Math 583B: *Topological Data Analysis*

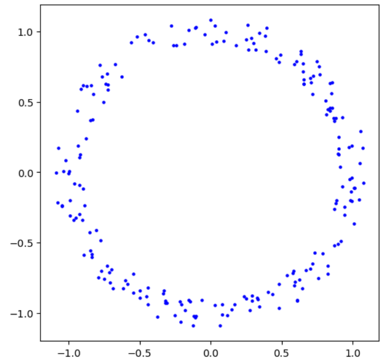
Kyle Ormsby

Spring 2024

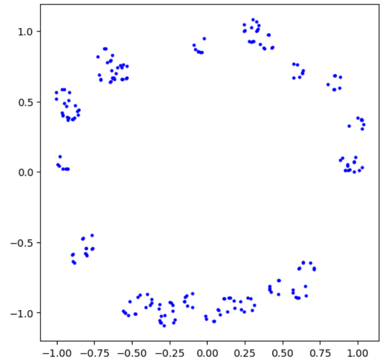
1 *Inferring the shape of data* — 25 March 2024

Imagine that you're running an experiment in which you measure a large number — say N — of real-valued variables with each observation. Each observation is then a point in $\mathbb{R}^N$, and if you make k total observations, then the data associated with your experiment is a *point cloud* $P = \{x_1, x_2, \dots, x_k\} \subseteq \mathbb{R}^N$.

If the system being observed is not purely random, then — up to issues of noise and accuracy — we expect P to be sampled from a subspace $M \subseteq \mathbb{R}^N$. How might we infer the structure and shape of M from P, at least under the assumption that k is relatively large? This is one of the questions that topological data analysis (TDA) aims to answer, at least for particular notions of "structure" and "shape". In the figure presented here, we see a point cloud P in $\mathbb{R}^2$ sampled with noise from the unit circle $S^1 \subseteq \mathbb{R}^2$, and we seek algorithmic methods that will recognize (features of) S^1 as the underlying space from which P is sampled. Of course, in practice, N might be very large, and it is unlikely that your visual cortex will rise to the challenge of guessing M.



But even for small N, we can still ask more from our methods. Consider the displayed point cloud $Q \subseteq \mathbb{R}^2$ which exhibits strikingly different structure at different scales. At small scales, points seem to be sampled from disjoint circles. After zooming out (so at a larger scale), those small circles seem to assemble into one big copy of S^1. The tools we will develop are *scale independent* and do not depend on parameter tuning. We will ultimately produce concise, interpretable summaries that capture the nature of data at all scales.



Our first and primary tool will be the *persistent homology* of the *Čech* or *Vietoris–Rips filtered complex* associated with a point cloud $P \subseteq \mathbb{R}^N$. We may view this as a two-step process:

$$\begin{tikzcd}\{\text{point clouds}\} \arrow[r, bend left, "\check{\mathbb{C}}"] \arrow[r, bend right, "\mathbb{VR}"'] & \{\text{filtered spaces}\} \arrow[r, "\mathrm{PH}"] & \{\text{persistence modules}\}.\end{tikzcd}$$

A *filtered space* $\mathscr{X} = \{X_s\}_{s \in \mathbb{R}}$ is a collection of spaces[1] X_s indexed by scales $s \in \mathbb{R}$ such that

$$s \le t \implies X_s \subseteq X_t.$$

For the purposes of this introduction, we will focus on the Čech filtered complex $\check{\mathbb{C}}(P)$ of our point cloud $P \subseteq \mathbb{R}^N$. At scale $s \in \mathbb{R}$, $\check{\mathbb{C}}_s(P)$ is

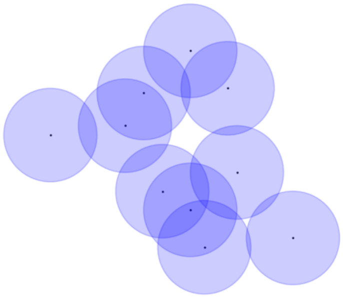
[1] By *space* we might mean topological space or (abstract) simplicial complex. If working with complexes, we take $\subseteq$ to mean subcomplex.

