

Spring semester extension to High Dimensional Inference and Random Matrices

Misha Belkin, Paul Bendich, Sayan Mukherjee, John Harer

April 9, 2007

Table of contents

- 1 Geometry and random matrices
 - Official information
 - Semester program
 - Week long workshop
 - Working groups
- 2 Persistence homology – theory
 - Persistence homology
 - Consistency of persistence diagrams
- 3 Persistence homology – (Bayesian) statistical methods
 - Uncertainty and inference
- 4 Open problems

Organizers

Organizing committee: Misha Belkin and Feng Liang

Local Scientific organizer: Sayan Mukherjee

Scientific committee: Mauro Maggioni, Yury Mileyko, and Herbert Edelsbrunner

SAMSI course instructors: Misha Belkin, Yury Mileyko, and Sayan Mukherjee

<http://www.samsi.info/programs/2006ranmatprogramext.shtml>

Research foci

Integration of:

- 1 statisticians involved in statistical inference of high-dimensional data;
- 2 computer scientists interested in computational geometry and topology;
- 3 mathematicians studying multiscale representations and diffusion processes.

Buzzwords

- 1 spectral clustering
- 2 nonlinear dimensionality reduction
- 3 manifold learning

Buzzwords

- 1 spectral clustering
- 2 nonlinear dimensionality reduction
- 3 manifold learning
- 4 learning homologies
- 5 topological persistence
- 6 diffusion models on manifolds

Buzzwords

- 1 spectral clustering
- 2 nonlinear dimensionality reduction
- 3 manifold learning
- 4 learning homologies
- 5 topological persistence
- 6 diffusion models on manifolds
- 7 priors and geometry
- 8 diffusion models on manifolds
- 9 random projections

Buzzwords

- 1 spectral clustering
- 2 nonlinear dimensionality reduction
- 3 manifold learning
- 4 learning homologies
- 5 topological persistence
- 6 diffusion models on manifolds
- 7 priors and geometry
- 8 diffusion models on manifolds
- 9 random projections
- 10 sparsity in high dimensions
- 11 model selection
- 12 integral geometry

Buzzwords

- 1 spectral clustering
- 2 nonlinear dimensionality reduction
- 3 manifold learning
- 4 learning homologies
- 5 topological persistence
- 6 diffusion models on manifolds
- 7 priors and geometry
- 8 diffusion models on manifolds
- 9 random projections
- 10 sparsity in high dimensions
- 11 model selection
- 12 integral geometry
- 13 Euler characteristic
- 14 covariance matrices

Week long workshop

- Geometry and sparsity

Week long workshop

- Geometry and sparsity
- Geometry and topology

Week long workshop

- Geometry and sparsity
- Geometry and topology
- Machine learning

Week long workshop

- Geometry and sparsity
- Geometry and topology
- Machine learning
- Random matrices and covariance

Speaker abstracts: http://www.samsi.info/workshops/geo_speakers_200701.pdf

Geometry and sparsity

- Sparsity in High Dimensional Learning Problems – Vladimir Koltchinskii

Geometry and sparsity

- Sparsity in High Dimensional Learning Problems – Vladimir Koltchinskii
- Grouped and Hierarchical Selection through Composite Absolute Penalties (CAPs) – Guilherme Rocha

Geometry and sparsity

- Sparsity in High Dimensional Learning Problems – Vladimir Koltchinskii
- Grouped and Hierarchical Selection through Composite Absolute Penalties (CAPs) – Guilherme Rocha
- On Kernels, Energy and Metrics – Steven Damelin

Geometry and topology

- Integral geometry of random sets – Jonathan Taylor

Geometry and topology

- Integral geometry of random sets – Jonathan Taylor
- The 3D Random Computed Tomography Structuring of Proteins and the Spectral Non-Linear ICA Algorithm – Amit Singer

Geometry and topology

- Integral geometry of random sets – Jonathan Taylor
- The 3D Random Computed Tomography Structuring of Proteins and the Spectral Non-Linear ICA Algorithm – Amit Singer
- Computing Homology of High-Dimensional Point Clouds – Yuriy Mileyko

