TOPOLOGICAL DATA ANALYSIS

Dinh Huu Nguyen, 10/2016

Abstract: notes on topological data analysis.

Contents

1. Introduction

In data science so far we have made one crucial assumption that a dataset is drawn from a random variable X. We then use statistics to get its invariants μ_X, σ_X as well as approximations f'_X to its probability density function f_X and use this information in many data science methods.

Example 1.1. (Kernel density estimation) We estimate the probability density function of a feature X

$$f'_X(x) = \frac{1}{h|X|} \sum_{x' \in X} K\left(\frac{d(x, x')}{h}\right)$$

where K is a kernel function and h is a smoothing parameter and use that estimate in many methods.

Example 1.2. (Bayes methods) We make the assumption that the distribution of $X \,|\, Y = 0$ and the distribution of $X \,|\, Y = 1$ are both normal. Then we go on to learn about $\mu_{X \,|\, Y=0}, \mu_{X \,|\, Y=1}, \Sigma_{X \,|\, Y=0}, \Sigma_{X \,|\, Y=1}$. A further assumption about $\Sigma_{X \,|\, Y=0}, \Sigma_{X \,|\, Y=1}$ leads to either naive Bayes, linear discriminant analysis or quadratic discriminant analysis.

Example 1.3. (Generalized linear models) We make the assumption that the distribution of Y is in the exponential family and

$$E(Y \,|\, X = x) = g^{-1}(x\beta)$$

for some link function g. If that assumption is normal distribution and $g(u) = u, g^{-1}(u) = u$ then we get regular regression. If that assumption is Bernoulli distribution and $g(u) = \log_e\left(\frac{u}{1-u}\right), g^{-1}(u) = \frac{1}{1+e^{-u}}$ then we get logistic regression. If that assumption is another distribution and another g then we get another regression.

Now we can make another assumption that a dataset is drawn from a topological space X. We then use topology to get its invariants as well as approximations to its shape and use this information in existing data science methods or in new ones.

Example 1.4. (Generalized linear models revisited) Assuming that $Y \,|\, X = x$ has normal distribution and $g(u) = u, g^{-1}(u) = u$ in example 1.3 is equivalent to assuming that (X, Y) is a hyperplane.

The reason why topology is a good choice for us to use is it has the following properties

- coordinate invariance
- deformation invariance
- compressed representation

2. Topological Data Analysis

Topological data analysis refers to applying topology to develop methods for learning about the aforementioned geometric shape of a dataset. Each method intakes a dataset D, regarded as possibly noisy observations from an unknown topological space X whose topology was lost during sampling and outputs topological objects and topological invariants that quantify the topological features of X. By topological features, we mean

- holes

- clusters
- tendrils
- ...

The first method is persistence homology. Given a topological space X, one builds a filtration of abstract simplicial complexes

$$\ldots \longrightarrow (V_{n-1}, \mathcal{S}_{n-1}) \longrightarrow (V_n, \mathcal{S}_n) \longrightarrow (V_{n+1}, \mathcal{S}_{n+1}) \longrightarrow \ldots$$

converging to the homotopy type of X and feed that into the machinery

$$ASC \xrightarrow{C_\bullet} CMod_R \xrightarrow{H_k} Mod_R$$

to get what is called a persistence homology $\bigoplus_{n\in\mathbb{N}} H_k(C_\bullet(V_n, \mathcal{S}_n))$ for X.

The second method is Mapper. Given a finite metric set X, one finds a map $X \longrightarrow Y$ where Y has a handy cover $\mathcal{U}$ and feed that into the machinery

$$\begin{array}{ccccc} Covers(Y) & \xrightarrow{f^*} & Covers(X) & \xrightarrow{c} & Covers(X) \\ & & & \searrow^{\mathscr{N}_c} & \downarrow N \quad \downarrow MV \\ & & & & ASC \end{array}$$

to get a visual representation for X.

As we vary our choices above, we get different approximations to X. We will develop the above objects and maps in the order they appear.

3. Covers

Let Top be the category whose objects are topological spaces X and whose morphisms are continuous maps between them.

Definition 3.1. An open cover $\mathcal{U}$ for X is a collection $\{U_\alpha\}_{\alpha\in A}$ of open subsets U_α such that $X = \bigcup_{\alpha\in A} U_\alpha$. A good open cover for X is an open cover where every nonempty finite intersection $U_{\alpha_1} \cap \ldots \cap U_{\alpha_n}$ is contractible.

While there are covers of closed subsets and of other types of subsets, the word *cover* is almost synonymous with *open cover* in topology. In this paper we only consider open covers so we will drop the word *open* and only say *cover* when we mean *open cover*.

Naturally a cover $\mathcal{U}$ for X is called a subcover of another cover $\mathcal{U}'$ if $\mathcal{U} \subset \mathcal{U}'$.

The set of all covers $\{U_\alpha\}_{\alpha\in A}$ for X form a category $Covers(X)$ whose morphisms $\{U_\alpha\}_{\alpha\in A} \xrightarrow{\varphi} \{U'_{\alpha'}\}_{\alpha'\in A'}$ are all maps of sets

$$\begin{aligned} A &\xrightarrow{\varphi} A' \\ \alpha &\mapsto \varphi(\alpha) \end{aligned}$$

such that $U_\alpha \subset U'_{\varphi(\alpha)}$ for all $\alpha \in A$. If such φ exists then $\mathcal{U}$ is called a refinement of $\mathcal{U}'$. Surely every subcover is a refinement while the converse may not be true.

The set of all good covers for X form a full subcategory of the category $Covers(X)$

$$CoversGood(X) \longrightarrow Covers(X)$$

Definition 3.2. For each cover $\mathcal{U}$, we define its dimension $\dim(\mathcal{U}) = \max\{k,$ some distinct $U_{\alpha_0} \cap \ldots \cap U_{\alpha_k}$ is nonempty$\}$.

Equivalently, $\dim(\mathcal{U}) = \min\{k,$ all distinct $U_{\alpha_0} \cap \ldots \cap U_{\alpha_{k+1}}$ are empty$\}$.

Example 3.3. For the real line $\mathbb{R}$ and fixed $r > 0, \epsilon > 0$, the collection $\mathcal{U}^{r,\epsilon} = \{U_z^{r,\epsilon}\}_{z\in\mathbb{Z}}$ where $U_z^{r,\epsilon}$ is the open interval $(zr - \epsilon, zr + r + \epsilon)$ form a cover for $\mathbb{R}$. Surely $U_z^{r,\epsilon} \cap U_{z+1}^{r,\epsilon} \neq \emptyset$, so $\dim(\mathcal{U}^{r,\epsilon}) \geq 1$. And if $\epsilon < \frac{r}{2}$ then $U_{z_0}^{r,\epsilon} \cap \ldots \cap U_{z_{k+1}}^{r,\epsilon} = \emptyset$ whenever $k \geq 1$, so $\dim(\mathcal{U}^{r,\epsilon}) \leq 1$. In that case $\mathcal{U}^{r,\epsilon}$ has dimension 1. The Cartesian product of these covers form a corresponding cover for $\mathbb{R}^n$.

Example 3.4. Since $(zr - \epsilon, zr + r + \epsilon) \subset (zr - \epsilon', zr + r + \epsilon')$ for $\epsilon < \epsilon'$, the map

$$\begin{aligned} \mathbb{Z} &\xrightarrow{\varphi} \mathbb{Z} \\ z &\mapsto z \end{aligned}$$

is a map of covers $\{U_z^{r,\epsilon}\}_{z\in\mathbb{Z}} \xrightarrow{\varphi} \{U_z^{r,\epsilon'}\}_{z\in\mathbb{Z}}$ for $\mathbb{R}$.

Example 3.5. Since $(zr - \epsilon, zr + r + \epsilon) \subset (z2r - \epsilon, z2r + 2r + \epsilon)$, the map

$$\begin{aligned} \mathbb{Z} &\xrightarrow{\varphi} \mathbb{Z} \\ z &\mapsto z/2 \end{aligned}$$

is a map of covers $\{U_z^{r,\epsilon}\}_{z\in\mathbb{Z}} \xrightarrow{\varphi} \{U_z^{2r,\epsilon}\}_{z\in\mathbb{Z}}$ for $\mathbb{R}$.

Example 3.6. (open ball cover) If X is a metric space then the collection $\mathcal{U}_{X,\epsilon} = \{U_{v,\epsilon}\}_{v\in X}$ where $U_{v,\epsilon}$ is the open ball $\{x \in X, d(x, v) < \epsilon\}$ form a cover for X. Since $U_{v,\epsilon} \subset U_{v,\epsilon'}$ for $\epsilon < \epsilon'$, the map

$$\begin{aligned} X &\xrightarrow{\varphi} X \\ v &\mapsto v \end{aligned}$$

is a map of covers $\{U_{v,\epsilon}\}_{v\in X} \xrightarrow{\varphi} \{U_{v,\epsilon'}\}_{v\in X}$.

Depending on ϵ, a smaller collection $\mathcal{U}_{V,\epsilon}$ where $V \subset X$ may cover X. In that case it is a subcover, hence a refinement for $\mathcal{U}_{X,\epsilon}$. Its dimension and goodness depend on X, V and ϵ.

Example 3.7. For the unit circle $X = \{e^{i\theta}, 0 \leq \theta \leq 2\pi\}$, the collection

$$\mathcal{U} = \{U_1, U_2\} \text{ where } U_i = \{e^{i\theta}, (i-1)\pi - \delta \leq \theta \leq i\pi + \delta\}$$

form a cover for X that is not good and of dimension 1 while the collection

$$\mathcal{U}' = \{U_1', U_2', U_3'\} \text{ where } U_i' = \{e^{i\theta}, (i-1)\frac{2\pi}{3} - \delta \leq \theta \leq i\frac{2\pi}{3} + \delta\}$$

form a cover for X that is good and of dimension 1.

If we define a metric $d(e^{i\theta}, e^{i\theta'}) = |\theta - \theta'|$ then

$$\mathcal{U} = \mathcal{U}_{V,\epsilon} \text{ where } V = \{e^{i\frac{\pi}{2}}, e^{i\frac{3\pi}{2}}\} \text{ and } \epsilon = \frac{\pi}{2} + \delta$$

while

$$\mathcal{U}' = \mathcal{U}_{V',\epsilon'} \text{ where } V' = \{e^{i\frac{\pi}{3}}, e^{i\frac{3\pi}{3}}, e^{i\frac{5\pi}{3}}\} \text{ and } \epsilon' = \frac{\pi}{3} + \delta$$

If we increase ϵ' to $\frac{2\pi}{3}$ then $U_1' \cap U_2' \cap U_3'$ has 3 connected components so now $\mathcal{U}'$ is not good and of dimension 2.

