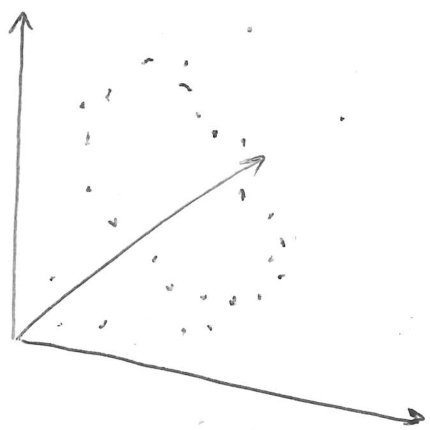
Topology, Geometry & Data - PhD Seminar 22/11/19

N. SALE
Swansea University

Motivation

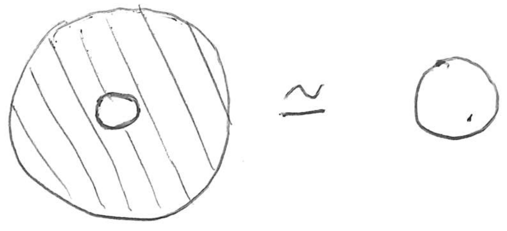
- data often comes in the form of a pointcloud $X \subseteq \mathbb{R}^N$, or as data on a lattice - e.g. images, physics models
- in many situations, this data is sparse and may be lying (roughly) on some submanifold of $\mathbb{R}^N$
- Traditional methods of dimensionality reduction are linear, or assume gaussian distributions, or depend on some choice of scale parameter.



- TDA is one approach to look at more complex, non-linear features in a scale-invariant way.

References:
- Chazal + Michel 2017 - Intro to TDA
- Ghrist 2008 - Barcodes: The persistent topology of data
- Carlsson + Zomorodian 2009 - Theory of multiparameter persistence

topology

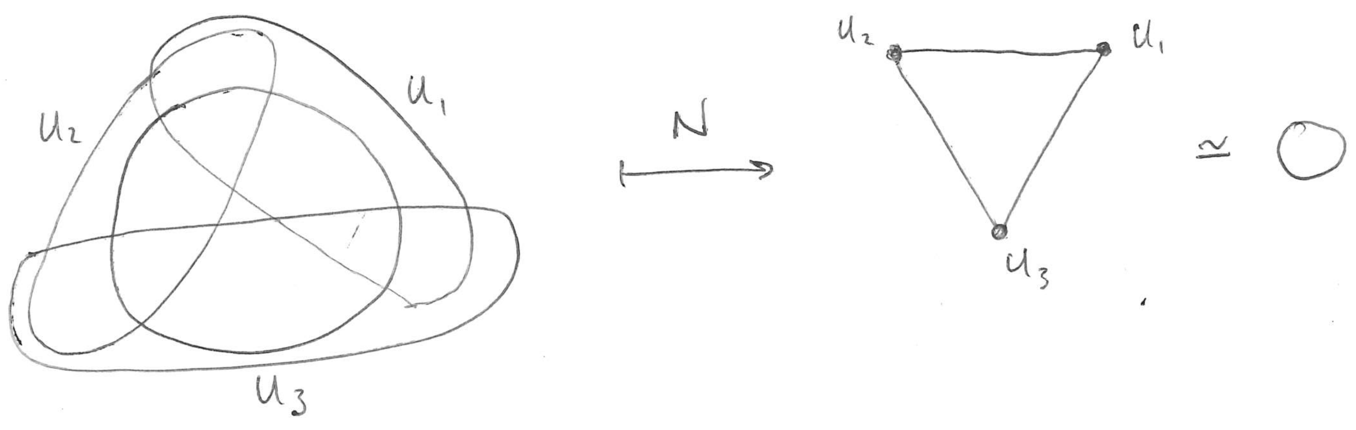
- We recall that topology studies topological spaces and continuous maps between them.
  We'll only really consider spaces which are also metric spaces, with the topology given by that metric.
- So a space X will consist of a set of points, as well as a function $d: X \times X \to \mathbb{R}^{\geq 0}$ giving the distance between any two points.
- A map $f: X \to Y$ is then continuous if
  $\forall x \in X, \forall \varepsilon > 0, \exists \delta > 0$ s.t. $f(B(x, \delta)) \subseteq B(f(x), \varepsilon)$.
- X and Y are homeomorphic ($\cong$) if $\exists f: X \to Y$ and $g: Y \to X$ continuous s.t. $fg = id_Y$ and $gf = id_X$.
- $f: X \to Y$, $g: X \to Y$ are homotopic ($\simeq$) if $\exists H: X \times [0,1] \to Y$ continuous s.t.
  $H(x, 0) = f(x)$ and $H(x, 1) = g(x)$ $\forall x \in X$.
- X, Y homotopy equivalent ($\simeq$) if $\exists f: X \to Y$ and $g: Y \to X$ cont. s.t. $fg \simeq id_Y$ and $gf \simeq id_X$.
- $X \cong Y \Rightarrow X \simeq Y$ 

