Persistent Homology: an approach for high dimensional data analysis

Rui Hu, Zhichun Zhai, Bei Jiang, and Linglong Kong

Giseon Heo

Mathematical and Statistical Challenges in Neuroimaging Data Analysis

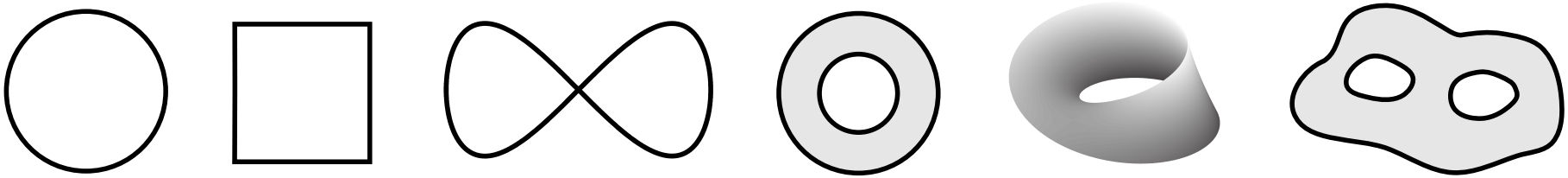
Outline

Quiz

This problem can be solved by pre-school children in five to ten minutes, by programmers in an hour and by people with higher education... well, check it yourself!

8809 = 6	5555 = 0
7111 = 0	8193 = 3
2172 = 0	8096 = 5
6666 = 4	1012 = 1
1111 = 0	7777 = 0
3213 = 0	9999 = 4
7662 = 2	7756 = 1
9313 = 1	6855 = 3
0000 = 4	9881 = 5
2222 = 0	5531 = 0
3333 = 0	2581 = ???

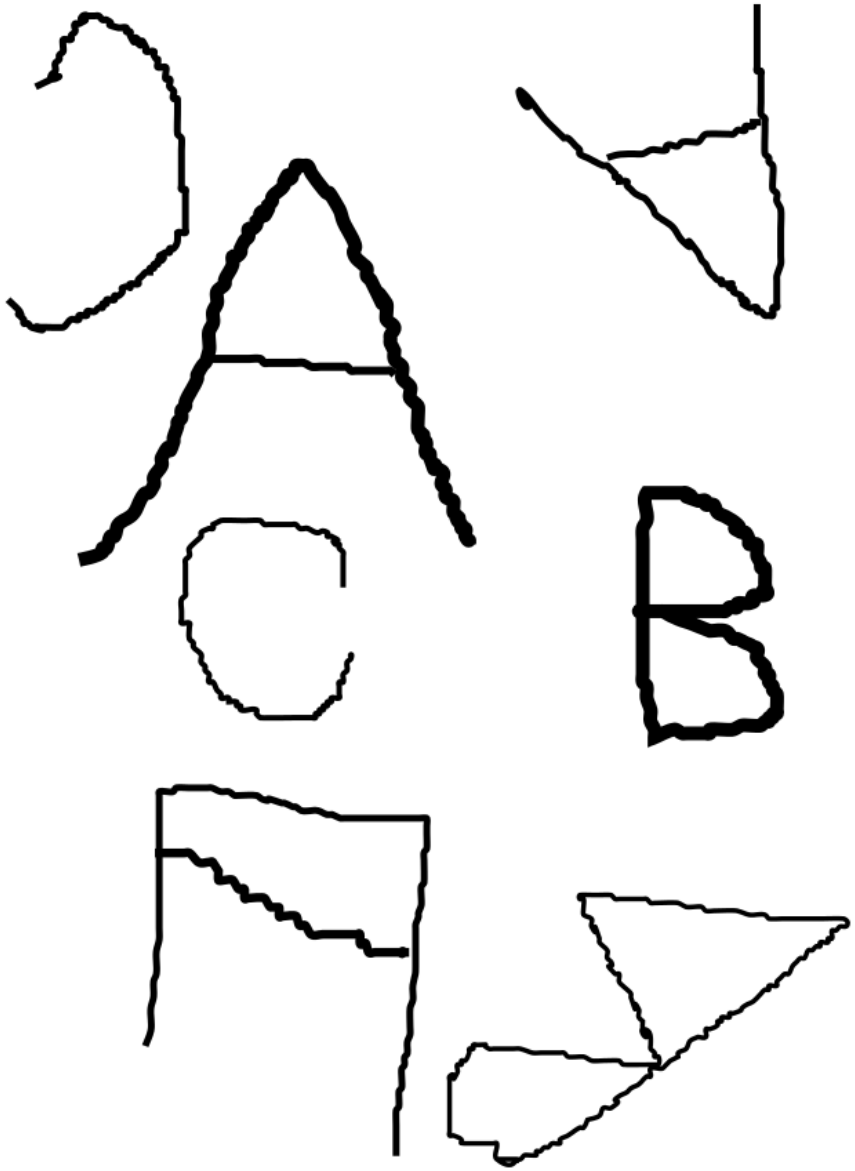
How do we distinguish geometrical objects?



What is Topology?

One popular answer: It is the branch of mathematics which does not distinguish between a teacup and a bagel.

$\cong$



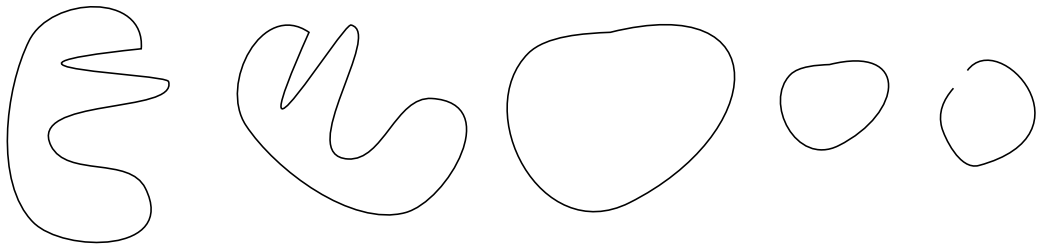
Classification of capital letters

- Equivalence classes by topological type
  $\{A, R\}, \{B\}, \{C, G, I, J, L, M, N, S, U, V, W, Z\},$
  $\{D, O\}, \{E, F, T, Y\}, \{H, K\},$
  $\{P\}, \{Q\}, \{X\}$

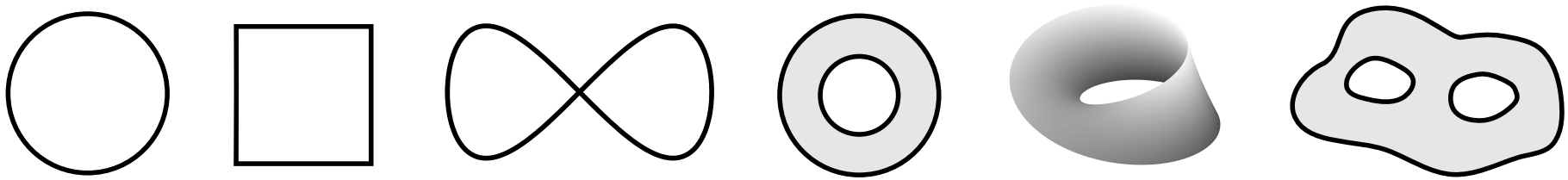
- Equivalence classes by homotopy type
  $\{A, R, D, O, P, Q\}, \{B\},$
  $\{C, I, L, M, N, S, U, V, W, Z, F, J, T, Y, G, H, K, X\}$

Topology cares connectivity

Topology allows the larger group of homeomorphisms that deform an object by stretching, shrinking, bending, but neither tearing nor gluing.



Homotopy equivalent



- A circle is homeomorphic to a square.
- **Homotopy equivalent** intuitively means that one can be continuously deformed to the other.
- The figure eight and the island with two lakes are equivalent although they are not homeomorphic to each other.
- Similarly the circle, the square, and Möbius band.

Algebraic Topology

- Algebraic topology associates to a topological space an algebraic system such as a group or a sequence of groups.
- There is a natural interplay between continuous maps $f : X \to Y$ between topological spaces and algebraic homomorphisms $f_* : G(X) \to G(Y)$ on their associated groups.
- There exist several ways to associate topological spaces with groups, but homology fits our interests the best.

Homology group

- The zero-dimensional homology group $H_0(X)$ is generated by elements representing **connected components** of X.
- $k \geq 1$ the k dimensional homology group $H_k(X)$ is generated by elements representing k-**dimensional holes** in X.

Betti numbers

- β_0 counts the number of connected components of a complex K.
- β_i, $i \geq 1$ counts the number of i dimensional holes of a complex K.

