Topology for Data Science 3

Peter Bubenik

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

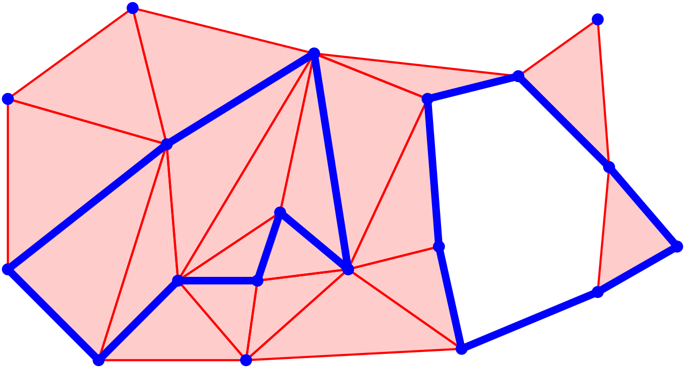
January 23, 2017

Tercera Escuela de Análisis Topológico de Datos
y Topología Estocástica
ABACUS, Estado de México

Homology

Definition

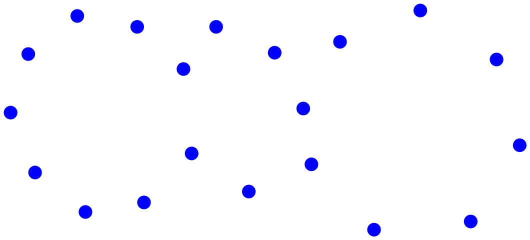
Homology in degree k is given by k-cycles modulo the k-boundaries.



Persistent homology

Main idea

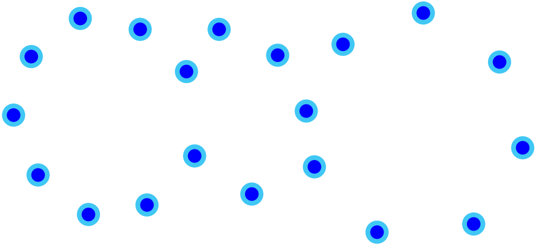
Vary a parameter and keep track of when homology appears and disappears.



Persistent homology

Main idea

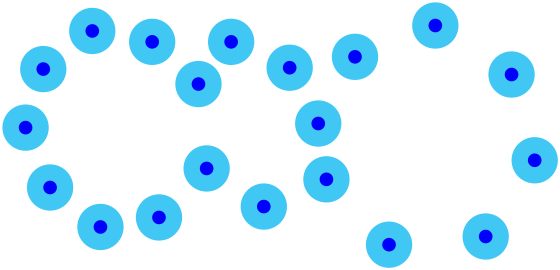
Vary a parameter and keep track of when homology appears and disappears.



Persistent homology

Main idea

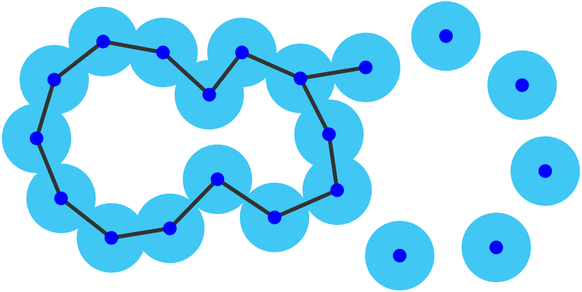
Vary a parameter and keep track of when homology appears and disappears.



Persistent homology

Main idea

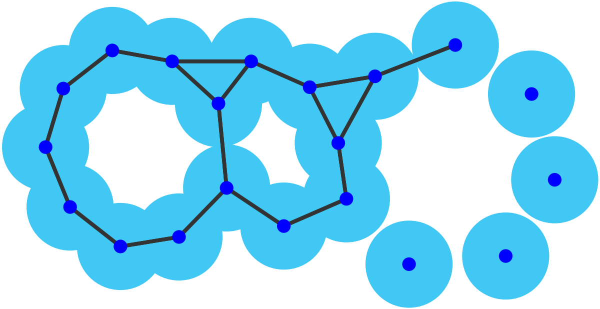
Vary a parameter and keep track of when homology appears and disappears.



Persistent homology

Main idea

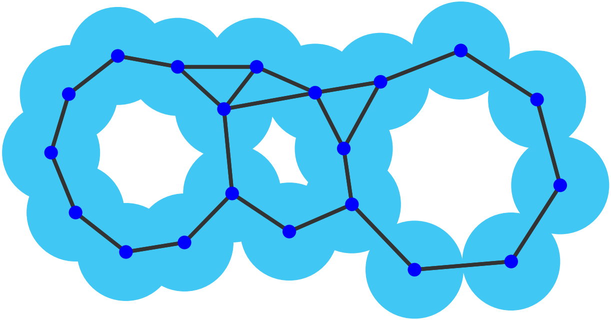
Vary a parameter and keep track of when homology appears and disappears.



Persistent homology

Main idea

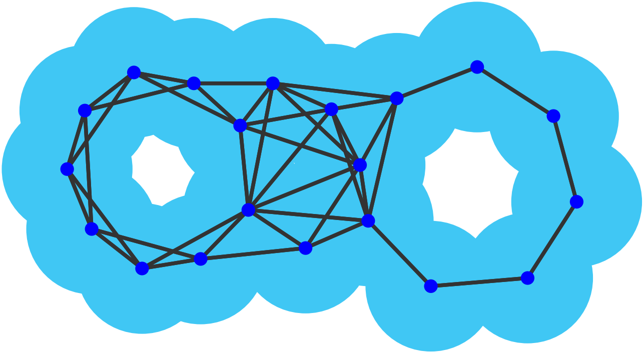
Vary a parameter and keep track of when homology appears and disappears.