Example 3.8. (open cell cover) If X is a metric space and $V \subset X$ then the collection $\mathcal{U}_V = \{U_v\}_{v \in V}$ where U_v is the open cell $\{x \in X, d(x,v) \leq d(x,v') \text{ for all } v' \in V\}$ form a cover for X. The open cells U_v are just clusters of points around the *landmark points* v. Their number can be much smaller than the number of open balls in an open ball cover.

The set of all categories $Covers(X), X \in Top$ form a category $Covers$ whose morphisms are covariant functors between those categories. There is a contravariant functor

$$\begin{array}{ccc} Top & \longrightarrow & Covers \\ X & \longmapsto & Covers(X) \\ \Big\downarrow f & & \Big\uparrow f^* \\ Y & \longmapsto & Covers(Y) \end{array}$$

where the covariant functor f^* does

$$\begin{array}{ccc} \{V_\beta\}_{\beta \in B} & \overset{f^*}{\longmapsto} & \{f^{-1}(V_\beta)\}_{\beta \in B} \\ \Big\downarrow \varphi & & \Big\downarrow \varphi \\ \{V'_{\beta'}\}_{\beta' \in B'} & \overset{f^*}{\longmapsto} & \{f^{-1}(V'_{\beta'})\}_{\beta' \in B'} \end{array}$$

This means whenever f is available and Y has some handy covers then X will get some covers. However, f^* does not map good covers to good covers

$$\begin{array}{ccc} Covers(Y) & \xrightarrow{f^*} & Covers(X) \\ \Big\uparrow & & \Big\uparrow \\ CoversGood(Y) & \not\longrightarrow & CoversGood(X) \end{array}$$

unless f is reasonable. In our setting, f is called a reference map and Y is called a reference space.

Example 3.9. For each cover $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ for X, we can define another cover $\mathcal{U}^{\pi_0} = \{U_{\alpha\beta}\}_{\alpha \in A, \beta \in B_\alpha}$ by breaking each $U_\alpha = \bigsqcup_{\beta \in B_\alpha} U_{\alpha\beta}$ up into its connected components. The superscript π_0 is used because the 0^{th} homotopy group $\pi_0(X)$ is the set of connected components of X.

The map

$$\begin{array}{c} A \times \bigsqcup_{\alpha \in A} B_\alpha \xrightarrow{\varphi} A \\ (\alpha, \beta) \mapsto \alpha \end{array}$$

is a map of covers $\mathcal{U}^{\pi_0} \xrightarrow{\varphi} \mathcal{U}$ for X. Thus this procedure defines a covariant functor

$$\begin{array}{ccc} Covers(X) & \xrightarrow{\pi_0} & Covers(X) \\ \uparrow & & \uparrow \\ CoversGood(X) & \xrightarrow{\pi_0} & CoversGood(X) \end{array} \qquad \mathcal{U} \longmapsto \mathcal{U}^{\pi_0}$$

Though it may not map covers to good covers, it does map good covers to good covers.

4. Abstract Simplicial Complexes

We consider the next set of objects.

Definition 4.1. An abstract simplicial complex is a pair $(V, \mathcal{S})$ of finite set $V = \{v_0, \ldots, v_n\}$ and collection $\mathcal{S} = \{\text{nonempty subsets } S \text{ of } V\}$ that is closed under inclusion.

Closure under inclusion means if $S \in \mathcal{S}$ and nonempty $R \subset S$ then $R \in \mathcal{S}$. Clearly $\mathcal{S} = \bigsqcup_{k=0}^{n} \mathcal{S}_k$ where $\mathcal{S}_k = \{S, |S| = k+1\}$ is the set of those subsets of size $k+1$. They are called k-simplices. The 0-simplices $\mathcal{S}_0$ is the set of vertices V.

Naturally an abstract simplicial complex $(V, \mathcal{S})$ is called a subcomplex of another abstract simplicial complex $(V', \mathcal{S}')$ if $V \subset V'$ and $\mathcal{S} \subset \mathcal{S}'$.

The set of all abstract simplicial complexes $(V, \mathcal{S})$ form a category ASC whose morphisms $(V, \mathcal{S}) \xrightarrow{\varphi} (V', \mathcal{S}')$ are all maps of sets

$$\begin{array}{c} V \xrightarrow{\varphi} V' \\ v \mapsto \varphi(v) \end{array}$$

such that $\varphi(S) \in \mathcal{S}'$ for all $S \in \mathcal{S}$. If such φ exists then $(V', \mathcal{S}')$ is called an enrichment of $(V, \mathcal{S})$. Surely every abstract simplicial complex is an enrichment of its subcomplexes while the converse may not be true.

Associated to each abstract simplicial complex $(V, \mathcal{S})$ is the concrete simplicial complex $|V, \mathcal{S}| = \bigcup_{S \in \mathcal{S}} c(S) \subset \mathbb{R}^n$ where $c(S)$ is the convex hull spanned by $e_i, v_i \in S$. In many places to come, when we mention $(V, \mathcal{S})$, we actually refer to $|V, \mathcal{S}|$.

Definition 4.2. For each abstract simplicial complex $(V, \mathcal{S})$, we define its dimension $\dim(V, \mathcal{S}) = \max\{k, \mathcal{S}_k \text{ is nonempty}\}$.

Equivalently, $\dim(V, \mathcal{S}) = \min\{k, \mathcal{S}_{k+1} \text{ is empty}\}$. This dimension is equal to the topological dimension of $|V, \mathcal{S}|$. We connect covers to abstract simplicial complexes.

Definition 4.3. For a cover $\mathcal{U} = \{U_v\}_{v\in V}$ for X, we define its nerve $N(\mathcal{U})$ to be the abstract simplicial complex $(V, \mathcal{S})$ where $S \in \mathcal{S}$ if and only if $\bigcap_{v\in S} U_v \neq \emptyset$.

One can verify that this nerve construction defines a covariant functor

$$\begin{array}{ccc} Covers(X) & \xrightarrow{N} & ASC \\ \mathcal{U} & \longmapsto & N(\mathcal{U}) \\ \downarrow{\scriptstyle\varphi} & & \downarrow{\scriptstyle\varphi} \\ \mathcal{U}' & \longmapsto & N(\mathcal{U}') \end{array}$$

where a map of cover $\mathcal{U} \xrightarrow{\varphi} \mathcal{U}'$ induces the same map of abstract simplicial complexes $N(\mathcal{U}) \xrightarrow{\varphi} N(\mathcal{U}')$. This means if $\mathcal{U}$ is a refinement of $\mathcal{U}'$ then $N(\mathcal{U}')$ is an enrichment of $N(\mathcal{U})$.

Example 4.4. The map of covers $\mathcal{U}_{V,\epsilon} \xrightarrow{\varphi} \mathcal{U}_{V,\epsilon'}$ in example 3.6 induces the same map of abstract simplicial complexes $N(\mathcal{U}_{V,\epsilon}) \xrightarrow{\varphi} N(\mathcal{U}_{V,\epsilon'})$

Here is one way a cover relates to its nerve.

Proposition 4.5. *The dimension of a cover $\{U_\alpha\}_{v\in V}$ is equal to the dimension of its nerve $(V, \mathcal{S})$.*

Proof. It follows immediately from the fact that if some $U_{\alpha_0} \cap \ldots \cap U_{\alpha_k}$ is nonempty then $\mathcal{S}_k$ is nonempty and if all $U_{\alpha_0} \cap \ldots \cap U_{\alpha_{k+1}}$ are empty then $\mathcal{S}_{k+1}$ is empty. □

4.1. **Čech Complex.** We make a distinction and call the nerves $N(\mathcal{U})$ of good covers Čech complexes and denote them by $\check{C}(\mathcal{U})$. That is

$$\begin{array}{ccc} Covers(X) & \xrightarrow{N} & ASC \\ \uparrow & \nearrow_{\check{C}} & \\ CoversGood(X) & & \end{array}$$

A reason for this distinction is the following theorem.

Theorem 4.6. *(Nerve Theorem) If $\mathcal{U}$ is a numerable good cover for X then $\check{C}(\mathcal{U})$ is homotopy equivalent to X.*

It follows that $H^{sin}(\check{C}(\mathcal{U})) \simeq H^{sin}(X)$. Therefore if we can find good covers $\mathcal{U}$ for X then we can compute its homology by computing the homology of $\check{C}(\mathcal{U})$. In the case X is a compact Riemannian manifold, they lie among the open ball covers.

Theorem 4.7. *For each compact Riemannian manifold X there exists an $\epsilon*$ such that $\mathcal{U}_{X,\epsilon^*}$ is a good over cover for X whenever $\epsilon < \epsilon^*$.*

Corollary 4.8. *For each compact Riemannian manifold X there exists an ϵ^* such that $\check{C}(\mathcal{U}_{X,\epsilon})$ is homotopy equivalent to X whenever $\epsilon < \epsilon^*$. Moreover, for such ϵ there exists a finite subset $V \subset X$ such that the subcomplex $\check{C}(\mathcal{U}_{V,\epsilon})$ is also homotopy equivalent to X.*

Proof. The first statement follows from theorem 4.7 and theorem 4.6. □

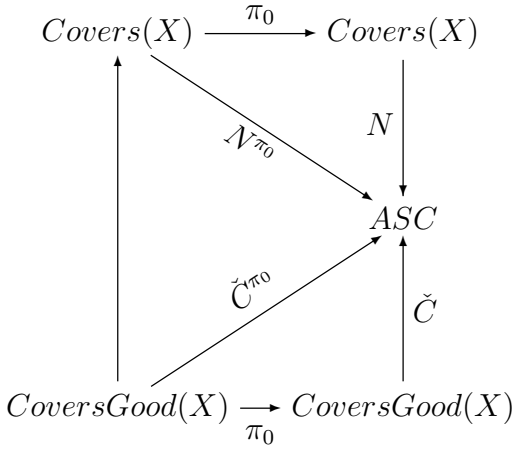
Example 4.9. From example 3.7 and proposition 4.5 we get

- $\mathcal{U}_{V',\frac{\pi}{3}}$ is a good cover for X of dimension 1.
- $\check{C}(\mathcal{U}_{V',\frac{\pi}{3}})$ has dimension 1 and is homotopy equivalent to X.
- $\mathcal{U}_{V',\frac{2\pi}{3}}$ is not a good cover for X of dimension 2.
- $N(\mathcal{U}_{V',\frac{2\pi}{3}})$ has dimension 2 and is not homotopy equivalent to X.

One may guess $\frac{\pi}{3} < \epsilon^* < \frac{2\pi}{3}$.

While straightforward to obtain, a Čech complex $\check{C}(\mathcal{U}_{V,\epsilon})$ may have a lot of vertices and as well as dimension higher than that of X. Thus it is computationally expensive.

