Statistical Topological Data Analysis

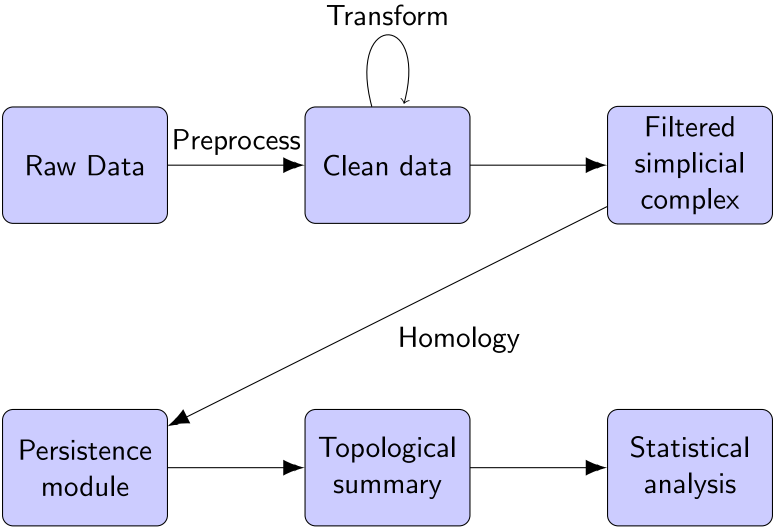
Peter Bubenik

Cleveland State University / University of Florida

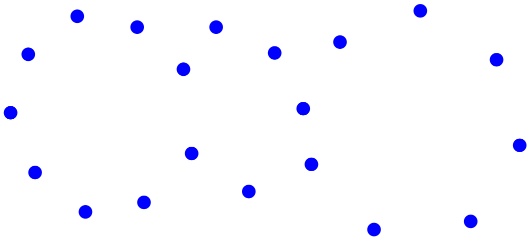
November 10, 2014

Discrete, Computational and Algebraic Topology
University of Copenhagen

Topological Data Analysis

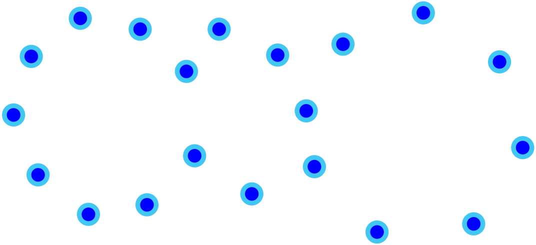


Filtered simplicial complex from points in $\mathbb{R}^2$



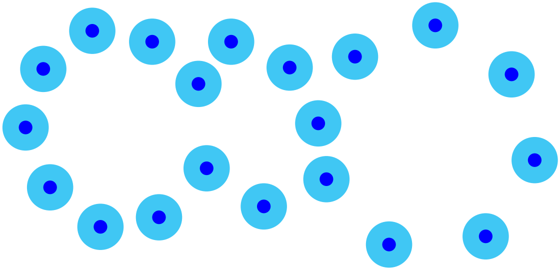
radius $= 0$

Filtered simplicial complex from points in $\mathbb{R}^2$



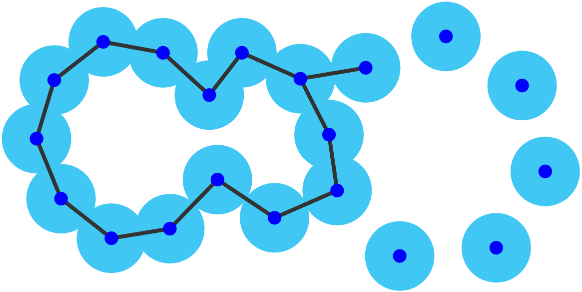
radius $= 1$

Filtered simplicial complex from points in $\mathbb{R}^2$