Persistent homology

Main idea

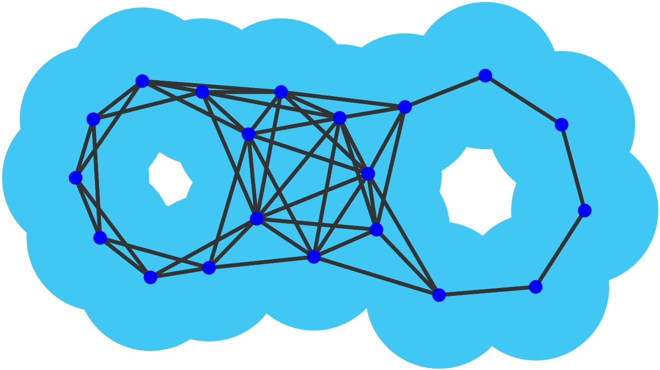
Vary a parameter and keep track of when homology appears and disappears.



Persistent homology

Main idea

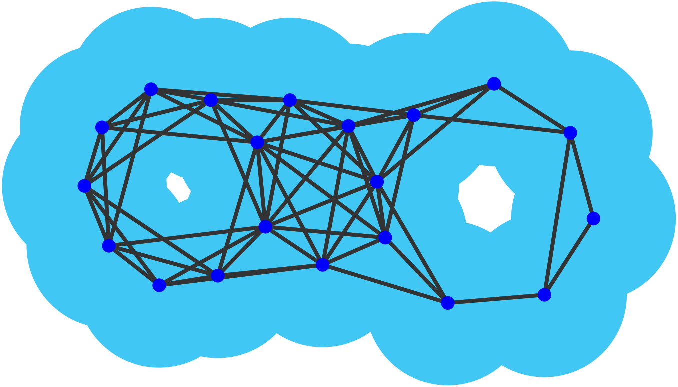
Vary a parameter and keep track of when homology appears and disappears.



Persistent homology

Main idea

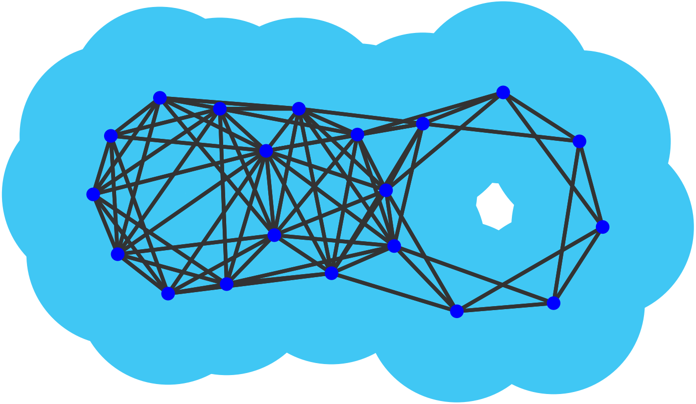
Vary a parameter and keep track of when homology appears and disappears.



Persistent homology

Main idea

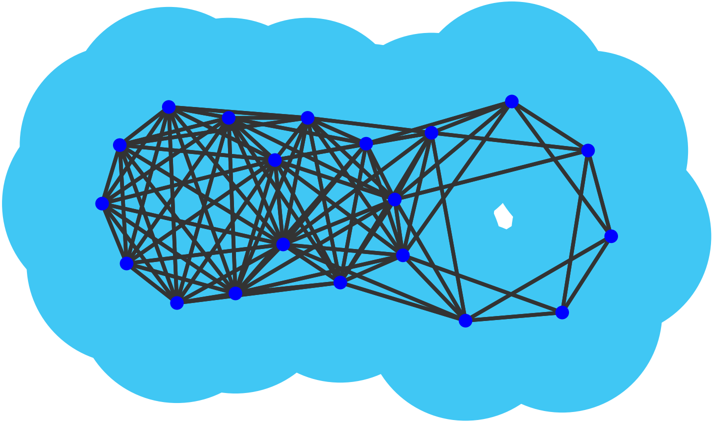
Vary a parameter and keep track of when homology appears and disappears.



Persistent homology

Main idea

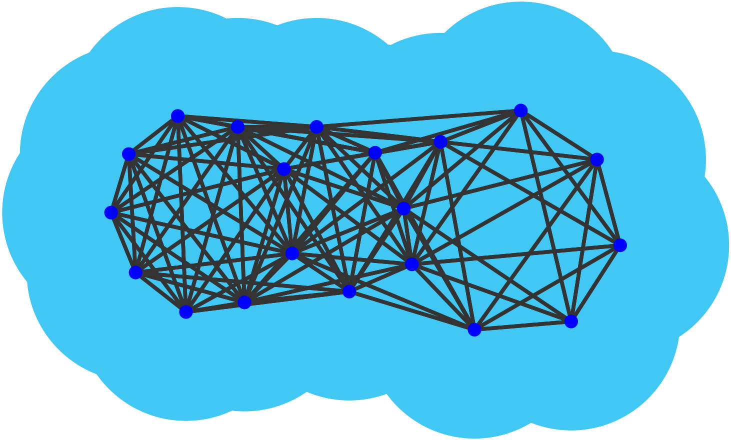
Vary a parameter and keep track of when homology appears and disappears.