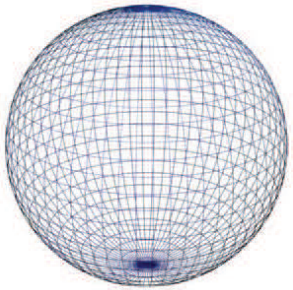
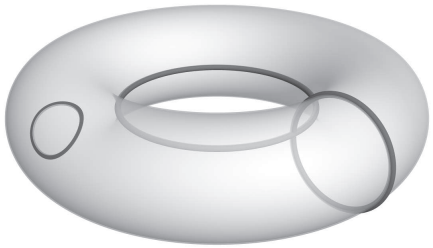
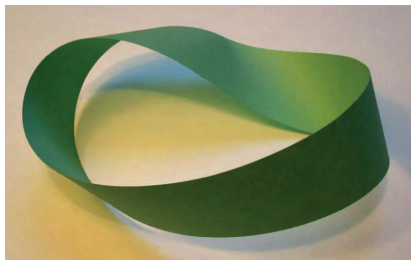
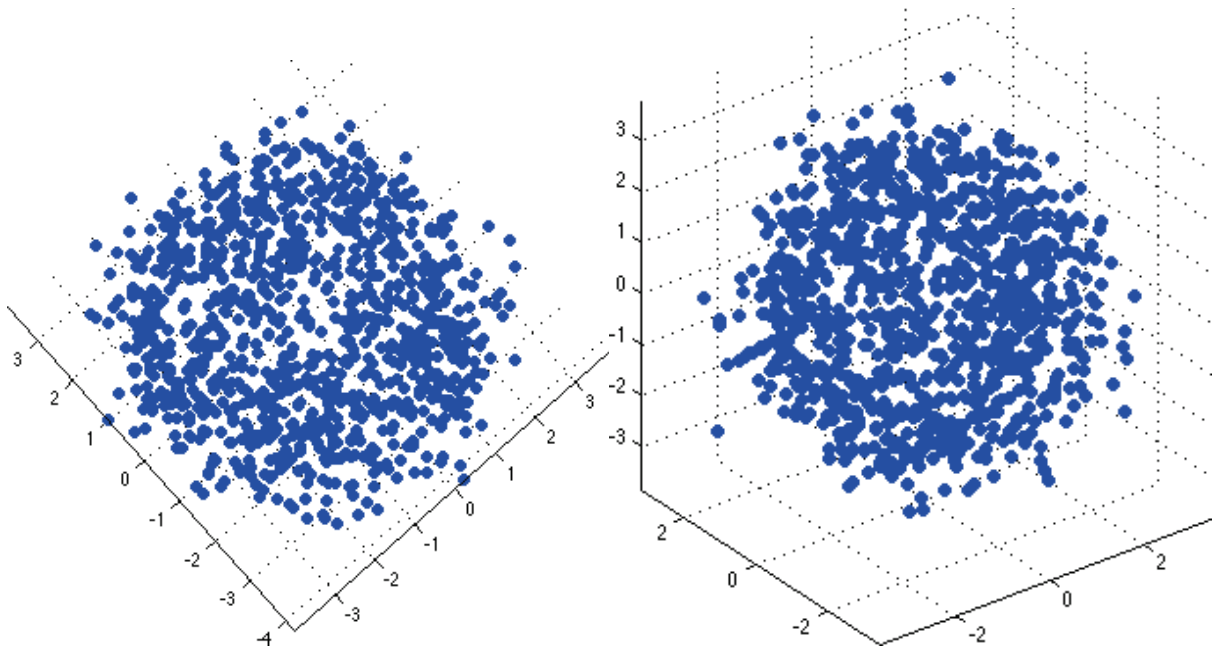




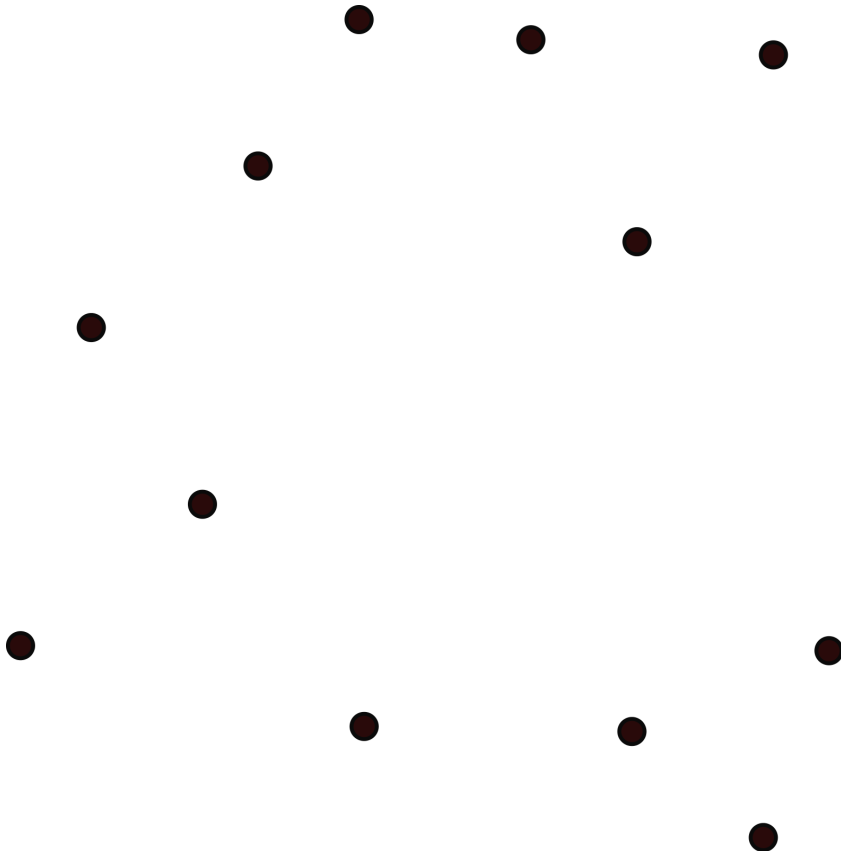

Figure: Sphere: $\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 1$. Torus : $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$.
Möbius band: $\beta_0 = 1, \beta_1 = 1$. Knot: $\beta_0 = 1, \beta_1 = 1$.

Noisy Torus vs. noisy Sphere

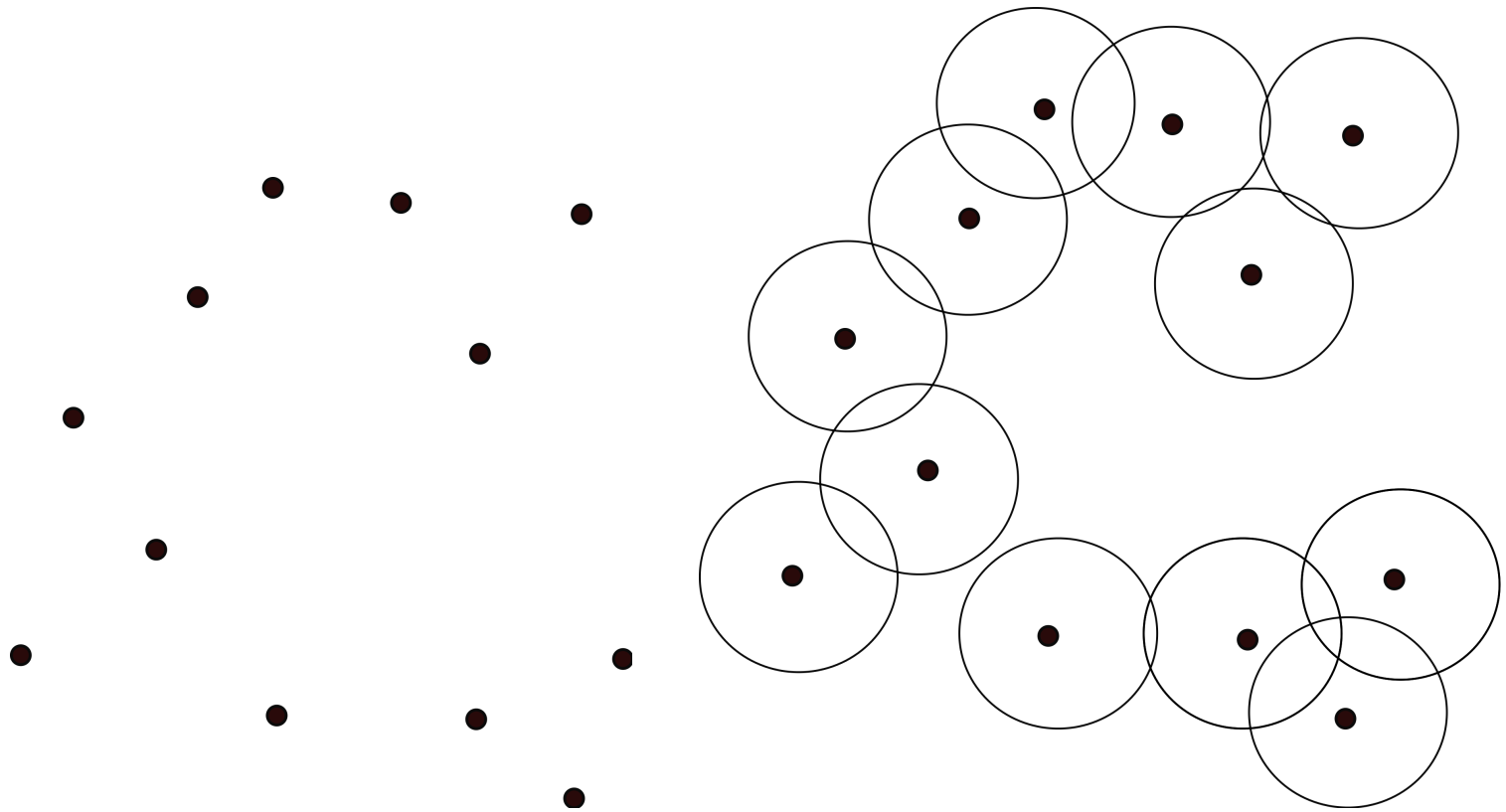


Mission: recover topology and geometry from point cloud data

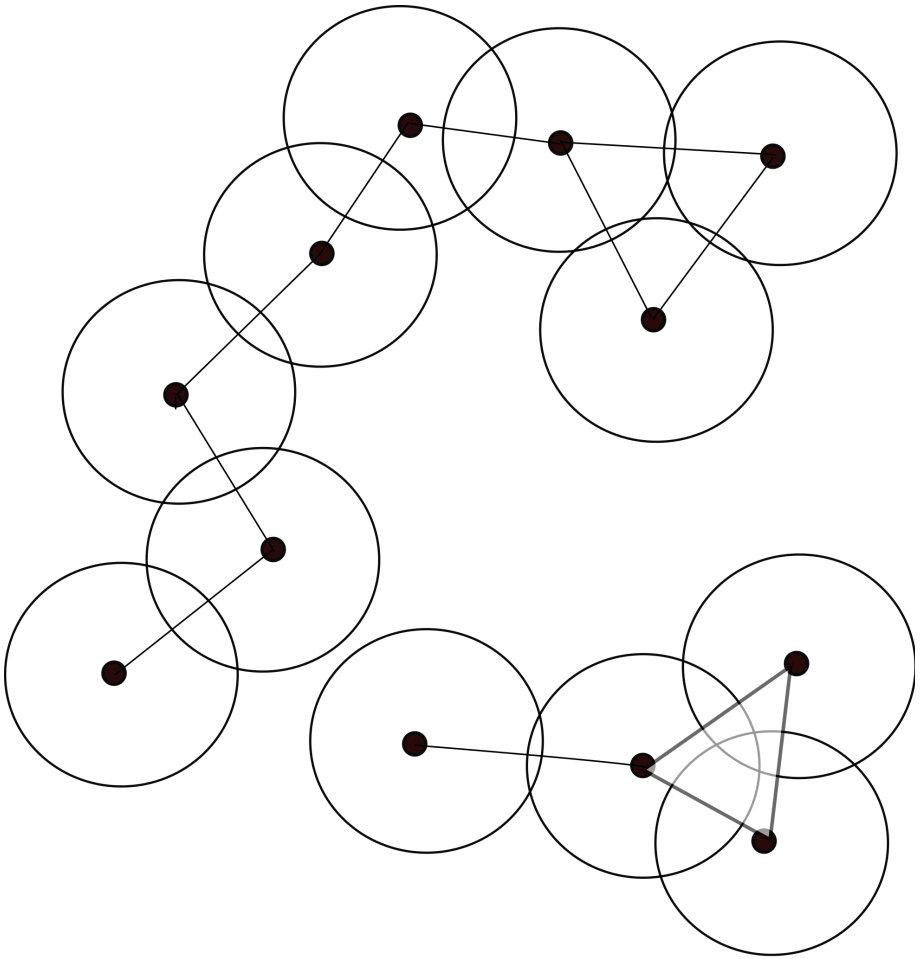
But.. set of points does not have interesting topology