Putting the above diagram and the diagram in example 3.9 together, we get



Moreover, the map of covers $\mathcal{U}^{\pi_0} \xrightarrow{\varphi} \mathcal{U}$ after example 3.9 induces maps of abstract simplicial complexes $N(\mathcal{U}^{\pi_0}) \xrightarrow{\varphi} N(\mathcal{U})$ and $\check{C}(\mathcal{U}^{\pi_0}) \xrightarrow{\varphi} \check{C}(\mathcal{U})$ by covariant functoriality of N. Hence we get the natural transformations $N^{\pi_0} \longrightarrow N$ and $\check{C}^{\pi_0} \longrightarrow \check{C}$.

The reason we mention $\check{C}^{\pi_0}$ is that it is even more sensitive than $\check{C}$. Specifically $\check{C}^{\pi_0}(\mathcal{U})$ is homeomorphic to X while $\check{C}(\mathcal{U})$ may just be homotopy equivalent to X. So if X has dimension d then $\mathcal{U}^{\pi_0}$ and $\check{C}^{\pi_0}(\mathcal{U})$ have dimension d as well.

Example 4.10. Writing the unit circle in example 3.7 as $X = \{(x, y), x^2 + y^2 = 1\} \subset \mathbb{R}^2$, one can define a third cover $\mathcal{U}'' = \{U_1'', U_2'', U_3''\}$ where

$$U_1'' = \{(x, y), y < 0\}$$
$$U_2'' = \{(x, y), y > 0\}$$
$$U_3'' = \{(x, y), x \neq 0\}$$

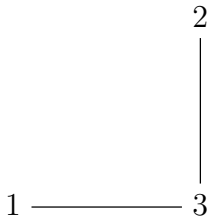
Then $\mathcal{U}''^{\pi_0} = \{U_1'', U_2'', U_{31}'', U_{32}''\}$ where

$$U_1'' = \{(x, y), y < 0\}$$

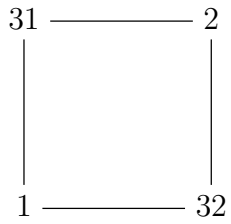
$$U''_2 = \{(x,y), y > 0\}$$
$$U''_{31} = \{(x,y), x < 0\}$$
$$U''_{32} = \{(x,y), x > 0\}$$

One can see that

- $\mathcal{U}''$ is not a good cover of dimension 1.
- $N(\mathcal{U}'') = (V, \mathcal{S})$ where $V = \{1,2,3\}$ and $\mathcal{S} = \{\{1\},\{2\},\{3\},\{1,3\},\{2,3\}\}$ is of dimension 1 and not homotopy equivalent to X. It looks like



- $\mathcal{U}''^{\pi_0}$ is a good cover of dimension 1.
- $N(\mathcal{U}''^{\pi_0}) = \check{C}(\mathcal{U}''^{\pi_0}) = \check{C}^{\pi_0}(\mathcal{U}'') = (A^{\pi_0}, S^{\pi_0})$ where $V^{\pi_0} = \{1,2,31,32\}$ and $\mathcal{S}^{\pi_0} = \{\{1\},\{2\},\{31\},\{32\},\{1,31\},\{1,32\},\{2,31\},\{2,32\}\}$ is of dimension 1 and homeomorphic to X. It looks like



- the map of covers $\mathcal{U}''^{\pi_0} \xrightarrow{\varphi} \mathcal{U}''$ that is

$$A^{\pi_0} \xrightarrow{\varphi} A$$
$$1 \mapsto 1$$
$$2 \mapsto 2$$
$$31 \mapsto 3$$
$$32 \mapsto 3$$

  induces the same map of abstract simplicial complexes $N(\mathcal{U}''^{\pi_0}) \xrightarrow{\varphi} N(\mathcal{U}'')$.

4.2. **Vietoris-Rips Complex.** If we can construct an abstract simplicial complex homotopy equivalent to X without going through a good cover then that will also be useful. To do that, we modify the Čech complex construction a bit. Again assume that X has a metric.

Definition 4.11. We define the Vietoris-Rips complex $VR(X,\epsilon)$ of X attached to ϵ to be the abstract simplicial complex $(X,\mathcal{S})$ where $S = \{v_0,\ldots,v_k\} \in \mathcal{S}$ if and only if $d(v_i,v_j) \leq \epsilon$ for all i,j.

Here is a comparison.

Proposition 4.12. *For $\epsilon < \epsilon'$, there exists a map of abstract simplicial complexes $VR(X,\epsilon) \xrightarrow{\varphi} VR(X,\epsilon')$.*

Proof. Surely $d(v_i, v_j) \leq \epsilon$ implies $d(v_i, v_j) \leq \epsilon'$, se we can take $X \xrightarrow{\varphi} X, x \mapsto x$. □

Here is another comparison.

Proposition 4.13. *For ϵ such that $\mathcal{U}_{X,\epsilon}$ is a good cover for X, we have*

$$\check{C}(X, \epsilon) \xrightarrow{\varphi_\epsilon} VR(X, 2\epsilon) \xrightarrow{\varphi'_\epsilon} \check{C}(X, 2\epsilon)$$

Proof. One can see this by looking at balls of radius ϵ in X and balls whose centers are 2ϵ apart. □

This means the Vietoris-Rips subcomplex is less computationally expensive than the Čech complex and we can use it as a substitute to get the same information on X, though both still have the same set of vertices $\mathcal{S}_0 = X$.

4.3. **Delaunay Complex.** From the open cell cover $\mathcal{U}_V$ in example 3.8, we get another abstract simplicial complex.

Definition 4.14. We define the Delaunay complex $D(V)$ of X attached to V to be the nerve $N(\mathcal{U}_V)$.

While the Čech complex construction and the Vietoris-Rips complex construction above often produce complexes of dimensions higher than that of X, this Delaunay complex construction often produces a complex of dimension equal to the dimension of X. So for finite metric space X, it often produces complexes of dimension 0, with $\mathcal{S}_k = \emptyset, k \geq 1$. The reason is $x \in U_v \cap U_{v'}$ nonempty requires that $d(x, v) = d(x, v')$ and this rarely ever happens.

4.4. **Strong Witness Complex.** We modify the Delaunay complex construction by introducing a little margin ϵ.

Definition 4.15. We define the strong witness complex $W^s(V, \epsilon)$ of X attached to V, ϵ to be the abstract simplicial complex $(V, \mathcal{S})$ where $S = \{v_0, \ldots, v_k\} \in \mathcal{S}$ if and only if there exists an $x \in X$ such that $\max_{v_i \in S}\{d(x, v_i)\} \leq \min_{v \in V}\{d(x, v)\} + \epsilon$.

Such point x acts as a witness to points in S, hence the name strong witness complex.

Definition 4.16. We define the Vietoris-Rips strong witness complex $W^s_{RV}(V, \epsilon)$ of X attached to V, ϵ to be the subcomplex of $W^s(V, \epsilon)$ where

$$\begin{aligned}
&\mathcal{S}_0(W^s_{RV}(V, \epsilon)) = \mathcal{S}_0(W^s(V, \epsilon)) \\
&\mathcal{S}_1(W^s_{RV}(V, \epsilon)) = \mathcal{S}_1(W^s(V, \epsilon)) \\
&\mathcal{S}_k(W^s_{RV}(V, \epsilon)) = \{\{v_0, \ldots, v_k\} \subset V, \{v_i, v_j\} \in \mathcal{S}_1(W^s(V, \epsilon)) \text{ for all } i, j\}
\end{aligned}$$

4.5. **Weak Witness Complex.** We can also weaken the witness requirement in the definition of a strong witness complex a bit.

Definition 4.17. We define the weak witness complex $W^w(V, \epsilon)$ of X attached to V, ϵ to be the abstract simplicial complex $(V, \mathcal{S})$ where $S = \{v_0, \ldots, v_k\} \in \mathcal{S}$ if and only if there exists an $x \in X$ such that $\max_{v_i \in S}\{d(x, v_i)\} \leq \min_{v \in V \setminus S}\{d(x, v)\} + \epsilon$.

Such point x acts as a weak witness to points in S, hence the name weak witness complex.

Definition 4.18. We define the Vietoris-Rips weak witness complex $W^w_{RV}(V,\epsilon)$ of X attached to V,ϵ to be the subcomplex of $W^w(V,\epsilon)$ where

$$\begin{aligned}
&\mathcal{S}_0(W^w_{RV}(V,\epsilon)) = \mathcal{S}_0(W^w(V,\epsilon))\\
&\mathcal{S}_1(W^w_{RV}(V,\epsilon)) = \mathcal{S}_1(W^w(V,\epsilon))\\
&\mathcal{S}_k(W^w_{RV}(V,\epsilon)) = \{\{v_0,\ldots,v_k\}\subset V, \{v_i,v_j\}\in \mathcal{S}_1(W^w(V,\epsilon)) \text{ for all } i,j\}
\end{aligned}$$

Here are more comparisons.

Proposition 4.19. *For $\epsilon < \epsilon'$ and $V \subset X$, there exist maps of abstract simplicial complexes*

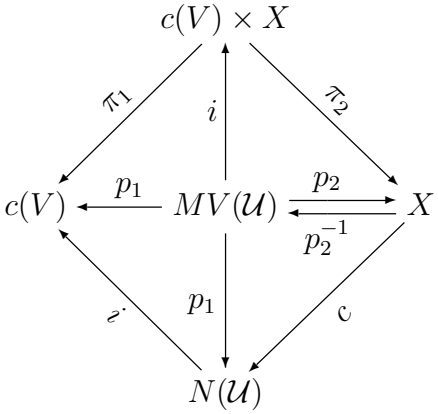
$$\begin{aligned}
W^s(V,\epsilon) &\longrightarrow W^s(V,\epsilon')\\
W^s_{RV}(V,\epsilon) &\longrightarrow W^s_{RV}(V,\epsilon')\\
W^w(V,\epsilon) &\longrightarrow W^w(V,\epsilon')\\
W^w_{RV}(V,\epsilon) &\longrightarrow W^w_{RV}(V,\epsilon')
\end{aligned}$$

Proof. Straightforward from their definitions. □

4.6. **Mayer-Vietoris Blowup.** For each topological space X and its finite cover $\mathcal{U} = \{U_\alpha\}_{v\in V}$, we have the Mayer-Vietoris blowup

$$MV(\mathcal{U}) = \bigcup_{\emptyset\neq S\subset V}\left(c(S)\times\left(\bigcap_{v\in S}U_v\right)\right)$$

where $c(S)\subset\mathbb{R}^n$ is the convex hull spanned by $e_v, v\in S$ as defined earlier. So we have

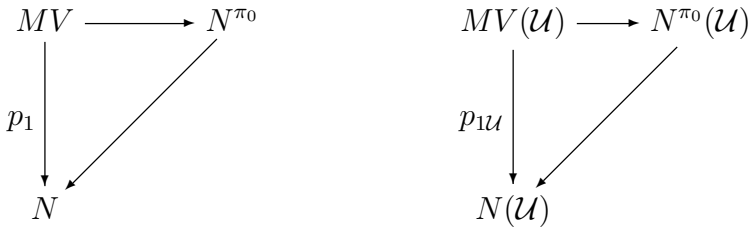


The map p_1 is a homotopy equivalence onto its image $N(\mathcal{U})$. The map p_2 is a homotopy equivalence when X has the homotopy type of a finite complex. Using a partition of unity subordinate to $\mathcal{U}$, one can give its inverse $X \xrightarrow{p_2^{-1}} MV(\mathcal{U})$. The composition $c = p_1p_2^{-1}$ is a kind of coordinatization of X, beside the usual coordinatizations $X \xrightarrow{c} \mathbb{R}^n$.

