An Introduction to Topological Data Analysis

Elizabeth Munch

Duke University :: Dept of Mathematics

Tuesday, June 11, 2013
“There is no branch of mathematics, however abstract, which may not some day be applied to phenomena of the real world.”

-Nikolai Lobachevsky
Outline

1. Homology & Persistent Homology
2. Statistics
3. Behavioral Clustering and Future Work
Outline

1. Homology & Persistent Homology
2. Statistics
3. Behavioral Clustering and Future Work
What is Homology?

A topological invariant which assigns a sequence of vector spaces, $H_k(X)$, to a given topological space $X$. 
Homology & Persistent Homology

\[ H_0(X) \quad H_1(X) \quad H_2(X) \quad H_3(X) \]
Homology & Persistent Homology

\[ H_0(X) \quad H_1(X) \quad H_2(X) \quad H_3(X) \]

\[ \mathbb{Z}_2 \quad \bullet \quad \bullet \quad \bullet \quad \bullet \]

\[ \vdots \]

\[ \text{Diagram of homology groups} \]
Homology & Persistent Homology

\[
\begin{array}{cccc}
H_0(X) & H_1(X) & H_2(X) & H_3(X) \\
\bullet & \mathbb{Z}_2 & \bullet & \bullet \\
\mathbb{Z}_2 & \mathbb{Z}_2 & \bullet & \bullet \\
\text{surface} & \text{handle} & \bullet & \bullet \\
\end{array}
\]
Homology & Persistent Homology

<table>
<thead>
<tr>
<th>$H_0(X)$</th>
<th>$H_1(X)$</th>
<th>$H_2(X)$</th>
<th>$H_3(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
</tr>
<tr>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
</tr>
<tr>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
</tr>
</tbody>
</table>
### Homology & Persistent Homology

<table>
<thead>
<tr>
<th>$H_0(X)$</th>
<th>$H_1(X)$</th>
<th>$H_2(X)$</th>
<th>$H_3(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
</tr>
<tr>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2 \times \mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
<td>$\mathbb{Z}_2$</td>
</tr>
</tbody>
</table>
Homology & Persistent Homology

\[
\begin{array}{cccc}
H_0(X) & H_1(X) & H_2(X) & H_3(X) \\
\bullet & \mathbb{Z}_2 & \bullet & \bullet \\
\circ & \mathbb{Z}_2 & \mathbb{Z}_2 & \bullet \\
\bigcirc & \mathbb{Z}_2 & \bullet & \mathbb{Z}_2 \\
\bigcirc & \mathbb{Z}_2 & \mathbb{Z}_2 \times \mathbb{Z}_2 & \mathbb{Z}_2 \\
\bigcirc & \mathbb{Z}_2 & \mathbb{Z}_2 \times \mathbb{Z}_2 & \mathbb{Z}_2 \\
\end{array}
\]
What is Persistent Homology?

A way to watch how the homology of a filtration (sequence) of topological spaces changes so that we can understand something about the space.
What is Persistent Homology?

A way to watch how the homology of a filtration (sequence) of topological spaces changes so that we can understand something about the space.
What is Persistent Homology?

A way to watch how the homology of a filtration (sequence) of topological spaces changes so that we can understand something about the space.
What is Persistent Homology?

A way to watch how the homology of a filtration (sequence) of topological spaces changes so that we can understand something about the space.
What is Persistent Homology?

A way to watch how the homology of a filtration (sequence) of topological spaces changes so that we can understand something about the space.
Homology & Persistent Homology

What is Persistent Homology?
A way to watch how the homology of a filtration (sequence) of topological spaces changes so that we can understand something about the space.
What is Persistent Homology?

A way to watch how the homology of a filtration (sequence) of topological spaces changes so that we can understand something about the space.
What is Persistent Homology?

A way to watch how the homology of a filtration (sequence) of topological spaces changes so that we can understand something about the space.
Filtration

Given topological space $K$ and filtration

$$K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_n$$

gives a sequence of maps on homology

$$H_2(K_0) \rightarrow H_2(K_1) \rightarrow H_2(K_2) \rightarrow \cdots \rightarrow H_2(K_n)$$
Birth and Death

**Birth**
A class $\gamma \in H_p(K_i)$ is born at $K_i$ if it is not in the image of

$$H_p(K_{i-1}) \rightarrow H_p(K_i)$$

**Death**
A class $\gamma$ dies entering $K_j$ if it merges with an older class.
Simplicial Complex

Simplices

Simplicial Complex
Boundary Matrix Data Structure

![Diagram of a triangle with labeled points and a matrix labeled D]
Large Data Sets
Large Data Sets
Large Data Sets
### Čech Complex

- Set of points $\chi \subset \mathbb{R}^n$
- Čech complex $\mathcal{C}$
- $\sigma \in \mathcal{C}$
  \[ \iff \bigcap B_r(V_i) \neq \emptyset \]