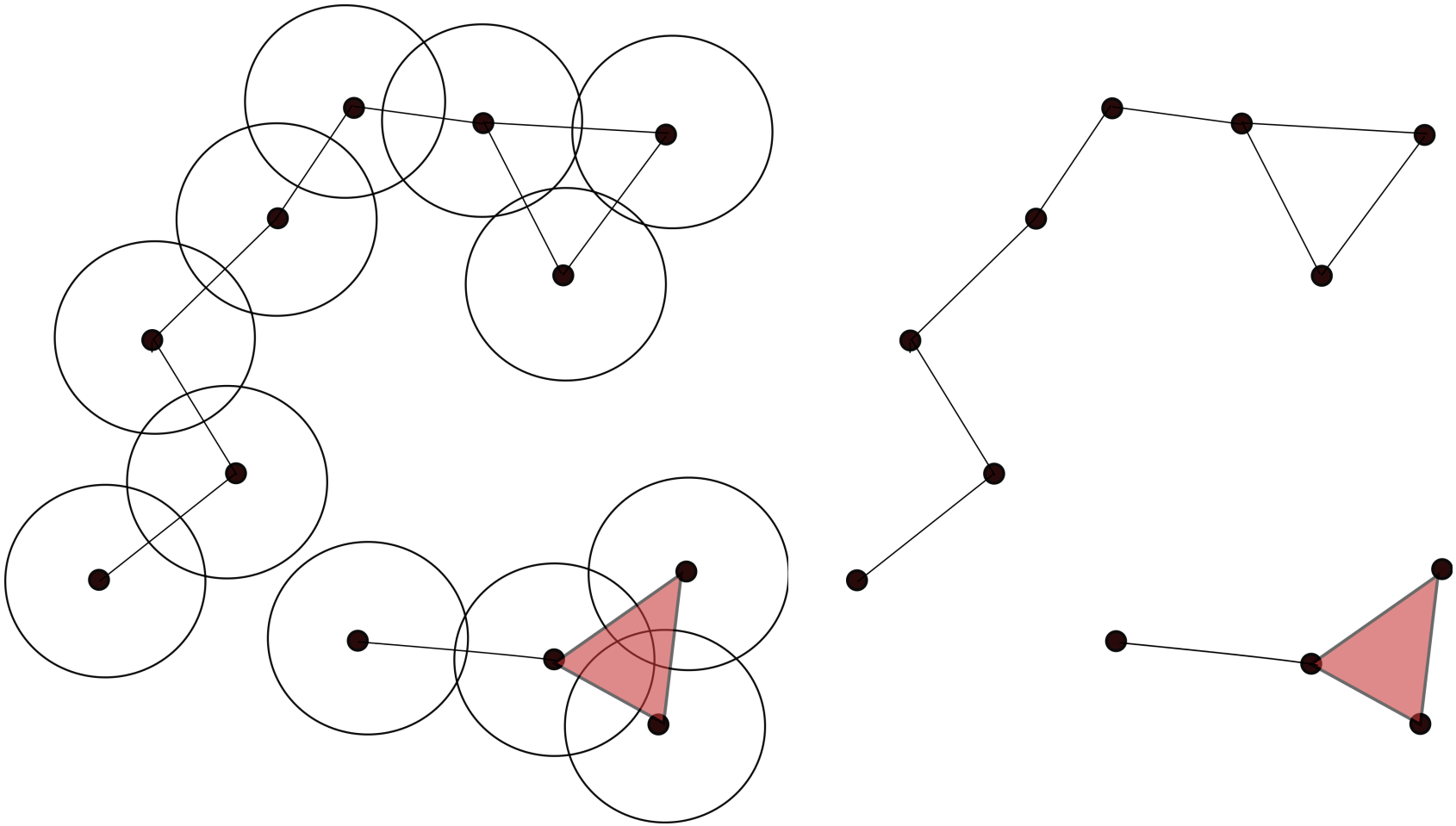
How about ϵ balls? No...



Answer: Simplicial complexes



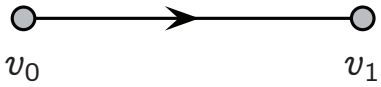
Čech Complex

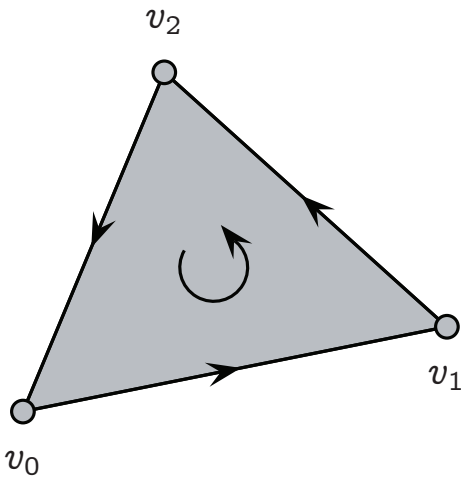


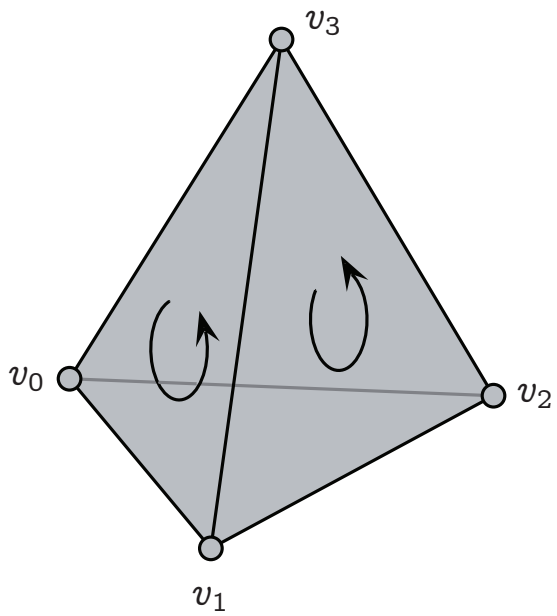
Čech (Nerve) Theorem Rotman (1988)

The *Nerve lemma* states that if the ambient space is $\mathbb{R}^{\mathrm{d}}(\mathbb{Y})$, then the Čech complex $\mathcal{C}_\epsilon(S)$ is homotopy equivalent to the union of balls, $\bigcup_i B_\epsilon(x_i)$.

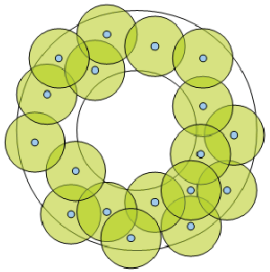
Simplexes

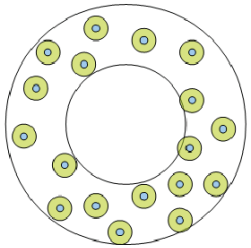


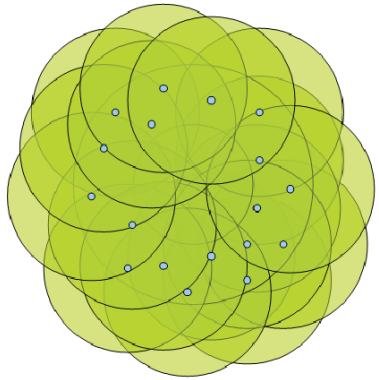




Which ε ?





