Multiparameter Persistent Homology

AMAT 840

Instructor: Michael Lesnick

https://www.albany.edu/~ML644186/

This class is about TDA, and in particular, multiparameter persistent homology (MPH).

- very active research area,
- rich theory,
- great practical promise.

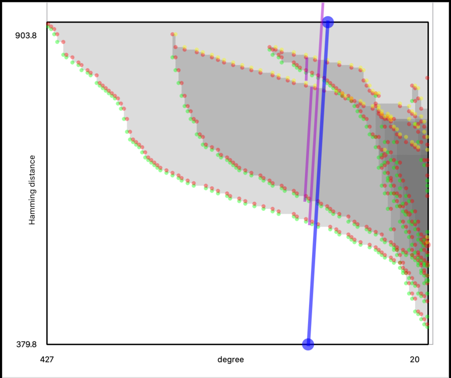
MPH arises naturally in applications:

- Noisy point cloud data
- Time-varying data
- Data equipped with an $\mathbb{R}$-valued function.

Yields richer but more complex invariants of data

- 1-parameter persistence theory / methodology doesn't extend naively,
- New ideas are needed.

Visualization of cluster structure in HIV genomic data using MPH:



Today:

- review of course logistics
- introductions
- intro to TDA and persistence

Course website:

- Just Google "UAlbany 840 2022"
- Beware: Website from 2019 looks similar!

This is the first course in a two semester sequence.

- This semester: August 22 - Dec. 5,
- Next Semester: TBD (late Jan - early May).

Course will be taught in hybrid format:

- live lectures,
- also broadcast on Zoom + recorded,
- UAlbany students are expected to come to the live lectures.

Office hours (tentative)

- M-W 4:30-5:30 (in person),
- T-Th 9:00-10:00 (Zoom only),
- By appointment.

Main reference is my course notes

- Will be updated throughout course,
- Suggestions/corrections welcome.

Prereqs:

- Topology: Topological spaces, homotopy equivalence, simplicial and singular homology,
- Abstract algebra: groups, rings,
- Solid understanding of linear algebra.

Homework

- Assigned semi-regularly (mostly theoretical stuff),
- I will try to grade it, provide solutions,
- Likely: One expository assignment on applications of persistence.

Grading (for UAlbany Students):

- Homework,
- Attendance/Participation,
- Midterm/Final

Regular attendance suffices to get a B.

Asynchronous UAlbany students:

- must take both exams,
- show evidence of effort / engagement.

Students outside of Albany:

- encouraged to participate actively (office hours, Discord).
- welcome to submit homework, take exams.

Finally, course feedback is welcome, by email or in real-time.

An introduction to TDA and (multiparameter) persistence

Topological Data Analysis (TDA)

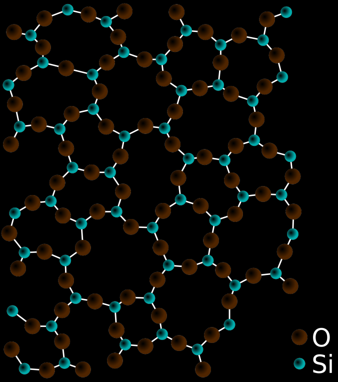
TDA is a branch of data science which uses topology to study the shape of data.

Types of data:

1. Point clouds, i.e., finite subsets of $\mathbb{R}^n$.
2. More generally, finite metric spaces.
3. Functions $f : T \to \mathbb{R}$, where T is a topological space.

Example of Low-Dimensional Point Cloud Data

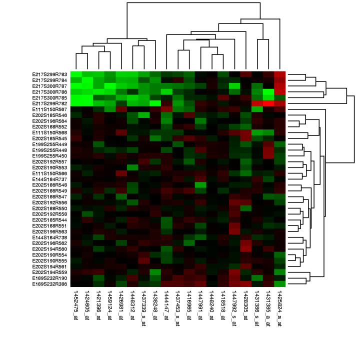
The atom centers of material (like a glass) or a biomolecule form a point cloud in $\mathbb{R}^3$:



Example of High-Dimensional Point Cloud Data

Gene expression data:

- Suppose we record the level of expression of each of 1500 genes in 300 breast cancer tumor samples, using RNA sequencing.
- This gives us a cloud of 300 points in $\mathbb{R}^{1500}$.