- A simplicial complex consists of a set of vertices $\{v_0, \dots, v_n\} = V$ and a set of simplices $\Sigma \subseteq \mathcal{P}(V)$ s.t. if $\sigma \in \Sigma$ and $\tau \subseteq \sigma$, then $\tau \in \Sigma$.

- Think of $[v_i]$ as a vertex, $[v_i, v_j]$ as an edge between vertices, $[v_i, v_j, v_k]$ as a filled in triangle, $[v_i, v_j, v_k, v_l]$ as a solid tetrahedron, etc...

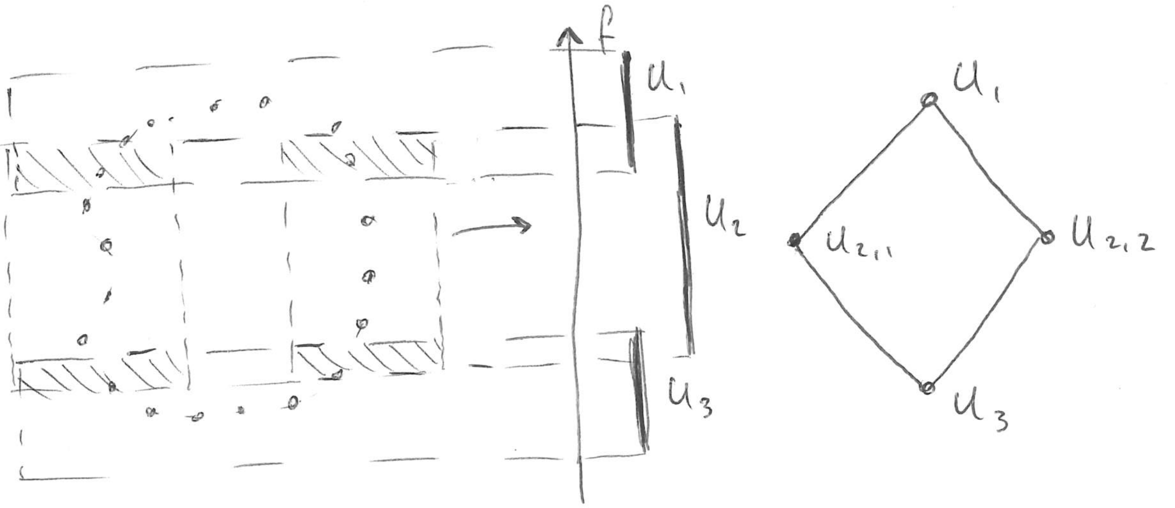
- Given ~~an~~ a cover $\mathcal{U} = \{U_i\}$ of a space X, the <u>nerve</u> $N(\mathcal{U})$ of this cover is the s.c. with $V = \{U_i\}$ and $[U_{i_0}, \dots, U_{i_k}] \in \Sigma$ iff $\bigcap_{j=0}^{k} U_{i_j} \neq \emptyset$.