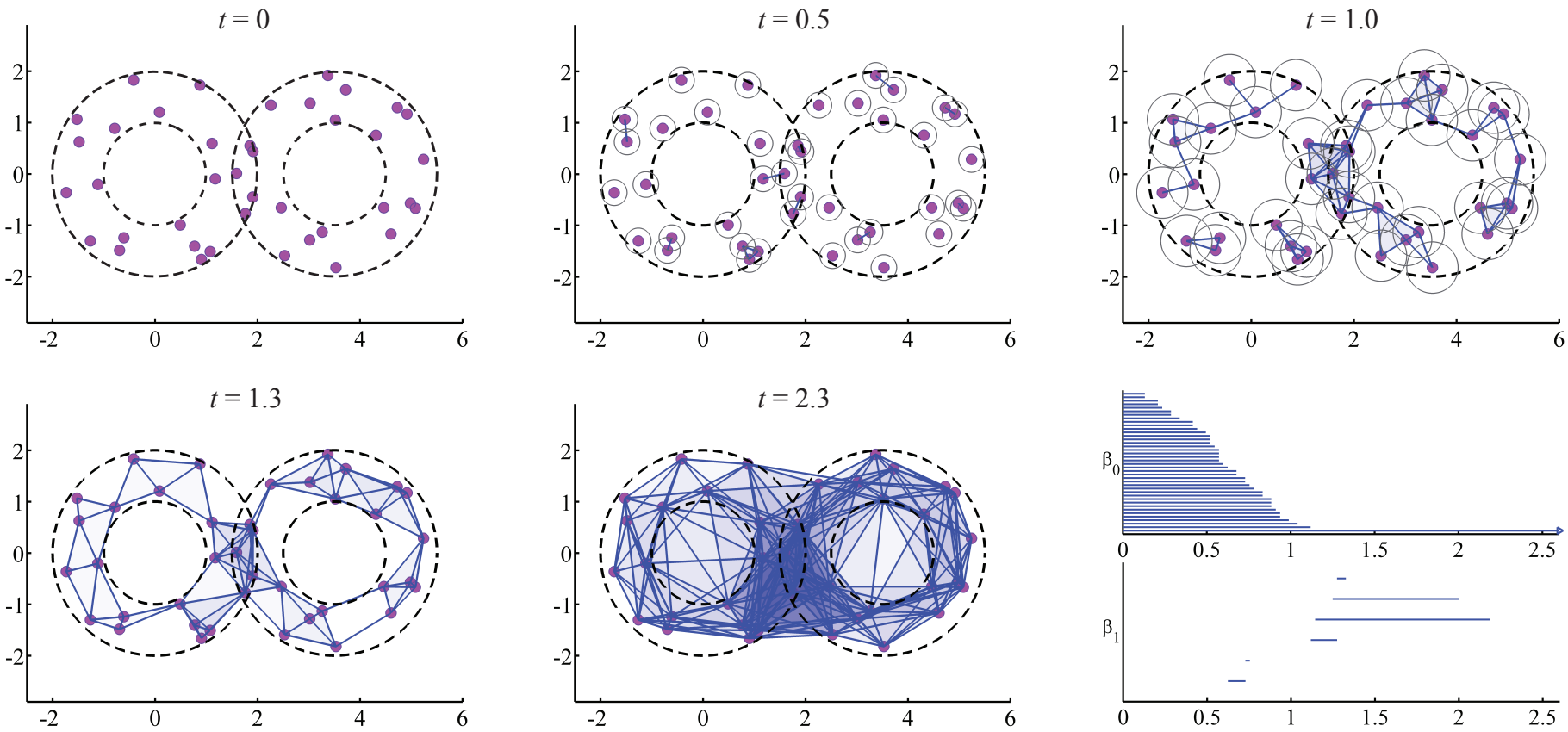


Computational topologists

Let $X_1, X_2, \ldots, X_n$ be n independent points randomly sampled from $\mathbb{X} \subseteq \mathbb{R}^{\mathrm{d}}$.

- Replace each point of discrete data set in $\mathbb{R}^{\mathrm{d}}$ by a ball of a fixed radius ε;
- Obtain **simplicial complexes** from the balls;
- Construct filtered simplicial complexes as ε increases;
- **Compute persistent homology (topological feature vs noise)**.

Persistent homology at glance



Structure Theorem [Zomorodian and Carlsson (2005)]

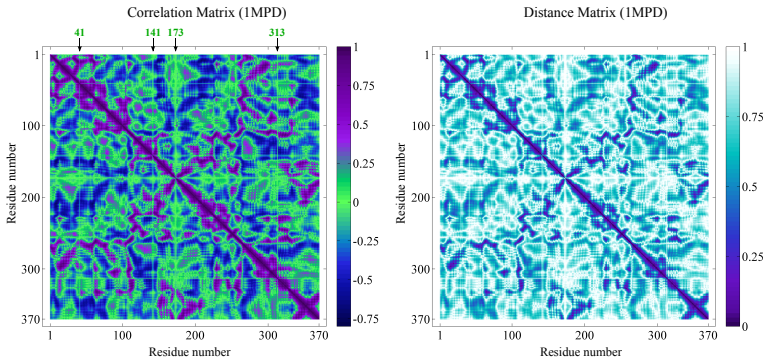
Theorem

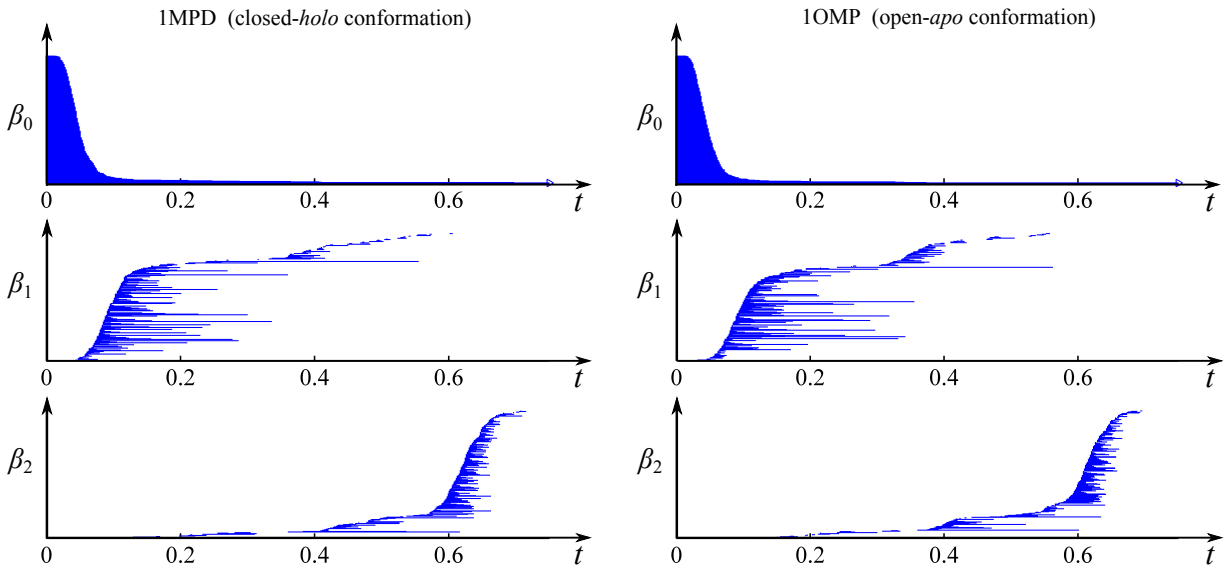
Taking a particular chain maps $t : C_*^i \to C_*^{i+1}$, for a finite persistence module $\mathcal{C} = \{C_*, t\}$ with field F coefficients,

$$H_*(\mathcal{C}, F) \cong \bigoplus_i t^{r_i} F[t] \oplus \left(\bigoplus_j t^{s_j} F[t]/(t^{n_j} F[t]) \right).$$

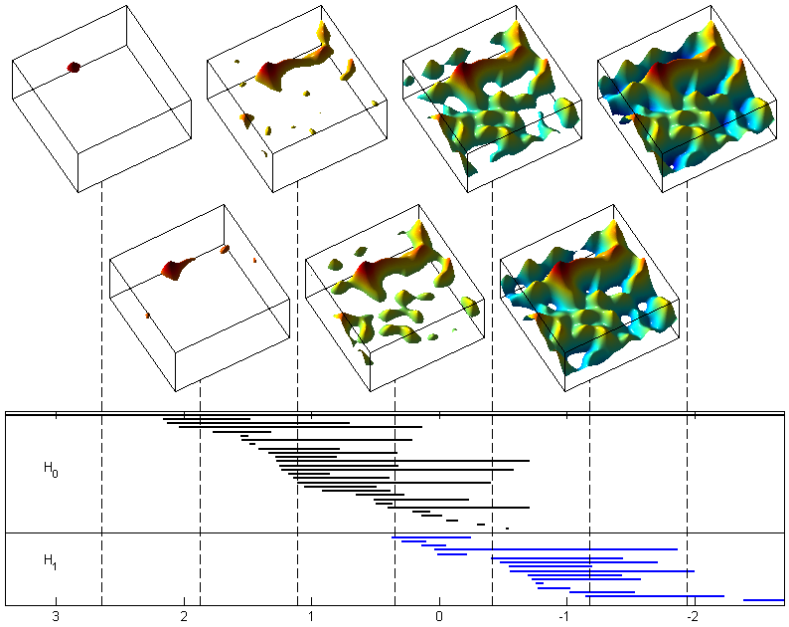
- The left sum provides $[r_i, \infty)$ corresponds to a topological feature that is created at time r_i and remains until the end of filtration.
- The right sum $[s_j, s_j + n_j)$ corresponds to a feature that is created at time s_j and dies after time n_j.
- The multiset of intervals is called a **barcode**.
- The longer the interval, the more significant the feature.

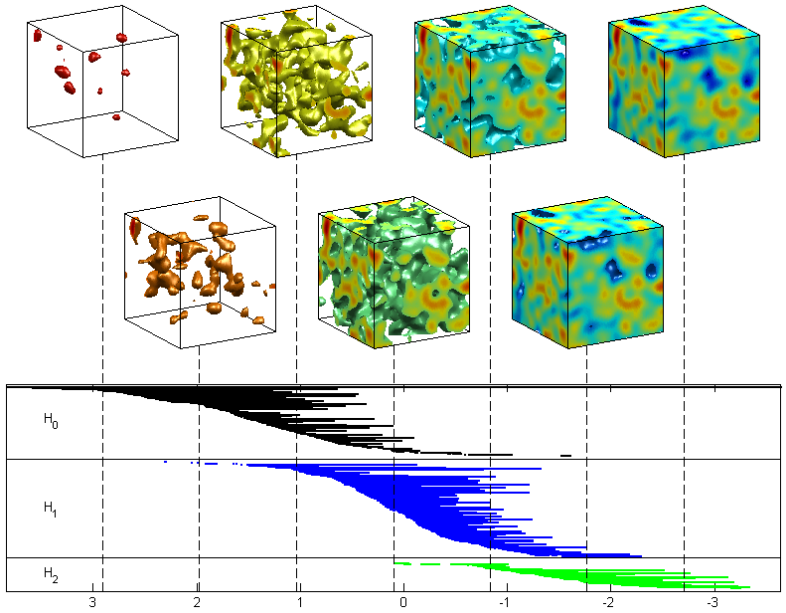
Types of Data: (dis)similarity measure





Random field [Adler et al. (2010)]