Example of Non-Euclidean Metric Data

The genome of an RNA virus is represented as a sequence of the letters A,U,C,G.

$$\begin{array}{cccccc} G & A & U & C & C & C \\ G & U & C & U & C & \end{array}$$

- We can view a set of genomes as metric space with the edit distance
- this is the minimum number of insertions, deletions, and replacements of a single letter needed to transform one sequence into the other.

The edit distance between the above sequences is 2.

$$\begin{array}{cccccc} G & A & U & C & C & C \\ G & - & U & C & U & C \end{array}$$

Examples of Functional Data

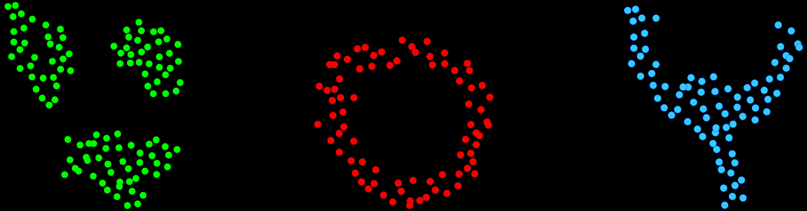
Greyscale image: T a rectangle, $f : T \to \mathbb{R}$ the pixel intensity.

fMRI image (at a fixed point in time): T the brain, $f : T \to \mathbb{R}$ measures oxygen level.

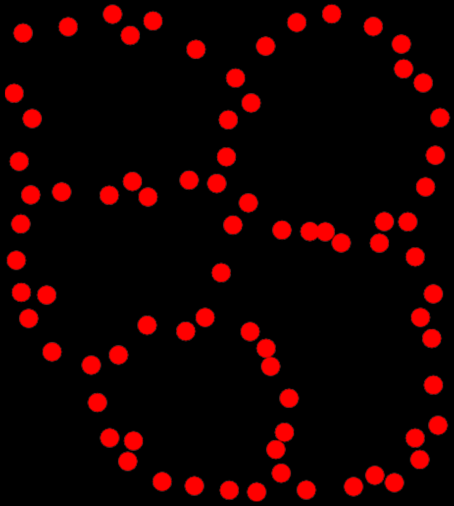
Shape of Data

Informally, shape of data =
coarse-scale, global, non-linear geometric features.

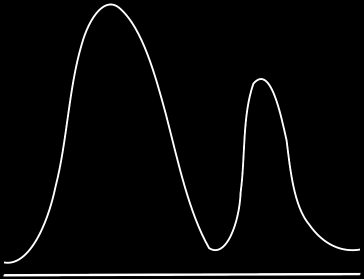
E.g., clusters, loops, and tendrils in point cloud data.

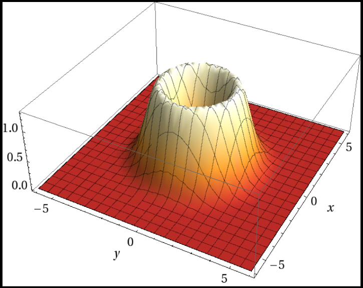


"Graph Structure"



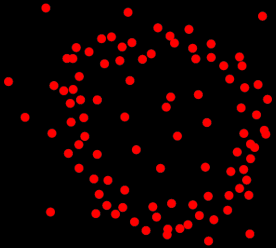
Shape features of functions: modes and ridges

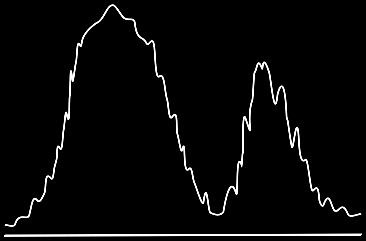




Computed by Wolfram Alpha

Noisy Shape Features





In TDA, we seek to develop:

- Formal definitions of such features
- Computational tools for detecting, visualizing such features
- Methodology for quantifying the statistical significance of such features.
- Applications.