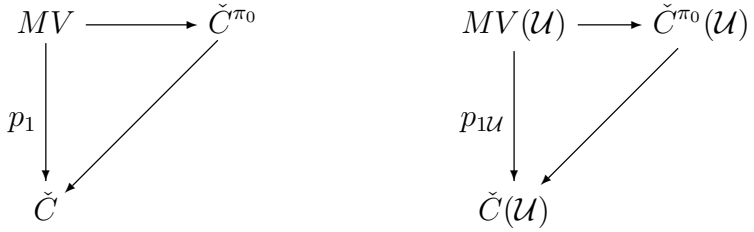
This construction defines a covariant functor

$$\begin{aligned}
Covers(X) &\xrightarrow{MV} ASC\\
\mathcal{U} &\mapsto MV(\mathcal{U})
\end{aligned}$$

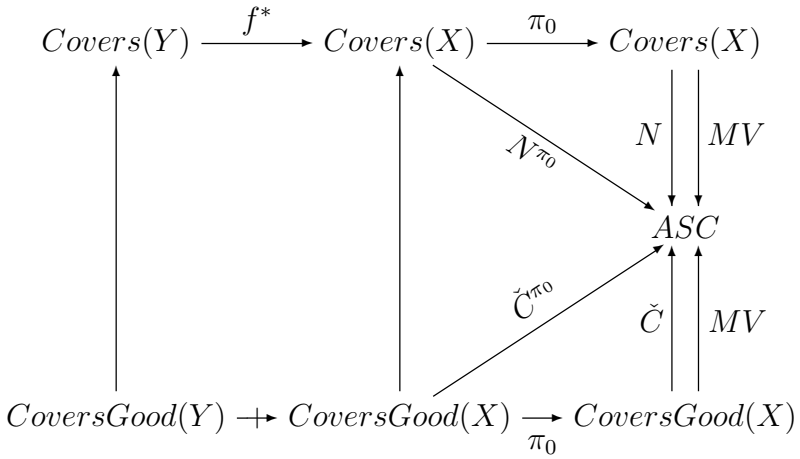
The maps $MV(\mathcal{U}) \xrightarrow{p_{1\mathcal{U}}} N(\mathcal{U})$ above now define a natural transformation $MV \xrightarrow{p_1} N$. Meanwhile, the maps $MV(\mathcal{U}) \longrightarrow N^{\pi_0}(\mathcal{U})$ induced by the projections $U_v \longrightarrow \pi_0(U_v), v \in V$ define a natural transformation $MV \longrightarrow N^{\pi_0}$. Together with the natural transformation $N^{\pi_0} \longrightarrow N$ in the previous section, we get a commutative diagram



When $\mathcal{U}$ is good then just replace N with $\check{C}$ and N^{π_0} with $\check{C}^{\pi_0}$



In summary, we have



5. Clusters

If X is a finite metric space with the discrete topology then a few things above become trivial and we will not get anything useful. For example, the discrete cover $\mathcal{U}^* = \{x\}_{x \in X}$ is the only good cover for X. Or if $\mathcal{U}$ is a cover for X then $\mathcal{U}^{\pi_0} = \mathcal{U}^*$ and $N^{\pi_0}(\mathcal{U}) = (X, \mathcal{S})$ where $\mathcal{S} = \mathcal{S}_0 = X$. So we repeat the above discussion with the following adjustments.

Let FMS be the subcategory of the category Top whose objects are finite metric spaces X and whose morphisms are liner preserving maps between them. Note that FMS is not a full subcategory of Top, and that liner preservation implies injectivity and so these maps are embeddings.

Definition 5.1. A partition $\mathcal{U}$ for X is a collection $\{U_\alpha\}_{\alpha\in A}$ of subsets U_α such that $X = \bigcup_{\alpha\in A} U_\alpha$ and $U_\alpha \cap U_{\alpha'} = \emptyset$ for all $\alpha, \alpha' \in A$.

Naturally a partition $\mathcal{U}$ for X is called a subpartition of another partition $\mathcal{U}'$ if $\mathcal{U} \subset \mathcal{U}'$.

The set of all partitions $\mathcal{U} = \{U_\alpha\}_{\alpha\in A}$ for X form a category $Parts(X)$ whose morphisms $\{U_\alpha\}_{\alpha\in A} \xrightarrow{\theta} \{U'_{\alpha'}\}_{\alpha'\in A'}$ are all maps of sets

$$\begin{aligned} A &\xrightarrow{\varphi} A' \\ \alpha &\mapsto \varphi(\alpha) \end{aligned}$$

such that $U_\alpha \subset U'_{\varphi(\alpha)}$ for all $\alpha \in A$. If such φ exists then $\mathcal{U}$ is called a refinement of $\mathcal{U}'$. Surely every subpartition is a refinement while the converse may not be true.

Under the discrete topology, every subset is open so each partition for X is a cover. Thus the category $Parts(X)$ is a full subcategory of the category $Covers(X)$

$$Parts(X) \longrightarrow Covers(X)$$

The set of all categories $Parts(X), X \in FMS$ form a category $Parts$ whose morphisms are covariant functors between those categories. There is a contravariant functor

$$\begin{array}{ccc} FMS & \longrightarrow & Parts \\ X & \longmapsto & Parts(X) \\ \Big\downarrow f & & \Big\uparrow f^* \\ Y & \longmapsto & Parts(Y) \end{array}$$

where the covariant functor f^* does

$$\begin{array}{ccc} \{V_\beta\}_{\beta\in B} & \xmapsto{f^*} & \{f^{-1}(V_\beta)\}_{\beta\in B} \\ \Big\downarrow \varphi & & \Big\downarrow \varphi \\ \{V'_{\beta'}\}_{\beta'\in B'} & \xmapsto{f^*} & \{f^{-1}(V'_{\beta'})\}_{\beta'\in B'} \end{array}$$

Though we never spoke about cover schemes that would create covers for all $X \in Top$, we speak about clustering schemes that create partitions for $X \in FMS$ here.

Definition 5.2. A map $FMS \xrightarrow{c} \bigsqcup_{X\in FMS} Parts(X), X \mapsto \{U_\alpha\}_{\alpha\in A} \in Parts(X)$ is called a clustering scheme. It is called functorial if for each morphism $X \xrightarrow{f} Y$ there exists a map of partitions $\{U_\alpha\}_{\alpha\in A} \xrightarrow{\varphi} \{f^{-1}(V_\beta)\}_{\beta\in B}$ for X.

So c is functorial if for each morphism $X \xrightarrow{f} Y$ there exists a map of sets $A \xrightarrow{\varphi} B, \alpha \mapsto \varphi(\alpha)$ such that $U_\alpha \subset f^{-1}(V_{\varphi(\alpha)})$, or equivalently $f(U_\alpha) \subset V_{\varphi(\alpha)}$ and the clusters of X map

to subsets of the clusters of Y

$$\begin{array}{ccccc} X & \xmapsto{c} & \{U_\alpha\}_{\alpha\in A} & \xrightarrow{\varphi} & \{f^{-1}(V_\beta)\}_{\beta\in B} \\ \Big\downarrow{\scriptstyle f} & & & & \Big\uparrow{\scriptstyle f^*} \\ Y & \xmapsto{c} & & & \{V_\beta\}_{\beta\in B} \end{array}$$

In particular, if $U \subset U'$ are subsets of X then the clusters of U stay intact inside the clusters of U'. We will need this in section 9.

Example 5.3. (Single Linkage Clustering) For $\epsilon \geq 0$, we define an equivalence relation

$$x \sim x' \text{ if } d(x, x') \leq \epsilon$$

to partition any finite metric space X into clusters $\{U_\alpha\}_{\alpha\in A}$. This equivalence relation defines a clustering scheme

$$\begin{aligned} FMS &\xrightarrow{c_\epsilon} \bigsqcup_{X\in FMS} Parts(X) \\ X &\mapsto \{U_\alpha\}_{\alpha\in A} \end{aligned}$$

It is clear that the clusters $\{U_\alpha\}_{\alpha\in A}$ correspond precisely to the connected components in $VR(X, \epsilon)$. Moreover, if $X \xrightarrow{f} Y$ is a morphism and $x \sim x'$ then $f(x) \sim f(x')$ and the clusters of X map to subsets of the clusters of Y. So this clustering scheme is functorial.

Example 5.4. (DBSCAN) For $\epsilon \geq 0, N > 0$, we say x can reach x' directly if $U_{x,\epsilon} \ni x'$ and $|U_{x,\epsilon}| > N$. Next we say x can reach x' if there exists a chain $x_1, \ldots, x_n$ such that $x_1 = x, x_n = x'$ and x_i can reach x_j directly. Now we define an equivalence relation

$$x \sim x' \text{ if there exists an } x^* \text{ that can reach both } x \text{ and } x'$$

to partition any finite metric space X into clusters $\{U_\alpha\}_{\alpha\in A}$. This equivalence relation defines a clustering scheme

$$\begin{aligned} FMS &\xrightarrow{c_{\epsilon,N}} \bigsqcup_{X\in FMS} Parts(X) \\ X &\mapsto \{U_\alpha\}_{\alpha\in A} \end{aligned}$$

One can verify that this clustering scheme is...

Example 5.5. (k-Means Clustering) One can describe k-means clustering as the clustering scheme

$$\begin{aligned} FMS &\xrightarrow{c} \bigsqcup_{X\in FMS} Parts(X) \\ X &\mapsto \operatorname*{argmin}_{\{U_\alpha\}_{\alpha=1}^k \in Parts(X)} \Big\{ \sum_{\alpha=1}^{k} \sum_{x\in U_\alpha} d(x - c_{U_\alpha})^2 \Big\} \end{aligned}$$

where c_{U_α} is the centroid of U_α. Defining centroid for a subset in $\mathbb{R}^n$ is easy and one can adapt it for general metric spaces as well. One can verify that this clustering scheme is...

Example 5.6. Let c be a functorial clustering scheme. For each cover $\{U_\alpha\}_{\alpha\in A}$ for X, we can define another cover $\mathcal{U}^c = \{U_{\alpha\beta}\}_{\alpha\in A,\beta\in B_\alpha}$ by breaking each $U_\alpha = \bigsqcup_{\beta\in B_\alpha} U_{\alpha\beta}$ into its clusters.