Persistent homology

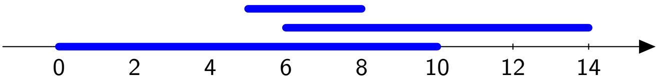
Main idea

Vary a parameter and keep track of when homology appears and disappears.

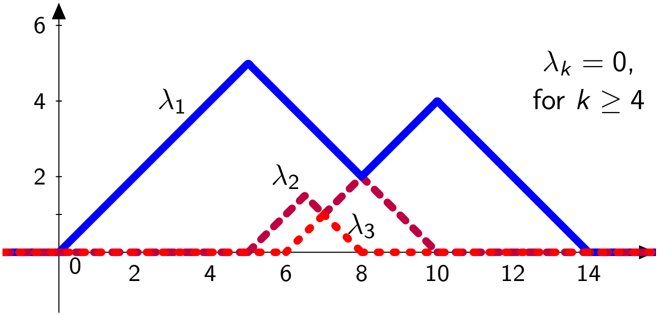


Barcode and Persistence Landscapes

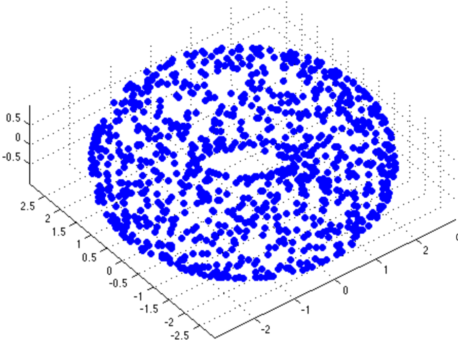
Barcode:

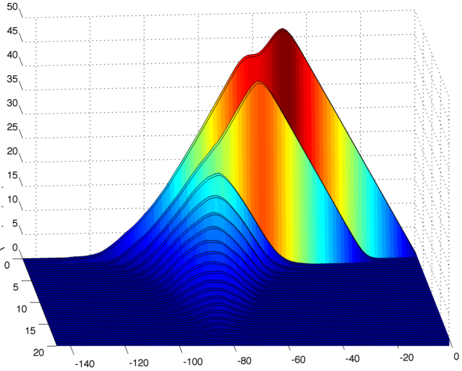


Convert to Persistence Landscape:



Persistent homology of sampled points





Short bars

Question

Can we understand the small bars in terms of the underlying geometry – specifically curvature?

This is joint work in progress with

- Dhruv Patel (Univ of Florida)
- Benjamin Whittle (Univ of Florida)

Metric geometry

Curvature in a metric space, M

- compare triangles in M with triangles in certain spaces

Model spaces of constant curvature K

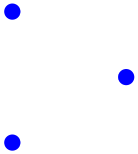
- $K = -1$: Hyperbolic plane
- $K = 0$: Euclidean plane
- $K = 1$: Sphere of radius 1

Assumptions:

- sample points independently
- from a uniform density
- on a unit disk of constant curvature

Čech complex

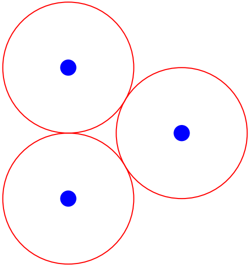
Acute triangles $\mapsto$ persistent H_1 in the Čech complex



Asymptotically almost all H_1 is of this form.

Čech complex

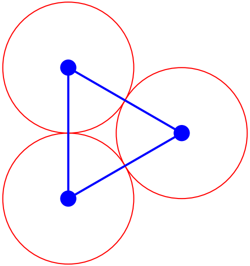
Acute triangles $\mapsto$ persistent H_1 in the Čech complex



Asymptotically almost all H_1 is of this form.

Čech complex

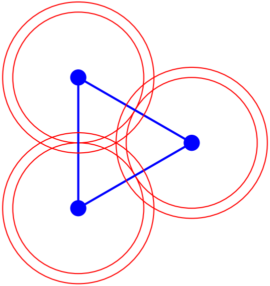
Acute triangles $\mapsto$ persistent H_1 in the Čech complex



Asymptotically almost all H_1 is of this form.

Čech complex

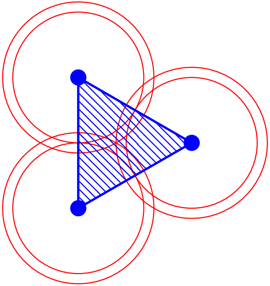
Acute triangles $\mapsto$ persistent H_1 in the Čech complex



Asymptotically almost all H_1 is of this form.

Čech complex

Acute triangles $\mapsto$ persistent H_1 in the Čech complex

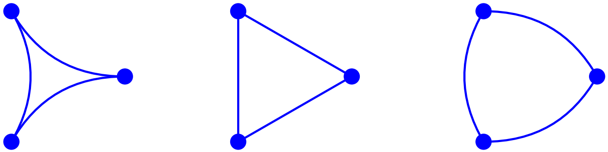


Asymptotically almost all H_1 is of this form.

Čech complex

The most persistent such H_1 arises from equilateral triangles.

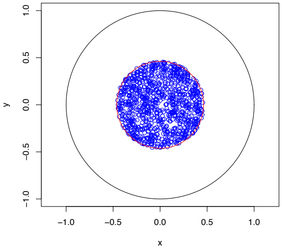
Consider equilateral triangles with circumcircle of radius 1.

