

Algebraic topology and statistics

Peter Bubenik

Cleveland State University

August 4, 2009

NSF/CBMS conference:

Algebraic topology in applied mathematics

Introduction

Goal

Compare topological and statistical approaches to analyzing data and show how they can be combined.

Outline:

- 1 Sampled points
- 2 Sampled points and values
- 3 Simplifying the calculations
- 4 An application to brain imaging

Sampled points

The setup

By experiment, we measure n points x_1, \dots, x_n (the **sample**) on a manifold.

Assumption

There is an underlying object (compact submanifold, probability density, ...) generating the data.

Goal

Recover some (global) information about this object from the sample.

Topological approach

Replace each point with a ball of some fixed radius.

Take the union of these balls.

Topological approach

Replace each point with a ball of some fixed radius.

Take the union of these balls.

Use these balls to construct a simplicial complex whose vertices are the sample.

Topological approach

Replace each point with a ball of some fixed radius.

Take the union of these balls.

Use these balls to construct a simplicial complex whose vertices are the sample.

Vary the radius to get a filtered simplicial complex whose vertices are the sample.

Calculate its persistent homology.

Statistical approach

Replace each point with a bump function, called a **kernel**.

Take the sum of the bump functions.

Statistical approach

Replace each point with a bump function, called a **kernel**.

Take the sum of the bump functions.

Calculate the persistent homology (this will be explained soon).

Advantages and disadvantages

Balls

Good: relatively easy to use theoretically and computationally

Bad: if errors are not bounded then as $n \rightarrow \infty$, outliers cause problems

Kernels

Good: outliers are not a problem as $n \rightarrow \infty$

Bad: harder to use theoretically and computationally

Sampled points and values

The setup

By experiment, we measure n pairs $(x_1, y_1), \dots, (x_n, y_n)$ (the **sample**) where x_i is a point on a manifold \mathcal{M} and $y_i \in \mathbb{R}$.

Assumption

There is an underlying object (a function on the manifold) generating the data $(y_i = f(x_i) + \varepsilon_i)$.

Goal

Recover some (global) information about this object from the data.

Topological approach

Use linear interpolation to extend the sample to a function on the manifold.

Calculate this function's persistent homology.

Small persistent homology classes are assumed to be due to experimental error.

Statistical approach

Replace each data point with a bump function (kernel).

Sum these kernel to get an **estimator** of the function.

Precisely: given kernel functions $K_{x_i}(x)$ centered at x_i , take the kernel weighted average

$$\tilde{f}(x) = \frac{\sum_i K_{x_i}(x)y_i}{\sum_i K_{x_i}(x)}.$$

Statistical approach

Replace each data point with a bump function (kernel).

Sum these kernel to get an **estimator** of the function.

Precisely: given kernel functions $K_{x_i}(x)$ centered at x_i , take the kernel weighted average

$$\tilde{f}(x) = \frac{\sum_i K_{x_i}(x)y_i}{\sum_i K_{x_i}(x)}.$$

Calculate this function's persistent homology.

Advantages and disadvantages

Interpolation

Good: Relatively easy to use computationally

Bad: If the errors are unbounded then outliers cause problems as $n \rightarrow \infty$.

Kernels

Good: Outliers are not a problem, get better and better estimator as $n \rightarrow \infty$.

Bad: Calculating critical points of the estimator is difficult.

Simplifying the calculations

Both topologists and statisticians have methods that simplify their constructions when the sample is large.

Topology: Landmark points

As the size of the sample increases, the Čech complex and Vietoris–Rips complex become increasingly expensive to compute.

One solution, is to choose a small set of points, called **landmark points** from which to build a smaller simplicial complex.

Statistics: Design points

As the size of the sample increases, the kernel estimator approaches the function. However, calculating the critical points of the estimator become increasingly expensive to compute.

One solution, is to use the kernel estimator \tilde{f} to construct a simpler estimator.