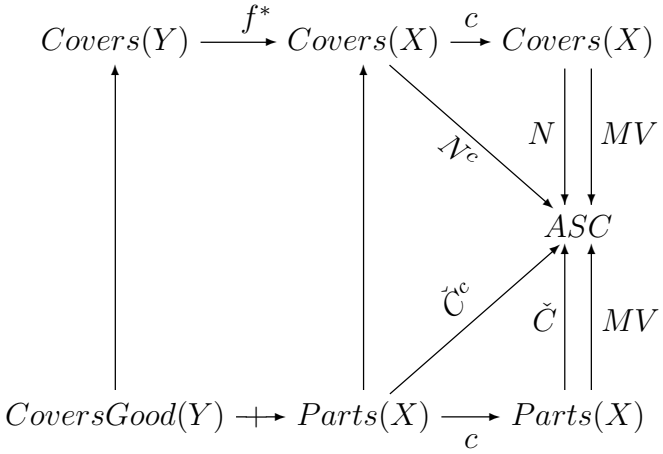
The map

$$\begin{array}{ccc} A\times \bigsqcup_{\alpha\in A} B_\alpha & \xrightarrow{\varphi} & A \\ (\alpha,\beta) & \mapsto & \alpha \end{array}$$

is a map of covers $\mathcal{U}^c \xrightarrow{\varphi} \mathcal{U}$ for X. Thus this procedure defines a covariant functor

$$\begin{array}{ccc} Covers(X) & \xrightarrow{c} & Covers(X) \\ \uparrow & & \uparrow \\ Parts(X) & \xrightarrow{c} & Parts(X) \end{array} \qquad\qquad \mathcal{U} \longmapsto \mathcal{U}^c$$

Though it may not map covers to partitions, it does map partitions to partitions. In summary, we have



6. Simplicial Homology

The theory of simplicial homology is already well developed.

$$\begin{array}{ccccc} ASC & \xrightarrow{C_\bullet} & CMod_R & \xrightarrow{H_k} & Mod_R \\ (V,\mathcal{S}) & \longmapsto & C_\bullet(V,\mathcal{S}) & \longmapsto & H_k(C_\bullet(V,\mathcal{S})) \\ \downarrow f & & \downarrow f & & \downarrow f_* \\ (V',\mathcal{S}') & \longmapsto & C_\bullet(V',\mathcal{S}') & \longmapsto & H_k(C_\bullet(V',\mathcal{S}')) \end{array}$$

Here

- ASC is the category of abstract simplicial complexes.
- $CMod_R$ is the category of chain complexes of R-modules.
- Mod_R is the category of R-modules.
- R is any commutative ring of coefficients.

We provide some details about the map $C_\bullet$ above. For each abstract simplicial complex $(V, \mathcal{S})$, one can construct a chain complex $C_\bullet(V, \mathcal{S})$ of R-modules

$$\ldots \longrightarrow C_k(V,\mathcal{S}) \xrightarrow{\partial_k} C_{k-1}(V,\mathcal{S}) \xrightarrow{\partial_{k-1}} C_{k-2}(V,\mathcal{S}) \longrightarrow \ldots \longrightarrow C_0(V,\mathcal{S}) \longrightarrow 0$$

as follows

- $C_k(V, \mathcal{S}) = F(\mathcal{S}_k, R)$ the free module generated by $\mathcal{S}_k$ over R.
- $\mathcal{S}_k \xrightarrow{\partial_k} \mathcal{S}_{k-1}, \{v_0, \ldots, v_k\} \mapsto \sum_{i=0}^{k} (-1)^i \{v_0, \ldots, v_k\} \backslash \{v_i\}$ and extended linearly.

One can verify that $\partial_k \partial_{k-1} = 0$ so $C_\bullet(V, \mathcal{S})$ is indeed a chain complex. Moreover, using basis $\mathcal{S}_k$ for $C_k(V, \mathcal{S})$ and basis $\mathcal{S}_{k-1}$ for $C_{k-1}(V, S)$ we can represent the linear maps ∂_k as matrices M_k and they will be used in the programmable computation of the homology groups $H_k(C_\bullet(V, \mathcal{S}))$.

Example 6.1. Consider $(V, \mathcal{S})$ where $V = \{v_0, v_1, v_2, v_3\}$ and $\mathcal{S} = \mathcal{S}_0 \bigsqcup \mathcal{S}_1 \bigsqcup \mathcal{S}_2$ where

$$\mathcal{S}_0 = \{v_0, v_1, v_2, v_3\}$$

$$\mathcal{S}_1 = \{\{v_0, v_1\}, \{v_0, v_2\}, \{v_0, v_3\}, \{v_1, v_2\}, \{v_1, v_3\}\}$$

$$\mathcal{S}_2 = \{\{v_0, v_1, v_2\}, \{v_0, v_1, v_3\}\}$$

Then we have

$$\begin{aligned}
\mathcal{S}_1 &\xrightarrow{\partial_1} \mathcal{S}_0 \\
\{v_0, v_1\} &\mapsto v_1 - v_0 \\
\{v_0, v_2\} &\mapsto v_2 - v_0 \\
\{v_0, v_3\} &\mapsto v_3 - v_0 \\
\{v_1, v_2\} &\mapsto v_2 - v_1 \\
\{v_1, v_3\} &\mapsto v_3 - v_1 \\
\mathcal{S}_2 &\xrightarrow{\partial_2} \mathcal{S}_1 \\
\{v_0, v_1, v_2\} &\mapsto \{v_1, v_2\} - \{v_0, v_2\} + \{v_0, v_1\} \\
\{v_0, v_1, v_3\} &\mapsto \{v_1, v_3\} - \{v_0, v_3\} + \{v_0, v_1\}
\end{aligned}$$

If we let $\mathcal{S}_0, \mathcal{S}_1$ and $\mathcal{S}_3$ be bases for $C_0(V, \mathcal{S}), C_1(V, \mathcal{S})$ and $C_2(V, \mathcal{S})$ then

$$M_{\partial_1} = \begin{pmatrix} -1 & -1 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & -1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

$$M_{\partial_2} = \begin{pmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and $M_{\partial_1} M_{\partial_2} = 0$ as expected.

We also provide some details about the map H_k. The fact that $\partial_{k-1}\partial_k = 0$ in any chain complex C means $im(\partial_k) \subset ker(\partial_k)$ and we define

$$H_k(C) = ker(\partial_{k-1})/im(\partial_k)$$

The elements in $ker(\partial_{k-1})$ are called cycles. The elements in $im(\partial_k)$ are called boundaries. The elements in $H_k(C)$ are called homology classes.

Theorem 6.2. *For each abstract simplicial complex $(V, \mathcal{S})$, $H_k(C_\bullet(V, \mathcal{S}))$ is isomorphic to the singular homology $H_k^{sin}(|V, \mathcal{S}|)$ of its associated space $|V, \mathcal{S}|$.*

Theorem 6.2 means if we want to compute the singular homology of X, build an abstract simplicial complex that is homotopy equivalent to X and compute the programmably computable last term

$$H_k^{sin}(X) \simeq H_k^{sin}(|V, \mathcal{S}|) \simeq H_k(C_\bullet(V, \mathcal{S}))$$

7. Persistence Objects

Let POS be the category whose objects are partially ordered sets $\mathcal{P}$ and whose morphisms are order preserving maps between them. Each partially ordered set $\mathcal{P}$ itself can be regarded as a category whose objects are the elements $p \in \mathcal{P}$ and there is a unique morphism $p \longrightarrow p'$ if $p \leq p'$. In this view, the order preserving maps between two partially ordered sets are the functors between them.

Example 7.1. Both $\mathbb{N}$ and $\mathbb{R}$ are partially ordered sets under the usual orderings. For $\epsilon > 0$, the map $\mathbb{N} \xrightarrow{f_\epsilon} \mathbb{R}, n \mapsto n\epsilon$ is an order preserving map between them.

Example 7.2. For a finite metric space X with cover $\mathcal{U}_{X,\epsilon}$, there are only finitely many ϵ such that $N(\mathcal{U}_{X,\epsilon'}) \not\simeq N(\mathcal{U}_{X,\epsilon''})$ for $\epsilon' < \epsilon < \epsilon''$. These are called transition values and indexed $\epsilon_0, \ldots, \epsilon_N$ in increasing order. The map

$$\mathbb{N} \xrightarrow{f} \mathbb{R}$$

$$n \mapsto \begin{cases} \epsilon_n \text{ if } n < N \\ \epsilon_N \text{ otherwise} \end{cases}$$

is an order preserving map.

Definition 7.3. A persistence object from a partially ordered set $\mathcal{P}$ to a category $\mathcal{C}$ is a covariant functor

$$\mathcal{P} \xrightarrow{P} \mathcal{C}$$

$$\begin{array}{ccc} p & \longmapsto & P(p) \\ \downarrow & & \downarrow \psi_{p,p'} \\ p' & \longmapsto & P(p') \end{array}$$

The name *persistence* refers to the perception that some elements in $P(p)$ persist as p increases. The set of all persistence objects from $\mathcal{P}$ to $\mathcal{C}$ form a category $Per(\mathcal{P}, Mod_R)$ whose morphisms are natural transformations between them. The word *object* at this time is just a place holder. Once the category $\mathcal{C}$ is known, the name of the objects in $\mathcal{C}$ will go there.