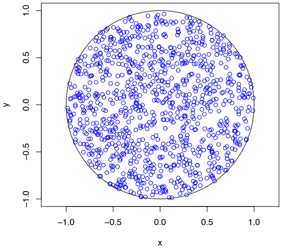


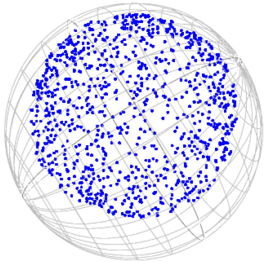
- Hyperbolic: death/birth ≈ 1.119
- Euclidean: death/birth $= 2/\sqrt{3} \approx 1.155$
- Spherical: death/birth ≈ 1.225

Points sampled from unit disks

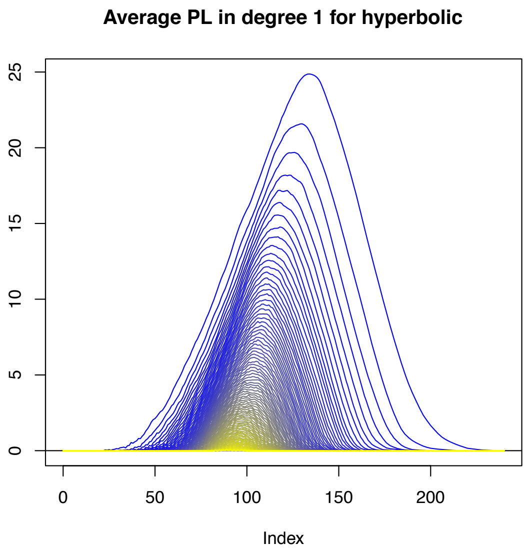
Sample 1000 points



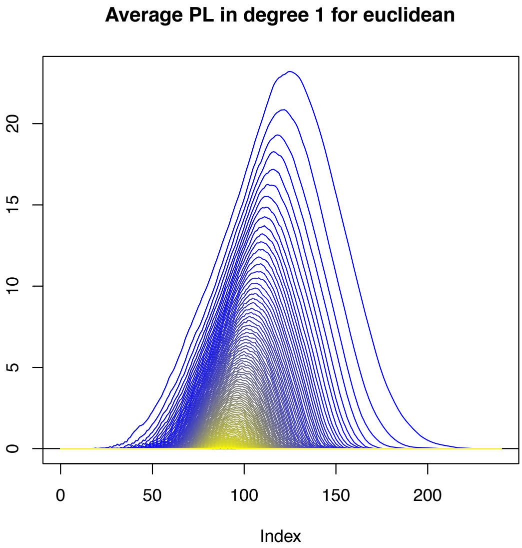




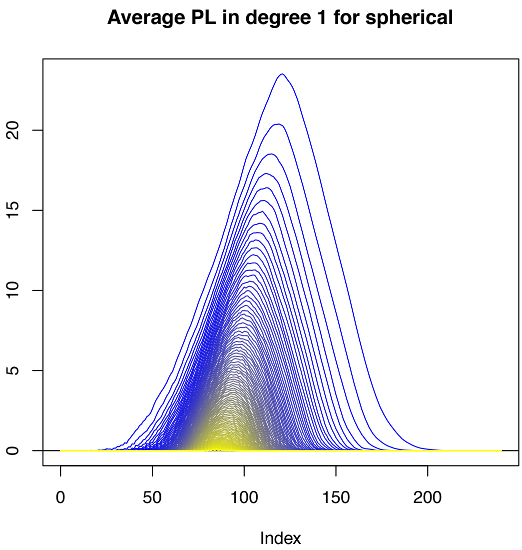
Average Landscapes



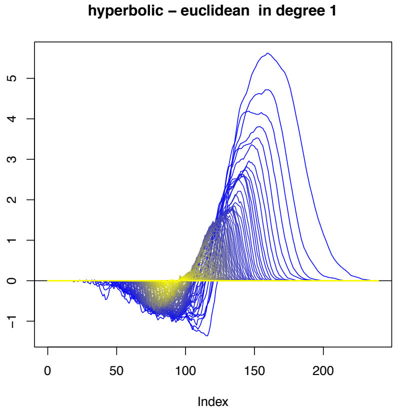
Average Landscapes

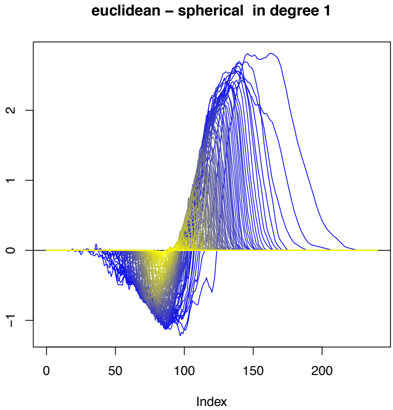


Average Landscapes



Differences in Average Landscapes





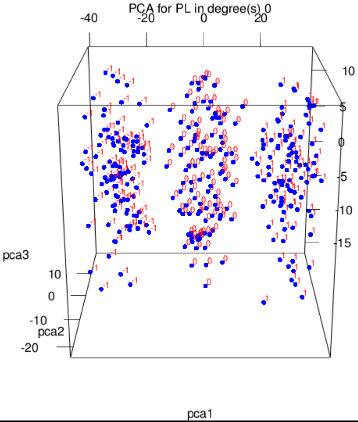
Classification

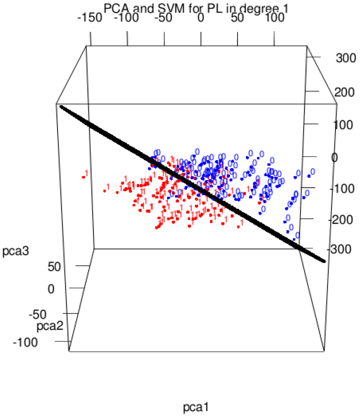
100 samples from each of hyperbolic, euclidean and spherical

