An introduction to Topological Data Analysis through persistent homology: Intro and geometric inference

Frédéric Chazal
INRIA Saclay - Ile-de-France
frederic.chazal@inria.fr

Some notes related to the course:
J-D. Boissonnat, F. Chazal, M. Yvinec, Computational Geometry and Topology for Data Analysis
http://geometrica.saclay.inria.fr/team/Fred.Chazal/
• Data often come as (sampling of) metric spaces or sets/spaces endowed with a similarity measure with, possibly complex, topological/geometric structure.

• Data carrying geometric information are becoming high dimensional.

• **Topological Data Analysis (TDA):**
  - infer relevant topological and geometric features of these spaces.
  - take advantage of topol./geom. information for further processing of data (classification, recognition, learning, clustering, parametrization...).
Introduction

**Problem:** how to compare topological properties of close shapes/data sets?

- Challenges and goals:
  - no direct access to topological/geometric information: need of intermediate constructions (simplicial complexes);
  - distinguish topological “signal” from noise;
  - topological information may be multiscale;
  - statistical analysis of topological information.
Question: Given an approximation $C$ of a geometric object $K$, is it possible to reliably estimate the topological and geometric properties of $K$, knowing only the approximation $C$?

Challenges:
- define a relevant class of objects to be considered (no hope to get a positive answer in full generality);
- define a relevant notion of distance between the objects (approximation);
- topological and geometric properties cannot be directly inferred from approximations.
Two strategies

1. **Reconstruction**
   + Full reconstruction of the underlying shape.
   + Strong topological and geometric information.
   + A well-developed theory.
   - Strong regularity assumptions.
   - Severe practical/algo issues in high dimensions.

1. **Topological inference**
   + Estimation of topological information without explicit reconstruction.
   + Lighter regularity assumptions.
   + A powerful theory that extends to general data.
   - Weaker information.

This course
A topology on a set $X$ is a family $\mathcal{O}$ of subsets of $X$ that satisfies the three following conditions:

i) the empty set $\emptyset$ and $X$ are elements of $\mathcal{O}$,

ii) any union of elements of $\mathcal{O}$ is an element of $\mathcal{O}$,

iii) any finite intersection of elements of $\mathcal{O}$ is an element of $\mathcal{O}$.

The set $X$ together with the family $\mathcal{O}$, whose elements are called open sets, is a topological space. A subset $C$ of $X$ is closed if its complement is an open set.

A map $f : X \rightarrow X'$ between two topological spaces $X$ and $X'$ is continuous if and only if the pre-image $f^{-1}(O') = \{x \in X : f(x) \in O'\}$ of any open set $O' \subset X'$ is an open set of $X$. Equivalently, $f$ is continuous if and only if the pre-image of any closed set in $X'$ is a closed set in $X$ (exercise).

A topological space $X$ is a compact space if any open cover of $X$ admits a finite subcover, i.e. for any family $\{U_i\}_{i \in I}$ of open sets such that $X = \bigcup_{i \in I} U_i$ there exists a finite subset $J \subseteq I$ of the index set $I$ such that $X = \bigcup_{j \in J} U_j$. 

Background mathematical notions

Metric space

A metric (or distance) on $X$ is a map $d : X \times X \to [0, +\infty)$ such that:

i) for any $x, y \in X$, $d(x, y) = d(y, x)$,

ii) for any $x, y \in X$, $d(x, y) = 0$ if and only if $x = y$,

iii) for any $x, y, z \in X$, $d(x, z) \leq d(x, y) + d(y, z)$.

The set $X$ together with $d$ is a metric space.

The smallest topology containing all the open balls $B(x, r) = \{y \in X : d(x, y) < r\}$ is called the metric topology on $X$ induced by $d$.

Example: the standard topology in an Euclidean space is the one induced by the metric defined by the norm: $d(x, y) = \|x - y\|$.

Compacity: a metric space $X$ is compact if and only if any sequence in $X$ has a convergent subsequence. In the Euclidean case, a subset $K \subset \mathbb{R}^d$ (endowed with the topology induced from the Euclidean one) is compact if and only if it is closed and bounded (Heine-Borel theorem).
Background mathematical notions

**Shapes and Hausdorff distance**

In the first lectures: shape = compact subset of $\mathbb{R}^d$

The **distance function** to a compact $K \subset \mathbb{R}^d$, $d_K : \mathbb{R}^d \to \mathbb{R}_+$ is defined by

$$d_K(x) = \inf_{p \in K} \|x - p\|$$

The **Hausdorff distance** between two compact sets $K, K' \subset \mathbb{R}^d$:

$$d_H(K, K') = \sup_{x \in \mathbb{R}^d} |d_K(x) - d_{K'}(x)|$$

**Exercise:** Show that

$$d_H(K, K') = \max \left( \sup_{y \in K'} d_K(y), \sup_{z \in K} d_{K'}(z) \right)$$
Distance functions and geometric inference

The distance function to a compact $K \subset \mathbb{R}^d$, $d_K : \mathbb{R}^d \to \mathbb{R}_+$ is defined by

$$d_K(x) = \inf_{p \in K} \|x - p\|$$

The Hausdorff distance between two compact sets $K, K' \subset \mathbb{R}^d$:

$$d_H(K, K') = \sup_{x \in \mathbb{R}^d} |d_K(x) - d_{K'}(x)|$$

The idea:

- Replace $K$ and $C$ by $d_K$ and $d_C$
- Compare the topology of the offsets

$$K^r = d_K^{-1}([0, r]) \text{ and } C^r = d_C^{-1}([0, r])$$
Comparing topological spaces

Homeomorphy and isotopy

- $X$ and $Y$ are **homeomorphic** if there exists a bijection $h : X \to Y$ s. t. $h$ and $h^{-1}$ are continuous.