Choose a small set of points called **design points**.

Evaluate the kernel estimator at these design points.

Use linear interpolation to obtain a simpler estimator \hat{f} .

A theorem

Let \mathcal{M} be a compact d -dimensional Riemannian manifold. Assume that there exists a function $f : \mathcal{M} \rightarrow \mathbb{R}$ such that

$$y = f(x) + \epsilon, \quad x \in \mathcal{M}$$

where ϵ is a normal random variable with mean zero and variance $\sigma^2 > 0$. Assume f is in a Lipschitz class of functions.

Theorem (B-P.Kim-Z.Luo)

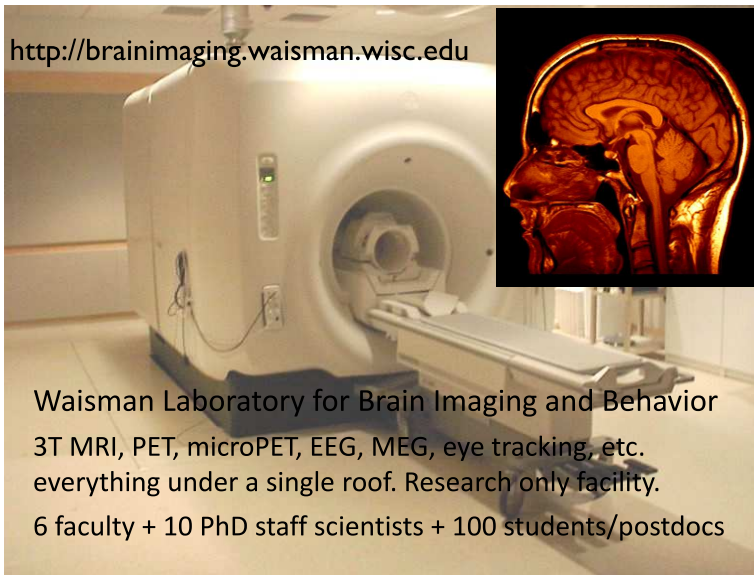
Given a sample $(x_1, y_1), \dots, (x_n, y_n)$, there is an estimator \hat{f} (constructed as above) such that

$$\mathbb{E}d_B(D_p(\hat{f}), D_p(f)) \leq C\psi_n$$

as $n \rightarrow \infty$.

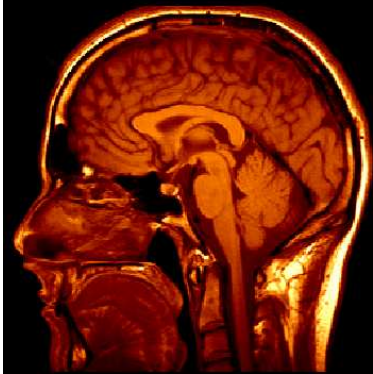
Application to Brain Imaging

<http://brainimaging.waisman.wisc.edu>



Waisman Laboratory for Brain Imaging and Behavior
3T MRI, PET, microPET, EEG, MEG, eye tracking, etc.
everything under a single roof. Research only facility.
6 faculty + 10 PhD staff scientists + 100 students/postdocs