Classify using SVM and 10-fold cross validation

Classification accuracy

- Using degree 0: 100%
- Using degree 1: 87%





Alzheimers Disease Neuroimaging Initiative (ADNI)

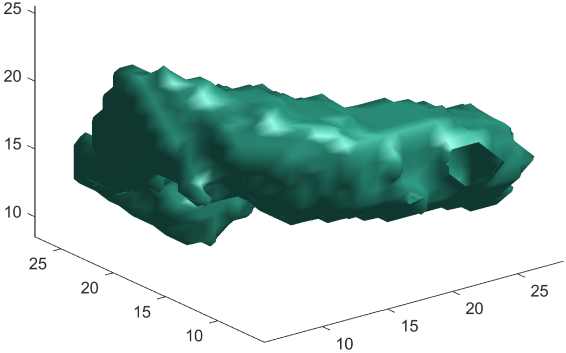
Joint work in progress with Ulrich Bauer (TU Munich), and Roland Kwitt (Salzburg).

The data:
995 left and right (paired) hippocampi consisting of

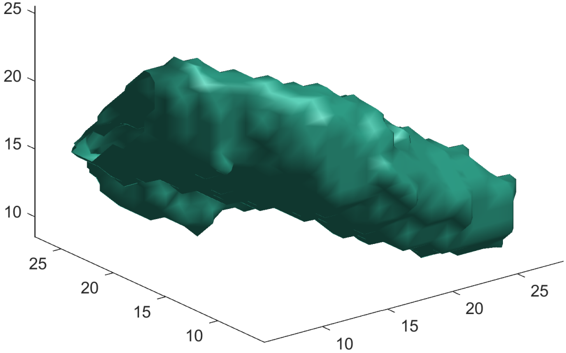
1. 284 Normal
2. 307 Mild Cognitive Impairment
3. 178 Late Mild Cognitive Impairment
4. 226 Alzheimer's Disease (AD)

Each hippocampus converted to a $32 \times 32 \times 32$ binary cubical grid.