- <u>Theorem</u>: (Nerve Theorem)
  If the intersection of any subcollection of the U_is is either empty or contractible ($\simeq \{*\}$), then $N(\mathcal{U}) \simeq X$.



Mapper

- given data X and a "lens function" $f : X \to \mathbb{R}$, we cover $\mathbb{R}$ with overlaping intervals U_i, pullback this cover to X, do some clustering, then compute the nerve.



- f could be density, centrality, coordinates from some dim-reduction technique.
- Mapper is very dependent on the choice of f and the cover U_i
- Most often used for exploratory data analysis.
- Pawel came up with an alternative idea:

  Ball mapper:

  - Take an ε-net C:
    - $c \neq c' \Rightarrow d(c, c') > \varepsilon$
    - $\forall x \in X, \exists c \in C$ s.t. $d(x, c) \leq \varepsilon$.
  - Take the nerve of $\{B(c, \varepsilon) \mid c \in C\}$.

persistent homology

- given a simplicial complex X, consider the ~~groups~~ modules

$$C_k(X) = \left\{ \sum_{i=0}^{n} a_i \sigma_i \;\middle|\; a_i \in R,\ \sigma_i \in \Sigma,\ |\sigma_i| = k+1 \right\}$$

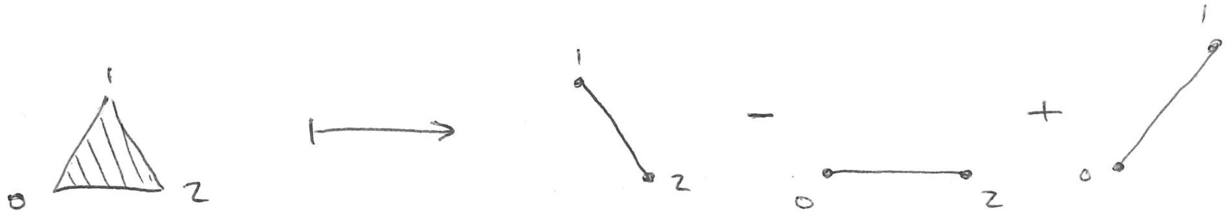
generated by the k-simplices (vertices are 0-simplices, edges are 1-simplices, etc ...). Simplicial chains.
For some ring R. Usually at least a PID. Often a field.

- There is a map $\partial_k : C_k(X) \longrightarrow C_{k-1}(X)$

$$[v_0, \dots, v_k] \longmapsto \sum_{i=0}^{k} (-1)^i [v_0, \dots, \hat{v_i}, \dots, v_k]$$

means deleted.

called the boundary map.



- It has the property that $\partial_k \circ \partial_{k+1} = 0 \quad \forall k$.
"the boundary of a boundary is zero"

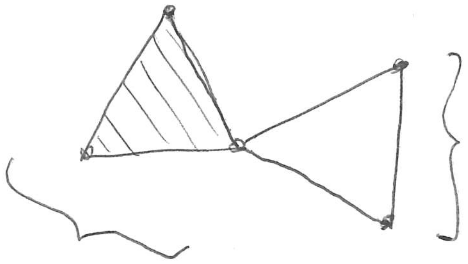
- If we think of $\ker \partial_k$ as ~~those~~ those k-chains with no boundary, "cycles" and $\operatorname{im} \partial_{k+1}$ as those chains which bound a higher dimensional one, "boundaries" then we also write this as

$$\operatorname{im} \partial_{k+1} \subseteq \ker \partial_k .$$

- Then we define the k^{th} homology group:

$$H_k(X) := \frac{\ker \partial_k}{\operatorname{im} \partial_{k+1}}.$$

thinking of it as those chains which form a cycle, but don't bound any higher-dimensional chains. i.e. a hole.



these 3 edges form a cycle, and are not the boundary of anything.

these 3 edges form a cycle, but are the boundary of the filled in bit.

- <u>theorem</u>: $X \simeq Y \Rightarrow H_k(X) \cong H_k(Y) \quad \forall k \in \mathbb{Z}$.