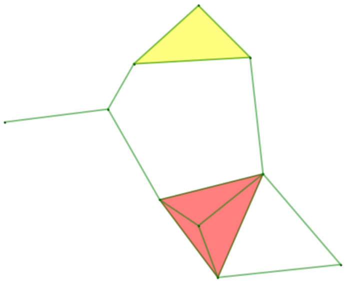
the simplicial complex with one n-simplex for each subset $A \subseteq P$ with $|A| = n+1$ and

$$\bigcap_{x \in A} \overline{B}_s(x) \neq \varnothing.$$

In other words, we get an n-simplex for each $(n+1)$-subset of P for which the closed Euclidean balls of radius s centered at points of A have nonempty common intersection. Since the intersection condition becomes less stringent as s gets larger, we have that $\check{\mathrm{C}}_s(P)$ is a subcomplex of $\check{\mathrm{C}}_t(P)$ when $s \leq t$. Later, we will encounter the Nerve Lemma, which roughly says that $\check{\mathrm{C}}_s(P)$ is homotopy equivalent to $\bigcup_{x \in P} \overline{B}_s(P)$ in reasonable scenarios. Note that the combinatorial nature of $\check{\mathrm{C}}_s(P)$ makes it much better adapted to computation than the filtered topological space $\{\bigcup_{x \in P} \overline{B}_s(P)\}_{s \in \mathbb{R}}$.

We write $\overline{B}_s(x)$ for the closed ball of radius s centered at x.





Black points are 0-simplices, green edges are 1-simplices, yellow shading is a 2-simplex, and red shading is a 3-simplex. Note that the bottom right triangle is not filled in yellow because the triple intersection of the balls around those vertices is empty.

Now that we have a filtered space $\mathscr{X} = \check{\mathrm{C}}(P)$, we aim to capture features of each space $X_s := \check{\mathrm{C}}_s(P)$ and how these features are related as the filtration parameter changes. Taking a cue from algebraic topology, we view $\mathrm{H}_*(X_s;\mathbb{F})$ — the homology[2] of X_s with coefficients in a field $\mathbb{F}$ — as a good summary of the features of X_s. Functoriality of homology then provides us with $\mathbb{F}$-linear transformations

$$(\iota_s^t)_* \colon \mathrm{H}_*(X_s;\mathbb{F}) \longrightarrow \mathrm{H}_*(X_t;\mathbb{F})$$

for $s \leq t$ and $\iota_s^t \colon X_s \subseteq X_t$, and these maps $(\iota_s^t)_*$ provide our comparisons of features. Packaging all of the homologies and comparisons maps together produces a *persistence module* $\mathrm{PH}_*(\mathscr{X};\mathbb{F})$, the $\mathbb{F}$*-persistent homology* of $\mathscr{X}$, which is our scale independent summary of the shape of our data.

The miracle here is that persistence modules admit a convenient and complete invariant called a *barcode* or (after a mild but tremendously beneficial transformation) *persistence diagram*. To give the flavor of barcodes, we will consider a simplified scenario in which we have $\mathbb{F}$-vector spaces $\{V_i\}_{i \in \mathbb{N}}$ and linear transformations $\iota_i^j \colon V_i \to V_j$ for $0 \leq i \leq j$ such that

(1) $\iota_i^i = \mathrm{id}_{V_i}$ for all i, and

(2) for $0 \leq i \leq j \leq k$, $\iota_j^k \circ \iota_i^j = \iota_i^k$.

The essential data here is of the form

$$V_0 \to V_1 \to \cdots \to V_i \to V_{i+1} \to \cdots$$

and we may view the persistence module $(\{V_i\}_{i \in \mathbb{N}}, \{\iota_i^j\}_{i \leq j})$ as a functor $\mathscr{V} = \{V_i\}_{i \in \mathbb{N}}$ from the category associated with the partially ordered set $(\mathbb{N}, \leq)$ to the category of $\mathbb{F}$-vector spaces and linear transformations. Such a persistence module might arise from a point cloud by considering Čech complexes at scales $s_0 < s_1 < \cdots$.

Let $\mathbb{F}[t]$ denote the ring of polynomials in variable t over $\mathbb{F}$, graded so that $|t| = 1$, and set

$$\Theta(\mathscr{V}) := \bigoplus_{i \in \mathbb{N}} V_i.$$

[2] We will review homology theory next lecture. It is a lie in the direction of truth to say that the dimension of the $\mathbb{F}$-vector space $\mathrm{H}_n(X_s;\mathbb{F})$ measures the number of n-dimensional "holes" in X_s.