Geometry and topology

- Integral geometry of random sets – Jonathan Taylor
- The 3D Random Computed Tomography Structuring of Proteins and the Spectral Non-Linear ICA Algorithm – Amit Singer
- Computing Homology of High-Dimensional Point Clouds – Yuriy Mileyko
- A Probabilistic View and Fundamental Limitations of Spectral Clustering – Boaz Nadler

Machine learning

- A Geometric Perspective on Learning – Partha Niyogi

Machine learning

- A Geometric Perspective on Learning – Partha Niyogi
- Random Geometry and Heat Kernels in Text Analysis – Guy Lebanon

Machine learning

- A Geometric Perspective on Learning – Partha Niyogi
- Random Geometry and Heat Kernels in Text Analysis – Guy Lebanon
- Projection Pursuit, Gaussian Scale Mixtures, and the EM Algorithm – Sanjoy Dasgupta

Random Matrices and Covariances

- Regularized Estimation of Large Covariance Matrices – Liza Levina

Random Matrices and Covariances

- Regularized Estimation of Large Covariance Matrices – Liza Levina
- Distributions for Random Positive Definite Matrices – Armin Schwartzman

Random Matrices and Covariances

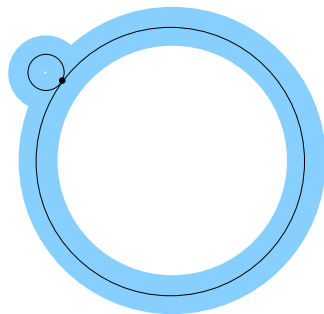
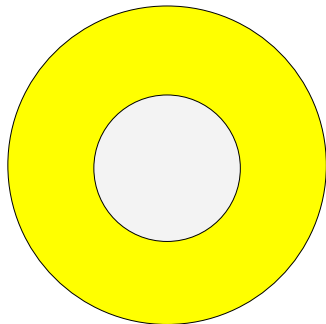
- Regularized Estimation of Large Covariance Matrices – Liza Levina
- Distributions for Random Positive Definite Matrices – Armin Schwartzman
- Multi-resolution Covariance Modelling in Spatial Random Effect Model – Tao Shi

Working group

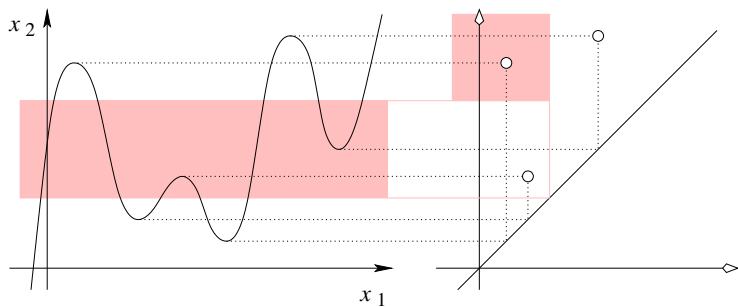
Statistical and probabilistic aspects of (computational) topology:

- Yury Mileyko – **CS and Math**
- Misha Belkin – CS
- Yoonkyung Lee – Stats
- Paul Bendich – **Math**
- Dmitriy Morozov – **CS**
- Bei Wang – CS
- Amit Patel – CS
- Sayan Mukherjee – Stats, CS, Comp Bio

Two pictures



Persistence



Persistence and high dimensional data

$$X \subset \mathbb{R}^d$$

- Is X a set of samples from a manifold? If so, how can we “learn” what the manifold is?
- What is its general shape? Does it look like a circle? A torus? A big X ?
- If it is something specific, how do we interpret it in terms of the original data?

α -complexes

Model the underlying space using simplicial complex.

1. Set X as a vertex set and add a simplex $[x_0, \dots, x_n]$ if there is a point P such that

$$d(P, x_0) = \dots = d(P, x_n) \leq d(P, y), \quad \forall y \in \mathbb{R}^d.$$

α -complexes

Model the underlying space using simplicial complex.