Example 7.4. The filtration of open intervals

$$\ldots \longrightarrow (-\infty, s) \longrightarrow (-\infty, s') \longrightarrow (-\infty, s'') \longrightarrow \ldots$$

defines a persistence set

$$\begin{array}{c} \mathbb{R} \xrightarrow{P} Open(\mathbb{R}) \\ s \mapsto (-\infty, s) \end{array}$$

If $X \xrightarrow{f} \mathbb{R}$ is a continuous map then it induces the filtration of lower level sets

$$\ldots \longrightarrow f^{-1}((-\infty, s)) \longrightarrow f^{-1}((-\infty, s')) \longrightarrow f^{-1}((-\infty, s'')) \longrightarrow \ldots$$

which defines a persistence set

$$\begin{array}{c} \mathbb{R} \xrightarrow{P} Open(\mathbb{R}) \xrightarrow{f^{-1}} Open(X) \\ s \mapsto (-\infty, s) \mapsto f^{-1}((-\infty, s)) \end{array}$$

Example 7.5. From example 3.5, the filtration of covers for $\mathbb{R}$

$$\ldots \longrightarrow \mathcal{V}^{r,\epsilon} \longrightarrow \mathcal{V}^{2r,\epsilon} \longrightarrow \mathcal{V}^{4r,\epsilon} \longrightarrow \ldots$$

defines a persistence cover

$$\begin{array}{c} \mathbb{N}^+ \xrightarrow{P} Covers(\mathbb{R})) \\ n \mapsto \mathcal{V}^{nr,\epsilon} \end{array}$$

Example 7.6. From example 3.6, the filtration of covers for X

$$\ldots \longrightarrow \mathcal{U}_{V,\epsilon} \longrightarrow \mathcal{U}_{V,\epsilon'} \longrightarrow \mathcal{U}_{V,\epsilon''} \longrightarrow \ldots$$

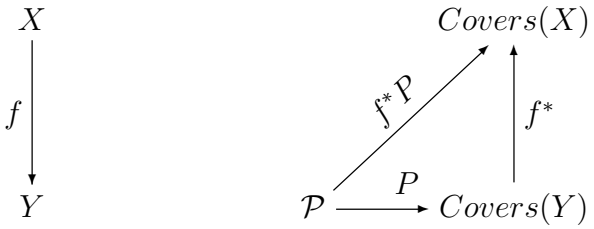
defines a persistence cover

$$\begin{array}{c} \mathbb{R}^+ \xrightarrow{P} Covers(X) \\ \epsilon \mapsto \mathcal{U}_{V,\epsilon} \end{array}$$

And if the ϵ above is less then the ϵ^* in theorem 4.7 then we are looking at a filtration of good covers and a persistence good cover

$$\begin{array}{c} (0, \epsilon^*] \xrightarrow{P} CoversGood(X) \\ \epsilon \mapsto \mathcal{U}_{V,\epsilon} \end{array}$$

Given a persistence cover $\mathcal{P} \xrightarrow{P} Covers(Y)$ and a continuous map $X \xrightarrow{f} Y$, we get a persistence cover f^*P



Example 7.7. From example 7.5, the persistence cover P and a continuous map $X \xrightarrow{f} \mathbb{R}$ give a persistence cover $\mathbb{N}^+ \xrightarrow{f^*P} Covers(X)$.

Given a persistence cover $\mathcal{P} \xrightarrow{P} Covers(X)$, we get persistence complexes

$$\mathcal{P} \xrightarrow{P} Covers(X) \underset{N^{\pi_0}}{\overset{N}{\rightrightarrows}} ASC$$

with natural transformation $N^{\pi_0}P \longrightarrow NP$.

Example 7.8. From the persistence cover P in example 7.5, we get persistence complexes

$$\begin{aligned} \mathbb{N}^+ &\xrightarrow{NP} ASC \\ \epsilon &\mapsto N(\mathcal{V}^{nr,\epsilon}) \end{aligned}$$

$$\begin{aligned} \mathbb{N}^+ &\xrightarrow{N^{\pi_0}P} ASC \\ \epsilon &\mapsto N^{\pi_0}(\mathcal{V}^{nr,\epsilon}) \end{aligned}$$

Example 7.9. From the persistence cover P in example 7.7, we get persistence complexes

$$\begin{aligned} \mathbb{R}^+ &\xrightarrow{NP} ASC \\ \epsilon &\mapsto N(\mathcal{U}_{V,\epsilon}) \end{aligned}$$

$$\begin{aligned} \mathbb{R}^+ &\xrightarrow{N^{\pi_0}P} ASC \\ \epsilon &\mapsto N^{\pi_0}(\mathcal{U}_{V,\epsilon}) \end{aligned}$$

or persistence complexes

$$\begin{aligned} (0,\epsilon^*] &\xrightarrow{\check{C}P} ASC \\ \epsilon &\mapsto \check{C}(\mathcal{U}_{V,\epsilon}) \end{aligned}$$

$$\begin{aligned} (0,\epsilon^*] &\xrightarrow{\check{C}^{\pi_0}P} ASC \\ \epsilon &\mapsto \check{C}^{\pi_0}(\mathcal{U}_{V,\epsilon}) \end{aligned}$$

Example 7.10. From proposition 4.12, the filtration of Vietoris-Rips complexes

$$\ldots \longrightarrow \mathcal{V}^{r,\epsilon} \longrightarrow \mathcal{V}^{r,\epsilon'} \longrightarrow \mathcal{V}^{r,\epsilon''} \longrightarrow \ldots$$

defines a persistence complex

$$\begin{aligned} \mathbb{R}^+ &\xrightarrow{Q} ASC \\ \epsilon &\mapsto \mathcal{V}^{r,\epsilon} \end{aligned}$$

From proposition 4.13, the maps φ_ϵ and φ'_ϵ define natural transformations $NP \xrightarrow{\{\varphi_\epsilon\}_{\epsilon\in\mathbb{R}^+}} Q$ and $NP \xleftarrow{\{\varphi'_\epsilon\}_{\epsilon\in\mathbb{R}^+}} Q$.

Example 7.11. From proposition 4.19, the filtration of strong witness complexes

$$\ldots \longrightarrow W^s(X,\epsilon) \longrightarrow W^s(X,\epsilon') \longrightarrow W^s(X,\epsilon'') \longrightarrow \ldots$$

defines a persistence complex

$$\begin{aligned} \mathbb{R}^+ &\xrightarrow{Q} ASC \\ \epsilon &\mapsto W^s(X,\epsilon) \end{aligned}$$

The same works for Vietoris-Rips strong witness complexes $W^s_{VR}(X,\epsilon)$, weak witness complexes $W^w(X,\epsilon)$, Vietoris-Rips weak witness complexes $W^w_{VR}(X,\epsilon)$.

Given a persistence complex $\mathcal{P} \xrightarrow{P} ASC$, we get a persistence chain complex

$$\mathcal{P} \xrightarrow{P} ASC \xrightarrow{C_\bullet} CMod_R$$

Continuing with this theme, we will get persistence chain complexes and persistence modules

$$\mathcal{P} \xrightarrow{P} Covers(X) \underset{N^{\pi_0}}{\overset{N}{\rightrightarrows}} ASC \xrightarrow{C_\bullet} CMod_R \xrightarrow{H_k} Mod_R$$

As the filtration of covers for X becomes coarser, the filtration of abstract simplicial complexes becomes richer, for our purpose approaching the homotopy type of X and the filtration of homology groups becomes closer to the singular homology group of X.

The set of all categories $Per(\mathcal{P},\mathcal{C}), \mathcal{P} \in POS$ form a category $Per(-,\mathcal{C})$ whose morphisms are covariant functors between those categories. There is a contravariant functor

$$\begin{array}{ccc} POS & \longrightarrow & Per(-,\mathcal{C}) \\ \mathcal{P} & \longmapsto & Per(\mathcal{P},\mathcal{C}) \\ \Big\downarrow{\scriptstyle f} & & \Big\uparrow{\scriptstyle f^*} \\ \mathcal{Q} & \longmapsto & Per(\mathcal{Q},\mathcal{C}) \end{array}$$

where the covariant functor f^* does

$$\begin{array}{ccc} Q & \xmapsto{f^*} & Qf \\ \Big\downarrow{\scriptstyle \nu} & & \Big\downarrow{\scriptstyle \mu} \\ Q' & \xmapsto{f^*} & Q'f \end{array}$$

In details, f^* maps any natural transformation $Q \xrightarrow{\{\nu_q\}_{q\in\mathcal{Q}}} Q'$ to the transformation $Qf \xrightarrow{\{\mu_p\}_{p\in\mathcal{P}}} Q'f$ where $\mu_p = \nu_{f(p)}$.

Example 7.12. Given a persistence object $\mathbb{R} \xrightarrow{Q} \mathcal{C}$ and an order preserving map $\mathbb{N} \xrightarrow{f} \mathbb{R}$, we get a persistence object $\mathbb{N} \xrightarrow{f} \mathbb{R} \xrightarrow{Q} \mathcal{C}$.

Among persistence objects, we care more about the persistence modules $Per(\mathcal{P}, Mod_R)$. In particular, we want to define

$$\begin{aligned} Per(\mathcal{P}, Mod_R) &\xrightarrow{M} Mod_R \\ P &\mapsto M_P = \bigoplus_{p\in\mathcal{P}} P(p) \end{aligned}$$

This M_P is generally incomputable. If $\mathcal{P} = \mathbb{N}$ and $M_P = \bigoplus_{n\in\mathbb{N}} P(n)$ then we can give it the obvious graded $R[x]$-module structure

$$\begin{aligned} P(n) &\xrightarrow{x\cdot} P(n+1) \\ a &\mapsto x \cdot a = \psi_{n,n+1}(a) \end{aligned}$$

where $\psi_{n,n+1}$ comes from

$$\begin{array}{ccc} n & \longmapsto & P(n) \\ \downarrow & & \downarrow{\scriptstyle \psi_{n,n+1}} \\ n+1 & \longmapsto & P(n+1) \end{array}$$

in definition 7.3.

Theorem 7.13. *The categories $Per(\mathbb{N}, Mod_R)$ and $Mod_{R[x]}$ are equivalent.*

Proof. The map

$$\begin{aligned} Per(\mathbb{N}, Mod_R) &\xrightarrow{M} Mod_{R[x]} \\ P &\mapsto \bigoplus_{n\in\mathbb{N}} P(n) \end{aligned}$$

has obvious inverse

$$\begin{aligned} Mod_{R[x]} &\xrightarrow{M^{-1}} Per(\mathbb{N}, Mod_R) \\ \bigoplus_{n\in\mathbb{N}} M_n &\mapsto P \text{ where } P(n) = M_n \end{aligned}$$

□

If $R = \mathbb{Z}$ then we are looking at a graded $\mathbb{Z}[x]$-module. It is still hard to compute because $\mathbb{Z}[x]$ is not a PID. But if $R = F$ a field then $R[x] = F[x]$ is a PID and the structure theorem for finitely generated modules over PID may come in handy.

Definition 7.14. A persistence module $\mathbb{N} \xrightarrow{P} Vect_F$ is called tame if each $P(n)$ is finite-dimensional and $P(n) \xrightarrow{\psi_{n,n+1}} P(n+1)$ is an isomorphism for sufficiently large n.

Proposition 7.15. *A persistence module $\mathbb{N} \xrightarrow{P} Vect_F$ is tame iff its associated graded $F[x]$-module M_P is finitely generated.*

Proof. □

So we have a diagram of equivalent categories and subcategories.

$$\begin{array}{ccc} Per(\mathbb{N}, Vect_F) & \xrightarrow{M} & Mod_{F[x]} \\ \uparrow & & \uparrow \\ PerTame(\mathbb{N}, Vect_F) & \xrightarrow{M|} & ModFin_{F[x]} \end{array}$$

8. Persistence Homology

There is not much left to do here except definition and computation. Given a topological space X, we can use one of the abstract simplicial complex constructions in section 4 to get

$$\mathbb{R} \xrightarrow{P} ASC \xrightarrow{C_\bullet} CMod_R \xrightarrow{H_k} Vect_F$$

The last ingredient is an order preserving map $\mathbb{N} \xrightarrow{f} \mathbb{R}$. If that is defined, we get a persistence module

$$\mathbb{N} \xrightarrow{f} \mathbb{R} \xrightarrow{P} ASC \xrightarrow{C_\bullet} CMod_R \xrightarrow{H_k} Vect_F$$

Definition 8.1. Given a topological space X with a persistence complex $\mathbb{R} \xrightarrow{P} ASC$ and an order preserving map $\mathbb{N} \xrightarrow{f} \mathbb{R}$, we define its persistence module P_X to be $H_k \circ C_\bullet \circ P \circ f$ and its persistence homology $H_k^{per}(X)$ to be M_{P_X}.