### Rips Complex

- Set of points $\chi \subset \mathbb{R}^n$
- Rips complex $\mathcal{R}$
- $\sigma \in \mathcal{R}$
  \[ \iff \| v_i - v_j \| \leq r \text{ for all } v_i, v_j \in \sigma \]
Persistence Algorithm

\[ D \]
Persistence Algorithm

\[ R = D \]
\[ V = I \]

for \( j = 1 \) to \( m \):
    while \( \exists j_0 < j \) with \( \text{low}(j_0) = \text{low}(j) \):
        add column \( j_0 \) to column \( j \)
Visualizing Birth and Death

Elizabeth Munch (Duke)
CompTop-SUNYIT
Tuesday, June 11, 2013
Large Data Sets
Wasserstein Distance on $D_p$

$p$-Wasserstein distance for diagrams

Given diagrams $X$ and $Y$, the distance between them is

$$W_p[L_q](X, Y) = \inf_{\varphi : X \to Y} \left( \sum_{x \in X} (\|x - \varphi(x)\|_q)^p \right)^{1/p}.$$
Stability

Theorem (Cohen-Steiner, Edelsbrunner, Harer, Mileyko)

Let \( \mathbb{X} \) be a triangulable metric space that implies bounded degree \( k \) total persistence, for \( k \geq 1 \), and let \( f, g : \mathbb{X} \to \mathbb{R} \) be two tame Lipschitz functions. Then

\[
W_p(f, g) \leq C^{1/p} \cdot \|f - g\|_\infty^{1-k/p}
\]

for all \( p \geq k \), where \( C = C_X \max\{\text{Lip}(f)^k, \text{Lip}(g)^k\} \).
Properties of $D_p$

**Mileyko, Mukherjee, Harer**
- Complete
- Separable
- Characterization of Compact Sets

**Turner, Mileyko, Mukherjee, Harer**
- Non-negatively curved Alexandrov space
- Geodesics
Homology & Persistent Homology

Statistics

Behavioral Clustering and Future Work
Statistics

How do we give a summary of the data?

Will it play nicely with time varying persistence diagrams?
How do we give a summary of the data?
Will it play nicely with time varying persistence diagrams?
Given a probability space \((D_p, \mathcal{B}(D_p), \mathcal{P})\), the quantity

\[
\text{Var}_\mathcal{P} = \inf_{X \in D_p} \left[ F_{\mathcal{P}}(X) = \int_{D_p} W_{\mathcal{P}}(X, Y)^2 \, d\mathcal{P}(Y) < \infty \right]
\]

is the Fréchet variance of \(\mathcal{P}\).

The set at which the value is obtained

\[
\mathbb{E}(\mathcal{P}) = \{X | F_{\mathcal{P}}(X) = \text{Var}_\mathcal{P}\}
\]

is the Fréchet expectation, also called Fréchet mean.
Properties of the Fréchet mean on $D_p$

**Theorem (Mileyko et al.)**

Let $\mathcal{P}$ be a probability measure on $(D_p, \mathcal{B}(D_p))$ with a finite second moment. If $\mathcal{P}$ has compact support, then $\mathbb{E}(\mathcal{P}) \neq \emptyset$.

**Theorem (Mileyko et al.)**

Let $\mathcal{P}$ be a tight probability measure on $(D_p, \mathcal{B}(D_p))$ with the rate of decay at infinity $q > \max\{2, p\}$. Then $\mathbb{E}(\mathcal{P}) \neq \emptyset$. 
Goal

Given diagrams $X_1, \ldots, X_N$, find a diagram $Y$ which minimizes the Fréchet function.
Example
Example
**Algorithm for Computation - Outline**

**Algorithm - Kate Turner et al.**

1. **Persistence Diagrams**
   \[ X_1, \cdots, X_N \]

2. Pick one at random to start,
   \[ Y = X_i \]

3. Repeat:
   - Find best matching for Wasserstein distance
     \[ W_p(Y, X_i) \]
   - Create matching \( G \) with a selection for each point \( y_j \in Y \) with the points
     \( x_i \in X_i \) for each \( i \) paired with \( y_j \)
   - Replace \( Y \) with \( \text{mean}_X(G) \)
Theorem (Turner et al.)

The algorithm terminates at a local minimum of the Fréchet function.
The Problem with Pointwise Fréchet means

![Diagram showing points a, 1, b, and 2 on a line.](image)
The Problem with Pointwise Fréchet means
Instead of a single diagram, return a distribution on diagrams!
Instead of a single diagram, return a distribution on diagrams!

Main Idea:

1. Perturb the diagrams and compute the matching.
The Solution

Instead of a single diagram, return a distribution on diagrams!

Main Idea:

1. Perturb the diagrams and compute the matching.
2. Do this repeatedly to get a distribution on matchings.
The Solution

Instead of a single diagram, return a distribution on diagrams!

Main Idea:

1. Perturb the diagrams and compute the matching.
2. Do this repeatedly to get a distribution on matchings.
3. Associate the probability of getting a particular matching to the mean of the matching.
The Solution

Instead of a single diagram, return a distribution on diagrams!