1. Set X as a vertex set and add a simplex $[x_0, \dots, x_s]$ if there is a point P such that

$$d(P, x_0) = \dots = d(P, x_n) \leq d(P, y), \quad \forall y \in \mathbb{R}^d.$$

2. Associate to the simplex the radius $r = d(P, y)$ that it first occurs.

α -complexes

Model the underlying space using simplicial complex.

1. Set X as a vertex set and add a simplex $[x_0, \dots, x_s]$ if there is a point P such that

$$d(P, x_0) = \dots = d(P, x_n) \leq d(P, y), \quad \forall y \in \mathbb{R}^d.$$

2. Associate to the simplex the radius $r = d(P, y)$ that it first occurs.
3. Vary r from 0 to ∞ adding simplices as they appear.

α -complexes

Model the underlying space using simplicial complex.

1. Set X as a vertex set and add a simplex $[x_0, \dots, x_s]$ if there is a point P such that

$$d(P, x_0) = \dots = d(P, x_n) \leq d(P, y), \quad \forall y \in \mathbb{R}^d.$$

2. Associate to the simplex the radius $r = d(P, y)$ that it first occurs.
3. Vary r from 0 to ∞ adding simplices as they appear.
4. Track how adding each simplex changes the homology.

α -complexes

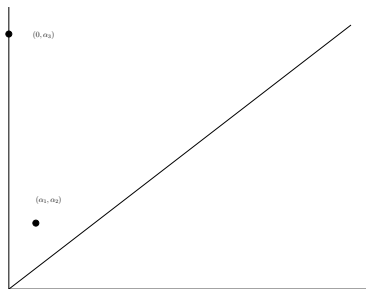
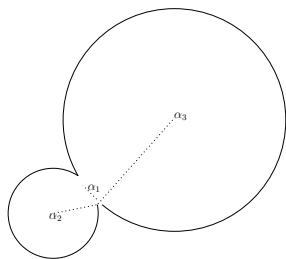
Model the underlying space using simplicial complex.

1. Set X as a vertex set and add a simplex $[x_0, \dots, x_s]$ if there is a point P such that

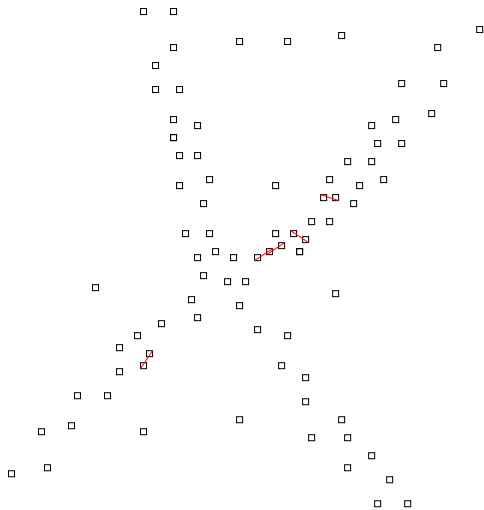
$$d(P, x_0) = \dots = d(P, x_n) \leq d(P, y), \quad \forall y \in \mathbb{R}^d.$$

2. Associate to the simplex the radius $r = d(P, y)$ that it first occurs.
3. Vary r from 0 to ∞ adding simplices as they appear.
4. Track how adding each simplex changes the homology.
5. See when each class is born and when it dies.

