Topological Data Analysis

July 25, 2013
What is Topological Data Analysis?

**Definition**

Topological data analysis (TDA) is a new mathematical theory with applications in areas such as data mining and computer vision. The main problems are:
What is Topological Data Analysis?

**Definition**

Topological data analysis (TDA) is a new mathematical theory with applications in areas such as data mining and computer vision. The main problems are:

- How one infers high-dimensional structure from low-dimensional representations; and
What is Topological Data Analysis?

Definition

Topological data analysis (TDA) is a new mathematical theory with applications in areas such as data mining and computer vision. The main problems are:

- How one infers high-dimensional structure from low-dimensional representations; and
- How one assembles discrete points into global structure.
What it does with pictures

<table>
<thead>
<tr>
<th>Number</th>
<th>Name</th>
<th>Type</th>
<th>SurfWeight</th>
<th>BulkWeight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>David Shaw</td>
<td>red</td>
<td>189</td>
<td>165</td>
</tr>
<tr>
<td>2</td>
<td>Amelia Serrano</td>
<td>yellow</td>
<td>145</td>
<td>124</td>
</tr>
<tr>
<td>3</td>
<td>Alan Price</td>
<td>red</td>
<td>220</td>
<td>192</td>
</tr>
<tr>
<td>4</td>
<td>Ravi Singh</td>
<td>yellow</td>
<td>194</td>
<td>177</td>
</tr>
<tr>
<td>5</td>
<td>Ashley McNight</td>
<td>red</td>
<td>127</td>
<td>128</td>
</tr>
<tr>
<td>6</td>
<td>Gil Brown</td>
<td>yellow</td>
<td>220</td>
<td></td>
</tr>
</tbody>
</table>

Original Data

Point Cloud

Barcode Metadata

VISSTA

TDA
Homology

Topological invariant from algebraic topology
Homology
Topological invariant from algebraic topology
Homology

Topological invariant from algebraic topology

\[ C_1 \]

\[ C_2 \]
Homology

Topological invariant from algebraic topology
The “first” homology of a manifold is connected to the “fundamental group” of a manifold (it is its abelianization). One can play similar tricks with higher dimensional “voids” instead of just 2-dimensional holes.
The “first” homology of a manifold is connected to the “fundamental group” of a manifold (it is its abelianization).
The “first” homology of a manifold is connected to the “fundamental group” of a manifold (it is its abelianization).

One can play similar tricks with higher dimensional “voids” instead of just 2 dimensional holes.
Simplicial Complex approximation of a space:
Simplicial Complex approximation of a space:

- We use “simple pieces” (simplices): points, edges, triangles, tetrahedra etc (0,1,2,3 dimensional simplices respectively),
Simplicial Complex approximation of a space:

- We use “simple pieces” (simplices): points, edges, triangles, tetrahedra etc (0,1,2,3 dimensional simplices respectively).

- It is completely described by specifying how to glue these pieces together (edges with vertices, triangles with edges, etc) and
Simplicial Complex approximation of a space:

- We use “simple pieces” (simplices): points, edges, triangles, tetrahedra etc (0,1,2,3 dimensional simplices respectively).
- It is completely described by specifying how to glue these pieces together (edges with vertices, triangles with edges, etc) and
- The “gluing” may be expressed algebraically with linear operators.
We construct vector spaces $C_k$ with basis the $k$ dimensional simplices over some field (reals for example).
Homology

- We construct vector spaces $C_k$ with basis the $k$ dimensional simplices over some field (reals for example).

- We define boundary operators $\partial_k : C_k \to C_{k-1}$, which are basically linear maps (algebra) corresponding to the “gluing” (topology).
We construct vector spaces $C_k$ with basis the $k$ dimensional simplices over some field (reals for example).

We define boundary operators $\partial_k : C_k \rightarrow C_{k-1}$, which are basically linear maps (algebra) corresponding to the “gluing” (topology).

ex: $\partial_1(e_1) = v_2 - v_1$ and $\partial_2(\sigma_1) = e_1 + e_2 + e_3$
Notice that $\partial_1 (e_1 + e_2 + e_3) = v_2 - v_1 + v_3 - v_2 + v_1 - v_3 = 0$. 