Persistence descriptors: Barcode and Persistence diagram

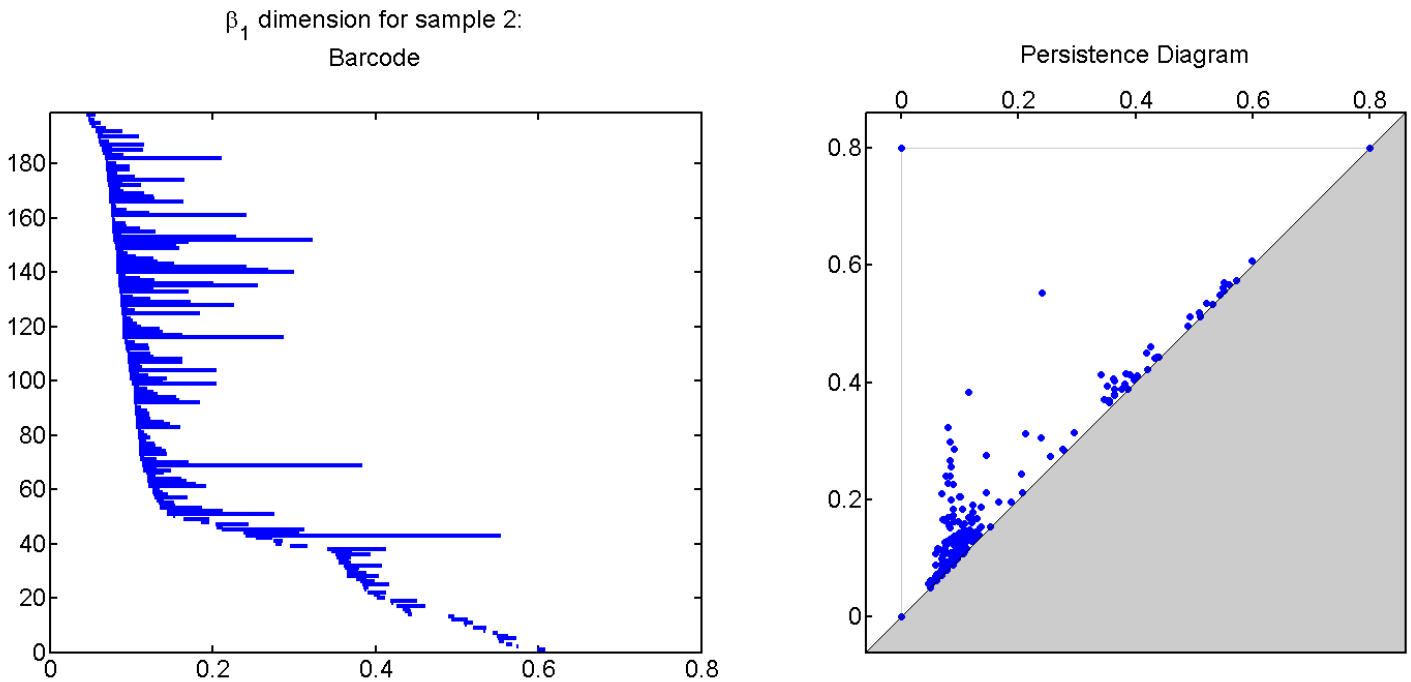
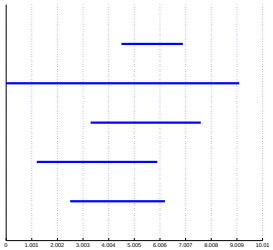
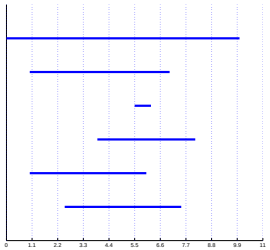
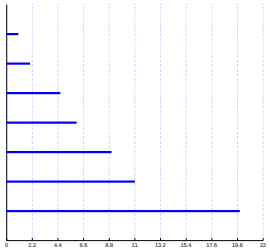
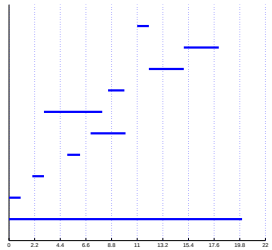
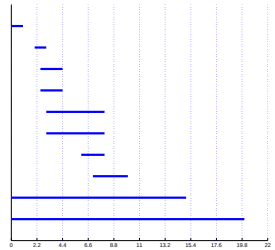


Figure: Space of barcodes (persistence diagrams) with Wasserstein distance is not a manifold, it is difficult to calculate Fréchet mean and variance.

Statistics with descriptors

- How do we calculate the mean and variance?
- Can we apply it to hypothesis testing?

New persistence descriptor Bubenik (2015)

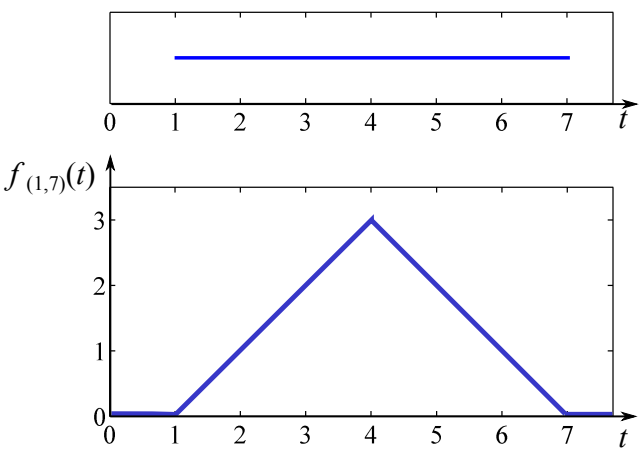
Given (a, b), with $a \leq b$, $f_{(a,b)} : \mathbb{R} \to \mathbb{R}$ by $f_{(a,b)}(t) = \min(t - a, b - t)_+$, where $x_+ = \max(x, 0)$.

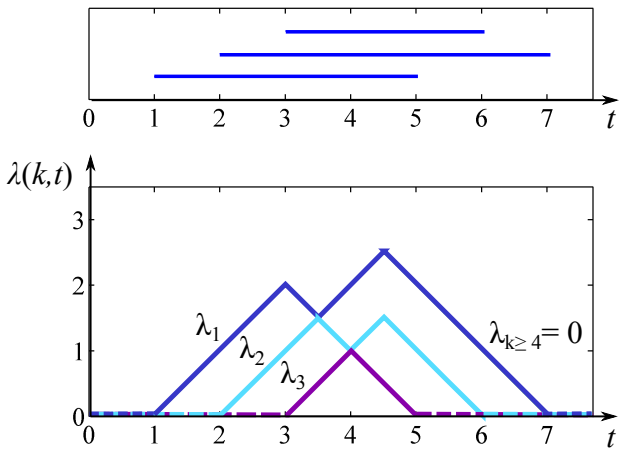
Definition

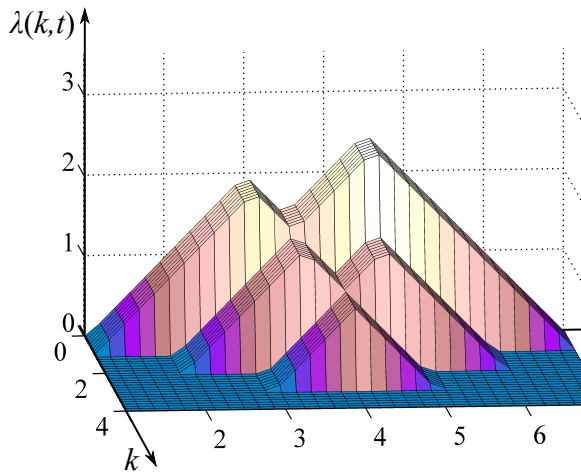
The **persistence landscape** of $\{(a_i, b_i)\}_{i=1}^m$ is defined as a set of functions, $\{\lambda(k, t) : \mathbb{N} \times \mathbb{R} \to \bar{\mathbb{R}}\}_{k \in \mathbb{N}}$, where $\lambda(k, t) = k^{th}$ largest value of $\{f_{(a_i,b_i)}(t)\}_{i=1}^m$, and $\lambda(k, t) = 0$, if $k > m$.

$\mathbb{N} = \{1, 2, 3, \cdots\}$ and $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$.

Persistence Landscape







Results in [Bubenik (2015)]

- Persistence landscapes belong to $L^{\mathrm{p}}(S), S = \mathbb{N} \times \mathbb{R}$, with the metric induced by p-integrable functions, which is a separable Banach space.
- Let $\lambda_1, \ldots, \lambda_n$ be a sample of persistence landscapes drawn from a probability measure with Fréchet mean μ.
- **Strong Law of Large Numbers**

$$\bar{\lambda}_n(k,t) \overset{a.s.}{\to} \mu(k,t) \quad \text{iff} \quad E(\lambda) \quad < \infty$$

- **Central Limit theorem**: Assume $p \geq 2$. If $E(\lambda) < \infty$ and $E(||\lambda||^2) < \infty$,

$$\sqrt{n}(\bar{\lambda}_n(k,t) - \mu(k,t))/\sigma \overset{d}{\to} N(0,1)$$

Functionals and Test

- $p \geq 2, E(\lambda) < \infty$ and $E(||\lambda||^2) < \infty$. For any $f \in L^q(S)$ with $\frac{1}{p} + \frac{1}{q} = 1$, let $X = \int_S f\lambda := ||f\lambda||_1$. Then

$$\sqrt{n}[\overline{X}_n - E(X)] \xrightarrow{d} N(0, \mathrm{var}(X))$$

- t-test: $\sum_{k=1}^{K} \int (\lambda_k^A(t) - \lambda_k^B(t))\, dt$
- Multivariate test (Hotellings T^2 test):

  Consider a vector, $(\int(\lambda_1^A - \lambda_1^B), \int(\lambda_2^A - \lambda_2^B), \ldots, \int(\lambda_k^A - \lambda_k^B))$,

  where k is chosen so that, $k << n_1 + n_2 - 2$.

