Metric spaces, Covers, and Simplicial Complexes

김지수 (Jisu KIM)

통계이론세미나 - 위상구조의 통계적 추정, 2023 가을학기

The lecture note is largely based on [2].

As topological and geometric features are usually associated with continuous spaces, data represented as finite sets of observations do not directly reveal any topological information. A natural way to highlight some topological structure out of data is to "connect" data points that are close to each other in order to exhibit a global continuous shape underlying the data. Quantifying the notion of closeness between data points is usually done using a distance (or a dissimilarity measure), and it often turns out to be convenient to consider data sets as discrete metric spaces or as samples of metric spaces. This lecture note introduces general concepts for geometric and topological inference

Definition ([6, Section 20])**.** A metric on a set X is a function $d : X \times X \to \mathbb{R}$ having the following properties:

1. $d(x, y) \geq 0$ for all $x, y \in X$; equality holds if and only if $x = y$.
2. $d(x, y) = d(y, x)$ for all $x, y \in X$.
3. (Triangle inequality) $d(x, y) + d(y, z) \geq d(x, z)$ for all $x, y, z \in X$.

Given a metric d on X, the number $d(x, y)$ is often called the distance between x and y. Given $\epsilon > 0$, consider the set $B_X(x, \epsilon) = \{y : d(x, y) < \epsilon\}$ of all points y whose distance from x is less than ϵ. It is called the ϵ-ball centered at x. Sometimes we omit X and write $B(x, \epsilon)$.

Distance between sets on metric spaces

When topological information of the underlying space is approximated by the observed points, it is often needed to compare two sets with respect to their metric structures. Here we present two distances on metric spaces, Hausdorff distance and Gromov-Hausdorff distance.

The *Hausdorff distance* is on sets embedded in the same metric spaces. This distance measures how two sets are close to each other in the embedded metric space. When $S \subset \mathbb{X}$, we denote by $U_r(S)$ the r-neighborhood of a set S in a metric space, i.e. $U_r(S) = \bigcup_{x \in S} \mathbb{B}_{\mathbb{X}}(x, r)$.

Definition (Hausdorff distance [1, Definition 7.3.1])**.** Let $\mathbb{X}$ be a metric space, and $X, Y \subset \mathbb{X}$ be a subset. The *Hausdorff distance* between X and Y, denoted by $d_H(X, Y)$, is defined as

$$d_H(X, Y) \coloneqq \inf \{r > 0 : \ X \subset U_r(Y) \text{ and } Y \subset U_r(X)\}.$$

The Hausdorff distance quantifies the proximity between different data sets issued from the same ambient metric space. However, sometimes one has to compare data sets that are not sampled from the same ambient space. The notion of the Hausdorff distance can be generalized to the comparison of any pair of metric spaces. The *Gromov-Hausdorff distance* measures how two sets are far from being isometric to each other.

Definition ([1, Definition 1.1.3])**.** Let X and Y be two metric spaces. A map $f : X \to Y$ is called distance-preserving if $d_Y(f(x), f(y)) = d_X(x, y)$ for all $x, y \in X$. A bijective distance-preserving map is called an isometry. Two spaces are isometric if there exists an isometry from one to the other.

Definition ([1, Definition 7.3.10])**.** Let X and Y be two metric spaces. The *Gromov-Hausdorff distance* between X and Y, denoted by $d_{GH}(X, Y)$, is defined as

$$d_{GH}(X, Y) \coloneqq \inf \{d_H(X', Y') : \text{there exists a metric space } Z \text{ and } X', Y' \subset Z \text{ with } X, Y \text{ isometric to } X', Y', \text{ respectively.}\}$$