- moreover, homology is <u>functorial</u>: given a map $f: X \to Y$, there is an induced map $H_k(f): H_k(X) \to H_k(Y)$, and this assignment respects identity and composition.

- So, given some data points, how do we look at homology?

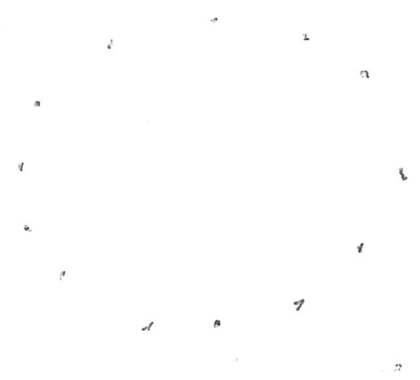
Build a simplicial complex on top.

- Given $\varepsilon > 0$ and X a pointcloud in a metric space, the Vietoris-Rips complex at ε is the simplicial complex with $V = X$ and

$$\Sigma = \{ [x_{i_1}, \dots, x_{i_k}] \mid d(x_{i_a}, x_{i_b}) \leq \varepsilon \ \forall a, b \}$$

denoted $VR_\varepsilon(X)$. Then we can compute $H_k(VR_\varepsilon(X))$.

- How do we pick ε?



too small, don't capture the cycle.

too big, we fill it in.

Say we get it just right. What if our data looked like:

with multiple scales?

- idea: don't pick ε. Let it vary from 0 to ∞.

- note that $\varepsilon \leq \varepsilon'$ implies

$$VR_\varepsilon(X) \subseteq VR_{\varepsilon'}(X).$$

in particular, there's an inclusion map

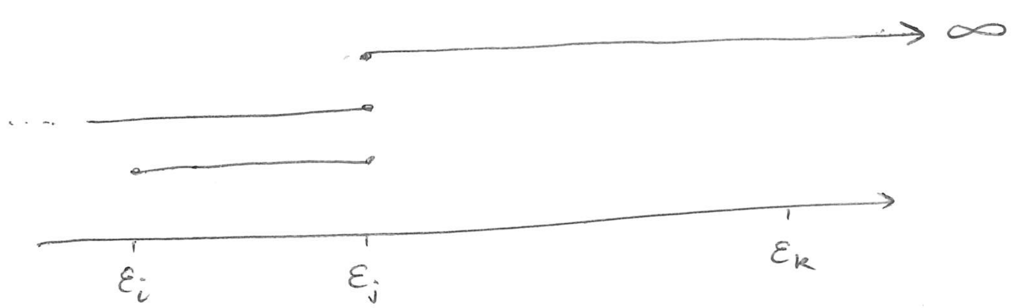
$$VR_\varepsilon(X) \hookrightarrow VR_{\varepsilon'}(X).$$

- say we have $\varepsilon_0 \leq \varepsilon_1 \leq \dots \leq \varepsilon_N$ where the complex changes. Then applying the functoriality of H_k, we have a sequence

$$H_k(VR_{\varepsilon_0}(X)) \longrightarrow H_k(VR_{\varepsilon_1}(X)) \longrightarrow \dots \longrightarrow H_k(VR_{\varepsilon_N}(X))$$

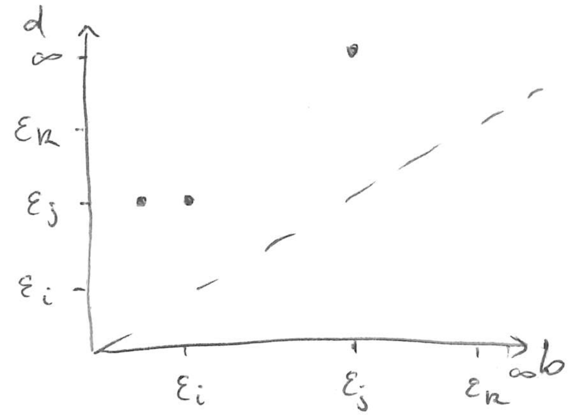
we track when homology classes are born and when they die as we walk through the sequence.