Open and closed proteins

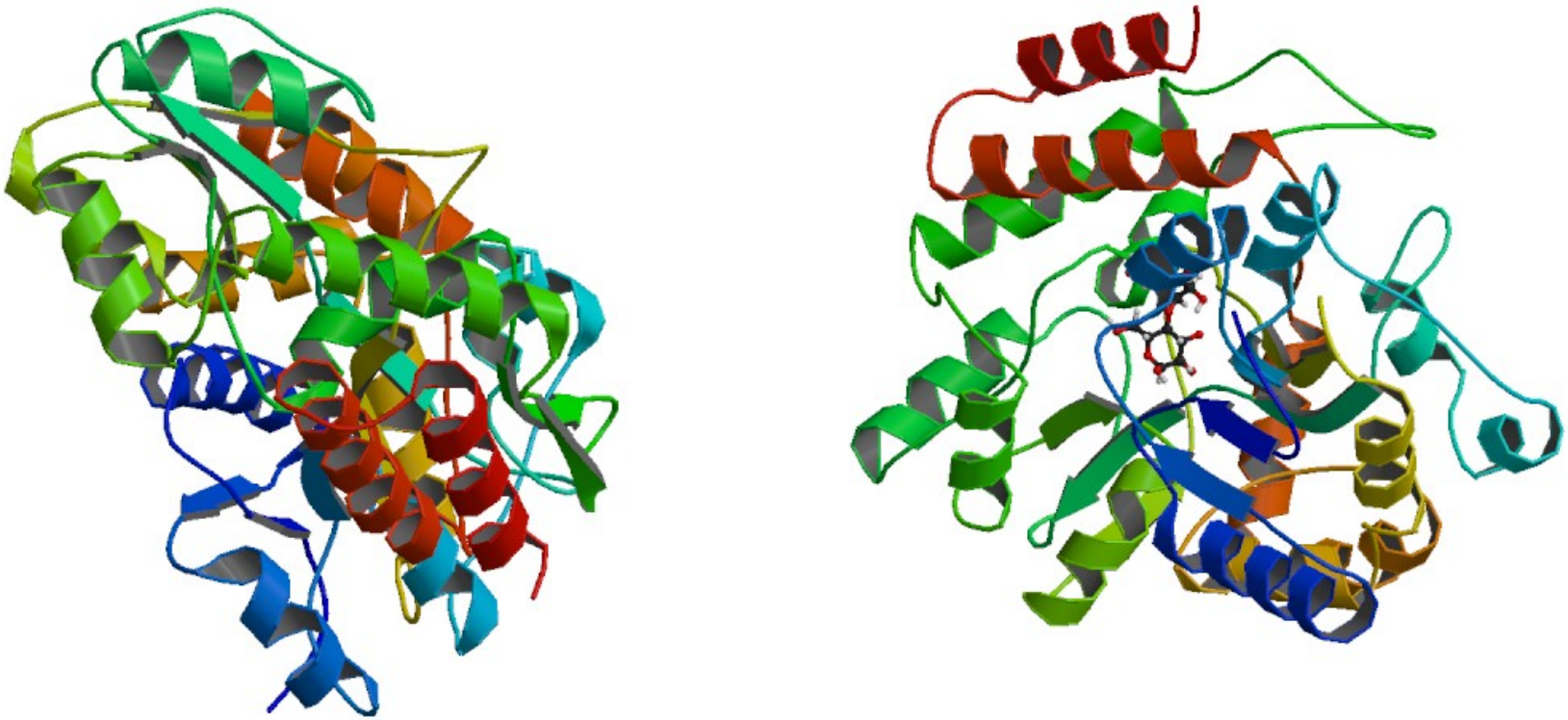


Figure: Maltose binding proteins: (left) 1OMP open and ligand-free; (right) 1MPD closed and ligand-bound. Each protein consists of 370 amino acid residues. DATA:**Correlations between 370 amino acid residues**.

Allosteric path with 1MPD [Kovacev-Nikolic et al. (2015)]

The allosteric path is the set of highly correlated and mutually interacting amino acid residues that do not lie close to each other in the space.

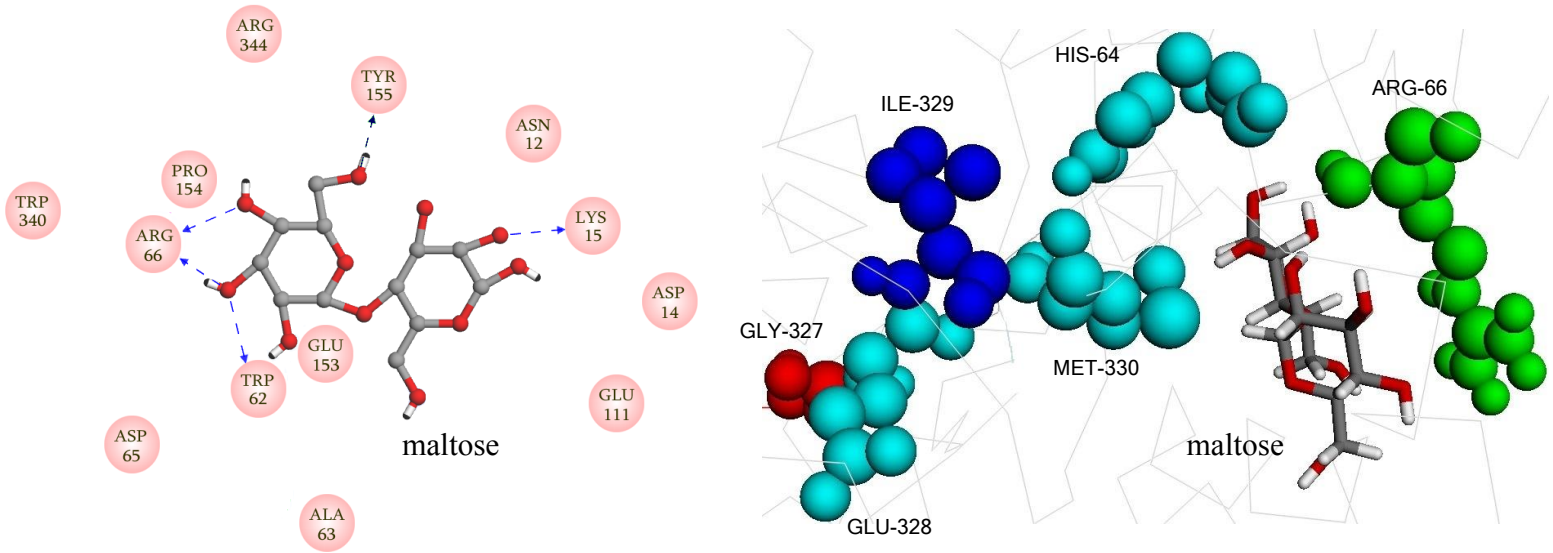
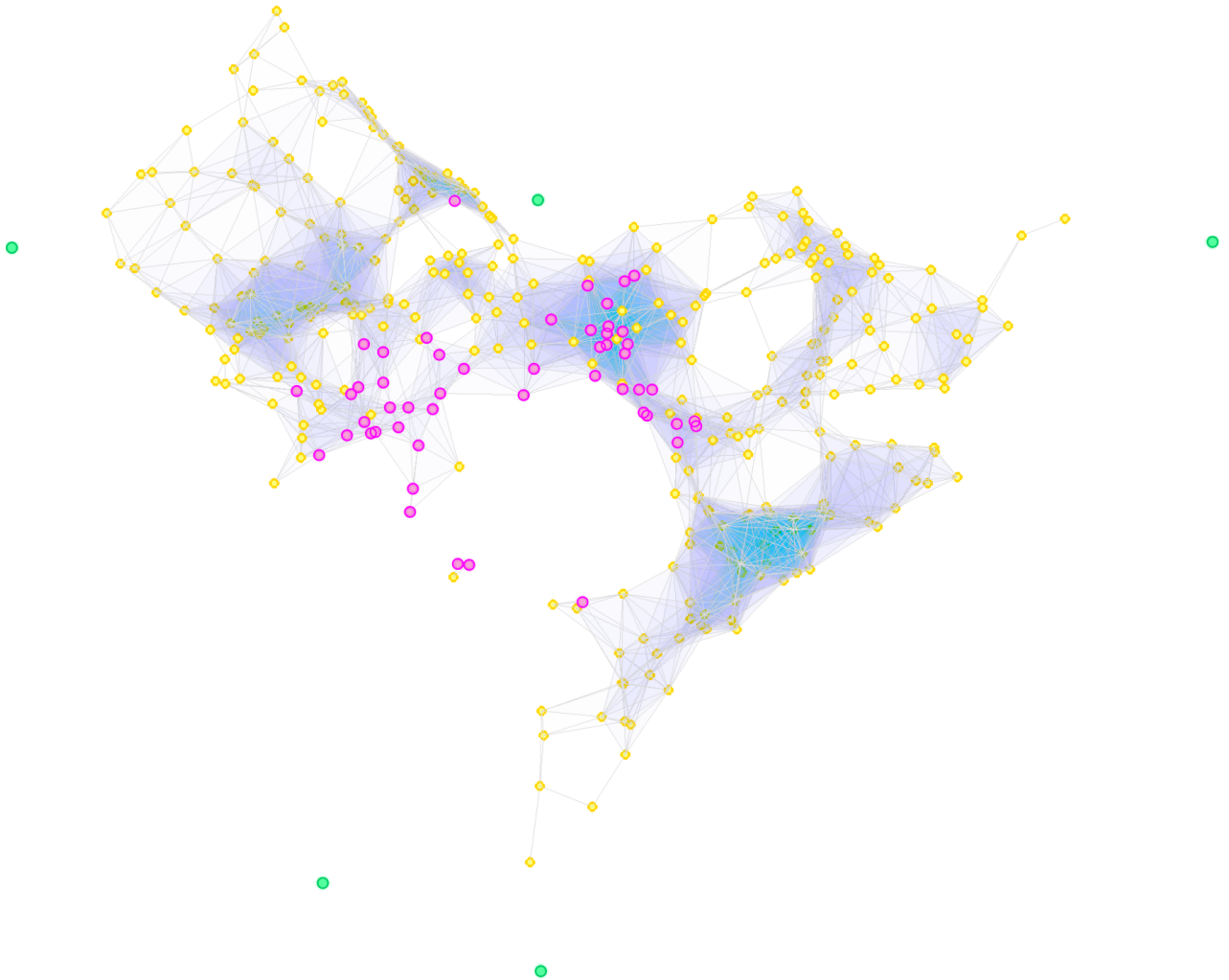
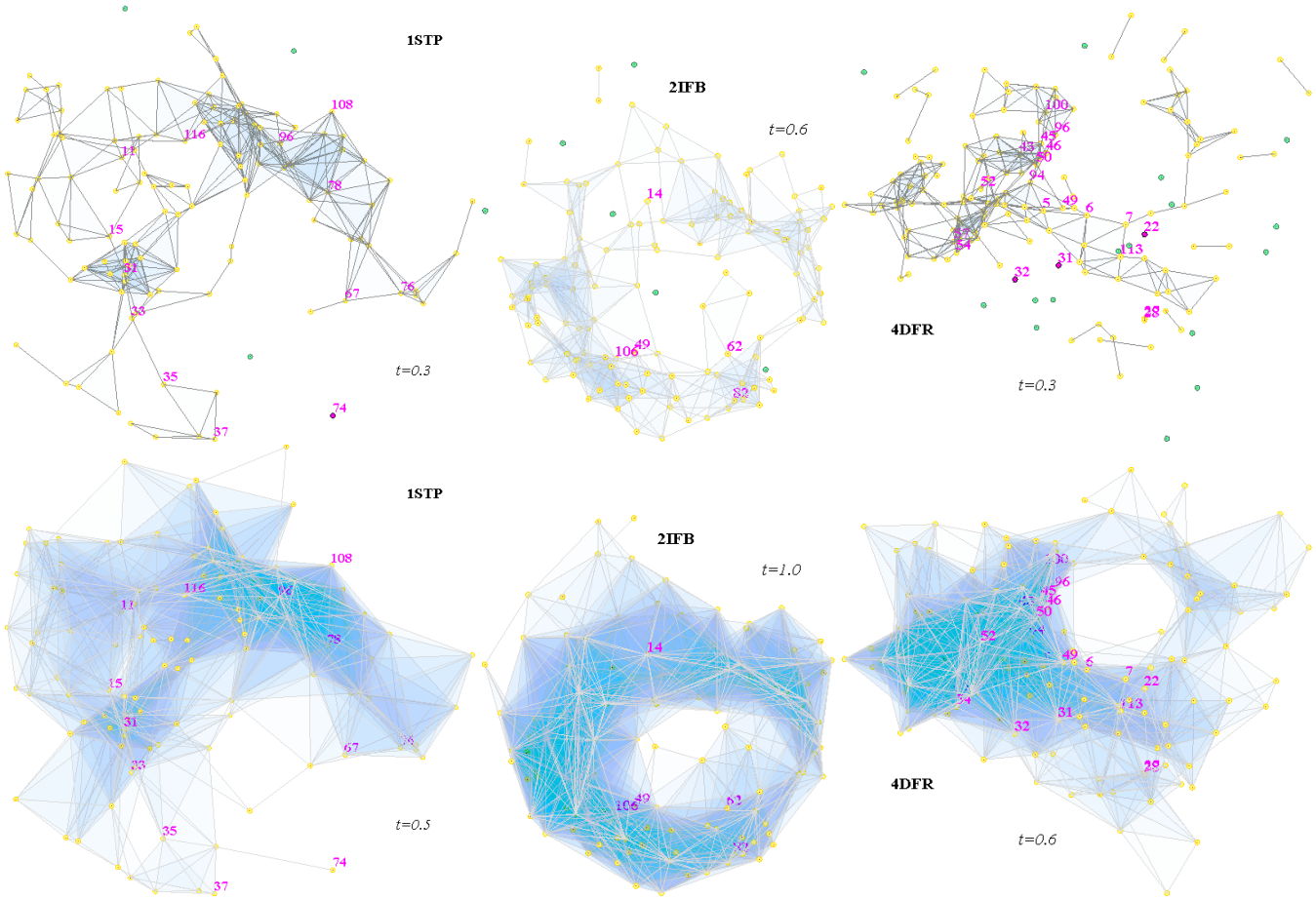


