Topological Data Analysis

Peter Bubenik

University of Florida
Department of Mathematics,
peter.bubenik@ufl.edu
http://people.clas.ufl.edu/peterbubenik/

British Applied Mathematics Colloquium
University of Oxford
April 6, 2016
Idea

Use topology to summarize and learn from the “shape” of data.
Simplicial complexes from point data

The Čech construction
Simplicial complexes from point data

The Čech construction
Simplicial complexes from point data

The Čech construction
Simplicial complexes from point data

The Čech construction
Simplicial complexes from point data

The Čech construction
Homology of simplicial complexes

Definition

Homology in degree $k$ is given by $k$-cycles modulo the $k$-boundaries.
Homology of simplicial complexes

Definition

Homology in degree $k$ is given by $k$-cycles modulo the $k$-boundaries.
Main idea

Vary a parameter and keep track of when features appear and disappear.
Persistence

Main idea

Vary a parameter and keep track of when features appear and disappear.
Persistence

Main idea

Vary a parameter and keep track of when features appear and disappear.
**Persistence**

**Main idea**

Vary a parameter and keep track of when features appear and disappear.
Main idea

Vary a parameter and keep track of when features appear and disappear.
Main idea

Vary a parameter and keep track of when features appear and disappear.
Main idea
Vary a parameter and keep track of when features appear and disappear.
Persistence

Main idea

Vary a parameter and keep track of when features appear and disappear.
Persistence

Main idea

Vary a parameter and keep track of when features appear and disappear.
**Main idea**

Vary a parameter and keep track of when features appear and disappear.
**Persistence**

**Main idea**

Vary a parameter and keep track of when features appear and disappear.
Main idea

Vary a parameter and keep track of when features appear and disappear.
Mathematical encoding

We have an increasing sequence of simplicial complexes

\[ X_0 \subseteq X_1 \subseteq X_2 \subseteq \cdots \subseteq X_m \]

called a filtered simplicial complex.

Apply homology.

We get a sequence of vector spaces and linear maps

\[ V_0 \to V_1 \to V_2 \to \cdots \to V_m \]

called a persistence module.
Persistence module to Barcode

\[ V_0 \to V_1 \to V_2 \to V_3 \to V_4 \to V_5 \to V_6 \to V_7 \to \cdots \to V_m \]

**Fundamental Theorem of Persistent Homology**

There exists a choice of bases for the vector spaces \( V_i \) such that each map is determined by a bipartite matching of basis vectors.

Get a **barcode**:
Barcode to Persistence Landscape

Barcode:

Persistence Landscape:

\[ \lambda_k = 0, \quad \text{for } k \geq 4 \]
Maltose Binding Protein, two conformations

Maltose Binding Protein, two conformations

Maltose Binding Protein Data

The Data

Fourteen MBP structures from the Protein Data Bank.
- 7 closed conformations
- 7 open conformations

X-ray crystallography: coordinates of atoms.

Represent each amino acid residue by its $C\alpha$ atom.

Have 14 sets of 370 points in $\mathbb{R}^3$.

The Goal

Can we use topological data analysis to distinguish the open and closed conformations?
Filtered simplicial complex
Filtered simplicial complex
Filtered simplicial complex
Filtered simplicial complex
Filtered simplicial complex
Average persistence landscapes

\(H_1\) closed

\(H_1\) open
Figure 1: Distance analysis based on the $2$-Landscape distance shows a separation between the closed (blue) and the open (red) MBP conformation for degree $0$ (left) and degree $1$ (right) persistent homology. Similar results hold for degree $2$. The projection of the $14 \times 14$ distance matrix onto the $3$D space is attained via Isomap.

4.3 Statistical Inference

To measure the statistical significance of visually observed differences between the closed and the open conformation we use a permutation test. For each degree, we calculate fourteen sample values of the random variable $X$ from Equation (3). The permutation test carried out at the significance level of $0.05$ yields a $p$-value of $5 \times 10^{-4}$ for both homology in degree $0$ and in degree $1$. We obtain the same $p$-value since in both cases the observed statistic was the most extreme among all $1716$ possible permutations. Hence, at...
Classification of protein conformations

- Closed
- Open
- Support vector

SVM on Isomap embedded 3D coordinates in the metric space induced by the 2-Landscape distance
Software

Persistent Homology software:
- JavaPlex
- PHAT, DIPHA
- Perseus
- Dionysus
- CHOMP
- GUDHI

Persistence Landscape software:
- The Persistence Landscape Toolbox
- the R package TDA
Topological Data Analysis Summary

1. Raw Data
2. Preprocess
3. Clean data
4. Filtered simplicial complex
5. Transform
6. Homology
7. Persistence module
8. Topological summary
9. Statistics and Machine Learning