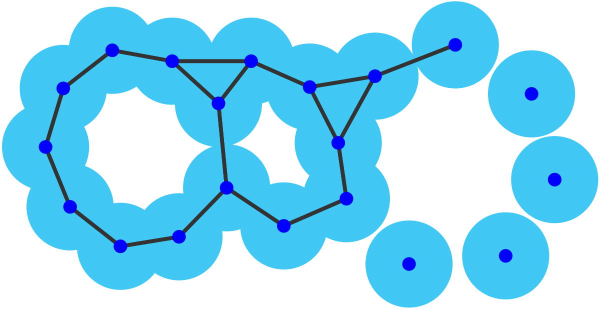
radius $= 2$

Filtered simplicial complex from points in $\mathbb{R}^2$



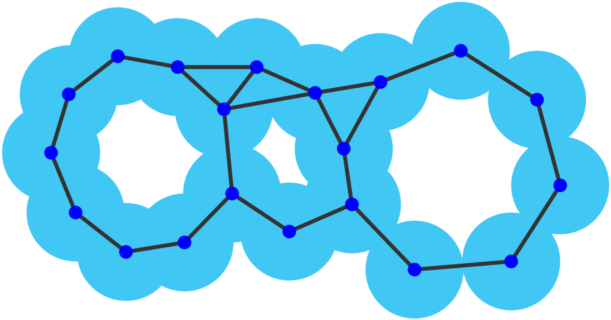
radius = 3

Filtered simplicial complex from points in $\mathbb{R}^2$



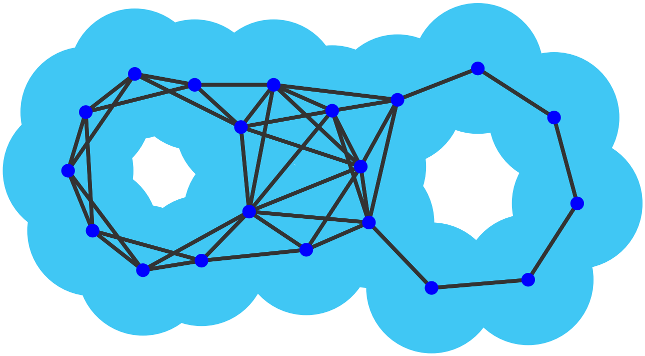
radius $= 4$

Filtered simplicial complex from points in $\mathbb{R}^2$



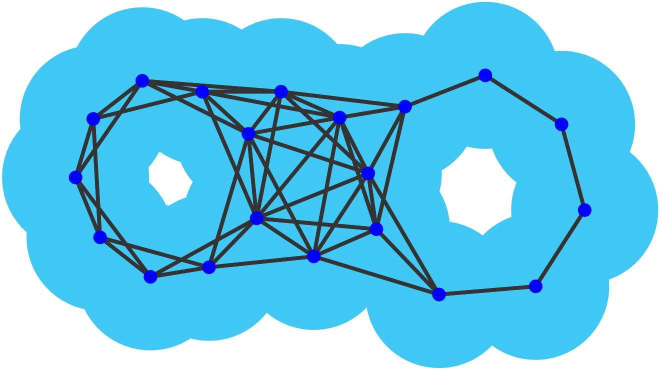
radius $= 5$

Filtered simplicial complex from points in $\mathbb{R}^2$



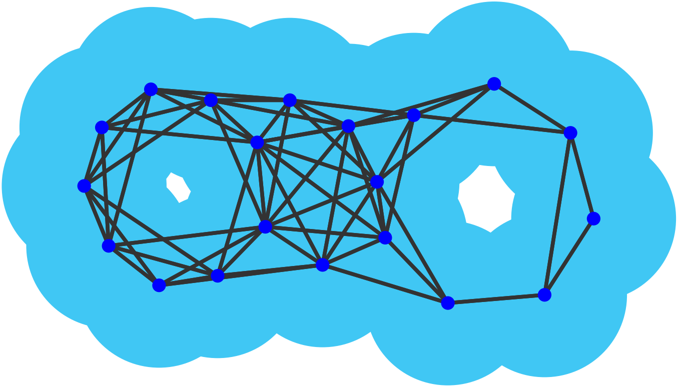
radius = 6

Filtered simplicial complex from points in $\mathbb{R}^2$



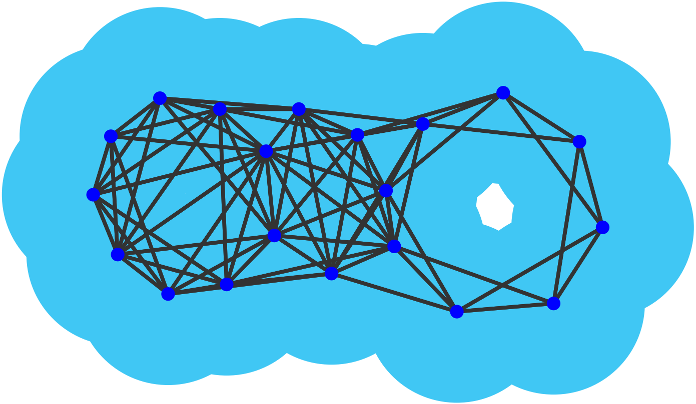
radius $= 7$