Figure: Two shortest allosteric paths, 66-64-330-328-327/66-64-329-328-327. AlloPathFinder software Tang et al (2007).

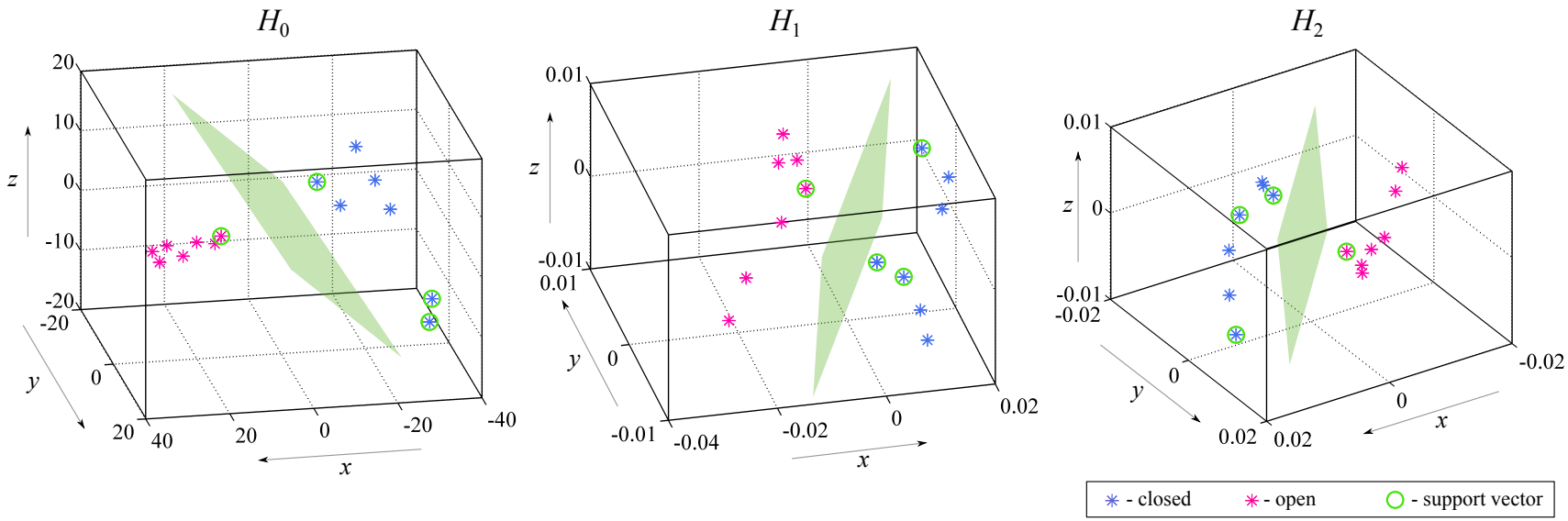
Binding sites and allosteric path (β_1) [Kovacev-Nikolic et al. (2015)]



Kovacev-Nikolic et al. (2015)



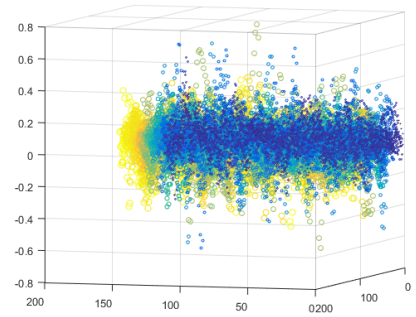
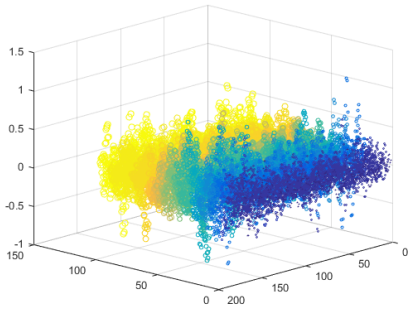
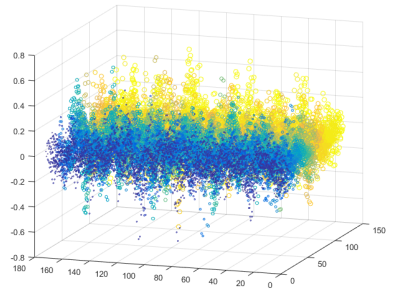
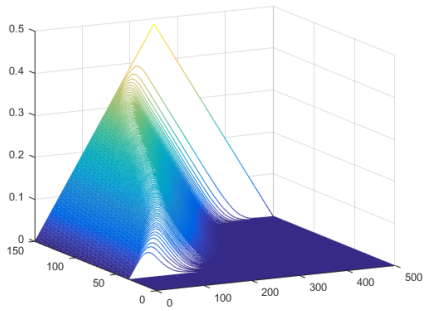
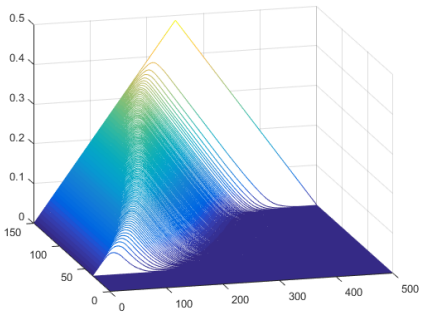
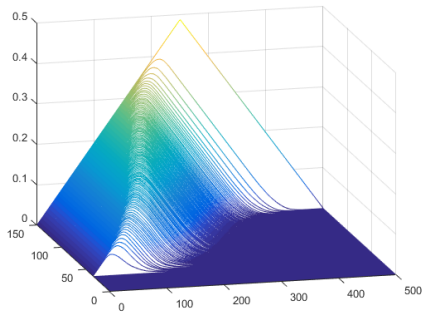
Classification by support vector machine



