Persistent Homology

김지수 (Jisu KIM)

통계이론세미나 - 위상구조의 통계적 추정, 2023 가을학기

When analyzing data, we prefer robust features where features of the underlying manifold can be inferred from features of finite samples. As I introduced before, homology is counting holes. But there is a problem of using homology to data: that homology of finite sample is very differnt from homology of underlying manifold. See Figure .

We first recall the homology.

Definition. Let K be a simplicial complex, $k \geq 0$ be a nonnegative integer, and G be an abelian group. The space of k-chains on K, $C_k(K;G)$, is the set whose elements are a finite formal sum of k-simplices of K with coefficients from G, i.e.,

$$C_k(K;G) = \left\{ \sum_i n_i \sigma_i : n_i \in G, \sigma_i \in K_k \right\},$$

where $K_k \subset K$ is the set of k-simplices of K. We write $C_k(K)$ if the coefficient group G is understood from the context.

For an integer $k \leq -1$, we define $C_k(K) = 0$ for convenience.

Remark. Typical examples of G are $G = \mathbb{Z}$ and $G = \mathbb{Z}_2 = \mathbb{Z}/2\mathbb{Z}$. For $G = \mathbb{Z}_2$, $C_k(K;\mathbb{Z}_2)$ becomes a vector space.

Remark. $C_k(K;G)$ has an abelian group structure as for $\sum_i n_i \sigma_i, \sum_i n_i' \sigma_i \in C_k(K;G)$,

$$\left(\sum_i n_i \sigma_i \right) + \left(\sum_i n_i' \sigma_i \right) := \sum_i (n_i + n_i') \sigma_i.$$

When G is a field, $C_k(K;G)$ has a natural vector space structure as for $\sum_i n_i \sigma_i \in C_k(K;G)$ and $\lambda \in G$,

$$\lambda \cdot \left(\sum_i n_i \sigma_i \right) = \sum_i (\lambda \cdot n_i) \sigma_i.$$

To relate chain groups of different dimensions, we define the boundary map as sending each -simplex to the sum of its $(k-1)$-dimensional faces. We write $\sigma = [v_0, \ldots, v_k]$ for an ordered simplex, i.e., $[v_0, v_1] = -[v_1, v_0]$.

Definition. A boundary map $\partial_k : C_k(K) \to C_{k-1}(K)$ is defined for each simplex as (see Figure)

$$\partial_k [v_0, \ldots, v_k] = \sum_{j=0}^{k} (-1)^j [v_0, \ldots, \hat{v}_j, \ldots, v_k],$$

where $[v_0, \ldots, \hat{v}_j, \ldots, v_k] = [v_0, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k] \in K_{k-1}$, i.e., $\hat{v}_j$ means that v_j is omitted. The definition is extended to entire k-chain as

$$\partial_k \left(\sum_i n_i \sigma_i \right) = \sum_i n_i \partial_k \sigma_i.$$

Remark. ∂_k satisfies that for $c, c' \in C_k(K)$, $\partial_k(c + c') = \partial_k c + \partial_k c'$, so $\partial_k : C_k(K) \to C_{k-1}(K)$ is a homomorphism.

Lemma ([2, Lemma 2.1])**.** $\partial_{k-1} \circ \partial_k = 0$.

Definition. Cycles and boundaries

(a) A k-cycle group $Z_k = Z_k(K)$ is the k-cycle whose boundary is 0, i.e.,

$$Z_k(K) = \ker \partial_k = \{ c \in C_k(K) : \partial_k c = 0 \}.$$

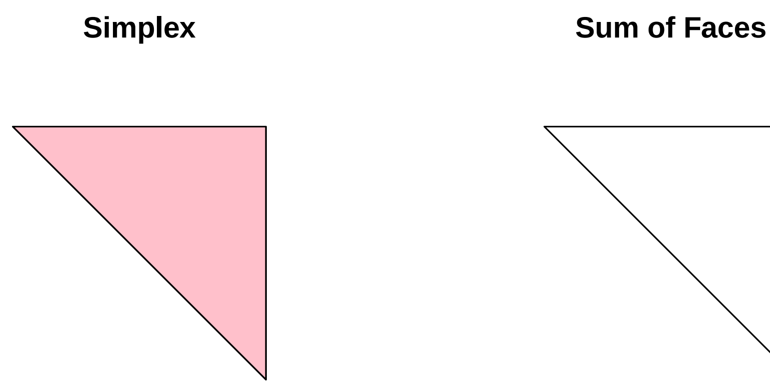


Figure 1: Boundary map.