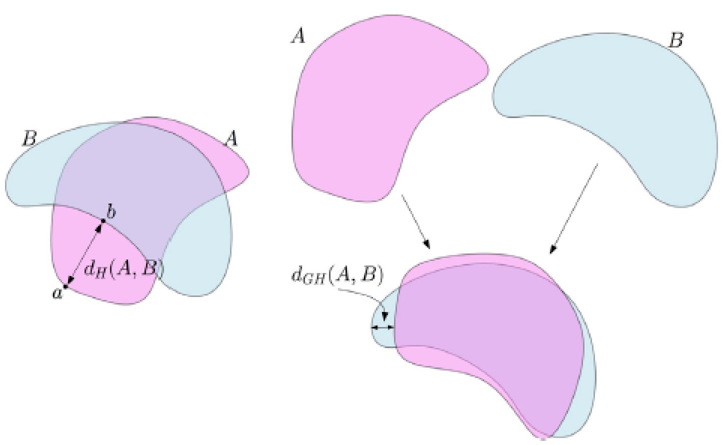

Figure 1: [2, Figure 1] Hausdorff distance $d_H(A, B)$ (left) and Gromov-Hausdorff distance $d_{GH}(A, B)$ between A and B.

See Figure 1 to compare the Hausdorff distance and the Gromov-Hausdorff distance. The Gromov-Hausdorff distance requires constructing a new metric space Z, which is many times cumbersome. More convenient way to compute $d_{GH}(X, Y)$ is by comparing the distance structures of X and Y. For this approach, we first define a relation between two sets called *correspondence*. Roughly speaking, having a correspondence between two sets X and Y means that for every point of X there are one or more "corresponding" points in Y, and vice versa.

Definition ([1, Definition 7.3.17])**.** Let X and Y be two sets. A *correspondence* between X and Y is a set $C \subset X \times Y$ whose projections to both X and Y are both surjective, i.e. for every $x \in X$, there exists $y \in Y$ such that $(x, y) \in C$, and for every $y \in Y$, there exists $x \in X$ with $(x, y) \in C$.

For a correspondence, we define its *distortion* by how the metric structures of two sets differ by the correspondence.

Definition ([1, Definition 7.3.21])**.** Let X and Y be two metric spaces, and C be a correspondence between X and Y. The *distortion* of C is defined by

$$dis(C) = \sup \left\{ |d_X(x, x') - d_Y(y, y')| : \, (x, y), (x', y') \in C \right\}.$$

Now the Gromov-Hausdorff distance is defined as the smallest possible distortion between two sets.

Definition (Gromov-Hausdorff distance [1, Theorem 7.3.25])**.** (equivalent definition) Let X and Y be two metric spaces. The *Gromov-Hausdorff distance* between X and Y, denoted as $d_{GH}(X, Y)$, is defined as

$$d_{GH}(X, Y) = \frac{1}{2} \inf_C dis(C),$$

where the infimum is over all correspondences between X and Y.

Simplicial complex

When inferring topological properties of a metric space $(\mathbb{X}, d)$ (usually a subset of a Euclidean space) from a finite collection $\mathcal{X}$ of observed points from it, we rely on the notion of a *simplicial complex*. A simplicial complex can be seen as a high dimensional generalization of a graph. Given a set V, an *(abstract) simplicial complex* is a set K of finite subsets of V such that $\alpha \in K$ and $\beta \subset \alpha$ implies $\beta \in K$. Each set $\alpha \in K$ is called its *simplex*. The *dimension* of a simplex α is $\dim \alpha = \mathrm{card}\alpha - 1$, and the dimension of the simplicial complex is the maximum dimension of any of its simplices. Note that a simplicial complex of dimension 1 is a graph.

One common choice is the *Vietoris-Rips complex* (or *Rips complex*), where simplexes are built based on pairwise distances among its vertices.

Definition (Vietoris-Rips complex)**.** Let $\mathcal{X}$ be a set of points and $r > 0$. The *Vietoris-Rips complex* $\mathrm{Rips}_{\mathcal{X}}(r)$ is the simplicial complex

$$\mathrm{Rips}_{\mathcal{X}}(r) := \{\sigma \subset \mathcal{X} : d(x_i, x_j) < 2r, \forall x_i, x_j \in \sigma\}. \tag{1}$$

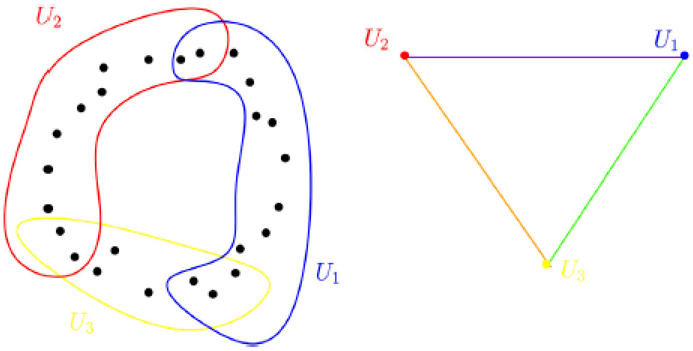

Figure 2: [2, Figure 3] Point cloud and an open cover (left), and the nerve of this cover (right).