In general for any "cycle" $\sigma$ 

$\partial_1 (\sigma) = 0$ 

The range of $\partial_2$ gives us the boundaries of "filled-in" spaces. Holes are "cycles which are not boundaries", algebraically captured as homology spaces: 

$H_k = \ker(\partial_k) / \text{Img}(\partial_{k+1})$ 

For instance $\alpha = (e_4 + e_5 - e_1)$ is a cycle that is not a boundary. One can show that this is the only element thus $H_1 = \langle \alpha \rangle$. 

\[ v_2 - v_1 + v_3 - v_2 + v_1 - v_3 = 0 \]
Notice that $\partial_1(e_1 + e_2 + e_3) = v_2 - v_1 + v_3 - v_2 + v_1 - v_3 = 0$

In general for any “cycle” $\sigma$
$$\partial_1(\sigma) = 0$$
Notice that $\partial_1(e_1 + e_2 + e_3) = v_2 - v_1 + v_3 - v_2 + v_1 - v_3 = 0$

In general for any “cycle” $\sigma$

$\partial_1(\sigma) = 0$

The range of $\partial_2$ gives us the boundaries of “filled-in” spaces
Notice that $\partial_1 (e_1 + e_2 + e_3) = v_2 - v_1 + v_3 - v_2 + v_1 - v_3 = 0$

In general for any “cycle” $\sigma$

$\partial_1 (\sigma) = 0$

The range of $\partial_2$ gives us the boundaries of “filled-in” spaces

Holes are “cycles which are not boundaries”, algebraically captured as homology spaces:

$H_k = \ker(\partial_k)/\text{img}(\partial_{k+1})$
Notice that $\partial_1 (e_1 + e_2 + e_3) = v_2 - v_1 + v_3 - v_2 + v_1 - v_3 = 0$

In general for any “cycle” $\sigma$

$\partial_1 (\sigma) = 0$

The range of $\partial_2$ gives us the boundaries of “filled-in” spaces

Holes are “cycles which are not boundaries”, algebraically captured as homology spaces:

$H_k = \ker(\partial_k) / \text{Img}(\partial_{k+1})$

For instance $\alpha = (e_4 + e_5 - e_1)$ is a cycle that is not a boundary. One can show that this is the only element thus $H_1 = \langle \alpha \rangle$
Summing up!

Definition

Simplicial Homology with field coefficients $\mathbb{F}$, is a “functor”, $H : \{\text{Simplicial Complexes}\} \rightarrow \{\text{Vector Spaces over } \mathbb{F}\}$. In other words, the input is a simplicial complex and the output is a family of vector spaces, one for each dimension.
Filtration

A simplicial complex is a collection of simplices with the property that two simplices are disjoint or intersect on a simplex (common face).

A subcomplex of $K$ is a subset $L \subseteq K$ which is also a simplicial complex.

A filtration of a complex $K$ is a nested sequence of subcomplexes $\emptyset = K_{-1} \subseteq K_0 \subseteq K_1 \subseteq ... \subseteq K_m = K$.

VISSTA TDA
Filtration

A simplicial complex is a collection of simplices with the property that two simplices are disjoint or intersect on a simplex (common face).
Filtration

A simplicial complex is a collection of simplices with the property that two simplices are disjoint or intersect on a simplex (common face).

A subcomplex of $K$ is a subset $L \subseteq K$ which is also a simplicial complex.
A simplicial complex is a collection of simplices with the property that two simplices are disjoint or intersect on a simplex (common face).

A subcomplex of $K$ is a subset $L \subseteq K$ which is also a simplicial complex.

