Analyzing the Shape of Data

Construction of Complexes for Persistent Homology

Students: Caroline Ding, Zongze Li

Advisors: Zhixu Su, Chengyuan Ma

Autumn Quarter, 2022
“Shape” of a data set

Here are two data sets with the same mean and covariance, but different “shapes”.

Examples of 3D data set sampled from surfaces:

- torus
- double torus
- the inner ear

How to characterize the “shape” of a data set in terms of its **connectivity** and **hole structures**?
Homology of Simplicial complex

- In topology, the $n$-th **homology** characterizes the $n$-dim hole structure of a space.
- A **simplicial complex** is a collection of simplicies. A $n$-**simplex** is the smallest convex set containing $n + 1$ points, $\sigma = [v_0, \cdots, v_n]$.
- **Simplicial homology** identifies non-trivial $n$-dim holes as $n$-cycles that are not boundary of any $n + 1$ simplicies.

**Betti number** $b_n = \# \text{ of } n$-dim holes

- $b_0 = 2$, 2 connected components
- $b_1 = 3$, 3 loop (1-dim hole)
- $b_2 = 1$, 1 void (2-dim hole)
Persistent homology of filtered simplicial complex

Given a data set $S$, generate a sequence of simplicial complexes $\{K_t\}$ that capture the topological features at different scales $t$. (generated by our Matlab implementation)

Then compute homology of the filtered simplicial complexes and identify the $t$ interval in which each homology cycle persists. (computed by existing implementation JavaPlex)
Vietoris-Rips complex

The Vietoris-Rips complex of a set of points $S$ at scale $t$ is

$$VR_t(S) = \{\sigma \subseteq S : d(v_i, v_j) \leq 2t \text{ for all } v_i, v_j \in \sigma\}.$$

- Two points are connected by a 1-simplex if their distance is $\leq 2t$.
- Three points are connected by a 2-simplex if the distance between every pair of points is $\leq 2t$.

Example. VR complexes of 15 points drawn from a 2D Gaussian distribution (plot generated by our Matlab implementation)

Cons of VR cx: computationally expensive, it generates large number of simplices.
Witness Complex - landmark points

**Motivation:** Generate smaller number of simplices to speed up the construction of filtered simplicial complexes.

**Idea:** Choose a subset of data points, called **landmarks**, that can still capture the shape of the original data set.

**Algorithm:** Sequential MaxMin Method (Farthest-first traversal).

Example. 100 points synthesized from a figure 8 curve, generate 6 landmark points.

\[
\ell_1 = \text{RANDOMIZED-SELECTED-POINT}
\]

\[
\text{for } i = 2, \cdots, k \text{ do}
\]

\[
\ell_i = \arg\max_{v \in S} \left( \arg\min_{j \in \{1, \cdots, i-1\}} d(v, \ell_j) \right)
\]
Witness complex - construction

At each filtration value $t$, two landmarks $\ell_i$ and $\ell_j$ are connected by a 1-simplex if there exists a **witness** point $w$ such that:

$$\max\{d(\ell_i, w), d(\ell_j, w)\} \leq t + \nu(w)$$

where $\nu(w)$ is the distance between $w$ and its nearest landmark point. Three landmarks are connected by a 2-simplex if every pair has been connected.

- When $t = 0$, two landmarks $\ell_i$ and $\ell_j$ are connected by a 1-simplex if there exists a witness point $w$ such that $d(\ell_i, w) = d(\ell_j, w) = \nu(w)$.

Example. Witness complexes of 15 points and 5 landmarks, $t = 0.15, 0.2, 0.25, 0.3$. 
From witness complex to persistent homology

Example. 100 data points synthesized from a figure 8 curve, 6 landmarks points, $t = 0.01, 0.2, 0.4, 0.6$. 
Boundary Matrix - 1-simplices

\[ B(k,j) = 1 \text{ if } [v_k] \in \partial[e_j] \]

\[
\begin{pmatrix}
  e1 & e2 & e3 & e4 & e5 & e6 & e7 & e8 \\
  v1 & 1 & 1 & & & & & \\
  v2 & 1 & & 1 & 1 & & & \\
  v3 & 1 & & & 1 & 1 & & \\
  v4 & 1 & & & 1 & 1 & & \\
  v5 & 1 & 1 & & 1 & & 1 & \\
  v6 & 1 & 1 & & & 1 & & \\
\end{pmatrix}
\]