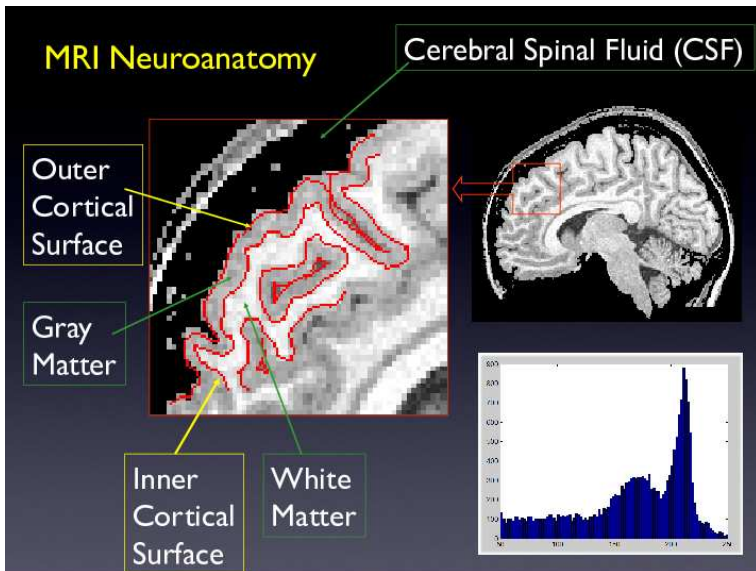
3 Tesla Magnetic Resonance Imaging



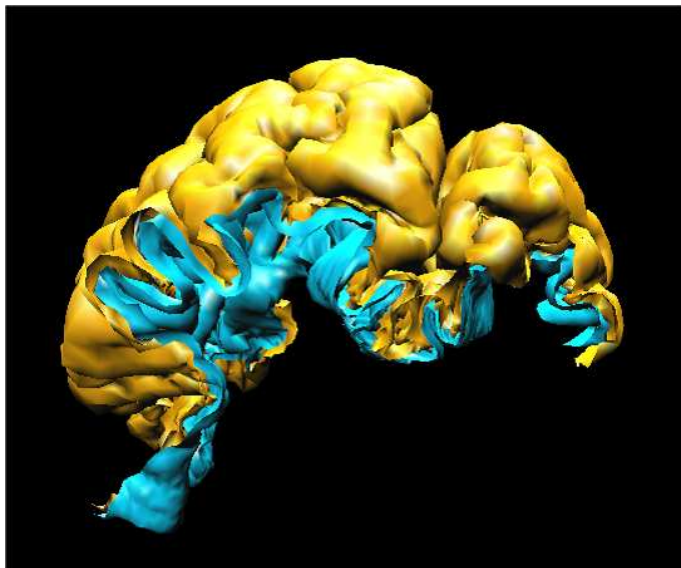
1 brain image
= $200 \times 200 \times 100$ array
= 4million measurements