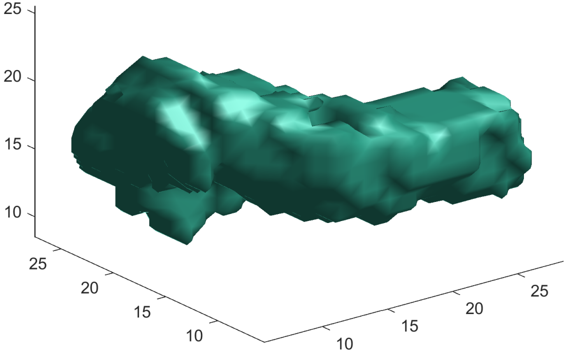
Left Hippocampi



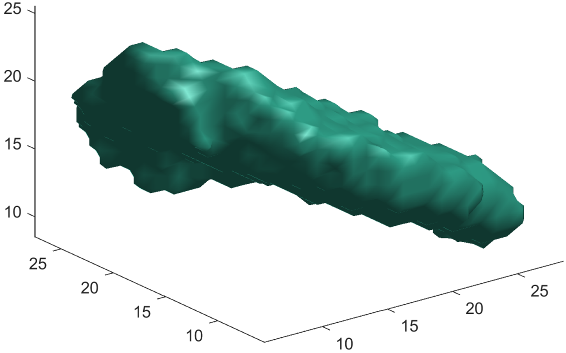
Left Hippocampi



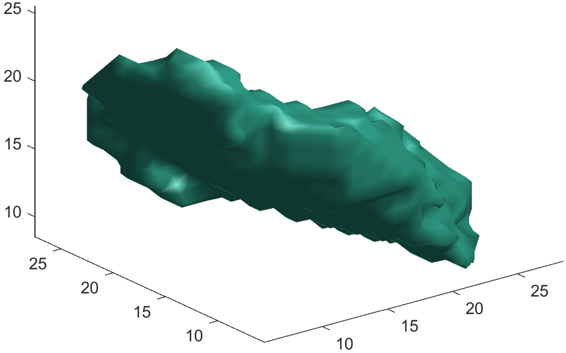
Left Hippocampi



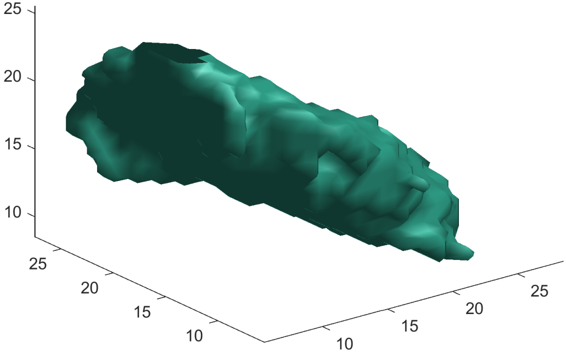
Left Hippocampi



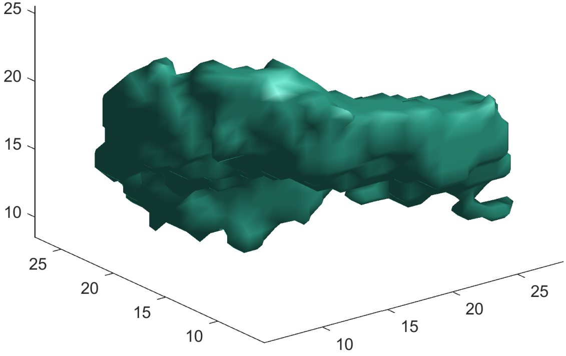
Left Hippocampi



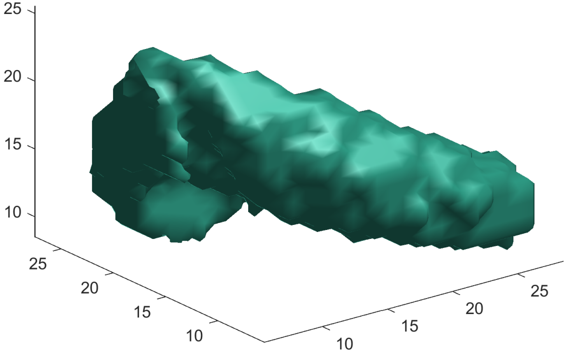
Left Hippocampi



Left Hippocampi



Left Hippocampi



Persistent homology transform

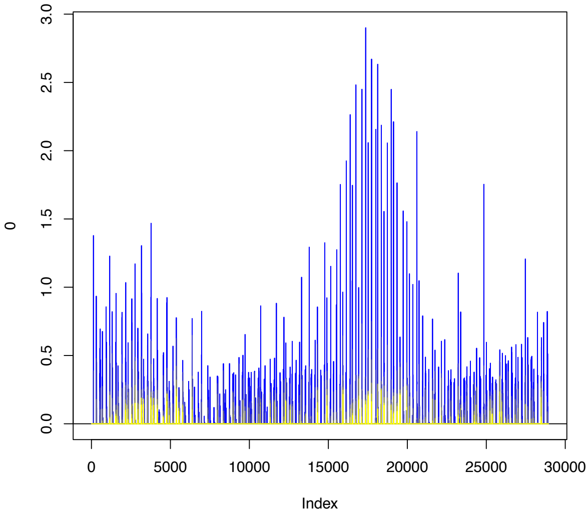
Theorem (Turner, Mukherjee, Boyer (2014))

For a surface in $\mathbb{R}^3$, persistent homology of sublevel sets in all directions is a sufficient statistic.

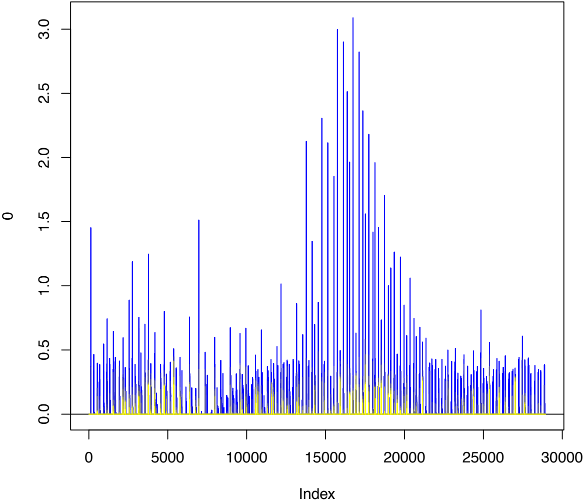
Our approach:

- filter each hippocampus in 144 directions
- calculate persistent homology
- convert to persistence landscape
- concatenate