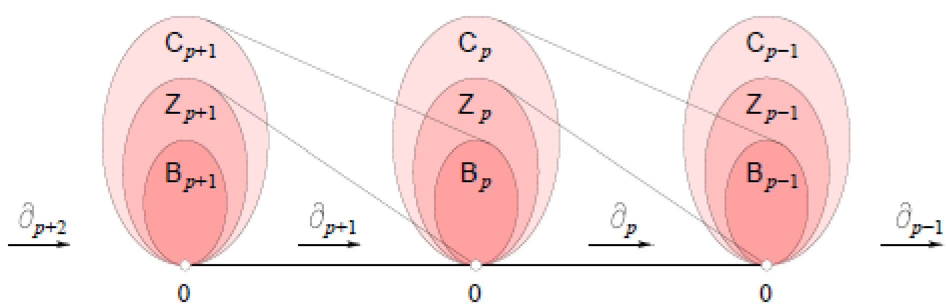


Figure 2: [1, Figure IV.1] Interleaving relations between cycle groups and boundary groups via boundary map.

(b) A k-boundary group $B_k = B_k(K)$ is the k-cycle that is a boundary of $(k+1)$-chain,

$$B_k(K) = \text{im}\partial_{k+1} = \{\partial_{k+1}d \in C_k(K) : d \in C_{k+1}(K)\}.$$

Then the above Lemma implies that $B_k(K)$, $Z_k(K)$, $C_k(K)$ are interleaved as subgroups (see Figure):

$$B_k(K) \subset Z_k(K) \subset C_k(K).$$

Definition. The k-th homology group is the k-th cycle group modulo the k-th boundary group,

$$H_k = H_k(K) \coloneqq Z_k(K)/B_k(K).$$

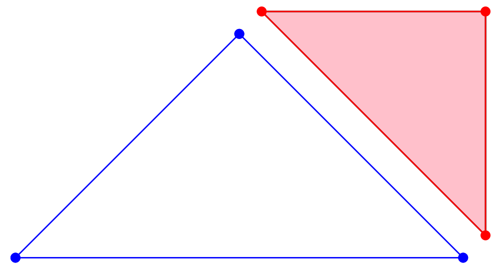
The k-th Betti number is the rank of this group, $\beta_k = \text{rank}H_k$.

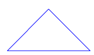
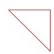
Example. Suppose K is given as the right of Figure **??**, and use $G = \mathbb{Z}$. Then for $k = 1$, its cycle group, boundary group, homology group, and betti number is computed as in Figure .

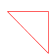
Persistent homology is a multiscale approach to represent topological features.

A *filtration* $\mathcal{F}$ is a collection of objects (subcomplexes or subsets) approximating the data points at different resolutions, formally defined as follows.

Definition (Filtration)**.** A *filtration* $\mathcal{F}$ is a collection of increasing objects:



- $Z_1(K) = \ker \partial_1 = \mathbb{Z}^2 =<$  ,  $>$

- $B_1(K) = \mathrm{im}\partial_2 = \mathbb{Z} =<$  $>$

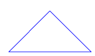
- $H_1(K) = Z_1(K)/B_1(K) = \mathbb{Z} =<$  $>$, $\beta_1(K) = 1$

Figure 3: Homology example for Figure **??**.

- Let K be a simplicial complex. A (subcomplexes) *filtration* $\mathcal{F} = \{\mathcal{F}_a \subset K\}_{a \in \mathbb{R}}$ is a collection of subcomplexes of K such that $a \leq b$ implies that $\mathcal{F}_a \subset \mathcal{F}_b$.
- A (subsets) *filtration* $\mathcal{F} = \{\mathcal{F}_a \subset \mathbb{X}\}_{a \in \mathbb{R}}$ is a collection of subsets of a topological space $\mathbb{X}$ such that $a \leq b$ implies that $\mathcal{F}_a \subset \mathcal{F}_b$.

A typical way of setting the filtration is through a real-valued function.

Definition. For a simplicial complex K, a real function $f : K \to \mathbb{R}$ is monotonic if $f(\sigma) \leq f(\tau)$ whenever σ is a face of τ.

