Diffusion Maps and Topological Data Analysis

Melissa R. McGuirl

OVERVIEW

Topological Data Analysis

The use of algebraic topology to develop tools that extract qualitative features from high-dimensional, noisy data.

Diffusion Maps

A non-linear dimension reduction technique aimed at discovering the underlying manifold that the data has been sampled from.

Main Question

Can we combine diffusion maps and topological data analysis to extract extract qualitative features from high-dimensional, noisy data that lie on complicated manifolds?

Overview

Topological Data Analysis

The use of algebraic topology to develop tools that extract qualitative features from high-dimensional, noisy data.

Diffusion Maps

A non-linear dimension reduction technique aimed at discovering the underlying manifold that the data has been sampled from.

Main Question

Can we combine diffusion maps and topological data analysis to extract extract qualitative features from high-dimensional, noisy data that lie on complicated manifolds?

Overview

Topological Data Analysis

The use of algebraic topology to develop tools that extract qualitative features from high-dimensional, noisy data.

Diffusion Maps

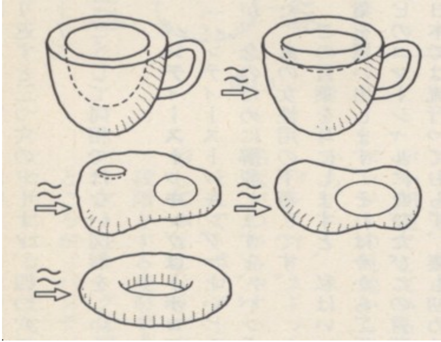
A non-linear dimension reduction technique aimed at discovering the underlying manifold that the data has been sampled from.

Main Question

Can we combine diffusion maps and topological data analysis to extract extract qualitative features from high-dimensional, noisy data that lie on complicated manifolds?

TOPOLOGY OVERVIEW

- Topology focuses on studying invariants under continuous deformation
- Algebraic topology looks at "connectedness" of spaces
- Homotopy and Homology
- Computational topology focuses on homology, specifically persistent homology



Source: http://jwilson.coe.uga.edu/EMAT6680/

Simplicial complexes

Simplex

A k-simplex is the convex hull of $k+1$ affinely independent points,

$$\sigma = \text{conv}(u_0, \ldots, u_k\}.$$

Simplicial Complex

A simplicial complex is a finite collection of simplices K such that

1. $\sigma \in K$ and $\tau \leq \sigma$ implies $\tau \in K$
2. $\sigma_1, \sigma_2 \in K$ implies either (1) $\sigma_1 \cap \sigma_2 = \emptyset$ or (2) $\sigma_1 \cap \sigma_2$ is a face of both σ_1 and σ_2.

Chains, Cycles, Boundaries

Chains

Let K be a simplicial complex. A p-chain is

$$c = \sum a_i \sigma_i,$$

where a_i are coefficients (we usually use $\mathbb{Z}/2\mathbb{Z}$ coefficients), and σ_i are p-simplices.

Boundary Map

Let $\sigma = [u_0, \ldots u_p]$ be a p-simplex. $\partial_p : C_p \to C_{p-1}$ is a map defined by

$$\partial_p(\sigma) = \sum_{j=0}^{p} [u_0, \ldots, \hat{u}_j, \ldots, u_p].$$

The boundary map is a group homomorphism.

Chains, Cycles, Boundaries

Chains

Let K be a simplicial complex. A p-chain is

$$c = \sum a_i \sigma_i,$$

where a_i are coefficients (we usually use $\mathbb{Z}/2\mathbb{Z}$ coefficients), and σ_i are p-simplices.

Boundary Map

Let $\sigma = [u_0, \ldots u_p]$ be a p-simplex. $\partial_p : C_p \to C_{p-1}$ is a map defined by

$$\partial_p(\sigma) = \sum_{j=0}^{p} [u_0, \ldots, \hat{u}_j, \ldots, u_p].$$

The boundary map is a group homomorphism.

Homology Groups

Chain Complex

$$\dots \overset{\partial_{p+2}}{\rightarrow} C_{p+1} \overset{\partial_{p+1}}{\rightarrow} C_p \overset{\partial_p}{\rightarrow} C_{p-1} \overset{\partial_{p-1}}{\rightarrow} \dots$$

Cycles and Boundaries

A p-cycle is $Z_p = \mathrm{Ker}(\partial_p)$. A p-boundary is $B_p = \mathrm{Im}(\partial_{p+1})$.

Homology group

The p-th homology group is the quotient group

$$H_p = Z_p / B_p$$

Homology Groups

Chain Complex

$$\cdots \overset{\partial_{p+2}}{\rightarrow} C_{p+1} \overset{\partial_{p+1}}{\rightarrow} C_p \overset{\partial_p}{\rightarrow} C_{p-1} \overset{\partial_{p-1}}{\rightarrow} \cdots$$

Cycles and Boundaries

A p-cycle is $Z_p = \mathrm{Ker}(\partial_p)$. A p-boundary is $B_p = \mathrm{Im}(\partial_{p+1})$.