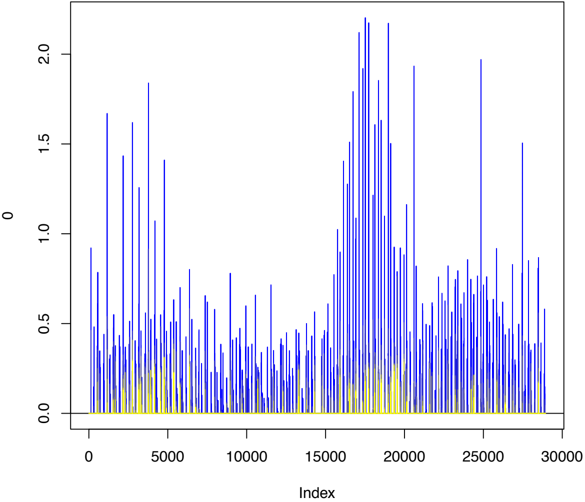
Persistence Landscape Transform



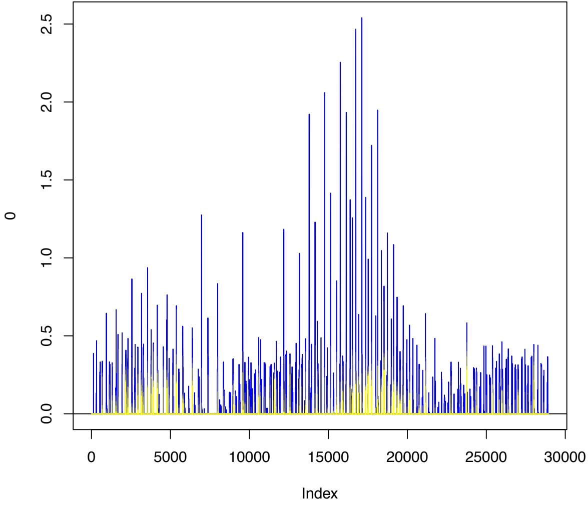
Persistence Landscape Transform



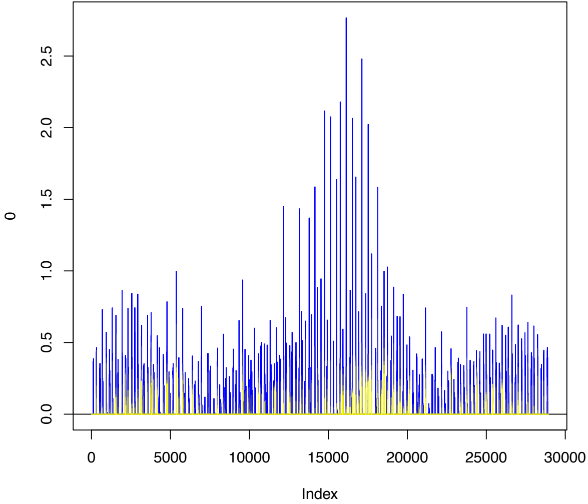
Persistence Landscape Transform



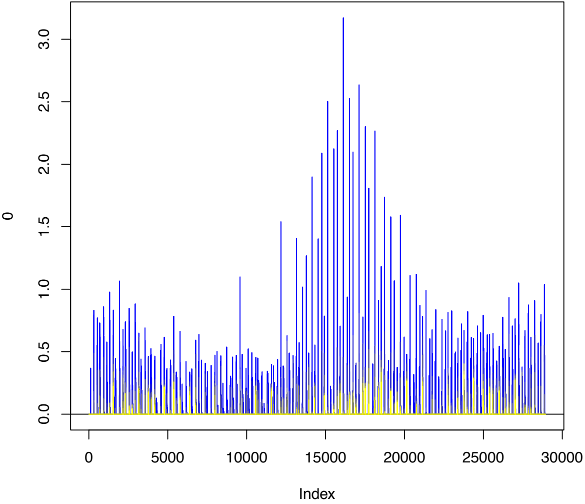
Persistence Landscape Transform



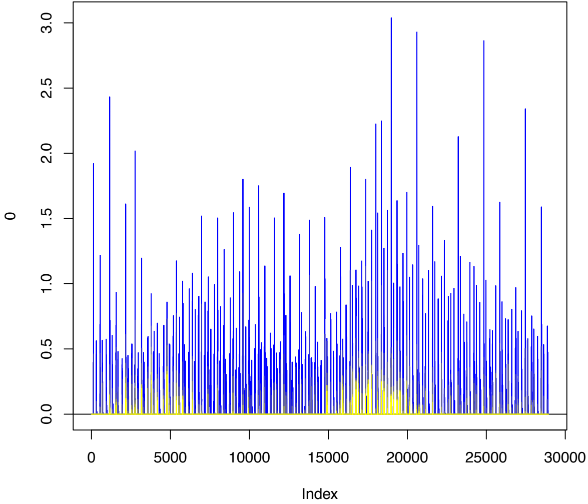
Persistence Landscape Transform



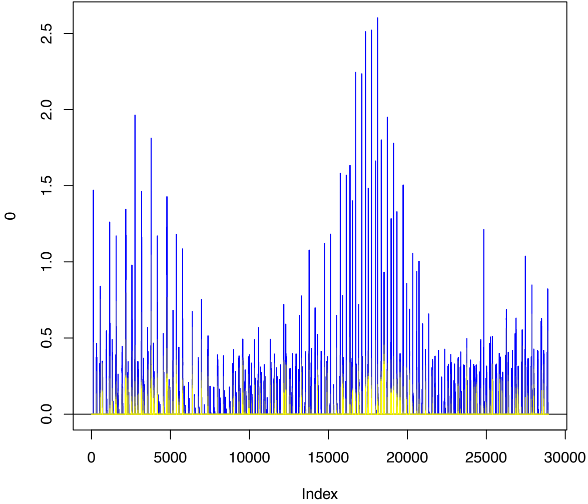
Persistence Landscape Transform



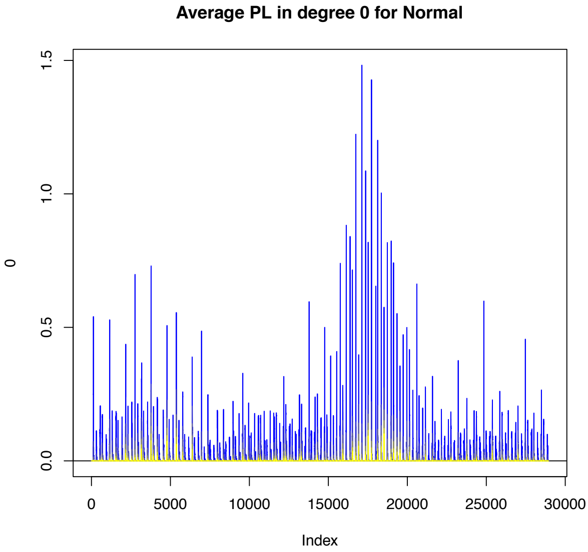
Persistence Landscape Transform



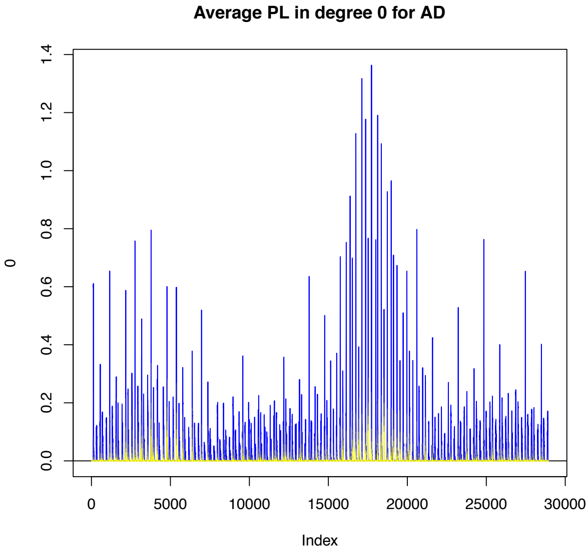
Persistence Landscape Transform



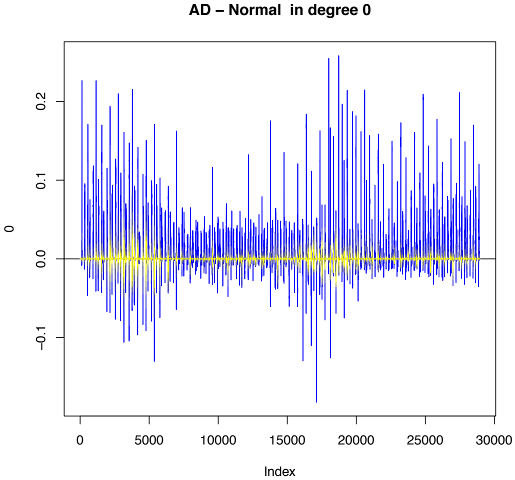
Average Landscapes