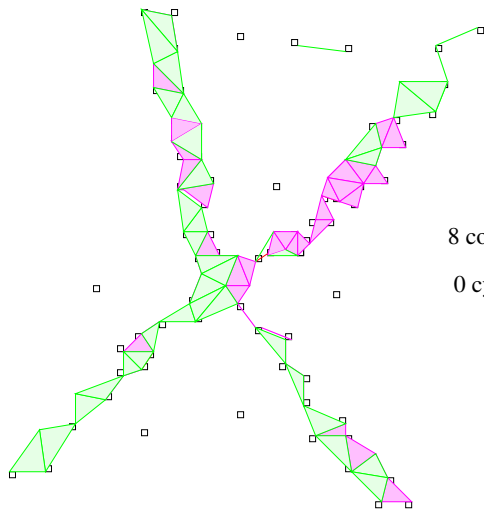
An example



α -complexes



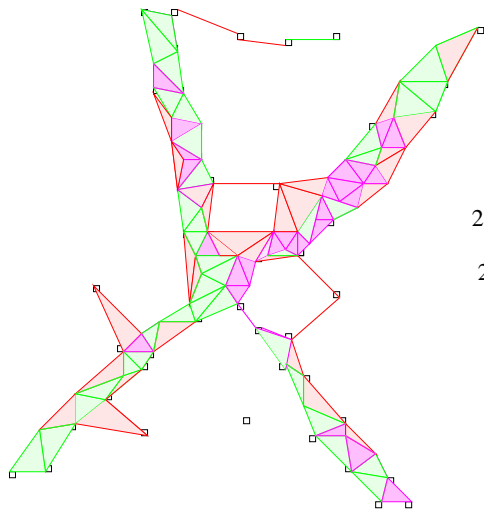
α -complexes



8 components

0 cycles

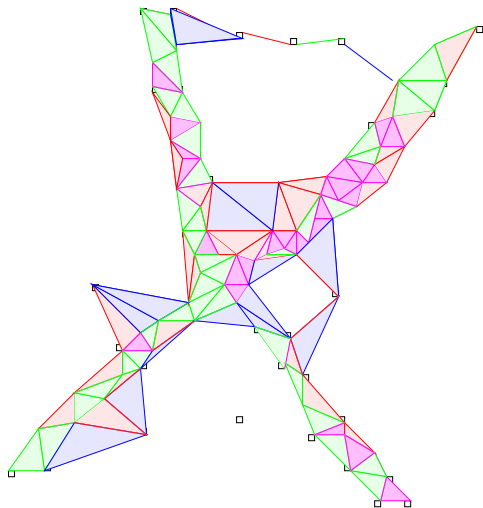
α -complexes



2 components

2 cycles

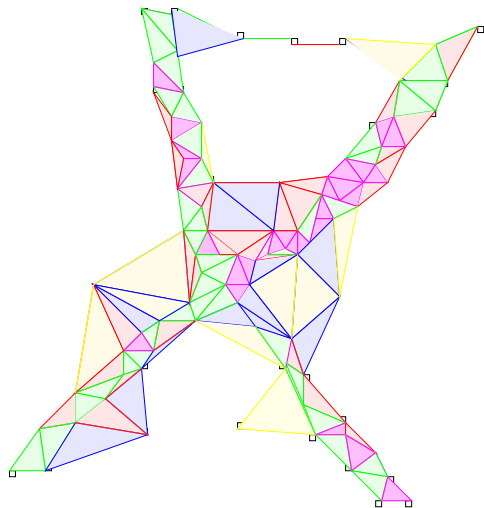
α -complexes



2 components

2 cycles

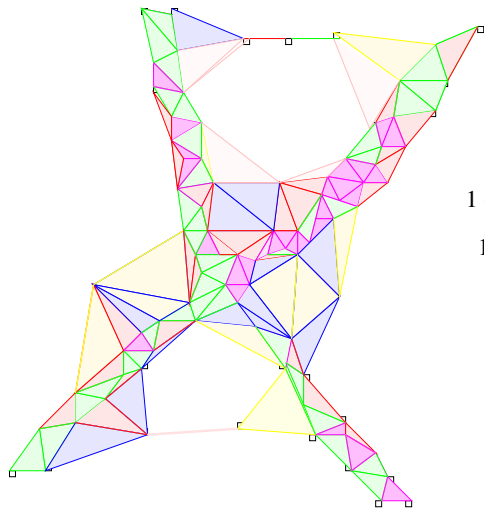
α -complexes



1 component

1 cycle

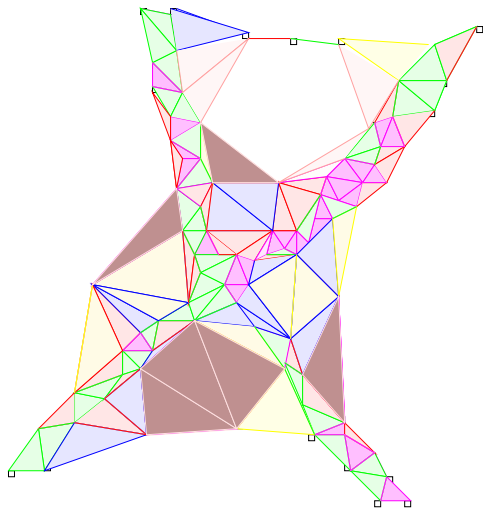
α -complexes



1 component

1 cycle

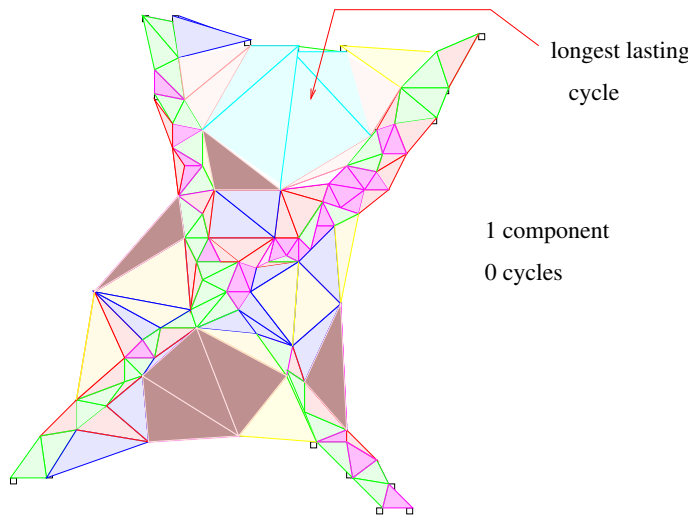
α -complexes



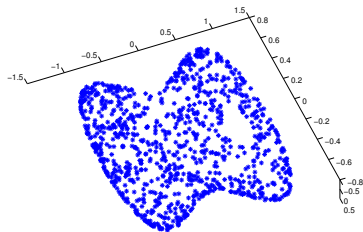
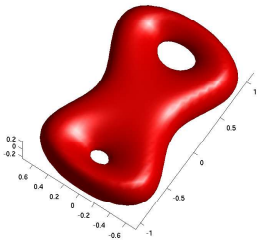
1 component

1 cycle

α -complexes



Manifolds and samples



Related work

Finding the Homology of Submanifolds with High Confidence from Random Samples. P. Niyogi, S. Smale, and S. Weinberger. to appear, Discrete and Computational Geometry, 2006.

A theorem

Given a manifold $\mathcal{M} \in \mathbb{R}^d$ with uniform measure ρ we draw a point sample P of n points i.i.d. from ρ . The persistence diagram of this point set and manifold are $D(P)$ and $D(\mathcal{M})$.

Theorem (Probabilistic Persistent Homology Inference)

Given n points then with probability at least $1 - \delta$ the points in $D(P)$ which are at least ϵ away from the diagonal represent actual persistent homological features. Moreover, for every one of those points $x \in D(P)$, the point of $D(\mathcal{M})$ representing the true persistent homology class lies in an ϵ -box around x . Under mild assumptions

$$\epsilon = \mathcal{O} \left(\log\left(\frac{1}{\delta}\right) n^{-1/d} \right).$$