The basic TDA pipeline: Given a data set X, we

1. Construct a diagram of topological spaces $F(X)$.
2. Analyze topological structure of $F(X)$ with classical tools.

$$\begin{tikzcd}
\vdots & \vdots & \vdots & \\
F_{1,3} \arrow[u, hook] \arrow[r, hook] & F_{2,3} \arrow[u, hook] \arrow[r, hook] & F_{3,3} \arrow[u, hook] \arrow[r, hook] & \cdots \\
F_{1,2} \arrow[u, hook] \arrow[r, hook] & F_{2,2} \arrow[u, hook] \arrow[r, hook] & F_{3,2} \arrow[u, hook] \arrow[r, hook] & \cdots \\
F_{1,1} \arrow[u, hook] \arrow[r, hook] & F_{2,1} \arrow[u, hook] \arrow[r, hook] & F_{3,1} \arrow[u, hook] \arrow[r, hook] & \cdots
\end{tikzcd}$$

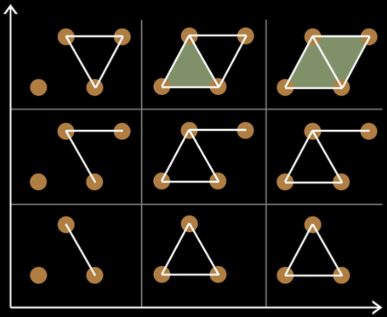


fig: Wright 2015

Each map is assumed to be an inclusion.

1-parameter persistent homology

Persistent Homology

- Provides invariants of data called barcodes
- Barcode is a collection of intervals $[b, d)$ in $\mathbb{R}$
- Each interval represents a geometric feature of the data
- Interval length is a measure of size

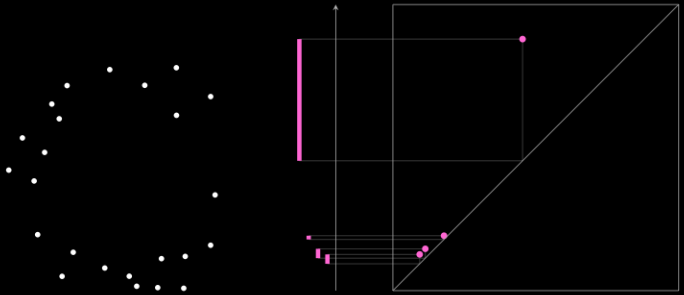
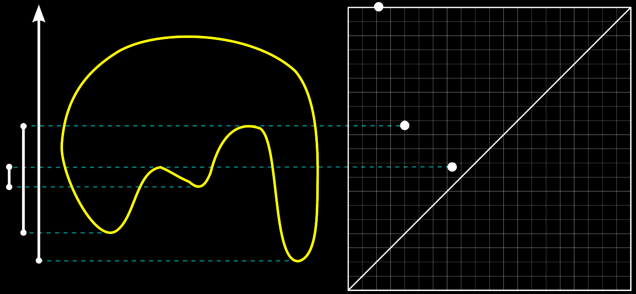


fig: Matthew Wright

Barcodes of Functions



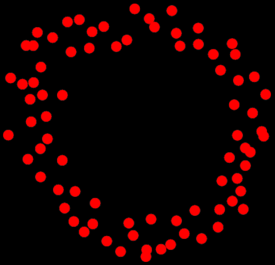
Barcodes of functions detect modes, and give information about the size of the modes.

They also detect higher order information.

Applications of Persistent Homology

- Shape/image classification
- Neuroscience: Representation of visual/spatial information in cortex
- Biophysics of proteins
- Atomic structure of glasses
- Virus evolution
- Coverage in sensor networks
- Detection of (near)-periodicity in gene expression data
- Clustering w/ theoretical guarantees

Model example:



Goal: Use homology to detect the loop.

- Fix a field K, say $K = \mathbb{Q}$ or $K = \mathbb{Z}/2\mathbb{Z}$.
- For each $i \in \mathbb{N}$ and topological space X, homology w/ K-coefficients gives a K-vector space $H_i X$.
- $\mathbf{dim}\, H_i X$ is the number of i-dimensional holes in X.

For $X \subset \mathbb{R}^n$ finite,

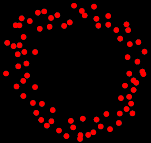
$$\mathbf{dim}\, H_0 X = |X|, \quad \mathbf{dim}\, H_i X = 0 \text{ for } i > 0,$$

so homology tells us nothing interesting.

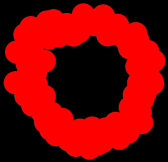
Naive Idea

For $X \subset \mathbb{R}^n$, let $O(X)_r$ be the r-offset of X.

r-offset = union of balls of radius r centered at points of X.



