PERSISTENCE IMAGES

An Alternative Persistent Homology Representation

June 23, 2015

Lori Ziegelmeier*, Tegan Emerson, Eric Hanson, Rachel Neville, Sofya Chepushtanova, Francis Motta, Chris Peterson, Michael Kirby, Patrick Shipman

SOCG 2015
Minisymposium on Computational Topology
Statistical Approaches to Topological Data Analysis

OVERVIEW

INTRODUCTION

MOTIVATION:

- Topological data analysis allows extraction of coarse topological features from data

- Persistent homology is a key technique to extract topological features

- Extending the list of tools from machine learning which can apply to persistent homology features is desirable

Goal:

Develop an alternative representation of persistent homology that ‘vectorizes’ topological information.

PERSISTENT HOMOLOGY

TOPOLOGICAL DATA ANALYSIS

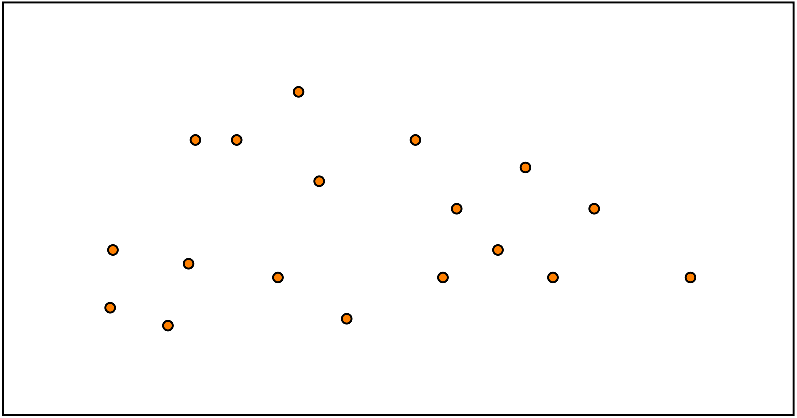
1. Envision data as a point cloud

2. Create connections between proximate points
   - build simplicial complex

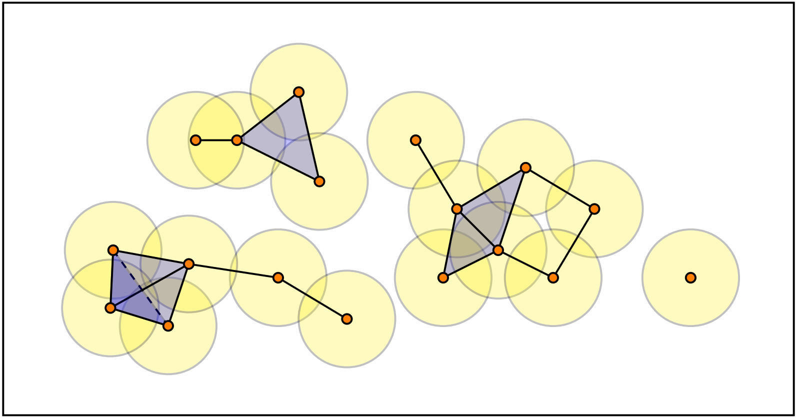
3. Determine topological structure of complex
   - compute homology