This definition of persistence module and persistence homology for X depends on the persistence complex P. If we use one of those filtrations from the abstract simplicial complex constructions in section 4, we may get something better related to X because the filtration is related to the homotopy type of X.

This definition also depends on the order preserving map f. If we use f as in example 7.1 then we are sampling vector spaces at equal intervals. This sampling is finer for smaller ϵ. If we use f as in example 7.2 then we may get something better related to X because the map is related to the structure of X.

Lastly, this definition depends on the field F. In practice, $F = F_2$ is chosen.

8.1. Computation.

Now comes the matter of computation.

Proposition 8.2. *Given a finite metric space X with a persistence complex P and an order preserving map f, its persistence module P_X defined in 8.1 is tame.*

8.1.1. *via Structure Theorem.* Together, proposition 8.2 and proposition 7.15 imply $H_k^{per}(X)$ is a finitely generated graded $F[x]$-module. The structure theorem for finitely generated modules over principal ideal domains then implies

$$H_k^{per}(X) \simeq \bigoplus_{i=1}^{N} x^{t_i} F[x] \oplus \left(\bigoplus_{j=1}^{N'} x^{s_j} F[x]/(x^{r_j}) \right)$$

Of course, this decomposition is unique up to permutation of the summands.

8.1.2. *via Intervals.* Another way to compute a tame persistence module is via intervals. For any interval $I \subset \mathbb{R}$, we define the persistence module

$$\mathbb{R} \xrightarrow{P_I} Vect(F)$$
$$r \mapsto \begin{cases} F \text{ if } r \in I \\ 0 \text{ otherwise} \end{cases}$$

Similarly, for any interval $0 \leq s \leq t \leq \infty$, we define the persistence module

$$\mathbb{N} \xrightarrow{P_{s,t}} Vect(F)$$
$$n \mapsto \begin{cases} F \text{ if } s \leq n \leq t \\ 0 \text{ otherwise} \end{cases}$$

These persistence modules $P_{s,t}$ are tame and their associated modules $M_{P_{s,t}}$ are irreducible finitely generated graded $F[x]$-modules. Moreover, we have the following result.

Proposition 8.3. *Every tame persistence module $P \in PerTame(\mathbb{N}, Vect_F)$ has the following decomposition $P \simeq \bigoplus_{i=1}^{N} P_{s_i,t_i}$ as persistence modules. Correspondingly, $M_P \simeq \bigoplus_{i=1}^{N} M_{P_{s_i,t_i}}$ as graded $F[x]$-modules.*

Together, proposition 8.2 and proposition 8.3 imply

$$H_k^{per}(X) \simeq \bigoplus_{i=1}^{N} M_{P_{s_i,t_i}}$$

Again this decomposition is unique up to permutation of the summands. At this point, one can compute the $M_{P_{s_i,t_i}}$ by using the Smith normal forms related to the matrices M_k in section 6, though I have not actually seen it.

8.2. **Representation by Persistence Barcodes.** The intervals $\{(s_i, t_i)\}_{i=1}^{N}$ in proposition 8.3 draw our persistence barcode. A wide bar corresponds to a homology class that lasts for a long time in the filtration of homology groups, which in turn corresponds to a significant hyperhole in X. On the other hand, a narrow bar corresponds to a homology class that lasts for a short time in the filtration of homology groups, which in turn corresponds to an insignificant hyperhole in X, perhaps one due to noise.

8.3. **Multidimensional Persistence Homology.** It is useful in some cases to consider $P \in Per(\mathbb{N}^m, Vect_F), m > 1$ as well. Then M_P is a a finitely generated m-graded $F[x_1, \ldots, x_m]$-module. There is no structure theorem for finitely generated m-graded $F[x_1, \ldots, x_m]$-modules. In fact, their classification depends on F. However, there is the rank invariant, which has to do with Gröbner basis.

8.4. **Representation by Persistence Diagrams.** some way to represent the set of rank invariants above.

8.5. **Other Persistence Homology.** Still it is useful in other cases to consider $P \in Per(\mathcal{P}, Vect_F)$ where $\mathcal{P}$ could be $\mathbb{N}$ with a zigzag ordering, or where $\mathcal{P}$ could be S^1.

9. Mapper

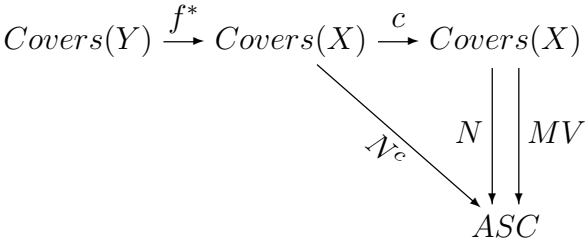
There is not much left to do here except definition and computation.

Definition 9.1. Given a topological space X with cover $\mathcal{U}$, we define its topological Mapper $M^{\pi_0}(\mathcal{U})$ to be $N^{\pi_0}(\mathcal{U})$.

If $\mathcal{U}$ is a numerable good cover for X then theorem 4.6 tells us that $\check{C}(\mathcal{U})$ is homotopy equivalent to X and $M^{\pi_0}(\mathcal{U}) = \check{C}^{\pi_0}(\mathcal{U})$ is homeomorphic to X. And if X is a finite metric space then $M^{\pi_0}(\mathcal{U}) = N^{\pi_0}(\mathcal{U}) = (X, \mathcal{S})$ where $\mathcal{S} = \mathcal{S}_0 = X$, also homeomorphic to X but offering no additional way to learn about X. So we switch to a functorial clustering c defined in definition 5.2.

Definition 9.2. Given a finite metric space X with cover $\mathcal{U}$, we define its statistical Mapper $M^c(\mathcal{U})$ to be $N^c(\mathcal{U})$.

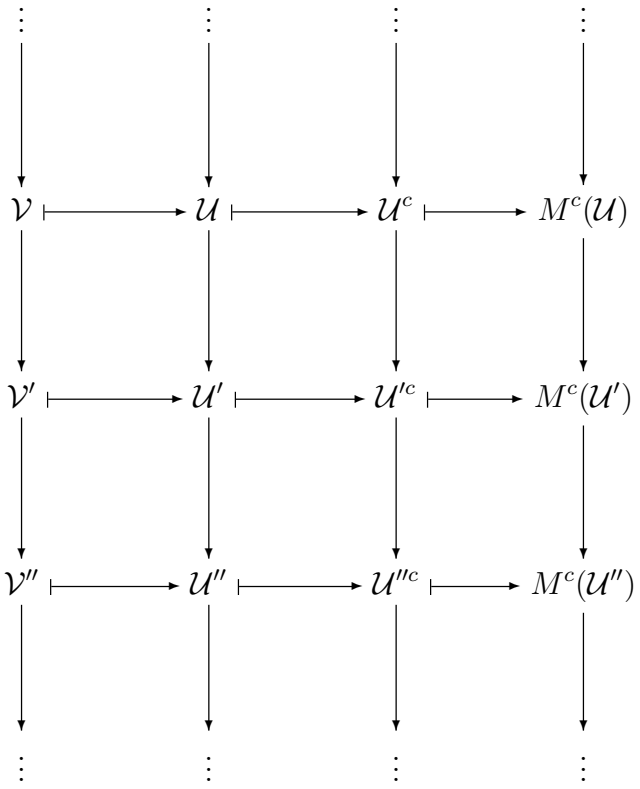
If $\mathcal{U}$ is the good cover $\mathcal{U}^*$ then again $M^c(\mathcal{U}) = N^c(\mathcal{U}) = N(\mathcal{U}^c) = N(\mathcal{U})$ is always homeomorphic to X. If $\mathcal{U}$ is a general cover then whether $M^c(\mathcal{U})$ is relevant to the homeomorphism class or the homotopy type of X or of the topological space $\mathbb{X}$ that X is sampled from depends on both $\mathcal{U}$ and c. As we choose different c, and for each c choose different $\mathcal{U}$, we hope to get different approximations to X. To do this, we return to the diagram



Given a reference map $X \xrightarrow{f} Y$ and a cover $\mathcal{V}$ for the reference space Y, we get a statistical mapper $M^c(\mathcal{U})$

$$\mathcal{V} \xmapsto{f^*} \mathcal{U} \xmapsto{c} \mathcal{U}^c \xmapsto{N} M^c(\mathcal{U})$$

Better yet, given a filtration of covers for the reference space Y, we get a filtration of statistical mappers



As the filtration of covers for Y becomes coarser, the filtration of mappers for X presumably becomes finer as well and we hope to get finer approximations to X. Note that going from the second column to the third column uses functoriality of the clustering algorithm c.

If the filtration of covers for Y is indexed by a partially ordered set $\mathcal{P}$ then we get a persistence cover

$$\mathcal{P} \xrightarrow{P} Covers(Y)$$

which induces a persistence cover

$$\mathcal{P} \xrightarrow{P} Covers(Y) \xrightarrow{f^*} Covers(X) \xrightarrow{c} Covers(X)$$

and a persistence complex

$$\mathcal{P} \xrightarrow{P} Covers(Y) \xrightarrow{f^*} Covers(X) \xrightarrow{c} Covers(X) \xrightarrow{N^c} ASC$$

We hope that the features that persist through these complexes correspond to significant features in X.

Example 9.3. Example 7.9 gives us a persistence complex to study X, with a functorial clustering scheme c replacing π_0. Recall that it comes from a persistence cover for $\mathbb{R}$ and a reference map $X \xrightarrow{f} \mathbb{R}$.

9.1. **Reference Maps for Data.** Since reference maps play an important role in this Mapper method, it is worth considering a few of them.

Example 9.4. Any projection

$$\begin{aligned} X &\longrightarrow \mathbb{R}^k \\ x &\mapsto (x_{i_1}, \ldots, x_{i_k}) \end{aligned}$$

can serve as a reference map.

Example 9.5. Any principal component projection

$$\begin{aligned} X &\longrightarrow \mathbb{R}^k \\ x &\mapsto (x'_{i_1}, \ldots, x'_{i_k}) \end{aligned}$$

can serve as a reference map.

Example 9.6. Any k^{th}-nearest neighbor liner

$$\begin{aligned} X &\xrightarrow{f_k} \mathbb{R} \\ x &\mapsto \frac{k}{|X| d_k(x)} \end{aligned}$$

where

$$\begin{aligned} X &\xrightarrow{d_k} \mathbb{R} \\ x &\mapsto \underset{x \neq x' \in X}{k^{\text{th}}\text{-min}} \{d(x, x')\} \end{aligned}$$

can serve as a nonnegative reference map.