16 autistic &
12 normal controls
age matched right-handed
males

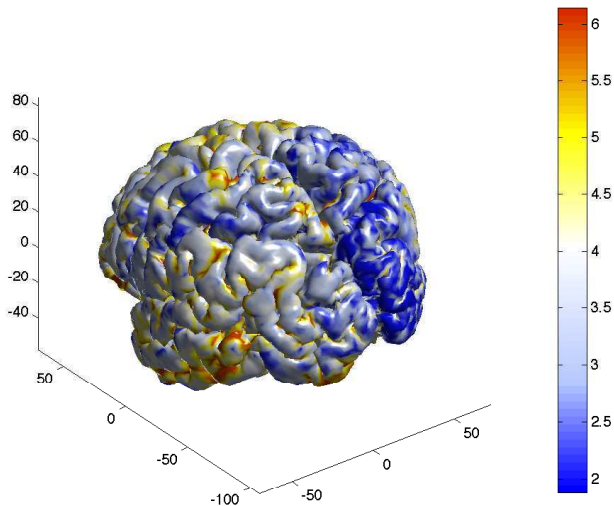
MRI



Cortical surface

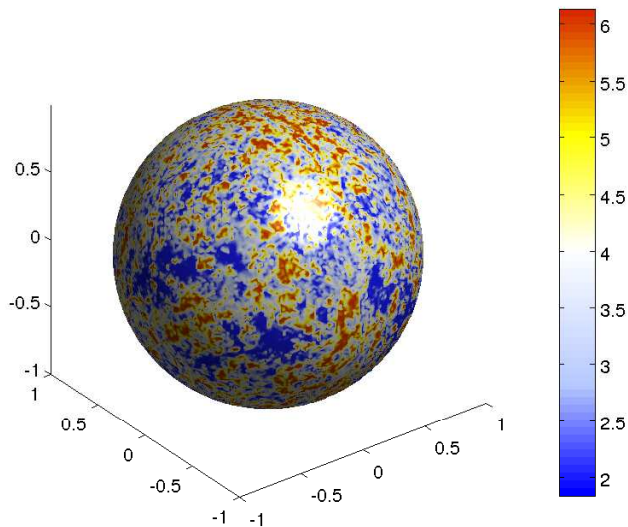


Cortex thickness



Cortex thickness

The data



Constructing our estimator

Construct an estimator:

First, smooth the data on S^2 using the kernel

$$K_{x_i}(x) = \max(1 - \kappa \arccos(x_i^t x), 0),$$

and the kernel function estimator

$$\tilde{f}(x) = \frac{\sum_i y_i K_{x_i}(x)}{\sum_i K_{x_i}(x)}. \quad (1)$$

Constructing our estimator

Construct an estimator:

First, smooth the data on S^2 using the kernel

$$K_{x_i}(x) = \max(1 - \kappa \arccos(x_i^t x), 0),$$

and the kernel function estimator

$$\tilde{f}(x) = \frac{\sum_i y_i K_{x_i}(x)}{\sum_i K_{x_i}(x)}. \quad (1)$$

Next, choose design points from a triangulation of the sphere: take an iterated subdivision of the icosahedron, which has 1280 faces and 642 vertices.

Constructing our estimator

Construct an estimator:

First, smooth the data on S^2 using the kernel

$$K_{x_i}(x) = \max(1 - \kappa \arccos(x_i^t x), 0),$$

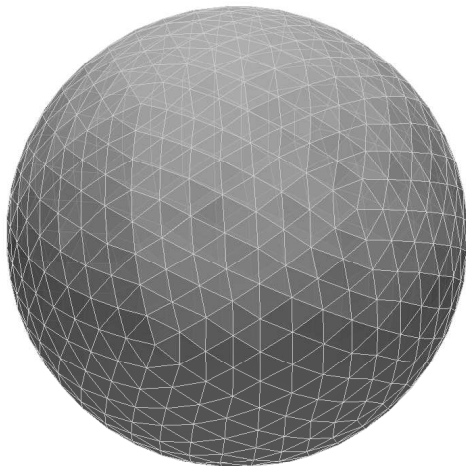
and the kernel function estimator

$$\tilde{f}(x) = \frac{\sum_i y_i K_{x_i}(x)}{\sum_i K_{x_i}(x)}. \quad (1)$$

Next, choose design points from a triangulation of the sphere: take an iterated subdivision of the icosahedron, which has 1280 faces and 642 vertices.

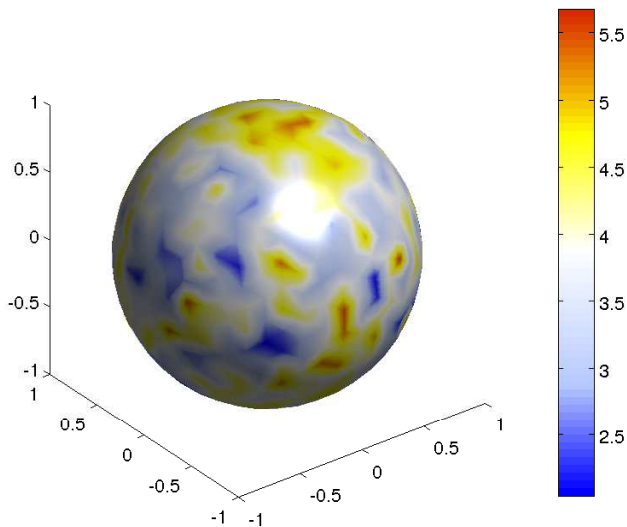
Define \hat{f} on vertices using (1) and extend by linear interpolation.