4. Vary proximity parameter to assess different scales
   - calculate persistent homology

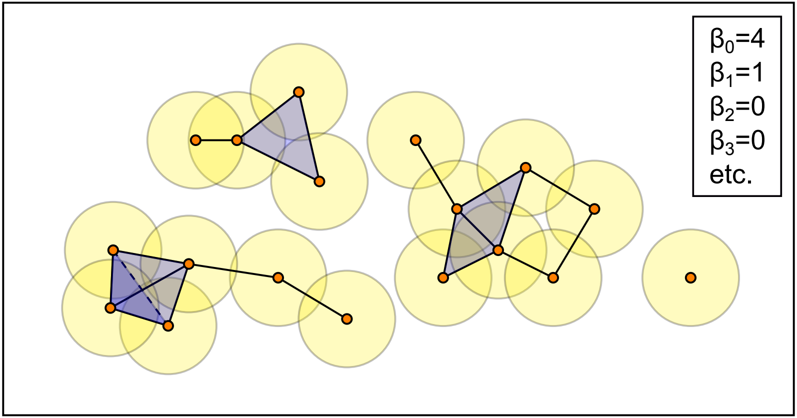
1. ENVISION DATA AS A POINT CLOUD



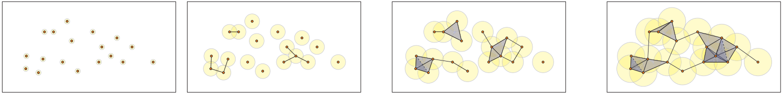
2. BUILD A SIMPLICIAL COMPLEX



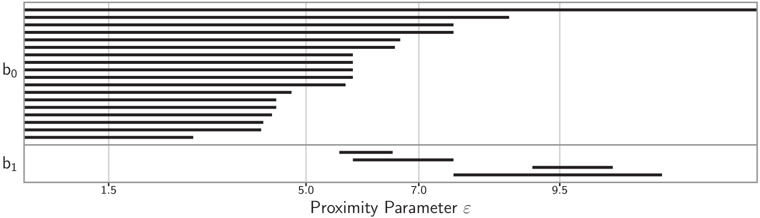
3. COMPUTE BETTI NUMBERS

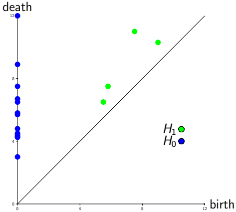


4. COMPUTE PERSISTENT HOMOLOGY



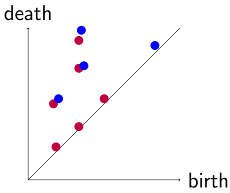
$\varepsilon = 1.5$ $\varepsilon = 5.0$ $\varepsilon = 7.0$ $\varepsilon = 9.5$





PERSISTENCE DIAGRAMS AS A METRIC SPACE

The space of Persistence Diagrams (PDs) can be endowed with a metric.



Definition

The Bottleneck distance between two PDs X and Y is given by

$$d_B(X, Y) := \inf_{\gamma : X \to Y} \sup_{x \in X} ||x - \gamma(x)||_\infty,$$

where $||\cdot||_\infty$ is the L_∞-distance and γ ranges over bijections between X and Y.

MACHINE LEARNING TASKS ON PD

Any machine learning algorithm that only requires a distance matrix as input can be implemented on the space of PDs.

Many other techniques do not fall into this category:

- Support vector machines
- Decision tree classification
- Neural networks
- Feature selection

Need a ‘feature vector’ representation to analyze data in these algorithms.

PERSISTENCE IMAGES

PERSISTENCE IMAGES (PI)

Goal:

Develop an alternative representation of persistent homology that ‘vectorizes’ topological information while maintaining an interpretable connection to the original PD.

PERSISTENCE IMAGES (PI)

Goal:

Develop an alternative representation of persistent homology that ‘vectorizes’ topological information while maintaining an interpretable connection to the original PD.

1. For each point (b_x, b_y) in PD **B**, center a Gaussian.

PERSISTENCE IMAGES (PI)

Goal:

Develop an alternative representation of persistent homology that 'vectorizes' topological information while maintaining an interpretable connection to the original PD.

1. For each point (b_x, b_y) in PD **B**, center a Gaussian.
2. Overlay a grid onto the PD.

PERSISTENCE IMAGES (PI)

Goal:

Develop an alternative representation of persistent homology that ‘vectorizes’ topological information while maintaining an interpretable connection to the original PD.

1. For each point (b_x, b_y) in PD $\mathbf{B}$, center a Gaussian.
2. Overlay a grid onto the PD.
3. The image value at pixel p, a square in the grid, is the sum of all Gaussians over the area in that square

$$I(p) = \iint_p \sum_{(b_x, b_y) \in \mathbf{B}} \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{(x-b_x)^2}{\sigma_x^2} + \frac{(y-b_y)^2}{\sigma_y^2}\right)} dydx$$

where σ_x and σ_y are variances in the Gaussian.

WEIGHTING A PERSISTENCE IMAGE

May desire to weight points further from the diagonal more and suppress points closer to the diagonal.

Modify definition of a pixel as follows:

$$I(p) = \iint\limits_{p} \sum_{(b_x, b_y) \in \mathbf{B}} f(|\mathbf{b}|) \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{(x-b_x)^2}{\sigma_x^2} + \frac{(y-b_y)^2}{\sigma_y^2}\right)} dydx$$

where the weighting function $f(|\mathbf{b}|)$ depends on the distance from the diagonal, $|\mathbf{b}| = b_y - b_x$.

WEIGHTING A PERSISTENCE IMAGE

May desire to weight points further from the diagonal more and suppress points closer to the diagonal.