Filtered simplicial complex from points in $\mathbb{R}^2$



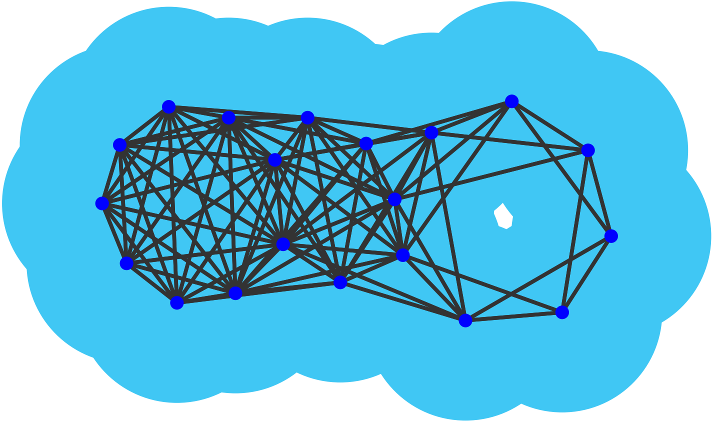
radius $= 8$

Filtered simplicial complex from points in $\mathbb{R}^2$



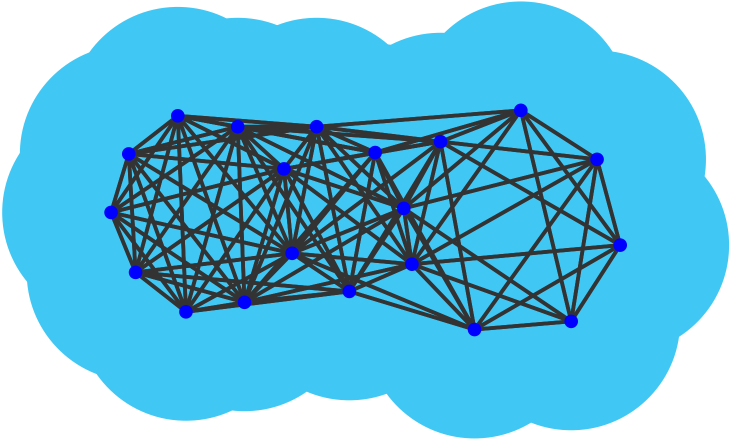
radius $= 9$

Filtered simplicial complex from points in $\mathbb{R}^2$



radius $= 10$

Filtered simplicial complex from points in $\mathbb{R}^2$



radius = 11

Filtered simplicial complexes

Discrete: simplicial complexes, $X_0 \subseteq X_1 \subseteq X_2 \subseteq \cdots \subseteq X_n$

Continuous:
- simplicial complexes $\{X_t\}_{t \in \mathbb{R}}$
- inclusions $X_t \subseteq X_{t'}$, for $t \leq t'$