A filtration of a complex $K$ is a nested sequence of subcomplexes

$$\emptyset = K^{-1} \subseteq K^0 \subseteq K^1 \subseteq \ldots \subseteq K^m = K$$
Start with any data set

<table>
<thead>
<tr>
<th></th>
<th>Feat 1</th>
<th>Feat 2</th>
<th>Feat 3</th>
<th>Feat 4</th>
<th>Feat 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u_1 )</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>( u_2 )</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>( u_3 )</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>( u_4 )</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( u_5 )</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>( u_6 )</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>23</td>
</tr>
<tr>
<td>( u_7 )</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>( u_8 )</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>( u_9 )</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>( u_{10} )</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>
## Calculate the distance matrix

### Example

<table>
<thead>
<tr>
<th></th>
<th>$u_1$</th>
<th>$u_2$</th>
<th>$u_3$</th>
<th>$u_4$</th>
<th>$u_5$</th>
<th>$u_6$</th>
<th>$u_7$</th>
<th>$u_8$</th>
<th>$u_9$</th>
<th>$u_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>9</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>$u_2$</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>$u_3$</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>$u_4$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>$u_5$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>$u_6$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>$u_7$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Choose a threshold \((M=3)\) and “connect” vertices as follows:
Choose a threshold \((M=3)\) and “connect” vertices as follows:

If two vertices are closer than \(M\) connect them with an edge.
Choose a threshold \((M=3)\) and “connect” vertices as follows:

- If two vertices are closer than \(M\) connect them with an edge.
- If three vertices are pairwise closer than \(M\) “fill” in the triangle they create.
Choose a threshold \((M=3)\) and “connect” vertices as follows:

- If two vertices are closer than \(M\) connect them with an edge.
- If three vertices are pairwise closer than \(M\) “fill” in the triangle they create.
- Don’t stop there ... do it for all combinations by increasing the dimensions of simplices to tetrahedra etc...
Homology

One then can use the "homology" tool that we described before to analyze this point cloud. Each basis element for the vector space $H_n(X; F)$ corresponds to a different $n$-dimensional hole of your simplicial complex $X$. The problem is which threshold should we use?
One then can use the “homology” tool that we described before to analyze this point cloud.
Homology

- One then can use the “homology” tool that we described before to analyze this point cloud.
- Each basis element for the vector space $H_n(X; \mathbb{F})$ corresponds to a different $n$-dimensional hole of your simplicial complex $X$. 
Homology

- One then can use the “homology” tool that we described before to analyze this point cloud.
- Each basis element for the vector space $H_n(X; \mathbb{F})$ corresponds to a different $n$-dimensional hole of your simplicial complex $X$.
- The problem is which threshold should we use?
And the answer is ...
And the answer is ... 

ALL OF THEM!
What happens if we increase $M$?

For a fixed value of $M$ we have the corresponding simplicial complex as described before. If we increase $M$ we increase the number of potential connections and get a new simplicial complex. This simplicial complex contains all the previous connections (simplices) and maybe some more.
What happens if we increase $M$?

- For a fixed value of $M$ we have the corresponding simplicial complex as described before.
What happens if we increase $M$?

- For a fixed value of $M$ we have the corresponding simplicial complex as described before.
- If we increase $M$ we increase the number of potential connections and get a new simplicial complex.
What happens if we increase $M$?

- For a fixed value of $M$ we have the corresponding simplicial complex as described before.
- If we increase $M$ we increase the number of potential connections and get a new simplicial complex.
- This simplicial complex contains all the previous connections (simplices) and maybe some more.
Definition of persistence

**Definition**

Suppose that $K_M$ is a nested sequence of simplicial indexed by a parameter $M$. i.e., there exist inclusion maps $f : K_M \rightarrow K_{M+1}$ creating the following sequence:

$$K_0 \xrightarrow{f} K_1 \xrightarrow{f} \ldots \xrightarrow{f} K_M \xrightarrow{f} K_{M+1} \xrightarrow{f} \ldots \xrightarrow{f} K_N$$
These maps give rise to natural homomorphisms for the p-dimensional homologies:

\[ 0 \rightarrow H_p(K_0) \xrightarrow{f_*} H_p(K_1) \xrightarrow{f_*} \ldots \xrightarrow{f_*} H_p(K_M) \xrightarrow{f_*} \ldots \xrightarrow{f_*} H_p(K_N) \]
Example using our pointcloud

M=4
M=5
M=6
Example using our point-cloud

M=4

M=5

M=6

Each simplicial complex is a subcomplex of the next!
Follow the “holes”, mark “births” and “deaths”!
Follow the “holes”, mark “births” and “deaths”!