X



$O(X)_r$

$\mathbf{dim}\, H_1(O(X)_r) = 1$, which is the number of loops in X.

Counting loops via the map $X \mapsto \mathbf{dim}\, H_1(O(X)_r)$ is a rudimentary form of TDA.

Problems with this approach:

Counting loops via the map $X \mapsto \mathbf{dim}\, H_1(O(X)_r)$ is a rudimentary form of TDA.

Problems with this approach:

1. No canonical choice of r.

Counting loops via the map $X \mapsto \mathbf{dim}\, H_1(O(X)_r)$ is a rudimentary form of TDA.

Problems with this approach:

1. No canonical choice of r.
2. Invariant is unstable with respect to perturbation of data or small changes in r.

Counting loops via the map $X \mapsto \mathbf{dim}\, H_1(O(X)_r)$ is a rudimentary form of TDA.

Problems with this approach:

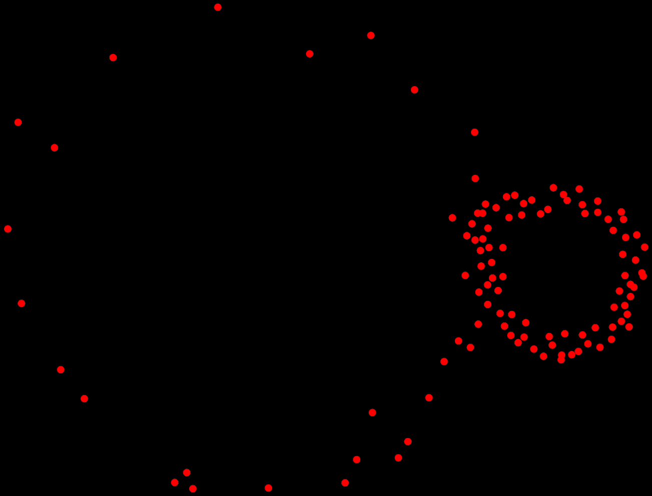
1. No canonical choice of r.
2. Invariant is unstable with respect to perturbation of data or small changes in r.
3. Doesn't distinguish small holes from big ones

Counting loops via the map $X \mapsto \mathbf{dim}\, H_1(O(X)_r)$ is a rudimentary form of TDA.

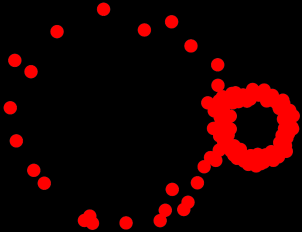
Problems with this approach:

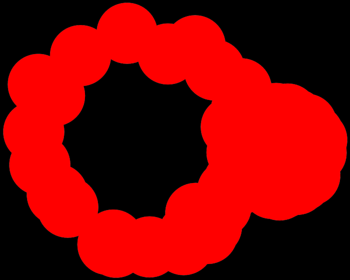
1. No canonical choice of r.
2. Invariant is unstable with respect to perturbation of data or small changes in r.
3. Doesn't distinguish small holes from big ones
4. Very sensitive to outliers.

Example: No Good Choice of r

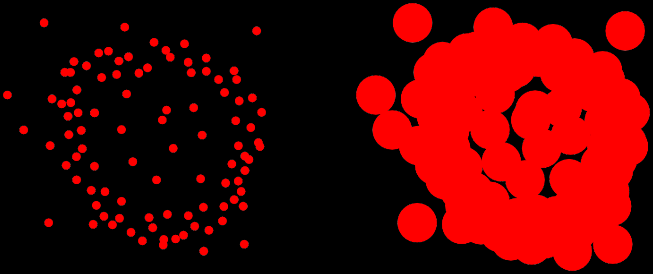


Example: No Good Choice of r





Example: Sensitivity to Outliers



$$B_1(U(X,r)) = 7;$$

Problems with this Descriptor

1. No canonical choice of r.
2. Invariant is unstable with respect to perturbation of data or small changes in r.
3. Doesn't distinguish small holes from big ones
4. Invariant is very sensitive to outliers.

Persistent homology provides a good solution to problems 1-3.

Multiparameter persistence provides a good solution to problem 4.