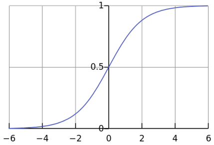
Modify definition of a pixel as follows:

$$I(p) = \iint_p \sum_{(b_x, b_y) \in \mathbf{B}} f(|\mathbf{b}|) \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{(x-b_x)^2}{\sigma_x^2} + \frac{(y-b_y)^2}{\sigma_y^2}\right)} dydx$$

where the weighting function $f(|\mathbf{b}|)$ depends on the distance from the diagonal, $|\mathbf{b}| = b_y - b_x$.

Options for f could include:

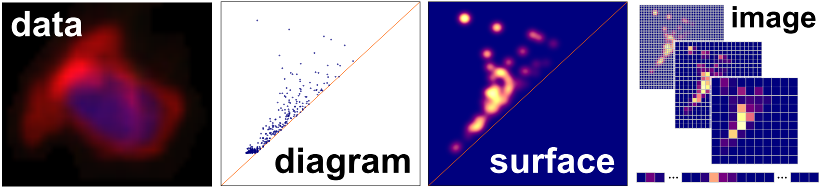
- Exponential
- Bump function
- Piecewise linear
- Sigmoidal



PARAMETERS FOR PERSISTENCE IMAGES

- Resolution of the image (*i.e.* choice of grid)
  - As resolution tends to infinity, converges to a continuous representation of the PD.
- Variance of the Gaussian
  - Corresponds to filtration step in PH computation
  - Related to confidence in location of points in PD
- Weighting function *f*
  - Suppress the effects of noise and amplify signal

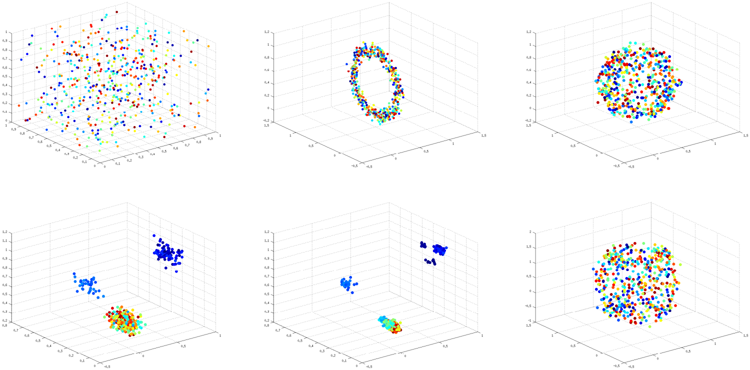
PERSISTENCE IMAGE PIPELINE



DATA ANALYSIS

TOY DATA

Points sampled from six topological spaces: the solid cube, a circle, a sphere, three clusters, three clusters within three clusters, and a torus



25 point clouds from each space, consisting of 500 points, 2 levels of noise $\eta = 0.05, 0.1$

COMPARISON OF K-MEDOIDS CLASSIFICATION

Goal:

Compare classification accuracy of toy data in the PD framework equipped with the Bottleneck distance and the PI framework equipped with Euclidean distance.

COMPARISON OF K-MEDOIDS CLASSIFICATION

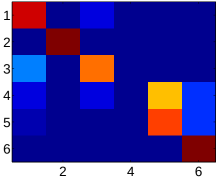
Goal:

Compare classification accuracy of toy data in the PD framework equipped with the Bottleneck distance and the PI framework equipped with Euclidean distance.

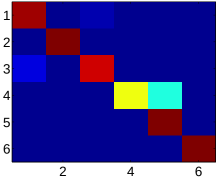
Use k-medoids:

- Iterative, clustering algorithm
- Takes as input a pairwise distance matrix and the number of clusters
- Chooses an existing datum, represented by an index in a distance matrix, as the center of each cluster so that the distance between each point and the center with which it is identified is minimized

CONFUSION MATRICES AND ACCURACY α, NOISE $\eta = 0.05$

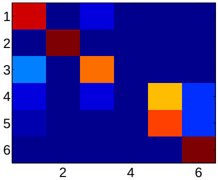


(a) PD, H_0, $\alpha = 66$

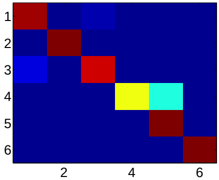


(b) PD, H_1, $\alpha = 90.7$

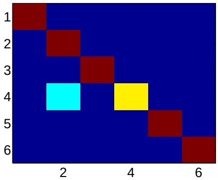
CONFUSION MATRICES AND ACCURACY α, NOISE $\eta = 0.05$