Then we may endow $\Theta(\mathscr{V})$ with the structure of a graded $\mathbb{F}[t]$-module by setting the action of the polynomial generator t to be

$$t \cdot (v_i)_{i\in\mathbb{N}} := (\iota_{i-1}^{i} v_{i-1})_{i\in\mathbb{N}}$$

where $v_{-1} := 0$. In fact, Θ is an equivalence of categories between $\mathbb{N}$-persistence modules and graded $\mathbb{F}[t]$-modules.[3]

A common capstone theorem of a first course in algebra is the classification of finitely generated modules over a principal ideal domain. A graded version of this theorem holds *mutatis mutandis*, and so it behooves us to understand which persistence modules correspond to finitely generated graded $\mathbb{F}[t]$-modules. Call a persistence module $\mathscr{V} = \{V_i\}_{i\in\mathbb{N}}$ *tame* when every V_i is finite-dimensional and ι_i^{i+1} is an isomorphism for sufficiently large i. One may prove that $\mathscr{V}$ is tame if and only if $\Theta(\mathscr{V})$ is finitely generated over $\mathbb{F}[t]$.

By the classification theorem for finitely generated graded modules over a PID, if $\mathscr{V}$ is tame then there are (essentially unique) integers $i_1, \ldots, i_m, j_1, \ldots, j_n, \ell_1, \ldots, \ell_n$ and an isomorphism

$$\Theta(\mathscr{V}) \cong \bigoplus_{s=1}^{m} \Sigma^{i_s}\mathbb{F}[t] \oplus \bigoplus_{t=1}^{n} \Sigma^{j_t}\mathbb{F}[t]/(t^{\ell_t})$$

where Σ^r denotes a grading shift upwards by r.[4] Translating this into the world of persistence modules, we learn that every tame persistence module decomposes (essentially uniquely) as

$$\mathscr{V} \cong \bigoplus_{j=0}^{N} \mathbb{I}[b_j, d_j]$$

where each b_j is a nonnegative integer, $d_j \in \mathbb{N} \cup \{\infty\}$, and $\mathbb{I}[b_j, d_j]$ is the *interval persistence module* with

$$\mathbb{I}[b_j, d_j]_i = \begin{cases} \mathbb{F} & \text{if } b_j \le i \le d_j, \\ 0 & \text{otherwise,} \end{cases}$$

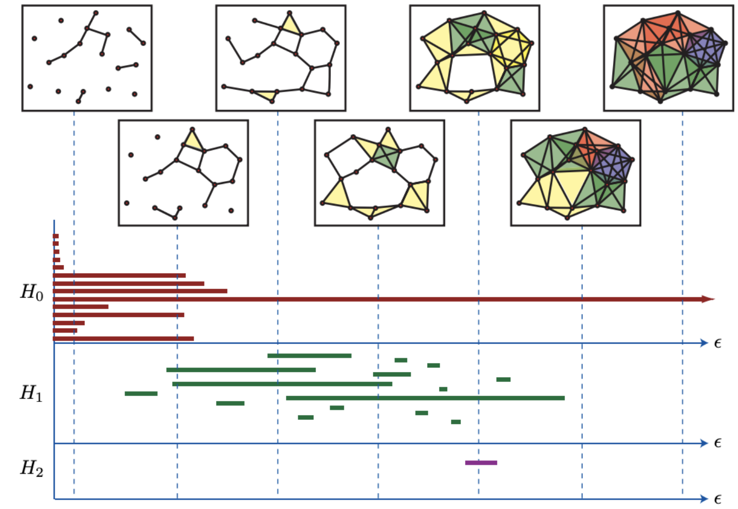
and $\iota_i^{i'} = \mathrm{id}_{\mathbb{F}}$ for $b_j \le i \le i' \le d_j$.

For an interval persistence module $\mathbb{I}[b, d]$, we refer to b as the *birth* and d as the *death* scale. We may then visualize the decomposition of $\mathscr{V}$ as a multiset of intervals $[b_j, d_j]$ called the *barcode* of $\mathscr{V}$. The following illustration is taken from Ghrist.[5] Beware, though, that it uses the Vietoris–Rips filtration instead of the Čech filtration; we will study $\mathbb{VR}$ in detail later.