For each column \( e_j \), \( L(e_j) = \) largest row index of nonzero entry in column \( e_j \)

\[
\text{for column } j = 1 \text{ to } n \text{ do}
\]

\[
\text{while } i < j \text{ with } L(i) = L(j) \text{ do}
\]

\[
\text{add column } i \text{ to column } j
\]

\[
\text{end while}
\]

\[
\text{end for}
\]

\[
\partial[e_3] \xrightarrow{+\partial[e_1]} [v_1] + [v_5] + [v_2] + [v_5]
\]

\[
\partial[e_6] \xrightarrow{+\partial[e_5]} [v_1] + [v_4] + [v_2] + [v_4]
\]

\[
\partial[e_6] \xrightarrow{+\partial[e_5]} [v_1] + [v_2] + [v_1] + [v_2]
\]
**Boundary matrix - reduction and interpretation**

**Reduced boundary matrix:**

\[
L(e_j) = v_i \iff \text{The occurrence of 1-simplex } [e_j] \text{ at time } t \text{ kills the 0-dim cycle (connected component) of } v_i \text{ by connecting it with an earlier point.}
\]

At \( t = 0.01 \), \([e_2]\) kills the component of \([v_6]\) by connecting \([v_6]\) with \([v_1]\).

\[
L(e_j) = \emptyset \iff \text{The occurrence of 1-simplex } [e_j] \text{ at time } t \text{ creates a 1-dim cycle.}
\]

At \( t = 0.2 \), \([e_6]\) creates the 1-dim cycle \([e_1] + [e_3] + [e_5] + [e_6]\) by closing up the loop.
Boundary matrix - 2-simplices

\[
\begin{bmatrix}
  F_1 & F_2 \\
  e_2 & 1 \\
  e_3 & 1 \\
  e_4 & 1 \\
  e_7 & 1 \\
  e_8 & 1 \\
\end{bmatrix}
\rightarrow
\begin{bmatrix}
  F_1 & F_2 \\
  e_2 & 1 \\
  e_3 & 1 \\
  e_4 & 1 \\
  e_7 & 1 \\
  e_8 & 1 \\
\end{bmatrix}
\]

\[L(F_j) = e_i \text{ and } L(e_i) = \emptyset \iff \text{The occurrence of 2-simplex } [F_j] \text{ at time } t \text{ kills the 1-dim cycle created by } [e_i] \text{ by covering the loop.}\]

At \( t = 0.6 \), \([F1]\) kills the 1-dim cycle \([e_2] + [e_4] + [e_8]\) created by \([e_8]\).

At \( t = 0.6 \), \([F2]\) kills the 1-dim cycle \([e_2] + [e_3] + [e_4] + [e_7]\) created by \([e_7]\).
Persistent homology

The “barcode” of each cycle illustrates the time interval $[t_b, t_d]$ from its birth to death. The longer it persists, the more significant the feature is.
Experiment with 3D data synthesized from genus $g$ surfaces

A compact orientable surface of genus $g$ is a connected sum of $g$ copies of tori.

Betti numbers of genus $g$ surface: $b_0 = 1$, $b_1 = 2g$, $b_2 = 1$

• sample synthesized data points from implicit surface equations.
• generate and plot Witness complexes of the data points using our Matlab/Python implementation.
• used JavaPlex to compute persistent homology within “appropriate” maximum filtration value $t$. 
Witness complexes and persistent homology of 3D data points

Example. Witness complexes generated by 10000 data points sampled from

genus 5 surface: \(3 + 8(x^4 + y^4 + z^4) = 8(x^2 + y^2 + z^2)\)

with 300 landmark points generated by Sequential Max-Min.

We choose maximum filtration value around \(t = 0.14\), observing that additional 2-cycles starts to form after \(t = 0.15\).
References