Look at each simplicial complex. Calculate its homology and find the elements that correspond to holes, voids etc. Mark the creation of holes (birth).
Follow the “holes”, mark “births” and “deaths”!

- Look at each simplicial complex. Calculate its homology and find the elements that correspond to holes, voids etc. Mark the creation of holes (birth).

- Track those changes as you move to the next simplicial complex. Calculate which holes “persist” and which ones get “filled in” (death).
Follow the “holes”, mark “births” and “deaths”!

- Look at each simplicial complex. Calculate its homology and find the elements that correspond to holes, voids etc. Mark the creation of holes (birth).
- Track those changes as you move to the next simplicial complex. Calculate which holes “persist” and which ones get “filled in” (death).
- Do this process for the whole filtration and obtain a “barcode” $\mathcal{B} = \{[b_i, d_i]\}$ of births and deaths.
What you end up with is:

**Example: Dimension 1**

**Example: Dimension 0**
One can then say ...

Looking at the zero dimensional homology (which represents the connected components—clusters) that your data forms one persistent cluster.

Looking at the first dimensional homology (which represents the 1 dimensional holes) that your data has 2 big holes!
One can then say ...

- Looking at the zero dimensional homology (which represents the connected components-clusters) that your data forms one persistent cluster.
One can then say ...

- Looking at the zero dimensional homology (which represents the connected components-clusters) that your data forms one persistent cluster.
- Looking at the first dimensional homology (which represents the 1 dimensional holes) that your data has 2 big holes!
One can then ...

- Distinguish this set of data from others.
- Discover clusters in the data that you were not aware of.
- Correlate data with each other.
- Who knows what else!

VISSTA

TDA
One can then ...

- Distinguish this set of data from others.
One can then ...

- Distinguish this set of data from others.
- Discover clusters in the data that you were not aware of.
One can then ...

- Distinguish this set of data from others.
- Discover clusters in the data that you were not aware of.
- Correlate data with each other.
One can then ...

- Distinguish this set of data from others.
- Discover clusters in the data that you were not aware of.
- Correlate data with each other.
- Who knows what else!
THE END!
THE END!
Or maybe the beginning...
Homology spaces

Detailed example

Basis for $C_0$:

$$\{ v_1, v_2, v_3, v_4 \}$$
Homology spaces
Detailed example

Basis for $C_0$:

$$\{v_1, v_2, v_3, v_4\}$$

Basis for $C_1$:

$$\{e_1, e_2, e_3, e_4, e_5\}$$
Basis for $C_0$:

$$\{v_1, v_2, v_3, v_4\}$$

Basis for $C_1$:

$$\{e_1, e_2, e_3, e_4, e_5\}$$

Basis for $C_2$:

$$\{\sigma_1\}$$
Homology spaces
Detailed example

\[ \partial_1 = \begin{pmatrix} e_1 & e_2 & e_3 & e_4 & e_5 \\ v_1 & 1 & 0 & -1 & 1 & 0 \\ v_2 & -1 & 1 & 0 & 0 & -1 \\ v_3 & 0 & -1 & 1 & 0 & 0 \\ v_4 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \]
## Homology spaces

### Detailed example

The boundary operator $\partial_1$ is given by:

$$
\begin{bmatrix}
1 & 0 & -1 & 1 & 0 \\
-1 & 1 & 0 & 0 & -1 \\
0 & -1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 1 \\
\end{bmatrix}
$$

After column reduction:

$$
\begin{bmatrix}
e_1 & e_2 & z_1 & e_4 & z_2 \\
v_1 & 1 & 0 & 0 & 1 & 0 \\
v_2 & -1 & 1 & 0 & 0 & 0 \\
v_3 & 0 & -1 & 0 & 0 & 0 \\
v_4 & 0 & 0 & 0 & -1 & 0 \\
\end{bmatrix}
$$