[3] The inverse functor takes M_* to $\{M_i\}_{i\in\mathbb{N}}$ with ι_i^j given by multiplication by t^{j-i}.

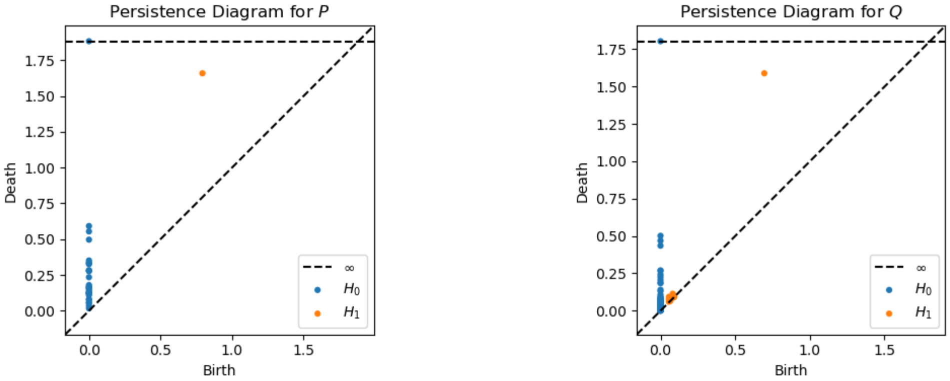
[4] That is, $(\Sigma^r M_*)_s = M_{s-r}$

[5] Ghrist, R. (2008). Barcodes: the persistent topology of data. *Bull. Amer. Math. Soc. (N.S.)*, 45(1):61–75



While barcodes prevailed in the early days of TDA, experience has shown that *persistence diagrams* are better suited to statistical analysis. The persistence diagram of $\mathscr{V}$ consists of the multiset of points (b_j, d_j) lying on or above the diagonal of $\mathbb{N} \times (\mathbb{N} \cup \{\infty\})$.

The Vietoris–Rips filtered complexes of the data sets P and Q from our initial discussion have the following persistence diagrams (with PH_0 in blue and PH_1 in orange):



Focusing on the blue PH_0 classes, we see that in both cases all connected components are born at time 0, and at scales above ≈ 0.7 there is a single connected component that persists to $+\infty$. This last class is analogous to the red bar of infinite length in the previous diagram.

Looking at orange PH_1 classes, we can readily observe significant differences between the point clouds. In each, there is a highly persistent class born around scale 0.75, but Q detects the small scale structure as well, giving a cluster of short-lived PH_1 classes born around scale 0.1. These classes witness the small radii circles (arranged around the unit circle) from which Q is sampled.

It is often claimed that classes with large persistence $d - b$ (*i.e.*, those high above the diagonal) represent the "true" topology of the data, while small persistence classes correspond to noise. The point clouds P and Q illustrate that this is not necessarily the case.

1.1 Future topics

One of our primary tasks will be the development of pseudometrics allowing us to compare persistence diagrams. We leave this to future development, along with the many foundational details elided or overlooked in this introduction. Once the foundations are established, the rest of the course will focus on the following:

(1) applications of persistent homology to particular data modalities,

(2) extending persistent homology to filtrations indexed by more exotic partially ordered sets, and

(3) refining PH_0 via hierarchical clustering.

See the syllabus for a detailed (but flexible) schedule of topics.

1.2 Notes

The content of this introduction was primarily drawn from the Oudot's textbook[6] and Carlsson's survey article.[7] The original images were produced in Python using the Ripser persistent homology package.[8] We will use Ripser extensively when exploring examples and applications, and you should follow the installation instructions at `https://ripser.scikit-tda.org/` to get it working on your personal computer. You can find the Jupyter notebook used to produce diagrams from this and future lectures at `https://github.com/kyleormsby/math583`.

[6] Oudot, S. Y. (2015). *Persistence theory: from quiver representations to data analysis*, volume 209 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI

[7] Carlsson, G. (2009). Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308

[8] Tralie, C., Saul, N., and Bar-On, R. (2018). Ripser.py: A lean persistent homology library for python. *The Journal of Open Source Software*, 3(29):925

2 *Spaces, complexes, and homology* — 27 March 2024

2.1 Topology

From the Kleinian perspective, geometry is the study of properties invariant under isometries, that is, distance-preserving transformations. Indeed, when a geometer says that two triangles are the same (or *congruent* or *isometric*), they do not mean that each triangle consists of exactly the same points, but rather that one may translate, rotate, and reflect one triangle until it matches the other.