For a real monotonic function $f : K \to \mathbb{R}$, if we let $\mathcal{F}_a \coloneqq f^{-1}(-\infty, a]$, then the monotonicity implies that $\mathcal{F}_a$ is a subcomplex of K and $\mathcal{F}_a \subset \mathcal{F}_b$ whenever $a \leq b$, so $\mathcal{F} = \{\mathcal{F}_a \subset K\}_{a \in \mathbb{R}}$ is a subcomplexes filtration. Similarly, for a real function $f : \mathbb{X} \to \mathbb{R}$ on a topological space $\mathbb{X}$ (not necessarily continuous), if we let $\mathcal{F}_a \coloneqq f^{-1}(-\infty, a]$, then $\mathcal{F} = \{\mathcal{F}_a \subset K\}_{a \in \mathbb{R}}$ is a subsets filtration. This filtration $\mathcal{F}$ is called a sublevel filtration (of f).

Remark. We also consider a superlevel filtration $\left\{f^{-1}[a, \infty)\right\}_{a \in \mathbb{R}}$, in particular when f is a density function. However, a

For a filtration $\mathcal{F}$ and for each $k \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$, the associated persistent homology $PH_k\mathcal{F}$ is an ordered collection of k-th dimensional homologies, one for each element of $\mathcal{F}$.

Definition (Persistent Homology)**.** Let $\mathcal{F}$ be a filtration and let $k \in \mathbb{N}_0$. The associated k-th *persistent homology* $PH_k\mathcal{F}$ is a collection of groups $\{H_k\mathcal{F}_a\}_{a \in \mathbb{R}}$ equipped with homomorphisms $\{\imath_k^{a,b}\}_{a \leq b}$, where $H_k\mathcal{F}_a$ is the k-th dimensional homology group of $\mathcal{F}_a$ and $\imath_k^{a,b} : H_k\mathcal{F}_a \to H_k\mathcal{F}_b$ is the homomorphism induced by the inclusion $\mathcal{F}_a \subset \mathcal{F}_b$. Write $H_k^{a,b} \coloneqq \mathrm{im}(\imath_k^{a,b})$. The corresponding k-th *persistent Betti numbers* are the ranks of these groups, $\beta_k^{a,b} = \mathrm{rank} H_k^{a,b}$.

The persistent homology groups $H_k^{a,b}$ consist of homology classes of $\mathcal{F}_a$ that are still alive at $\mathcal{F}_b$, or moreformally, $H_k^{a,b} = Z_k(\mathcal{F}_a)/(B_k(\mathcal{F}_b) \cap Z_k(\mathcal{F}_a))$.

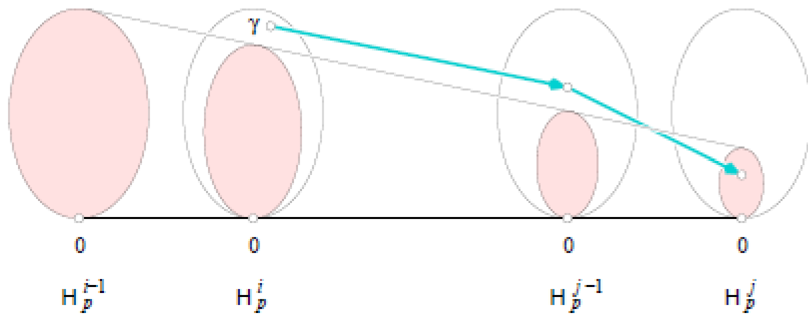


Figure 4: [1, Figure VII.2] The class is γ is born at i since it does not lie in the (shaded) image of H_k^{i-1}. Furthermore, γ dies entering j since this is the first time its image merges into the image of H_k^{i-1}.

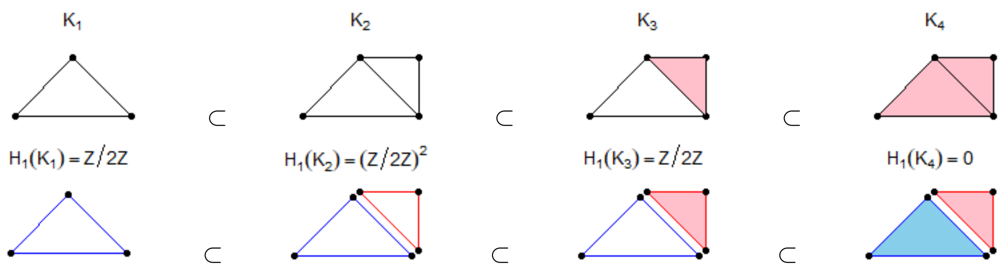


Figure 5: Persistent Homology from subcomplexes of a simplicial complex.