- $X, Y \subset \mathbb{R}^d$ are **ambient isotopic** if there exists a continuous map $F : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ s. t. $F(., 0) = Id_{\mathbb{R}^d}$, $F(X, 1) = Y$ and $\forall t \in [0, 1]$, $F(., t)$ is an homeomorphism of $\mathbb{R}^d$. 
Comparing topological spaces

Homotopy, homotopy type

- Two maps $f_0 : X \to Y$ and $f_1 : X \to Y$ are **homotopic** if there exists a continuous map $H : [0, 1] \times X \to Y$ s.t. $\forall x \in X$, $H(0, x) = f_0(x)$ and $H(1, x) = f_1(x)$.

- $X$ and $Y$ have the **same homotopy type** (or are **homotopy equivalent**) if there exists continuous maps $f : X \to Y$ and $g : Y \to X$ s.t. $g \circ f$ is homotopic to $Id_X$ and $f \circ g$ is homotopic to $Id_Y$. 

Comparing topological spaces

Homotopy, homotopy type

If $X \subset Y$ and if there exists a continuous map $H : [0, 1] \times X \rightarrow X$ s.t.:

i) $\forall x \in X, \ H(0, x) = x$,

ii) $\forall x \in X, \ H(1, x) \in Y$

iii) $\forall y \in Y, \ \forall t \in [0, 1], \ H(t, y) \in Y$,

then $X$ and $Y$ are homotopy equivalent. If one replaces condition iii) by $\forall y \in Y$, $\forall t \in [0, 1], \ H(t, y) = y$ then $H$ is a deformation retract of $X$ onto $Y$. 

\[
f_0(x) = x \quad \quad \quad f_1(x) = 0
\]

\[
f_t(x) = (1-t)x
\]
Medial axis and critical points

\[ \Gamma_K(x) = \{ y \in K : d_K(x) = d(x, y) \} \]

The Medial axis of \( K \):

\[ \mathcal{M}(K) = \{ x \in \mathbb{R}^d : | \Gamma_K(x) | \geq 2 \} \]

\( x \in \mathbb{R}^d \) is a critical point of \( d_K \) iff \( x \) is contained in the convex hull of \( \Gamma_K(x) \).

Exercise: What is the medial axis of a finite set of point \( K = \{ p_1, \ldots, p_n \} \subset \mathbb{R}^d \)?
What are the critical points of \( d_K \)?
The Medial axis of $K$:

$$\Gamma_K(x) = \{ y \in K : d_K(x) = d(x, y) \}$$

$x \in \mathbb{R}^d$ is a critical point of $d_K$ iff $x$ is contained in the convex hull of $\Gamma_K(x)$.

**Theorem:** [Grove, Cheeger,...] Let $K \subset \mathbb{R}^d$ be a compact set.

- if $r$ is a regular value of $d_K$, then $d_K^{-1}(r)$ is a topological submanifold of $\mathbb{R}^d$ of codim 1.

- Let $0 < r_1 < r_2$ be such that $[r_1, r_2]$ does not contain any critical value of $d_K$. Then all the level sets $d_K^{-1}(r)$, $r \in [r_1, r_2]$ are isotopic and

$$K^{r_2} \setminus K^{r_1} = \{ x \in \mathbb{R}^d : r_1 < d_K(x) \leq r_2 \}$$

is homeomorphic to $d_K^{-1}(r_1) \times (r_1, r_2)$.
Reach and weak feature size

The **reach** of $K$, $\tau(K)$, is the smallest distance from $\mathcal{M}(K)$ to $K$:

$$\tau(K) = \inf_{y \in \mathcal{M}(K)} d_K(y)$$

The **weak feature size** of $K$, $\text{wfs}(K)$, is the smallest distance from the set of critical points of $d_K$ to $K$:

$$\text{wfs}(K) = \inf \{ d_K(y) : y \in \mathbb{R}^d \setminus K \text{ and } y \text{ crit. point of } d_K \}$$
"Theorem:" Let $K \subset \mathbb{R}^d$ be such that $\tau = \tau(K) > 0$ and let $C \subset \mathbb{R}^d$ be such that $d_H(K, C) < c\tau$ for some (explicit) constant $c$. Then, for well-chosen (and explicit) $r$, $C^r$ is homotopy equivalent to $K$.

More generally, for compact sets with positive $\mu$-reach ($\text{wfs}(K) \leq r_\mu(K) \leq \tau(K)$):

- Topological/geometric properties of the offsets of $K$ are stable with respect to Hausdorff approximation:
  1. Topological stability of the offsets of $K$ (CCSL’06, NSW’06).
  2. Approximate normal cones (CCSL’08).
  3. Boundary measures (CCSM’07), curvature measures (CCSLT’09), Voronoi covariance measures (GMO’09).