The 1-parameter family of spaces

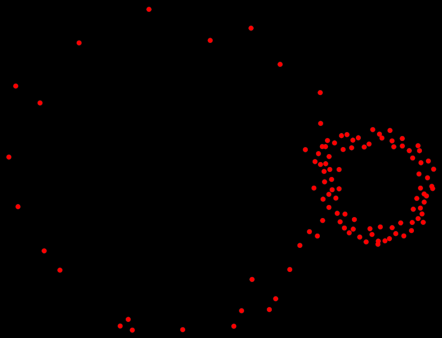
$$O(X) := (O(X)_r)_{r\in[0,\infty)}$$

is called the offset filtration of X.

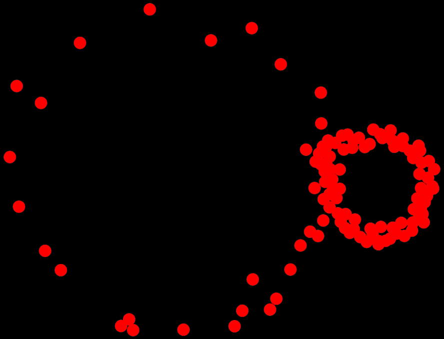
Key idea: Not only can we count holes in each space $O(X)_r$, we can track holes in a consistent way across the whole filtration at once.

The formalization of this idea is persistent homology.