Felix Klein (1849–1925)

Topology plays a similar game, but with a much coarser notion of "sameness". We say that two spaces — the objects of topology — are *homeomorphic* when there are continuous functions between them that are mutually inverse. In this sense, topology is the study of properties that are invariant under homeomorphism. Such properties include such notions as connectivity and compactness, but exclude more rigid properties such as angle, distance, or volume.

We will generally assume that the reader is familiar with point-set topology, but will quickly recall some of the basic definitions.

The standard reference for point-set topology is Munkres;   see also the recent graduate text of Bradley–Bryson–Terilla.

Munkres, J. R. (2000). *Topology*. Prentice Hall, Inc., Upper Saddle River, NJ, second edition; and Bradley, T.-D., Bryson, T., and Terilla, J. ([2020] ©2020). *Topology—a categorical approach*. MIT Press, Cambridge, MA

Definition 2.1. A *topological space* is a pair $(X, \mathscr{U})$ consisting of a set X and a collection of subsets $\mathscr{U} \subseteq 2^X$ called *open sets* such that

(1) $\varnothing$ and X are in $\mathscr{U}$,

(2) $\mathscr{U}$ is closed under arbitrary unions: $U_\alpha \in \mathscr{U}$ for $\alpha \in A$ implies $\bigcup_{\alpha \in A} U_\alpha \in \mathscr{U}$, and

(3) $\mathscr{U}$ is closed under finite intersections: $U_i \in \mathscr{U}$ for i in a finite set I implies $\bigcap_{i \in I} U_i \in \mathscr{U}$.

We will write X for $(X, \mathscr{U})$ when the topology $\mathscr{U}$ is clear from context. A subset $U \subseteq X$ is called *open* when it belongs to $\mathscr{U}$, and a subset $C \subseteq X$ is called *closed* when $X \smallsetminus C$ is open. These properties are not mutually exclusive, as exhibited by the *clopen* sets $\varnothing$ and X.

Example 2.2. In the *standard topology* on Euclidean space $\mathbb{R}^n$, a subset $U \subseteq \mathbb{R}^n$ is open if and only if it is a union of *open balls* $B_r(x) := \{y \in \mathbb{R}^n \mid |y - x| < r$. This is equivalent to saying that U is open if and only if for each $x \in U$ there exists $r > 0$ such that $B_r(x) \subseteq U$.

Definition 2.3. A function $f\colon X \to Y$ between topologyical spaces is *continuous* when the preimage $f^{-1}U$ over every open set $U \subseteq Y$ is open in X. A continuous function $f\colon X \to Y$ is a *homeomorphism* when it admits a continuous inverse $g\colon Y \to X$. In this case, we say that X and Y are *homeomorphic* and write $X \cong Y$.

The categorically inclined reader will note that topological spaces and continuous functions form a category, and the isomorphisms in this category are exactly the homeomorphisms.

Example 2.4. If X is a topological space and $f\colon X \to \mathbb{R}$ is continuous, then the *sublevel set* $f^{-1}(-\infty, u) = \{x \in X \mid f(x) < u\}$ is open in X since the interval $(-\infty, u) = \{t \in \mathbb{R} \mid t < u\}$ is open in $\mathbb{R}$.

2.2 Geometric and abstract simplicial complexes

2.3 Simplicial homology

References

Bradley, T.-D., Bryson, T., and Terilla, J. ([2020] ©2020). *Topology—a categorical approach.* MIT Press, Cambridge, MA.

Carlsson, G. (2009). Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308.

Ghrist, R. (2008). Barcodes: the persistent topology of data. *Bull. Amer. Math. Soc. (N.S.)*, 45(1):61–75.

Munkres, J. R. (2000). *Topology.* Prentice Hall, Inc., Upper Saddle River, NJ, second edition.

Oudot, S. Y. (2015). *Persistence theory: from quiver representations to data analysis*, volume 209 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI.

Tralie, C., Saul, N., and Bar-On, R. (2018). Ripser.py: A lean persistent homology library for python. *The Journal of Open Source Software*, 3(29):925.