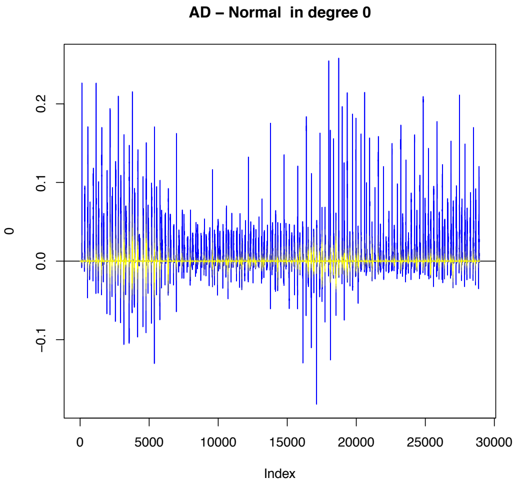
Average Landscapes



Average Landscape Difference

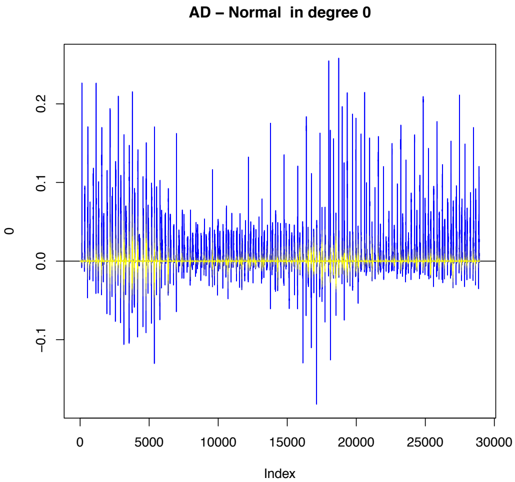


Average Landscape Difference



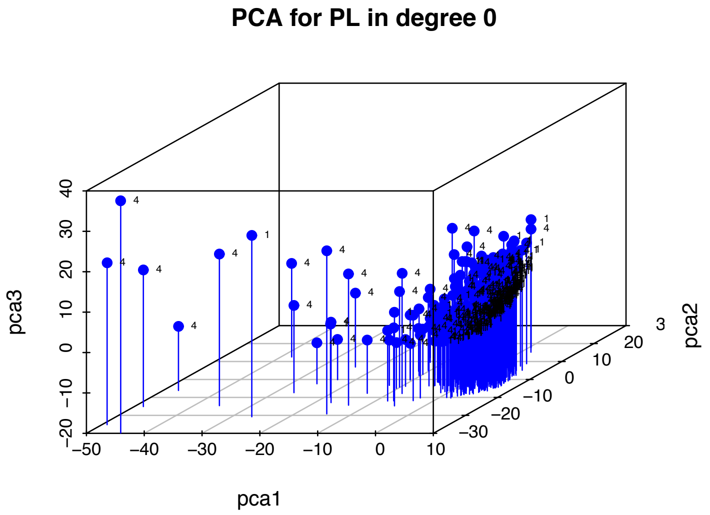
Is this difference significant?

Average Landscape Difference

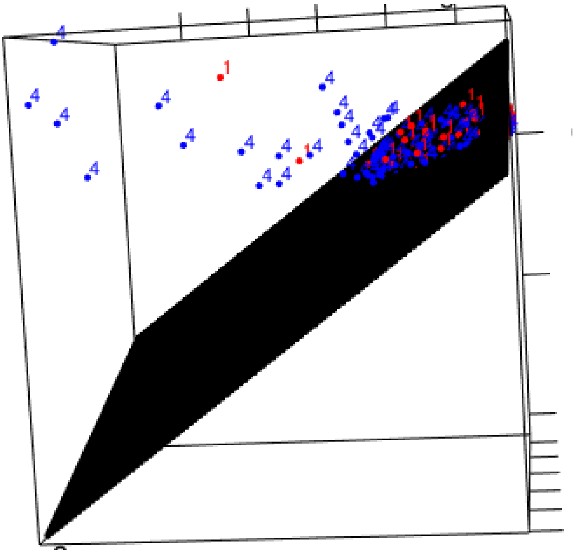


Is this difference significant? Permutation test: Yes (p val 0.000)

Principal Components Analysis



Support Vector Machine on PCA coordinates



Classification on Landscape coordinates

Support vector classification with 10-fold cross validation:

	true	
pred	Normal	Alzheimer's Disease
Normal	232	83
Alzheimer's Disease	52	143

Prediction accuracy: 73%

Topological Data Analysis Summary