$z_1 = e_1 + e_2 + e_3$, $z_2 = e_4 + e_5 - e_1$
Homology spaces

Detailed example

\[ \partial_1 = \begin{pmatrix} e_1 & e_2 & e_3 & e_4 & e_5 \\ v_1 & 1 & 0 & -1 & 1 & 0 \\ v_2 & -1 & 1 & 0 & 0 & -1 \\ v_3 & 0 & -1 & 1 & 0 & 0 \\ v_4 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \]

after column reduction:

\[ \begin{pmatrix} e_1 & e_2 & e_3 & e_4 & e_5 \\ v_1 & 1 & 0 & 0 & 1 & 0 \\ v_2 & -1 & 1 & 0 & 0 & 0 \\ v_3 & 0 & -1 & 0 & 0 & 0 \\ v_4 & 0 & 0 & 0 & -1 & 0 \end{pmatrix} \]

\[ \ker(\partial_1) \text{ is spanned by } \{ z_1 = e_1 + e_2 + e_3, z_2 = e_4 + e_5 - e_1 \} \]
**Homology spaces**

**Detailed example**

\[
\partial_1 = \begin{pmatrix}
1 & 0 & -1 & 1 & 0 \\
-1 & 1 & 0 & 0 & -1 \\
0 & -1 & 1 & 0 & 0 \\
0 & 0 & 0 & -1 & 1
\end{pmatrix}
\]

after column reduction:

\[
\begin{pmatrix}
1 & 0 & 0 & 1 & 0 \\
-1 & 1 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 0
\end{pmatrix}
\]

- \( \ker(\partial_1) \) is spanned by \( \{ z_1 = e_1 + e_2 + e_3, z_2 = e_4 + e_5 - e_1 \} \)
- \( \text{Img}(\partial_2) \) is spanned by \( \{ e_1 + e_2 + e_3 \} \)
Homology spaces

Detailed example

\[ \partial_1 = \begin{pmatrix} e_1 & e_2 & e_3 & e_4 & e_5 \\ v_1 & 1 & 0 & -1 & 1 & 0 \\ v_2 & -1 & 1 & 0 & 0 & -1 \\ v_3 & 0 & -1 & 1 & 0 & 0 \\ v_4 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \]

after column reduction:

\[ \begin{pmatrix} e_1 & e_2 & z_1 & e_4 & z_2 \\ v_1 & 1 & 0 & 0 & 1 & 0 \\ v_2 & -1 & 1 & 0 & 0 & 0 \\ v_3 & 0 & -1 & 0 & 0 & 0 \\ v_4 & 0 & 0 & 0 & -1 & 0 \end{pmatrix} \]

\[ z_1 = e_1 + e_2 + e_3, \quad z_2 = e_4 + e_5 - e_1 \]

- \( \ker(\partial_1) \) is spanned by \( \{z_1 = e_1 + e_2 + e_3, z_2 = e_4 + e_5 - e_1\} \)
- \( \text{Img}(\partial_2) \) is spanned by \( \{e_1 + e_2 + e_3\} \)
- \( H_1 = \ker(\partial_1)/\text{Img}(\partial_2) \cong \text{span}\{e_4 + e_5 - e_1\} \)
The cycles surrounding the same “hole” belong to the same equivalence class

\[ c_1, c_2 \in \ker(\partial_1), \quad c_1 = e_4 + e_5 - e_1, \quad c_2 = e_4 + e_5 + e_2 + e_3 \]
Homology spaces

Interpretation

The cycles surrounding the same “hole” belong to the same equivalence class

\[ c_1, c_2 \in \ker(\partial_1), \quad c_1 = e_4 + e_5 - e_1, \quad c_2 = e_4 + e_5 + e_2 + e_3 \]

\[ c_2 - c_1 = e_1 + e_2 + e_3 \in \text{Img}(\partial_2) \]
Homology spaces

Interpretation

The cycles surrounding the same “hole” belong to the same equivalence class

\[ c_1, c_2 \in \ker(\partial_1), \quad c_1 = e_4 + e_5 - e_1, \quad c_2 = e_4 + e_5 + e_2 + e_3 \]

\[ c_2 - c_1 = e_1 + e_2 + e_3 \in \text{Img}(\partial_2) \]

Thus the two cycles are equivalent in the quotient \( c_2 \sim c_1 \).