Its corollary

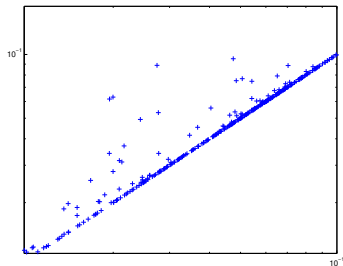
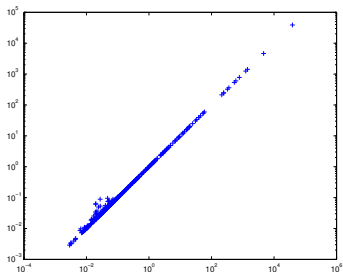
Corollary (Probabilistic Homology Inference Theorem)

Given n points and probability at least $1 - \delta$ and if for the same ϵ as above

$$\frac{\epsilon}{2} < \frac{hfs(\mathcal{M})}{8}$$

then with probability at least $1 - \delta$ for all dimensions i the dimension of $H_i(\mathcal{M})$ is the same as the number of points in $D_i(P)$ which lie above and to the left of the point $(\epsilon, 3\epsilon)$.

Persistence diagram for a double torus sample



What about uncertainty ?

A Bayesian formulation

The α shapes (complex) can be thought in a probabilistic framework. They can be thought of as parameters θ of a density giving rise to the point cloud P . In this context one can state a likelihood formulation

$$\text{Lik}(P|\theta),$$

where θ is a finite set of parameters that determine the simplices. Similarly priors can be placed on the density parameters θ

$$\pi(\theta).$$

Posterior probabilities on diagrams

Thus given a point cloud P we can compute a posterior probability for θ and the persistence diagram $D(\theta)$ corresponding to the simplices determined by θ

$$\Pr(\theta|P) \propto \text{Lik}(P|\theta)\pi(\theta)$$

$$\Pr(D(\theta)|P) \propto \text{Lik}(P|\theta)\pi(\theta).$$

This provides us with probabilities over diagrams.

Averaging diagrams

It is natural to want to compute the posterior mean diagram

$$\bar{D} = \int_{\theta} D(\theta|P) \Pr(D(\theta)|P) d\theta.$$

However how does one average diagrams.

Stability of persistence diagrams provides a natural answer

$$\text{avg}_{\epsilon}(D_1, D_2) = \frac{\epsilon^2}{2} (\chi_{\epsilon}(D_1) + \chi_{\epsilon}(D_2)),$$

where $\chi_{\epsilon}(D_1)$ places a characteristic function of size ϵ around each point in D_1 .

Persistence distributions

What size ϵ ?

Integrate out ϵ .

$$\bar{D} = \int_{\theta, \epsilon} \epsilon^2 \chi_{\epsilon}(D(\theta|P)) \Pr(D(\theta)|P) d\theta d\epsilon.$$

Details – boundary conditions, the diagonal.

What isn't

Probabilistic/statistical thinking about topological statistics is in its infancy.

- 1 More general cleaner consistency results

What isn't

Probabilistic/statistical thinking about topological statistics is in its infancy.

- 1 More general cleaner consistency results
- 2 Understanding the Bayesian procedure – likelihood functions and priors

What isn't

Probabilistic/statistical thinking about topological statistics is in its infancy.

- 1 More general cleaner consistency results
- 2 Understanding the Bayesian procedure – likelihood functions and priors
- 3 Computational aspects: Čech complex, witness complex, multiscale construction

What isn't

Probabilistic/statistical thinking about topological statistics is in its infancy.

- 1 More general cleaner consistency results
- 2 Understanding the Bayesian procedure – likelihood functions and priors
- 3 Computational aspects: Čech complex, witness complex, multiscale construction
- 4 Probabilistic construction of complex using stochastic processes
- 5 What is persistence – points not near the y -axis are not strictly topological invariants. What are they with respect to the embedding ?

What isn't

Probabilistic/statistical thinking about topological statistics is in its infancy.

- 1 More general cleaner consistency results
- 2 Understanding the Bayesian procedure – likelihood functions and priors
- 3 Computational aspects: Čech complex, witness complex, multiscale construction
- 4 Probabilistic construction of complex using stochastic processes
- 5 What is persistence – points not near the y -axis are not strictly topological invariants. What are they with respect to the embedding ?
- 6 Birth-death processes