For the k-th persistent homology $PH_k\mathcal{F}$, the set of filtration levels at which a specific homology appears is always an interval $[b,d) \subset [-\infty,\infty]$, i.e. a specific homology is formed at some filtration value b and dies when the inside hole is filled at another value $d > b$. To be more concrete about the classes counted by the persistent homology groups, let γ be a class in $H_k(\mathcal{F}_b)$. We say it is born at $b \in \mathbb{R}$ if $\gamma \notin H_k^{a,b}$ for all $a < b$. Furthermore, if γ is born at b then it dies entering d if it merges with an older class as we go from $\mathcal{F}_c$ to $\mathcal{F}_d$ for any $c \in [b,d)$, i.e., for any $c \in [b,d)$, there exists $a < b$ such that $i_k^{b,c}(\gamma) \notin H_k^{a,c}$ but $i_k^{b,d}(\gamma) \in H_k^{a,d}$. See Figure . In this sense,

We visualize the collection of persistent Betti numbers by drawing points in two dimensions.

Definition (Persistence Diagram)**.** Let $\mathcal{F}$ be a filtration and let $k \in \mathbb{N}_0$. The corresponding k-th persistence diagram $Dgm_k(\mathcal{F})$ is a finite multiset of $(\mathbb{R} \cup \{\infty\})^2$, consisting of all pairs (b,d) where $[b,d)$ is the interval of filtration values for which a specific homology appears in $PH_k\mathcal{F}$. b is called a birth time and d is called a death time.

Example. See Figure .

Example. See Figure . Suppose we want to find a loop structure of a circle, from 20 data points on a circle. We attach disks of radius r to each data point, and increase the radius r from 0 to ∞. When $r = 0.5$, the collection of disks form a loop, and this is the birth time of the loop. When $r = 1$, the inside hole is filled, and this is the death time of the loop. Then we collect birth time and death time of all possible loops, and this is the persistent homology / persistence diagram.

Example. See Figure . Suppose we consider superlevel sets of the kernel density estimator. We decrease level L from ∞ to 0. When $L = 0.15$, you can see that 1-dim hole is formed, and this is the birth time of this loop. And as you decrease L, the inside hole becomes smaller, and when $L = 0$, then inside hole is filled, and this is the death time of this loop. Then we collect birth time and death time of all possible loops, and this is the persistent homology / persistence diagram.

References

[1] Herbert Edelsbrunner and John L. Harer. *Computational topology*. American Mathematical Society, Providence, RI, 2010. An introduction.