Example 9.7. Any kernel density estimator

$$\begin{aligned} X &\xrightarrow{f} \mathbb{R} \\ x &\mapsto \frac{1}{h|X|} \sum_{x' \in X} K\left(\frac{d(x, x')}{h}\right) \end{aligned}$$

where K is a kernel and h is its smoothing parameter can serve as a nonnegative reference map.

Example 9.8. Any eccentricity measure

$$\begin{aligned} X &\xrightarrow{f_p} \mathbb{R} \\ x &\mapsto \left(\frac{1}{|X|} \sum_{x_i \in X} d(x, x')^p\right)^{\frac{1}{p}} \end{aligned}$$

for $1 \leq p < \infty$ and

$$\begin{aligned} X &\xrightarrow{f_\infty} \mathbb{R} \\ x &\mapsto \max_{x' \in X} \{d(x, x')\} \end{aligned}$$

for $p = \infty$ can serve as a nonnegative reference map. Those points with smallest eccentricity measurements can be thought of as the center of X, while those with large eccentricity measurements can be thought of as far way from the center, hence the name eccentricity measure. The Mapper output with this reference map will reflect this picture.

Example 9.9. Some invariants such as k^{th} moment or k^{th} central moment in statistics.

Choose a reference map that is relevant to your dataset X.

9.2. **Scales.** There is also a desire to vary ϵ by α in the single linkage clustering of $\{f^{-1}(V_\alpha)\}_{\alpha\in A}$ for each $X \xrightarrow{f} Y$ and cover $\{V_\alpha\}_{\alpha\in A}$ for Y in example 5.3. There are of course very many ways to do so. Below is one systematic way devised by Gunnar, based on the observation that the clusters in X correspond precisely to the connected components in $VR(X,\epsilon)$, which in turn relate to $H_0^{per}(C_\bullet(VR(X,\epsilon))) \simeq \bigoplus_{i=1}^{N} M_{P_{s_i,t_i}}$ and its persistence barcode.

Let $B = \{\epsilon_i\}_{i=0}^m$ be the endpoints s_i, t_i of the persistence barcode reordered increasingly and let $I = \{(\epsilon_{i-1}, \epsilon_i)\}_{i=1}^m$ be the intervals. For $\epsilon_{i-1} < \epsilon < \epsilon' < \epsilon_i$, the map $VR(X,\epsilon) \xrightarrow{\varphi} VR(X,\epsilon')$ induces an isomorphism $H_0^{per}(C_\bullet(VR(X,\epsilon))) \xrightarrow{\varphi} H_0^{per}(C_\bullet(VR(X,\epsilon')))$ of the same name, which induces a bijection between the connected components of $VR(X,\epsilon)$ and the connected components of $VR(X,\epsilon')$. For this reason, each $(\epsilon_{i-1},\epsilon_i)$ is called a stability interval for X.

Definition 9.10. For each reference map $X \xrightarrow{f} Y$ and a cover $\mathcal{U} = \{U_\alpha\}_{\alpha\in A}$ for Y, we define an abstract simplicial complex $SS(f,\mathcal{U}) = (V,\mathcal{S})$ where

- $V = A \times I$ where I is the set of all stability intervals for all $f^{-1}(U_\alpha), \alpha \in A$.
- $S = \{(\alpha_0,\iota_0),\ldots,(\alpha_k,\iota_k)\} \in \mathcal{S}$ if
  1. $U_{\alpha_0} \cap \ldots \cap U_{\alpha_k} \neq \emptyset$
  2. $\iota_0 \cap \ldots \cap \iota_k \neq \emptyset$

The map $A\times I \xrightarrow{p_1} A, (\alpha,\iota) \mapsto \alpha$ is a map of abstract simplicial complexes $SS(f,\mathcal{U}) \xrightarrow{p_1} \check{C}(\mathcal{U})$.

Definition 9.11. A scale choice for f and $\mathcal{U}$ is a section $\check{C}(\mathcal{U}) \xrightarrow{s} SS(f,\mathcal{U})$, that is a map $\check{C}(\mathcal{U}) \xrightarrow{s} SS(f,\mathcal{U})$ such that $p_1 s = id_{\check{C}(\mathcal{U})}$.

A scale choice s lets us define

$$\begin{aligned} A &\xrightarrow{\epsilon} \mathbb{R} \\ \alpha &\mapsto \epsilon_\alpha \text{ any } \epsilon_\alpha \in p_2 s(\alpha) \end{aligned}$$

and this is one way that ϵ can vary by α. Though there is a lot of variance in this definition, it has the following merits

- if $U_\alpha \cap U_{\alpha'} \neq \emptyset$ then $p_2 s(\alpha) \cap p_2 s(\alpha') \neq \emptyset$, so ϵ is continuous in a sense.
- one can compare two choices of scale s_1, s_2 by their stability intervals $p_2 s_1(\alpha), p_2 s_2(\alpha), \alpha \in A$.

10. TDA Applications

To turn a dataset into a finite metric space, we can use one of the following metrics.

- for continuous data: try Bray Curtis liner, Canberra liner, correlation liner, cosine liner, L^p liner, Manhattan liner.
- for boolean data: try binary Hamming liner, binary Jaccard dissimilarity, dice dissimilarity, matching dissimilarity, Rogers Tanimoto dissimilarity, Russell Rao dissimilarity.
- for string data: try categorical cosine liner, edit liner, Damerau Levenshtein liner, Hamming liner, Jaccard dissimilarity, Smith Waterman similarity.
- for image data: try image liner.
- for color data: try color liner.

Then we can apply any of the TDA methods to this finite metric space for any purpose we see fit. To apply the method Mapper, we can use one of the following reference maps.

- to see geometry: try projection, kernel density estimator, eccentricity measure.
- to differentiate groups: try principal component projection, k^{th} nearest neighbor liner.
- to identify anomaly: try statistical invariants.

10.1. **Invariants.** using persistence homology. It happens in a few cases, where by looking at the homology groups $H_k^p(X)$ one can tell the shape of our dataset X. It will take quite a bit of insight into the nature of the dataset, plus knowledge of the homology of the usual suspects, plus imagination.

Again the workflow to use persistence homology is

$$\begin{aligned}\text{dataset} &\longrightarrow \text{ finite metric space}\\ &\longrightarrow \text{filtration of complexes}\\ &\longrightarrow \text{persistence module as invariant}\\ &\longrightarrow \text{persistence barcode/persistence diagram for visualization}\end{aligned}$$

Many open source softwares implement this workflow. The C++ package DIPHA is documented in [6]. The R package TDA is documented in [2].

Example 10.1. See [1, example 2.4].

Example 10.2. Persistence homology for the Iris dataset, see [3, 2.2.4].

10.2. **Approximations.** using Mapper. It happens in many cases, where by looking at the filtration $M^c(\mathcal{U})$ one can tell the shape of our dataset X, its persistent features, its clusters and its tendrils.

Again the workflow to use Mapper is

$$\begin{aligned}\text{dataset} &\longrightarrow \text{ finite metric space}\\ &\longrightarrow \text{reference map and filtration of covers for reference space}\\ &\longrightarrow \text{filtration of covers for finite metric space}\\ &\longrightarrow \text{filtration of abstract simplicial complexes as approximations}\\ &\longrightarrow \text{1-skeletons of those abstract simplicial complexes as visualization}\end{aligned}$$

Many open source softwares implement this workflow. The Python package Mapper is documented in [4]. The R package TDAmapper is documented in [5].

If one uses the R package TDAmapper then one gives

- the reference map and its values
- number of intervals of the image
- percentage of overlapping of those intervals
- number of clusters of preimage of each interval

and one gets

- vertices
- samples in each vertex
- color of each vertex tells its level set
- vertices of a color / level set
- samples of a color / level set
- where samples of a label distribute

Example 10.3. cross.

Example 10.4. figure 8

Example 10.5. sphere. Find a bivariate filter for sphere that is more effective than previously tried.

Example 10.6. spirals.

Example 10.7. torus.

Example 10.8. trefoil knot.

11. TDA for Data Science

11.1. **Unsupervised.**

Example 11.1. One popular story is the Miller-Reaven diabetes study, in which Mapper decomposed cancer tumors into clusters, one of which had never been classified before.

Example 11.2. Use machine learning metrics to convert dataset X into a finite metric space.

Example 11.3. Apply persistence homology to some dataset X and see what happens. If we get something like

$$H_k^p(X) = \begin{cases} F \text{ for } k = 0, 2 \\ 0 \text{ otherwise} \end{cases}$$

$$B_k(X) = \begin{cases} 1 \text{ for } k = 0, 2 \\ 0 \text{ otherwise} \end{cases}$$

then X is the 2-sphere. In any case, we get some invariants for X.

Example 11.4. Apply Mapper to some dataset X, color the output by labels $A_1, \ldots, A_k$ and see what it shows.

Example 11.5. Match patterns of features with patterns of dependent feature, which could help select predictive features. See [3, 3.2].

Example 11.6. Can one represent each sample as a geometric object X, and use its persistent homology groups $H_k^p(X)$ as a feature?

11.2. **Supervised.**

Example 11.7. Apply Mapper to some dataset X of classes A, A', color the output by labels $A_1, \ldots, A_k, A'_1, \ldots, A'_{k'}$ and see what it shows.

Example 11.8. (model evaluation) Apply Mapper to some dataset X of class A, color the output by true labels and color it by predicted labels. Where the coloring schemes disagree is where the classifier does poorly.

Example 11.9. (model creation) Apply Mapper to some dataset X of class A, divide the output into regions and create local models for them.

12. Notes to Self

- the geometric shape mentioned in the introduction depends on a choice of a notion of liner/dissimilarity/similarity between the data points. Fortunately, TDA methods will usually produce something useful with just about any decent choice.
- how other clustering schemes than single linkage make a difference in Mapper.
- how to incorporate labels in TDA methods.
- more applications of TDA methods to machine learning.
- more methods in TDA.
- how to apply methods in algebraic geometry to data science, especially coordinate ring, function field, stalk.

References

[1] G. Carlsson, *Topology and data*, Bulletin of The American Mathematical Society **46** (2009), no. 2, 255–308.
[2] B. Fasy, J. Kim, F. Lecci, C. Maria, and V. Rouvreau, *Introduction to the R package TDA* (2014).
[3] H. E. Kim, *Evaluating Ayasdi's topological data analysis for big data* (2015).
[4] D. Müller, *Python Mapper documentation*, available at `danifold.net/mapper`.
[5] P. Pearson, *TDAmapper documentation*, available at `github.com/paultpearson/TDAmapper`.
[6] J. Reininghaus, *A distributed persistent homology algorithm*, available at `github.com/DIPHA/dipha`.