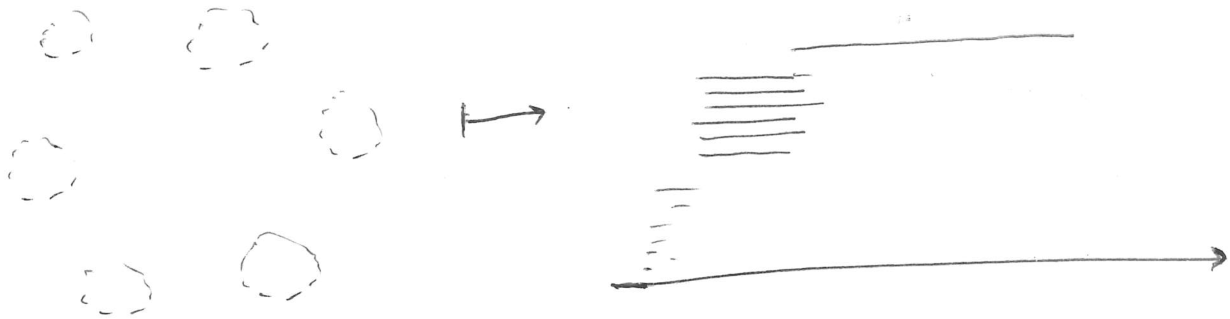
- we write this as a "barcode": each bar is a homology class.



where the longer a bar is, the longer that class persists through the filtration.

- We can also write this as a "persistence diagram"



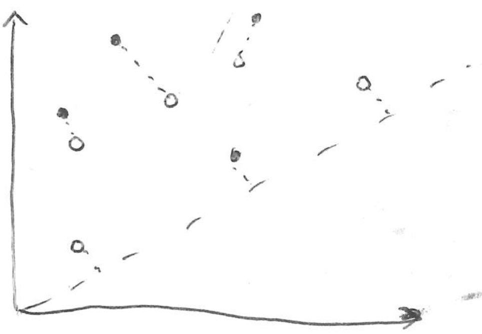
- 

- theorem (structure): Assuming some tameness conditions, every sequence of vector spaces $V_0 \to V_1 \to \cdots \to V_N$ can be decomposed into a direct sum of interval sequences of the form

$$0 \to \cdots \to 0 \to F \to F \to \cdots \to F \to 0 \to \cdots \to 0$$

Each such sequence can be thought of as a bar.

- how much does the PH change as we change the data?
- we can put metrics on persistence diagrams:



$$d_{\text{bottleneck}}(dgm_1, dgm_2) := \inf_{\substack{\text{matchings} \\ m}} \max_{(p,q) \in m} \|p-q\|_\infty .$$

- given $f: X \to \mathbb{R}$, and $a_0 \leq a_1 \leq \cdots \leq a_N$ we can have a filtration

$$f^{-1}((-\infty, a_0]) \subseteq f^{-1}((-\infty, a_1]) \subseteq \cdots \subseteq f^{-1}((-\infty, a_N]),$$

and we can consider the H_k persistence of this. Say the diagram is $dgm_k(f)$.

- ___theorem:___ (Stability) Given $f, g: X \to \mathbb{R}$ which yield tame sequences

$$d_{bottleneck}(dgm_k(f), dgm_k(g)) \leq \|f - g\|_\infty$$

$$\left(= \sup_{x \in X} \|f(x) - g(x)\|\right)$$

moving the data slightly only produces a slight change in the persistence diagram...

- directions of research:
  - using persistence diagrams / barcodes as features for machine learning: images, feature vectors, etc...
  - Statistics for persistence: say X is drawn as a sample from some ~~underlying~~ distribution supported on an underlying ~~[illegible]~~ space. Can we infer topological properties of that space from samples like X?
  - multiparameter persistence:
    - ___theorem___: There is not complete discrete invariant for multiparameter persistence.
  - other generalisations: Zigzag persistence, Circle persistence
  - Where can persistence be applied?: Medical imaging, material science, ...