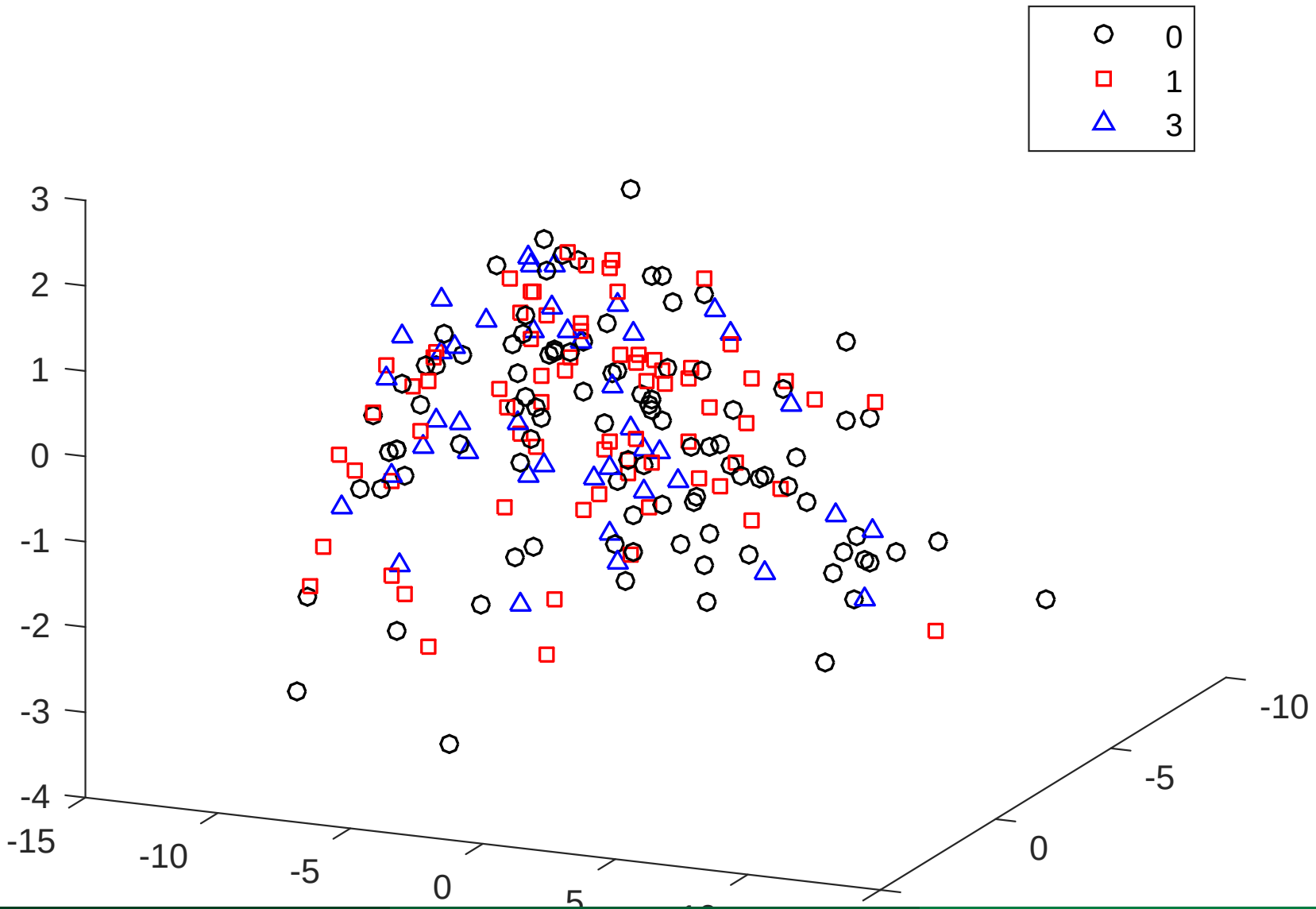
NYU ADHD data - 216 subjects

- BOLD signals at for resting status fMRI– 116 ROIs and 172 equally spaced time courses
- Diagnosis (DX-0, 1, 3)
- ADHD index
- Handedness
- Inattentive and Hyperimpulsive
- Vervall IQ, performance IQ, Full IQ
- MedStatus (medication or no medication) and QC(questionable or pass)

BOLD signals over time×ROI and Persistence Landscape

Clusters of diagnose ADHD



Propose Tensor Regression Model [Lu et al. (2008)]

- For each $\mathbf{X}_k \in \mathbb{R}^{t_1} \otimes \mathbb{R}^{t_2}$, apply the multilinear PCA.
  $\{\mathbf{U}^{(i)}, i = 1, 2\} : \mathbb{R}^{t_1} \otimes \mathbb{R}^{t_2} \longrightarrow \mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2}$ with $d_i < t_i, i = 1, 2$

$$\mathbf{M}_k = \mathbf{X}_k \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)},$$

  such that $\{\mathbf{M}_k \in \mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2}\}_{k=1}^n$ captures most of the variation observed in the original tensor objects.
- The variations are measured by the total tensor scatter.

$$\mathbf{U}^{(i)} = \text{argmax}_{\mathbf{U}^{(1)}, \mathbf{U}^{(2)}} \Phi_{\mathbf{M}}$$

  where $\Phi_{\mathbf{M}} = \sum_{k=1}^n \|\mathbf{M}_k - \bar{\mathbf{M}}\|_F^2$.

Rank-R generalized linear tensor regression model [Zhou et al. (2013)]

Let $\mathbf{v} = (\mathbf{u}, \mathbf{x})$, where $\mathbf{u}$ denotes covariates and $\mathbf{x}$ functional of PL in dim 0, 1, and 2. Response variable is ADHD index.

$$\mu_y = \alpha + \boldsymbol{\gamma}^T \mathbf{v} + \langle \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} (\boldsymbol{\beta}_2^{(r)})^T, \mathbf{M} \rangle, \quad \boldsymbol{\beta}_i^{(r)} \in \mathbb{R}^{d_i} \tag{1}$$

$$\mu_y = \alpha + \boldsymbol{\gamma}^T \mathbf{u} + \langle \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} (\boldsymbol{\beta}_2^{(r)})^T, \mathbf{M} \rangle, \quad \boldsymbol{\beta}_i^{(r)} \in \mathbb{R}^{d_i} \tag{2}$$

$$\mu_y = \alpha + \boldsymbol{\gamma}^T \mathbf{v} \tag{3}$$

$$\mu_y = \alpha + \boldsymbol{\gamma}^T \mathbf{u} \tag{4}$$

RMSE - Matlab toolbox [Acar et al. (2011), Bader et al. (2015), Zhou (2013)]

- Randomly choose 210 observation as training set and 6 as test set.
- Apply the regression model on training set to estimate the parameters and then make prediction on the test set.
- Repeat the process 100 times to find the mean of square root MSE (sum of training and predictive MSE) and its standard deviation.

P-value (sd) for coefficient in full model

	99 %	90 %	80 %
Gender	0.0762(0.2025)	0.2284(0.2725)	0.1755(0.1405)
Age	0.0569(0.1896)	0.2399(0.3338)	0.5631(0.2580)
handedness	0.0238(0.1038)	0.0675(0.1464)	0.0259(0.0528)
DX	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)
Inattentive	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)
HyperImpulsive	0.0524(0.1743)	0.1050(0.1861)	0.2134(0.2284)
Verbal IQ	0.0599(0.1749)	0.1870(0.2566)	0.5691(0.2548)
PerformanceIQ	0.0669(0.1877)	0.1393(0.2502)	0.3494(0.2256)
Full4IQ	0.0320(0.1335)	0.0308(0.1358)	0.0027(0.0143)
MedStatis	0.0864(0.2220)	0.2517(0.3213)	0.4793(0.2371)
QC	0.0803(0.1920)	0.1776(0.2587)	0.3242(0.2409)
0-D PL	0.0875(0.2119)	0.1882(0.2814)	0.4912(0.2355)
1-D PL	0.0591(0.1636)	0.2078(0.2578)	0.5786(0.2786)
2-D PL	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)
RMSE	1.4298(0.3212)	1.9695(0.2009)	2.3231(0.0835)

P-value (sd) for coefficient in model 2 (covr+tensor)

	99 %	90 %	80 %
Gender	0.1187(0.2588)	0.2007(0.2811)	0.4026(0.3340)
Age	0.0570(0.1741)	0.3035(0.3144)	0.6137(0.2554)
handedness	0.0494(0.1878)	0.0778(0.1818)	0.0332(0.0573)
DX	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)
Inattentive	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)
HyperImpulsive	0.0637(0.1829)	0.0760(0.1638)	0.2199(0.2128)
VerbalIQ	0.0744(0.1948)	0.1581(0.2512)	0.6469(0.2513)
PerformanceIQ	0.0707(0.1862)	0.1075(0.2082)	0.4219(0.2461)
Full4IQ	0.0242(0.0970)	0.0028(0.0154)	0.0014(0.0055)
MedStatis	0.0726(0.1771)	0.1806(0.2513)	0.3692(0.2547)
QC	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)
RMSE	1.4715(0.6302)	1.9682(0.1548)	2.3486(0.0849)

P-value (sd) for coefficient in all models

	full 90 %	covr+tensor 90%	covr+ PL	covr
Sex	0.2284(0.2725)	0.2007(0.2811)	0.3733(0.1127)	0.3141(0.1064)
Age	0.2399(0.3338)	0.3035(0.3144)	0.4868(0.0946)	0.5461(0.0896)
hand	0.0675(0.1464)	0.0778(0.1818)	0.4105(0.0834)	0.3211(0.0732)
DX	0.0000(0.0000)	0.0000(0.0000)	0.0034(0.0022)	0.0031(0.0018)
Inattn	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)	0.0000(0.0000)
Hyper	0.1050(0.1861)	0.0760(0.1638)	0.0000(0.0000)	0.0000(0.0000)
VIQ	0.1870(0.2566)	0.1581(0.2512)	0.5776(0.1169)	0.7021(0.1152)
PerfIQ	0.1393(0.2502)	0.1075(0.2082)	0.8335(0.0962)	0.7004(0.0949)
Full4IQ	0.0308(0.1358)	0.0028(0.0154)	0.8279(0.0990)	0.9100(0.0917)
Med	0.2517(0.3213)	0.1806(0.2513)	0.0432(0.0323)	0.0273(0.0234)
QC	0.1776(0.2587)	0.0000(0.0000)	0.3637(0.0715)	0.4786(0.0831)
0D PL	0.1882(0.2814)	na	0.6762(0.1219)	na
1D PL	0.2078(0.2578)	na	0.2133(0.0724)	na
2D PL	0.0000(0.0000)	na	0.4236(0.1010)	na
RMSE	1.9695(0.2009)	1.9682(0.1548)	3.1642(0.0265)	3.1670(0.0126)

Acknowledgment

- Banff NDA orginizers: **Hongtu Zhu** and **Linglong Kong**
- Violeta Kovacev-Nikolic, Dragan Nikolić, Peter Bubenik, and Jisu Kim
- NSERC
- McIntyre Memorial Funding in the School of Dentistry

References

Acar, E., Dunlavy, D. M., and Kolda, T. G. (2011), "A Scalable Optimization Approach for Fitting Canonical Tensor Decompositions," *Journal of Chemometrics*.

Adler, R. J., Bobrowksi, O., Borman, M. S., Subag, E., and Weinberger, S. (2010), "Persistent Homology for Random Fields and Complexes," *IMS collections*, 6, 124–143.

Bader, B. W., Kolda, T. G., et al. (2015), "MATLAB Tensor Toolbox Version 2.6," Available online.

Bubenik, P. (2015), "Statistical topological data analysis using persistence landscapes," *Journal of Machine Learning Research*, 16, 77–102.

Kovacev-Nikolic, V., Bubenik, B., Nikolić, D., and Heo, G. (2015), "Using persistent homology and dynamical distances to analyze protein binding," .

Lu, H., Plataniotis, K., and Venetsanopoulos, A. (2008), "MPCA: Multilinear Principal Component Analysis of Tensor Objects," *IEEE Transactions on Neural Networks*.

Rotman, J. (1988), *An Introduction to Algebraic Topology*, New York: Springer.

Zhou, H. (2013), "Matlab TensorReg Toolbox Version 0.0.2," .

Zhou, H., Li, L., and Zhu, H. (2013), "Tensor regression with applications in neuroimaging data analysis," *Journal of the American Statistical Association*, 108, 540–552.

Zomorodian, A. and Carlsson, G. (2005), "Computing persistent homology," *Discrete and Computational Geometry*, 33, 249–274.