Triangulated sphere



Cortex thickness estimator

The estimator



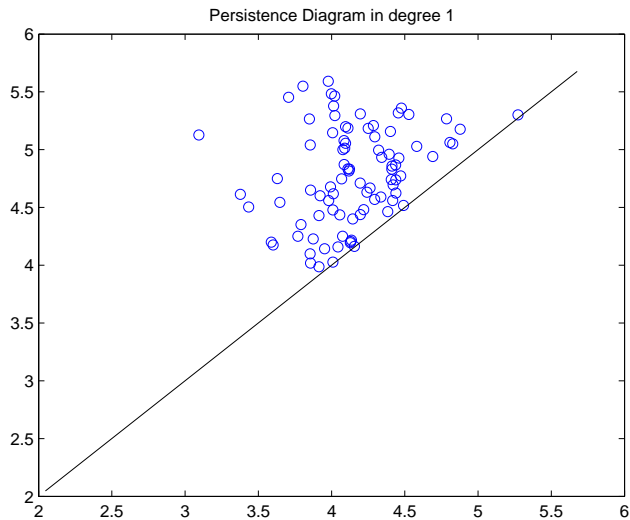
Calculating Persistent Homology

Remark

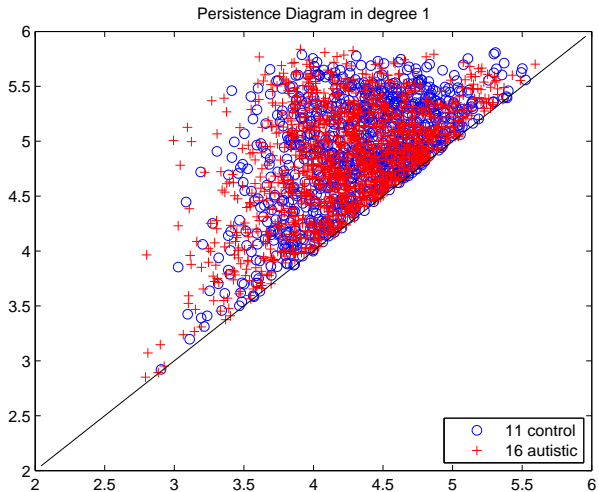
- Critical points only occur at vertices.
- The values of the estimator at the vertices, induce a filtration of the triangulation of the sphere.
- The persistent homology of this filtered complex is identical to the persistent homology of the estimator.

Use Plex to calculate the persistent homology of the filtered complex.

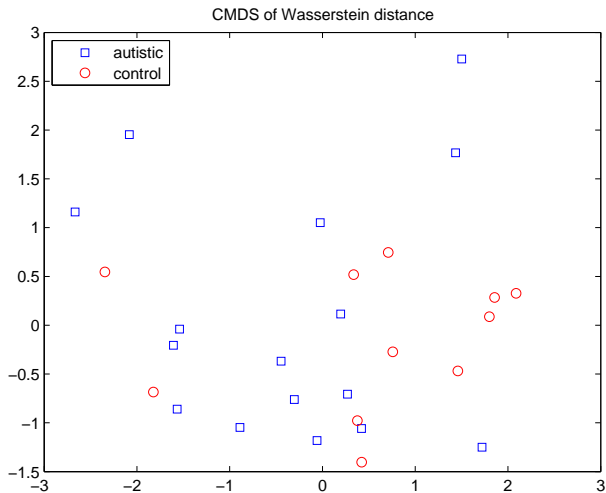
Persistence diagrams



Cumulative Persistence diagrams



Wasserstein distance



Summary

Both topologists and statisticians replace a point with an extended object (disk/kernel).

Topologists take unions, statisticians sum.

Both topologists and statisticians simplify their constructions by choosing a small set of points.

From a function on a manifold, we can consider the persistent homology of its lower excursion sets.

There is a statistical estimator for such functions from which one can calculate the persistent homology combinatorially.