**Main Idea:**

1. Perturb the diagrams and compute the matching.
2. Do this repeatedly to get a distribution on matchings.
3. Associate the probability of getting a particular matching to the mean of the matching.

- Limit to $S_{M,K} \subset D_p$: inside triangle of height $M$ with at most $K$ off-diagonal points.
The Drawing Procedure - Make $X'_i$

- Pick $\alpha$
- Pick distribution $\eta_x$ with mean at $x$ and support contained in $B_\alpha(x)$.
- For each $x \in X_i$, make $X'_i$ by:
  1. Draw point from $\eta_x$
  2. If contained in $B_{\|x-\Delta\|}(x)$, add it to $X'_i$.

![Diagram of the drawing procedure](image)
The Drawing Procedure - Determine Matching
The Drawing Procedure - Determine Matching

\[
\begin{pmatrix}
\Delta & y' & g' \\
\Delta & \Delta & h' \\
\Delta & z' & \Delta \\
\end{pmatrix}.
\]
The Drawing Procedure - Determine Matching

\[ \begin{pmatrix} d' & d' & d' \\ b' & x' & f' \\ \Delta & y' & g' \\ \Delta & \Delta & h' \\ \Delta & z' & \Delta \end{pmatrix} \]

\[ \begin{pmatrix} d' & d' & d' \\ b & x & f \\ \Delta & y & g \\ \Delta & \Delta & h \\ \Delta & z & \Delta \\ a & \Delta & \Delta \\ c & \Delta & \Delta \end{pmatrix} \]
The Drawing Procedure - Create Distribution

\[ \mu_X = \sum_G \mathbb{P}(G) \, \delta_{\text{mean}_X}(G) \]
Definition

\[ \mu_X = \sum_G \mathbb{P}(G) \delta_{\text{mean}_X}(G) \]
Continuity

**Theorem (Munch et al)**

The map

\[
S_{M,K} \times \cdots \times S_{M,K} \rightarrow \mathcal{P}(S_{M,K})
\]

\[
X_1, \cdots, X_N \mapsto \mu_X
\]

is continuous.
Examples
1 Homology & Persistent Homology

2 Statistics

3 Behavioral Clustering and Future Work
How do we automate behavioral analysis?
Behavior Vectors

Main Idea:
Quantify a particular behavior as a vector in $\mathbb{R}^D$ and cluster these points.

Example:
- Black: $\langle 6.52, 0, 0, 0 \rangle$
- Purple: $\langle 6.63, 0, 0, 0 \rangle$
- Orange: $\langle 5.01, 4.85, 0, 0 \rangle$
- Red: $\langle 8.20, 6.46, 0, 0 \rangle$
- Blue: $\langle 21.17, 17.60, 13.26, 7.54 \rangle$
- Green: $\langle 15.40, 11.78, 8.28, 7.12 \rangle$
Behavior Vectors

Main Idea:
Quantify a particular behavior as a vector in $\mathbb{R}^D$ and cluster these points.

Example:
Loop time vector: Put an entry into a list for each loop in a track, sort, and compare.

Black: $(6.52, 0, 0, 0)$
Purple: $(6.63, 0, 0, 0)$
Orange: $(5.01, 4.85, 0, 0)$
Red: $(8.20, 6.46, 0, 0)$
Blue: $(21.17, 17.60, 13.26, 7.54)$
Green: $(15.40, 11.78, 8.28, 7.12)$
Behavior Vectors

Main Idea:
Quantify a particular behavior as a vector in \( \mathbb{R}^D \) and cluster these points.

Example:
Loop time vector: Put an entry into a list for each loop in a track, sort, and compare.

Black: \((6.52, 0, 0, 0)\)
Purple: \((6.63, 0, 0, 0)\)
Orange: \((5.01, 4.85, 0, 0)\)
Red: \((8.20, 6.46, 0, 0)\)
Blue: \((21.17, 17.60, 13.26, 7.54)\)
Green: \((15.40, 11.78, 8.28, 7.12)\)
Dendrograms
Experimental Results

Experiment

- Tracks generated using open-source SUMO software.
- Additional tracks forced to make small circles, larger circles, and figure eights.
- Loops computed by:
  1. checking for crossings, looking at time needed to complete loop.
  2. computing persistence, using lifetime of class as score.
Experimental Results - NonPersistence

![Graph showing experimental results for NonPersistence with categories: Random, Small Loop, Large Loop, and Figure 8.](image-url)
Experimental Results - Persistence

![Graph showing persistence diagrams for Random, Small Loop, Large Loop, and Figure 8 categories.](image-url)
Group Formation
Acknowledgements

Funding:
DDR&E

Honors:
Jo Rae Wright Fellowship for Outstanding Women in Science, 2012-2013

Collaborators:
John Harer
Paul Bendich
Kate Turner (UChicago)
Sayan Mukherjee
Jonathan Mattingly
Ingrid Daubachies
Robert Calderbank
Peter Sang Chin (APL/JHU)