Discrete to continuous: $X_t := X_{\lfloor t \rfloor}$

Abstract: Let $[\mathbf{n}]$ denote the category $0 \to 1 \to 2 \to \cdots \to n$.

$$\begin{tikzcd}
[\mathbf{n}] \arrow[r] \arrow[d] & \mathbf{Simp} \text{ or } \mathbf{Top} \\
(\mathbb{R}, \leq) \arrow[ur] &
\end{tikzcd}$$

Example: Given $f : X \to \mathbb{R}$ define $X_t = f^{-1}(-\infty, t]$.

Persistence modules

Apply $H_k(-;\mathbb{F})$.

Discrete: vector sp and linear maps $V_0 \to V_1 \to V_2 \to \cdots \to V_n$

Continuous:
- vector spaces V_t
- linear maps $V_t \to V_{t'}$

Discrete to continuous: $V_t := V_{\lfloor t \rfloor}$

Abstract:

$$\begin{tikzcd} \mathbf{[n]} \arrow[r] \arrow[d] & \mathbf{vect} \\ (\mathbb{R}, \leq) \arrow[ur] & \end{tikzcd}$$

Example: Given $f : X \to \mathbb{R}$ define $X_t = H(f^{-1}(-\infty, t])$.

Summaries of persistence modules

Let $M = (M_t)_{t \in \mathbb{R}}$ be a persistence module.

Let $I \subseteq \mathbb{R}$ be an interval.
The indecomposable persistence modules are interval modules:

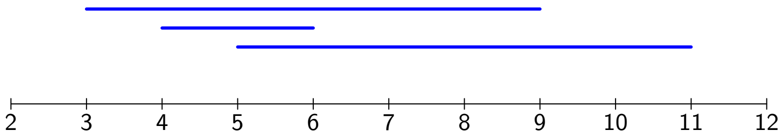
$$(A_I)_t = \begin{cases} \mathbb{F} & \text{if } t \in I \\ 0 & \text{if } t \notin I. \end{cases}$$

Theorem (Gabriel, Zomorodian-Carlsson, Crawley-Boevey)

M is a direct sum of interval modules, $M \cong \oplus_j A_{I_j}$.

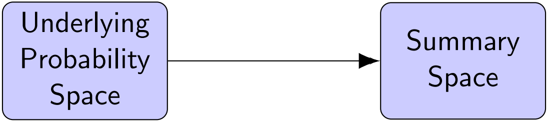
The set of intervals $\{I_j\}$ is called a barcode.

Barcode from our points



Statistical viewpoint

The barcode is a random variable; it is a summary statistic.



Challenges

For example:

- calculate averages
- understand variances
- test hypotheses
- cluster and classify

Statistics with barcodes/persistence diagrams

$$\text{Set of barcodes} \xrightarrow{\text{Metric}} \text{Statistics}$$

Easy:

- clustering
- doing permutation tests

Hard:

- calculating averages
- understanding variances

See work by Bendich, Harer, Mattingly, Mileyko, Mukherjee, Munch, Turner.

Making life easier

$$\text{Barcode space} \longrightarrow \mathbb{R}^n$$

Some constructions:

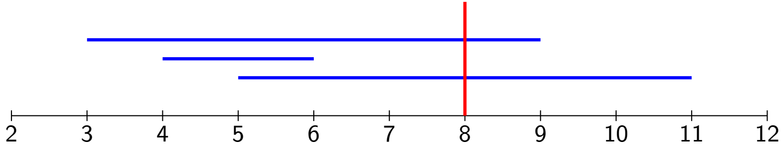
- lengths of the longest N bars
- coordinates of the longest N bars
- values of the functionals τ_{ij}
- the persistence landscape

Advantages of the persistence landscape:

- doesn't lose information
- is stable
- has a continuous version

Persistence landscape

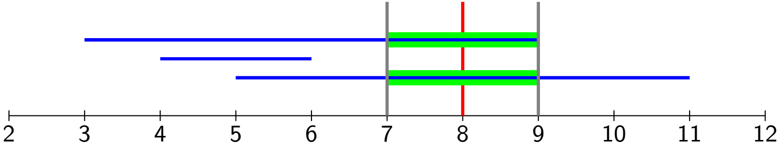
Continuous: $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ or $\lambda_k : \mathbb{R} \to \mathbb{R}$, $k = 1, 2, 3, \ldots$



$$\lambda_2(8) = ?$$

Persistence landscape

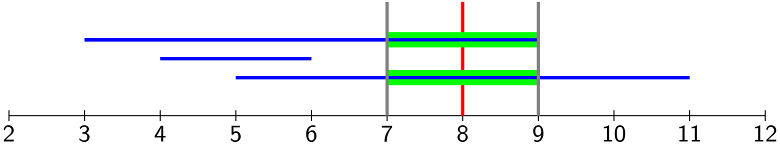
Continuous: $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ or $\lambda_k : \mathbb{R} \to \mathbb{R}$, $k = 1, 2, 3, \ldots$



$$\lambda_2(8) = 1$$

Persistence landscape

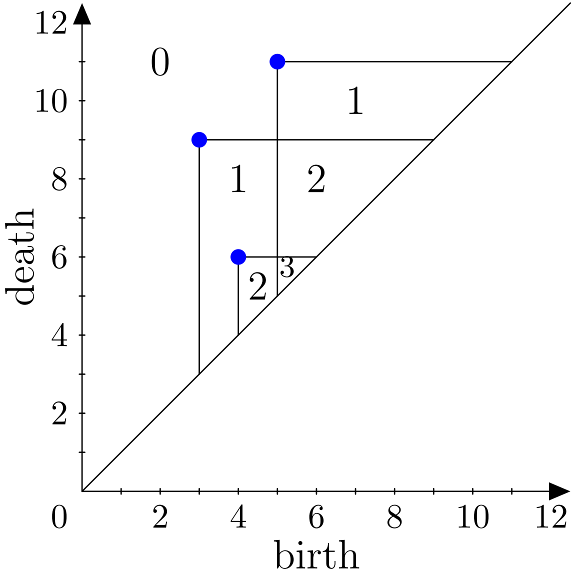
Continuous: $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ or $\lambda_k : \mathbb{R} \to \mathbb{R}$, $k = 1, 2, 3, \ldots$