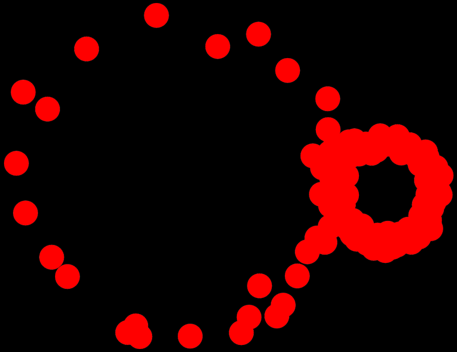
Example



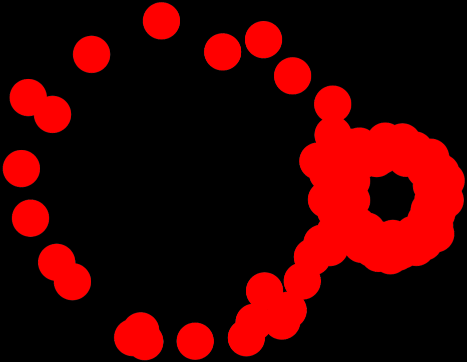
Example



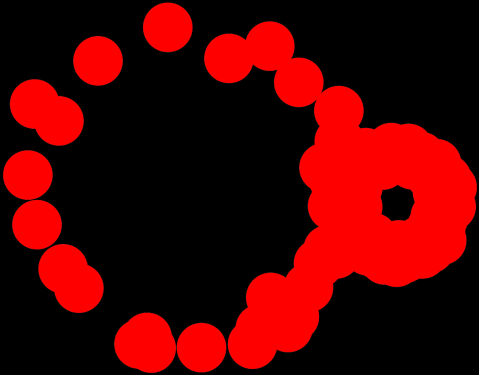
Example



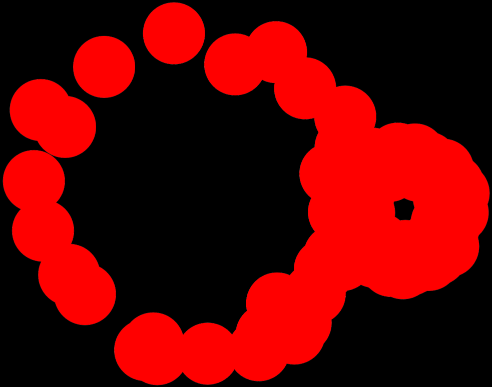
Example



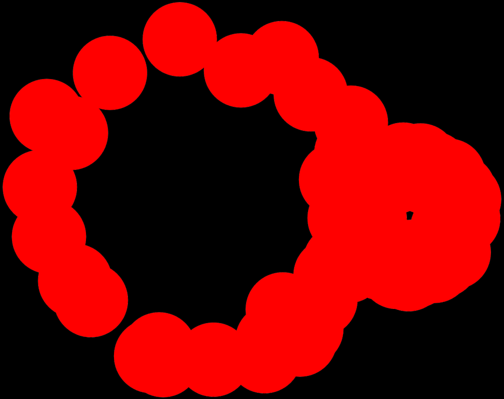
Example



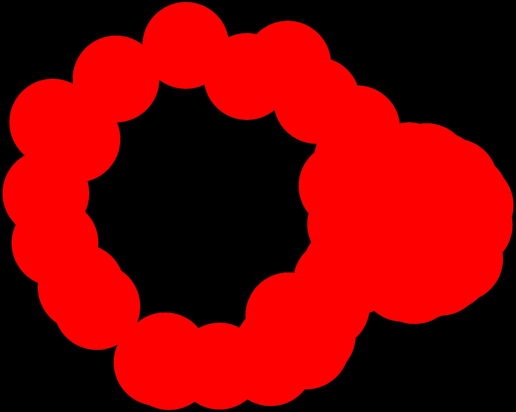
Example



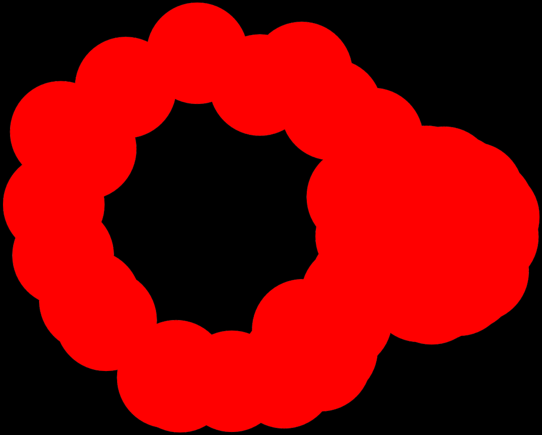
Example



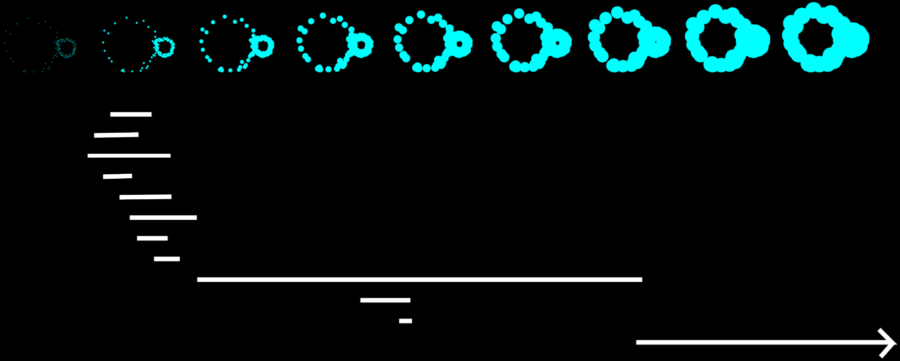
Example



Example



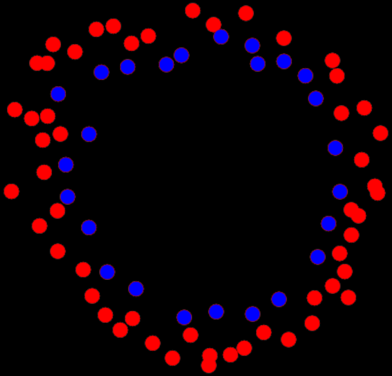
Barcode of the Filtration



- Each interval represents a hole in filtration,
- Left endpoint is index at which hole forms,
- Right endpoint is index at which hole closes up,
- Interval length is a measure of the size of the hole.

These barcodes are computable, using ideas from computational geometry and a variant of Gaussian elimination.

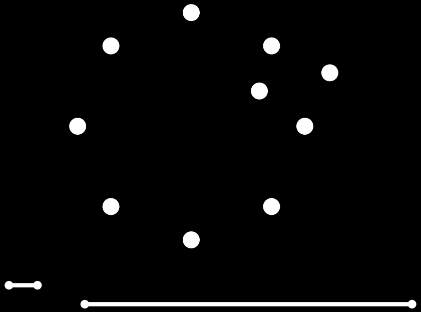
With some additional work, we can also find geometric representations of the holes.



The next figures were made using variant of the offset filtration called the Rips filtration.