Homology group

The p-th homology group is the quotient group

$$H_p = Z_p / B_p$$

Homology Groups

Chain Complex

$$\dots \overset{\partial_{p+2}}{\to} C_{p+1} \overset{\partial_{p+1}}{\to} C_p \overset{\partial_p}{\to} C_{p-1} \overset{\partial_{p-1}}{\to} \dots$$

Cycles and Boundaries

A p-cycle is $Z_p = \text{Ker}(\partial_p)$. A p-boundary is $B_p = \text{Im}(\partial_{p+1})$.

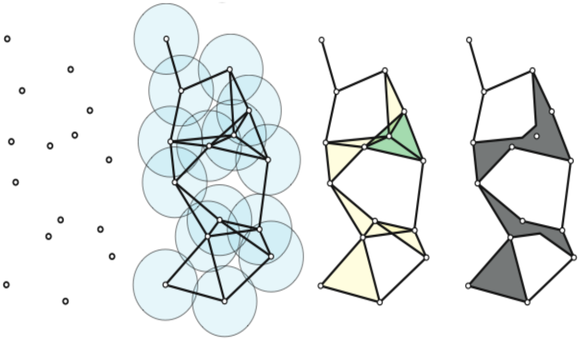
Homology group

The p-th homology group is the quotient group

$$H_p = Z_p / B_p$$

Computing homology of data

Step 1: Build a simplicial complex from data using open sets.
Why does this work? The Nerve theorem.



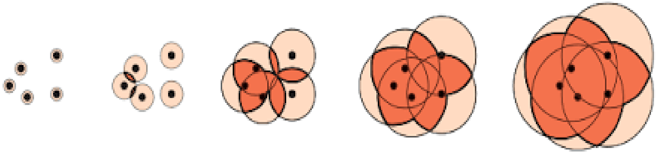
source:http://jeffe.cs.illinois.edu/pubs/rips.html

Computing homology of data

Step 2: Find all i-chains of simplicial complex and build a chain complex.
Step 3: Represent boundary maps ∂_p as matrices with $\mathbb{Z}/2\mathbb{Z}$ coefficients.
Step 4: Compute $Z_p = \text{Ker}(\partial_p)$ and $B_p = \text{Im}(\partial_{p+1})$.
Step 5: $H_p = Z_p/B_p$

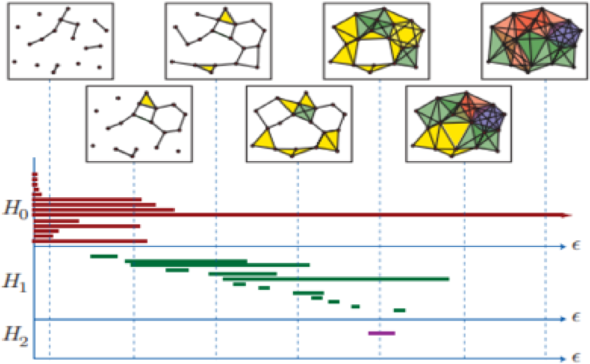
Persistent Homology

How big should we make the balls?



source: http://donsheehy.net/sheehy10multifiltering.html

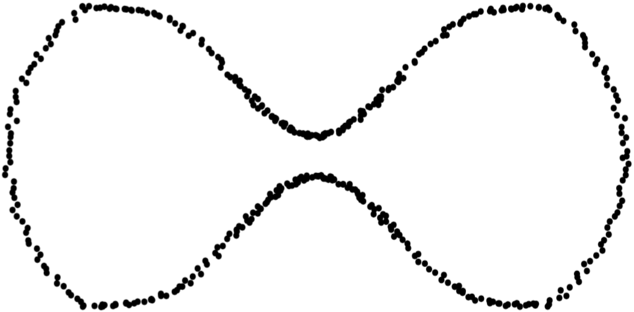
Bar Codes



source: http://xiangze.hatenablog.com/entry/2014/03/29/042627

Problematic data

$H_1 = 2?\ H_1 = 1?$



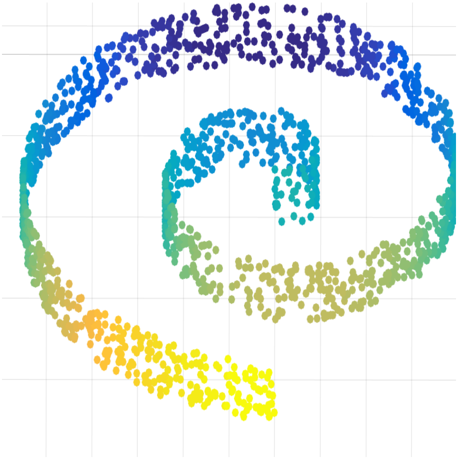
source: http://www.paulbendich.com/pubs/IP-DiffRips.pdf

Solution:

Diffusion Maps!