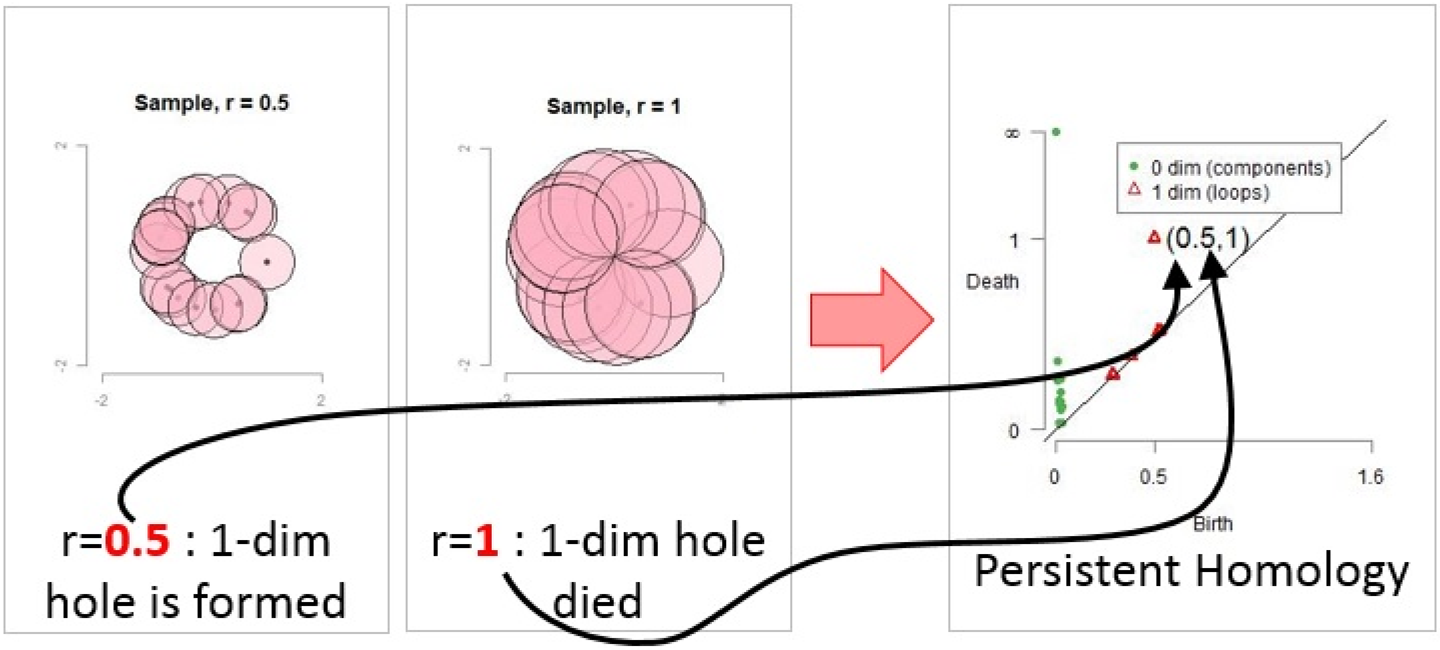


Figure 6: Persistent Homology from Cech filtration.

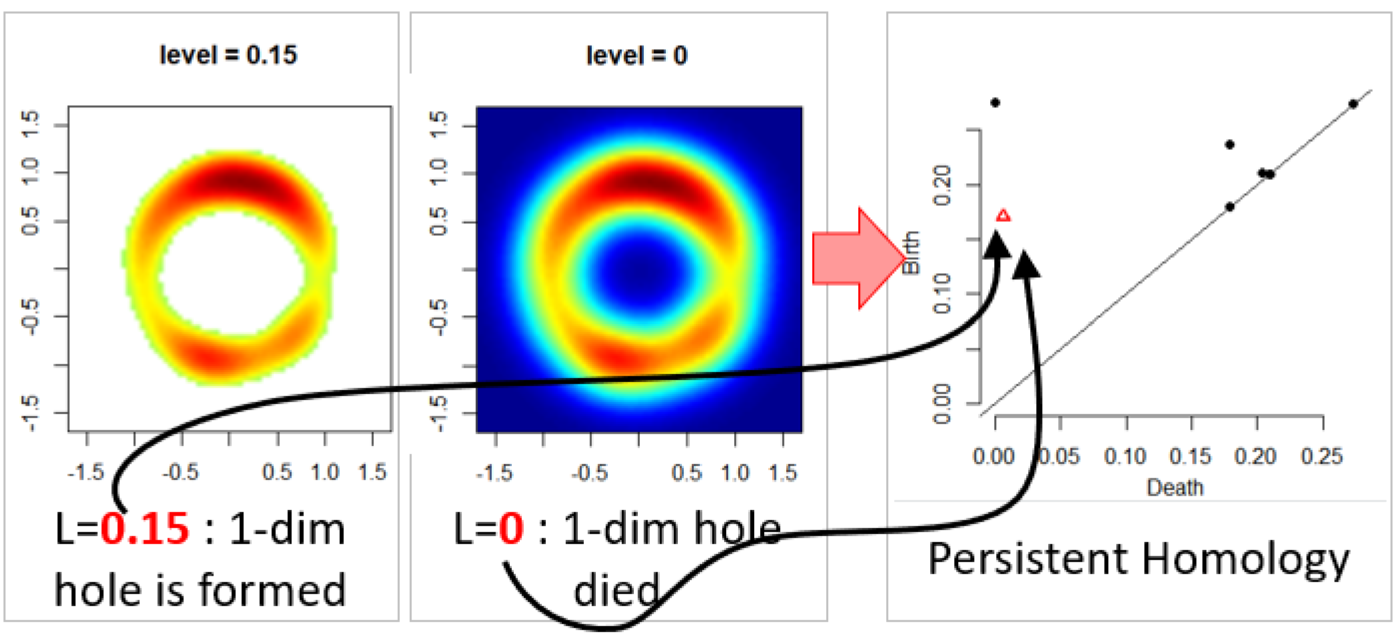


Figure 7: Persistent Homology from Kde filtration.

[2] Allen Hatcher. *Algebraic topology.* Cambridge University Press, Cambridge, 2002.