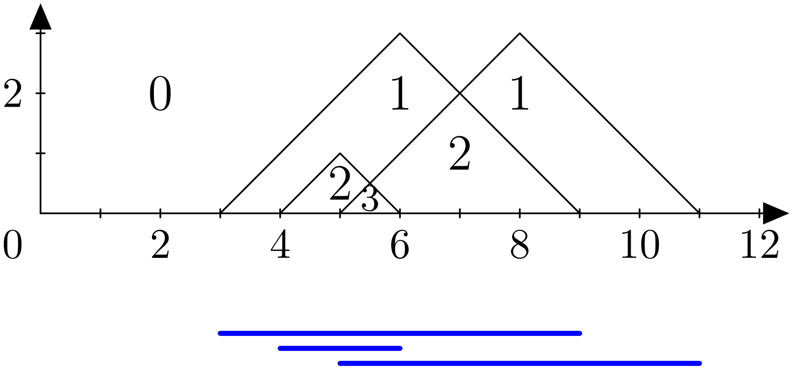
$$\lambda_2(8) = 1$$

Discrete: evaluate λ on a grid

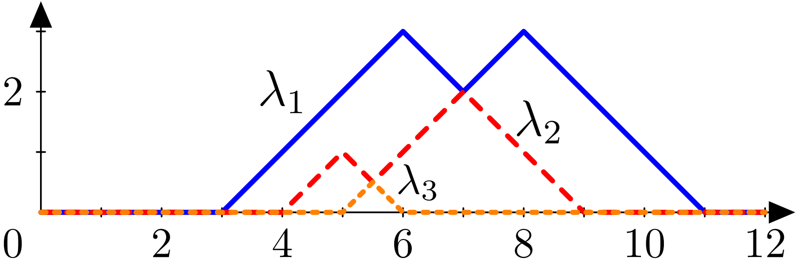
Persistence landscape from our points



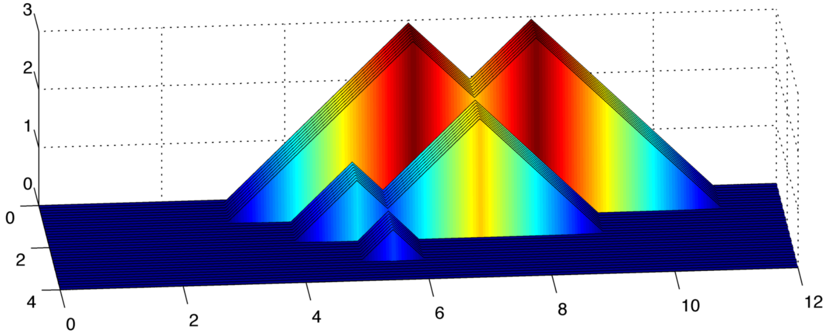
Persistence landscape from our points



Persistence landscape from our points



Persistence landscape from our points

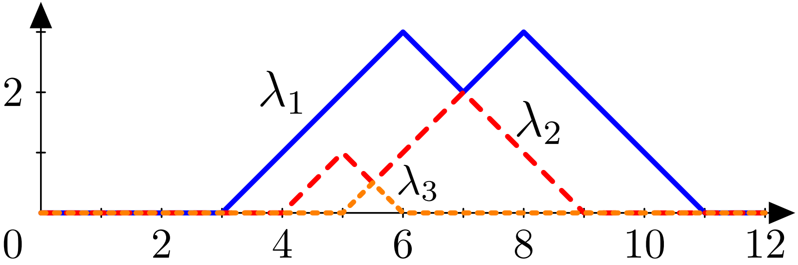


Norms

For a persistence landscape λ,
let (b_j, d_j) be the corresponding birth-death pairs.

Lemma

1. $\|\lambda\|_\infty = \frac{1}{2}\max_j(d_j - b_j)$, *and*
2. $\|\lambda\|_1 = \frac{1}{4}\sum_j(d_j - b_j)^2$.



Stability

Given $f : X \to \mathbb{R}$,
let $\lambda(f)$ the persistence landscape of sublevel sets of f.

Landscape Stability Theorem (B)

Let $f, g : X \to \mathbb{R}$.

$$\|\lambda(f) - \lambda(g)\|_\infty \leq \|f - g\|_\infty.$$

If X is nice and f and g are tame and Lipschitz then

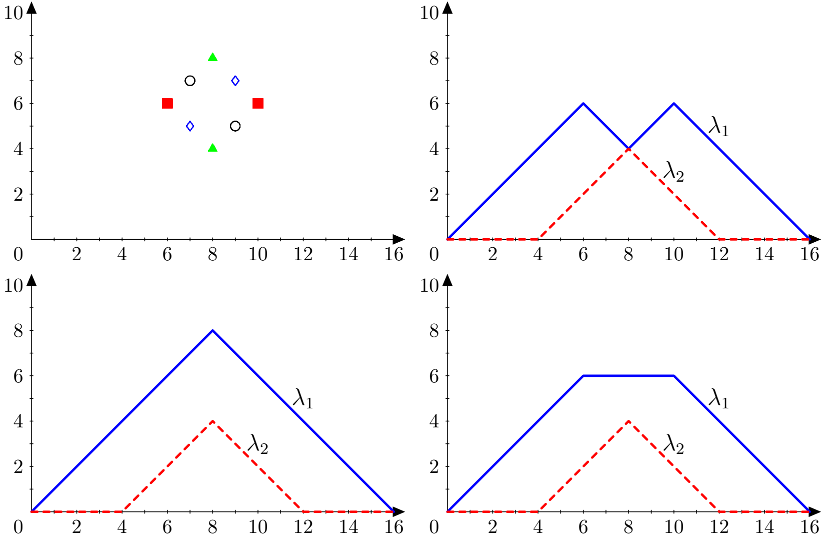
$$\|\lambda(f) - \lambda(g)\|_p^p \leq C\|f - g\|_\infty^{p-k}.$$