Another common choice is the *Čech complex,* defined as below:

Definition (Čech complex)**.** Let $(\mathbb{X}, d)$ be a metric space, $\mathcal{X} \subset \mathbb{X}$ and $r > 0$. The (weighted) Čech complex is the simplicial complex

$$\check{\mathrm{C}}\mathrm{ech}_{\mathcal{X}}^{\mathbb{X}}(r) := \{\sigma \subset \mathcal{X} :\ \cap_{x \in \sigma} \mathbb{B}_{\mathbb{X}}(x, r) \neq \emptyset\}, \tag{2}$$

The superscript $\mathbb{X}$ will be dropped when understood from the context.

Note that from (1) and (2), the Čech complex and Vietoris-Rips complex have the following interleaving inclusion relationship

$$\check{\mathrm{C}}\mathrm{ech}_{\mathcal{X}}(r) \subset R_{\mathcal{X}}(r) \subset \check{\mathrm{C}}\mathrm{ech}_{\mathcal{X}}(2r). \tag{3}$$

In particular, when $\mathbb{X}$ is a subset of $\mathbb{R}^d$, then the constant 2 can be tightened to $\sqrt{\frac{2d}{d+1}}$ [3, Theorem 2.5]:

$$\check{\mathrm{C}}\mathrm{ech}_{\mathcal{X}}(r) \subset R_{\mathcal{X}}(r) \subset \check{\mathrm{C}}\mathrm{ech}_{\mathcal{X}}\left(\sqrt{\frac{2d}{d+1}}r\right). \tag{4}$$

Cover and Nerve Theorem

Definition ([6, Section 26])**.** A collection $\mathcal{A}$ of subsets of a space X is said to cover X, or to be a covering of X, if the union of the elements of $\mathcal{A}$ is equal to X. It is called an open cover of X if its elements are open subsets of X.

The Čech complex is a particular case of a family of complexes associated with covers. We let $\mathcal{U} = \{U_i\}_{i \in I}$ be a cover of $\mathbb{X}$.

Definition. The nerve $Nrv_{\mathcal{U}}$ of $\mathcal{U}$ is the simplicial complex whose vertices are U_i's and

$$Nrv_{\mathcal{U}} := \left\{ \{U_0, \ldots, U_k\} \in \mathcal{U} : \bigcap_{i=0}^{k} U_i \neq \emptyset \right\}. \tag{5}$$

Given a cover of a data set, where each set of the cover can be, for example, a local cluster or a grouping of data points sharing some common properties, its nerve provides a compact and global combinatorial description of the relationship between these sets through their intersection patterns. See Figure 2.

The topology of the nerve is linked to underlying continuous spaces via Nerve Theorem. Under some assumptions, the nerve of a cover is homotopic equivalent to the topology of the union of sets of the cover by the following Nerve Theorem.

Theorem (Nerve Theorem [5, Corollary 4G.3][4, Section III.2])**.** *Let $\mathcal{U} = \{U_i\}_{i \in I}$ be an open cover of a space $\mathbb{X}$ such that for any finite subset $\{U_0, \ldots, U_k\} \subset \mathcal{U}$, the intersection $\bigcap_{i=0}^{k} U_i$ is either empty or contractible. Then, the nerve $Nrv_{\mathcal{U}}$ is homotopic equivalent to $\mathbb{X}$.*

References

[1] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001.

[2] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers Artif. Intell.*, 4:667963, 2021.

[3] Vin de Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7:339–358, 2007.

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

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

[6] James R. Munkres. *Topology.* Prentice Hall, Inc., Upper Saddle River, NJ, 2000. Second edition of [MR0464128].