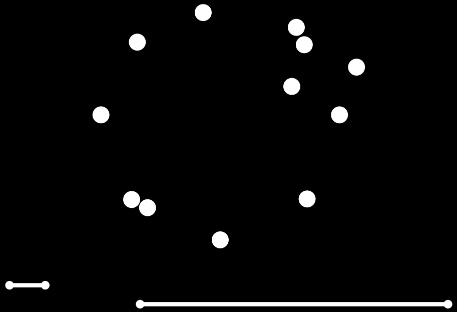
Stability

Persistent Homology of PCD is stable w.r.t. perturbations of points, addition of points near other points.

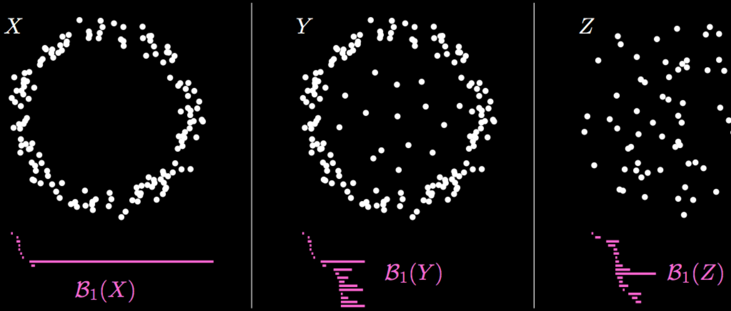


Stability

Persistent Homology of PCD is stable w.r.t. perturbations of points, addition of points near other points.



Limitations of 1-Parameter Persistence



- Persistent homology is not stable with respect to outliers,
- Can be insensitive to structure in high density regions of data.

This leads us to 2-parameter persistence:

- 2^{nd} parameter controls how aggressively we remove outliers.

1-parameter persistence: Build a filtration (1-parameter family of spaces) from data

2-parameter persistence: Build a bifiltration (2-parameter family of spaces).

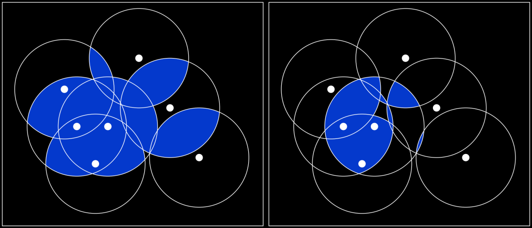
The are a number of density-sensitive bifiltration constructions for point cloud data, with different advantages.

I'll mention just one now, the multicover bifiltration, a 2-parameter extension of the union-of-balls filtration.

For $X \subset \mathbb{R}^n$, define

$$\tilde{\mathcal{M}}(X)_{k,r} = \{y \in \mathbb{R}^n \mid \exists\ k \text{ points } z \in X \text{ with } \|y - z\| \leq r\}.$$

allowing k and r to vary, we obtain the multicover bifiltration $\tilde{\mathcal{M}}(X)$.

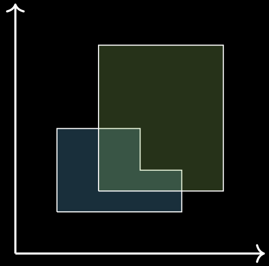


$k = 2$ $k = 3$

$\tilde{\mathcal{M}}(X)$ satisfies a strong robustness property (i.e., it is stable to outliers), and for fixed n, can be computed in polynomial time.

2-Parameter Barcodes?

Can we define a barcode of the multicover bifiltration as a collection of nice regions in $\mathbb{R}^2$?



Not in any good way.

However, it was recently discovered that there are good notions of a signed barcode for multiparameter persistence, where such regions are allowed to have positive and negative multiplicity.

Key theme of MPH: Many of the key ideas of 1-parameter persistence have very natural, yet non-obvious analogues in the 2-parameter (or multiparameter) setting [?].

	1-parameter	2-parameter
filtrations	offset Rips alpha	multicover subdivision (degree) rhomboid
metrics	Hausdorff Gromov-Hausdorff Bottleneck Barcode Wasserstein	Prohorov Gromov-Prohorov (Homotopy) Interleaving Presentation
structure thm.	interval decomp.	Krull-Schmidt-Azumaya
invariant	barcode	unsigned barcode
computation	barcode	minimal presentation
tool	persistent nerve thm.	multicover nerve thm.