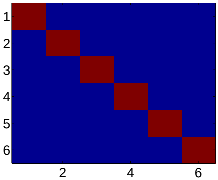
(f) PD, H_0, $\alpha = 66$



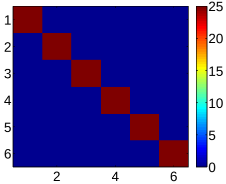
(g) PD, H_1, $\alpha = 90.7$



(h) PI, H_0, $\alpha = 94$

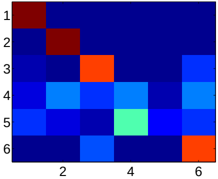


(i) PI, H_1, $\alpha = 100$

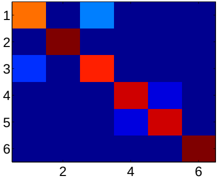


(j) PI, Both, $\alpha = 100$

CONFUSION MATRICES AND ACCURACY α, NOISE $\eta = 0.1$

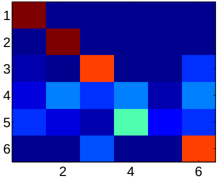


(a) PD, H_0, $\alpha = 74.7$

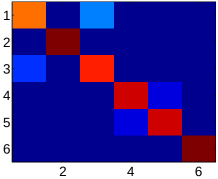


(b) PD, H_1, $\alpha = 91.3$

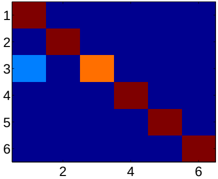
CONFUSION MATRICES AND ACCURACY α, NOISE $\eta = 0.1$



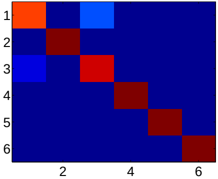
(f) PD, H_0, $\alpha = 74.7$



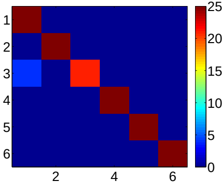
(g) PD, H_1, $\alpha = 91.3$



(h) PI, H_0, $\alpha = 96$



(i) PI, H_1, $\alpha = 95.3$



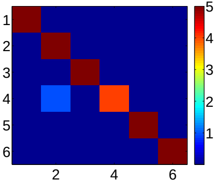
(j) PI, Both,$\alpha = 97.3$

BENEFITS OF PI

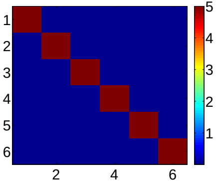
- Improved accuracy
- Time reduced
    - 1.9×10^5 seconds to generate a bottleneck distance matrix
    - Under 300 seconds to generate set of PIs and compute Euclidean distance
- Analyze multiple homology dimensions simultaneously by concatenating corresponding images
- Can implement more machine learning algorithms on PIs
    - *e.g.* Support Vector Machines, supervised binary classifier