Dimension Reduction

- Principal Component Analysis
  - Linear dimension reduction
  - Finds dimensions that capture most variability in the data
- Multidimensional scaling
  - Linear dimension reduction
  - Embeds data in a lower dimensional space while preserving pairwise distances between points
- Neither method can capture spiral behavior of Swiss roll



Diffusion Maps

- Non-linear dimension reduction algorithm introduced by [1]
- Main idea: embed data into a lower-dimensional space such that the Euclidean distance between points approximates diffusion distance data
- Diffusion distance between points is based on probability of jumping between points
- Random walk on data points

How it works

1. Diffusion distance:

$$K(x,y) = \exp\left(-\frac{|x-y|}{\alpha}\right)$$

2. Create distance/kernel matrix

$$K_{ij} = K(x_i, x_j)$$

3. Create diffusion matrix (Markov) M by normalizing so that sum over rows is 1
4. Calculate eigenvectors of M, sort by eigenvalues
5. Return *d* top eigenvectors, map original space into the d-eigenvectors

HOW IT WORKS

1. Diffusion distance:

$$K(x,y) = \exp\left(-\frac{|x-y|}{\alpha}\right)$$

2. Create distance/kernel matrix

$$K_{ij} = K(x_i, x_j)$$

3. Create diffusion matrix (Markov) M by normalizing so that sum over rows is 1
4. Calculate eigenvectors of M, sort by eigenvalues
5. Return *d* top eigenvectors, map original space into the d-eigenvectors

How it works

1. Diffusion distance:

$$K(x,y) = \exp\left(-\frac{|x-y|}{\alpha}\right)$$

2. Create distance/kernel matrix

$$K_{ij} = K(x_i, x_j)$$

3. Create diffusion matrix (Markov) M by normalizing so that sum over rows is 1
4. Calculate eigenvectors of M, sort by eigenvalues
5. Return *d* top eigenvectors, map original space into the d-eigenvectors

How it works

1. Diffusion distance:

$$K(x,y) = \exp\left(-\frac{|x-y|}{\alpha}\right)$$

2. Create distance/kernel matrix

$$K_{ij} = K(x_i, x_j)$$

3. Create diffusion matrix (Markov) M by normalizing so that sum over rows is 1
4. Calculate eigenvectors of M, sort by eigenvalues
5. Return *d* top eigenvectors, map original space into the d-eigenvectors

How it works

1. Diffusion distance:

$$K(x,y) = \exp\left(-\frac{|x-y|}{\alpha}\right)$$

2. Create distance/kernel matrix

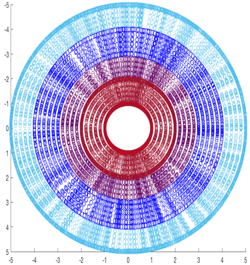
$$K_{ij} = K(x_i, x_j)$$

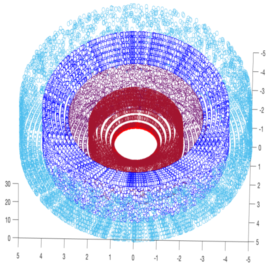
3. Create diffusion matrix (Markov) M by normalizing so that sum over rows is 1
4. Calculate eigenvectors of M, sort by eigenvalues
5. Return *d* top eigenvectors, map original space into the d-eigenvectors

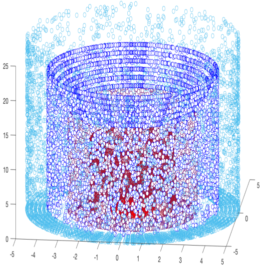
Research plan

1. Embed high dimension, noisy data into a lower-dimensional space using Diffusion map representation
2. Run topological data analysis on diffusion map
3. Compare results of TDA on original data to TDA on diffusion map
4. Apply approach to medical images to extract qualitative features from data

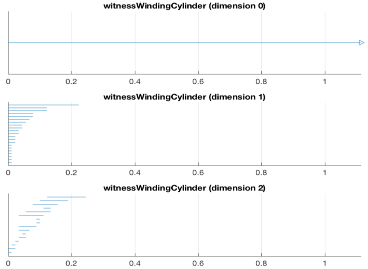
WINDING CYLINDER EXAMPLE

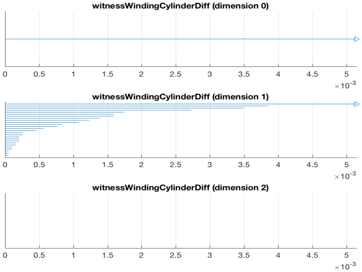






Winding Cylinder Bar Codes





References

[1] Coifman, Ronald and Stephane Lafon. Diffusion Maps. Applied and Computational Harmonic Analysis, *Elsevier* (2006).
[2] Hatcher, Allen. Algebraic Topology. Cambridge University Press. (2002)
[3] Herbert Edelsbrunner and John L. Harer. Computational Topology. American Mathematical Society, Providence (2009).