant – RIVET [LW15].

Zigzag persistence

Normal persistence requires a nested family of subspaces

$$K_0 \hookrightarrow K_1 \hookrightarrow K_2 \hookrightarrow \cdots \hookrightarrow K_n$$

Question

What if $K_i \not\subseteq K_j$?

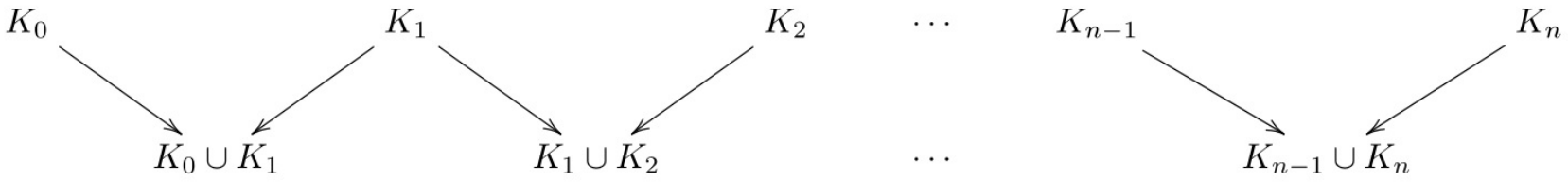
Example 1: $K_1, \cdots, K_n$ are different point clouds sampled from the same space $\mathcal{X}$.

Example 2: $K_1, \cdots, K_n$ are small random patches from a large image $\mathcal{X}$.

Zigzag persistence

Imperial College London

The idea of zigzag persistence is to add intermediate objects to connect K_i in the following way:



On homology level we have

$$H(K_0, F) \to H(K_0 \cup K_1, F) \leftarrow H(K_1, F) \to \cdots \leftarrow H(K_n, F)$$

Zigzag persistence

Imperial College London

- The structure of zigzag module is related to the representation of A_n-type quiver;
- A zigzag module is completely determined by its barcodes;
- Software for computing zigzag persistence – Javaplex [TVA14] and Dionysus [Mor].

References

Imperial College London

Reeb graphs and Mapper

[CO18] Mathieu Carrière and Steve Oudot. 'Structure and Stability of the One-Dimensional Mapper'. en. In: *Foundations of Computational Mathematics* 18.6 (Dec. 2018). Number: 6, pp. 1333–1396.

[NLC11] Monica Nicolau, Arnold J. Levine and Gunnar Carlsson. 'Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival'. In: *Proceedings of the National Academy of Sciences* 108.17 (2011), pp. 7265–7270.

References

Imperial College London

UMAP

[Dia+19] Alex Diaz-Papkovich et al. 'UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts'. In: *PLOS Genetics* 15.11 (Nov. 2019). Publisher: Public Library of Science, pp. 1–24.

[MHM20] Leland McInnes, John Healy and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020.

References

Imperial College London

Generalized persistence

[CZ09] Gunnar Carlsson and Afra Zomorodian. 'The theory of multidimensional persistence'. In: *Discrete & Computational Geometry* 42.1 (2009), pp. 71–93.

[LW15] Michael Lesnick and Matthew Wright. 'Interactive visualization of 2-D persistence modules'. In: *arXiv preprint arXiv:1512.00180* (2015).

[Mor] Dmitriy Morozov. In: URL: https://mrzv.org/software/dionysus/.

[TVA14] Andrew Tausz, Mikael Vejdemo-Johansson and Henry Adams. 'JavaPlex: A research software package for persistent (co)homology'. In: *Proceedings of ICMS 2014*. Ed. by Han Hong and Chee Yap. Lecture Notes in Computer Science 8592. 2014, pp. 129–136.