SVM ACCURACY ON PI

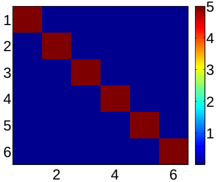
Noise $\eta = 0.05$



(a) PI, H_0, $\alpha = 96.7$



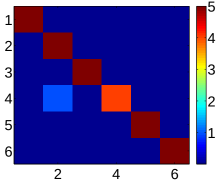
(b) PI, H_1, $\alpha = 100$



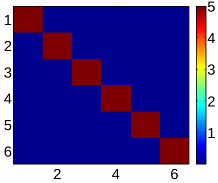
(c) PI, Both, $\alpha = 100$

SVM ACCURACY ON PI

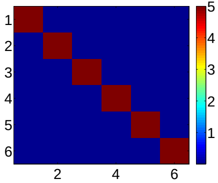
Noise $\eta = 0.05$



(g) PI, H_0, $\alpha = 96.7$

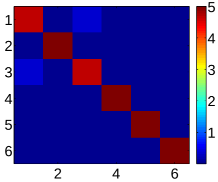


(h) PI, H_1, $\alpha = 100$

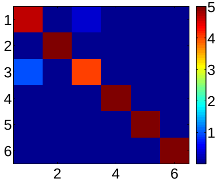


(i) PI, Both, $\alpha = 100$

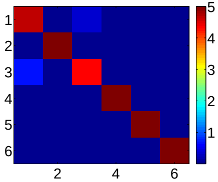
Noise $\eta = 0.1$



(j) PI, H_0, $\alpha = 97.8$



(k) PI, H_1, $\alpha = 95.6$



(l) PI, Both, $\alpha = 96.7$

TOY DATA PARAMETER SEARCH

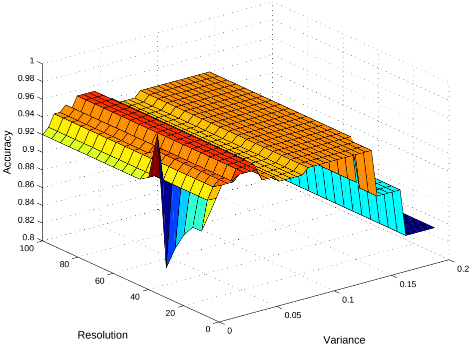
Parameter Search:

- 20 resolution choices from images of size 5×5 to 100×100
- 40 variance choices from 0.0001 to 0.2

TOY DATA PARAMETER SEARCH

Parameter Search:

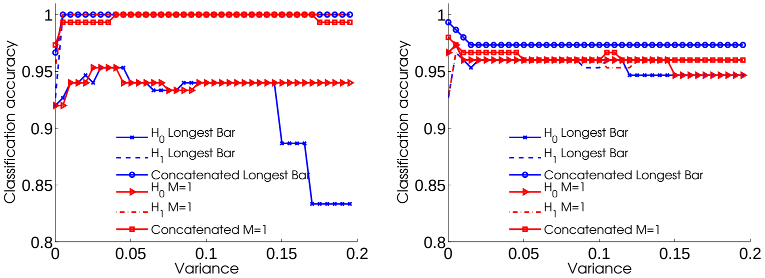
- 20 resolution choices from images of size 5×5 to 100×100
- 40 variance choices from 0.0001 to 0.2



Resolution had little effect on accuracy of toy data analysis.

TOY DATA PARAMETER SEARCH

For fixed resolution of 20 $\times$ 20, k-medoids accuracy:



(a) Noise=0.05 (b) Noise=0.1

LINKED-TWIST MAP

Dynamical system to model turbulent mixing in DNA microarrays (Hertzsch et. al.)

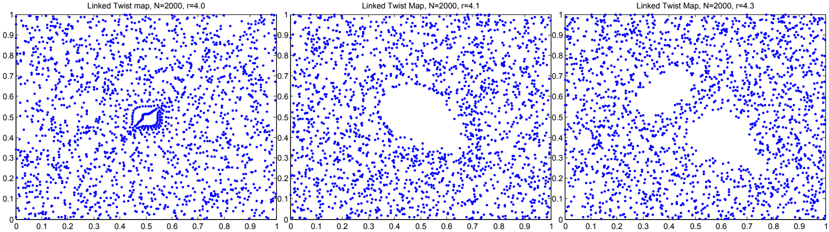
- Linked: coupled system
- Twist: mechanism of mixing

LINKED-TWIST MAP

Dynamical system to model turbulent mixing in DNA microarrays (Hertzsch et. al.)

- Linked: coupled system
- Twist: mechanism of mixing

Many parameter values exhibit interesting behavior.



LINKED-TWIST MAP CLASSIFICATION ACCURACY

- 3 parameter regimes for the Linked-Twist Map
- 25 samples of 500 points and 1000 points
- Each of the two sets were analyzed with persistence and put into PI framework
- Analyzed H_1 PIs with k-medoids

LINKED-TWIST MAP CLASSIFICATION ACCURACY

- 3 parameter regimes for the Linked-Twist Map
- 25 samples of 500 points and 1000 points
- Each of the two sets were analyzed with persistence and put into PI framework
- Analyzed H_1 PIs with k-medoids
  - 500 points: 92% accuracy
  - 1000 points: 96% accuracy

CONCLUSION

Persistence Images

PIs present a method for vectorization of topological characteristics of data that:

- have an interpretable connection to PDs
- yield higher classification accuracy than PDs equipped with the bottleneck distance
- speed up computations
- allow multiple homology dimensions to be analyzed simultaneously
- provide a wider access to a variety of metrics and machine learning tools

QUESTIONS?

Thank you!

Lori Ziegelmeier
lziegel1@macalester.edu
Department of Mathematics, Statistics, and Computer Science
MACALESTER
COLLEGE