Average landscapes

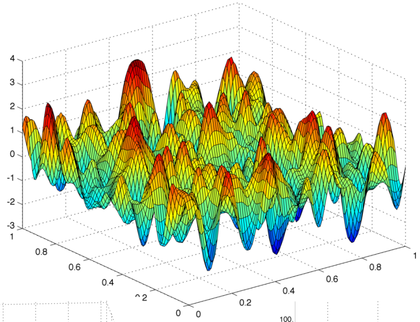
Persistence landscapes, $\lambda^{(1)}, \ldots, \lambda^{(n)}$, have pointwise average,

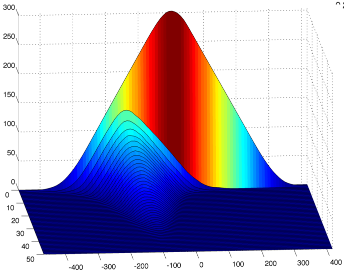
$$\overline{\lambda}(k,t) = \frac{1}{n}\sum_{i=1}^{n} \lambda^{(i)}(k,t)$$

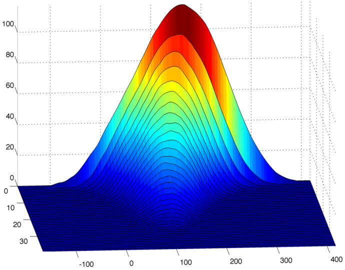
Average diagram vs average landscape



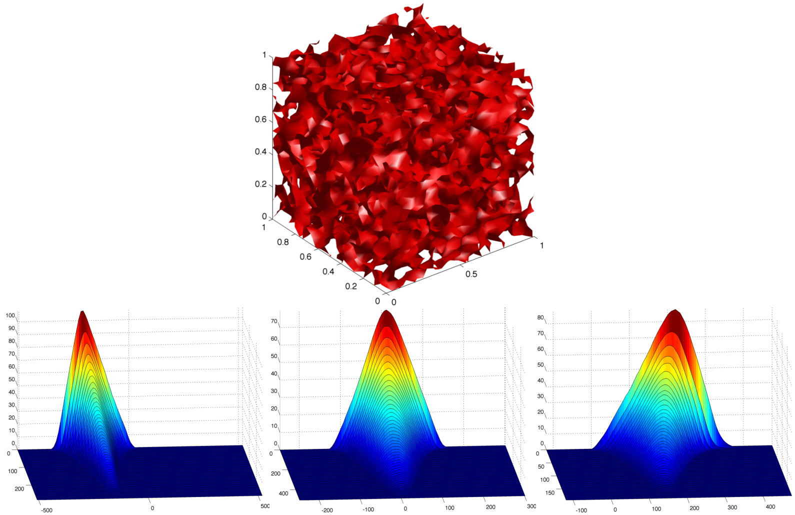
Average landscapes for Gaussian random fields







Average landscapes for Gaussian random fields



Summary space

Let $1 \leq p < \infty$. Then $\|\lambda\|_p = \left(\sum_k \int {\lambda_k}^p \right)^{\frac{1}{p}}$.

We assume $\|\lambda\| := \|\lambda\|_p < \infty$. That is, $\lambda \in L^p(\mathbb{N} \times \mathbb{R})$.

So λ is a random variable with values in a Banach space.

Asymptotics for persistence landscapes

$\lambda \in L^p(\mathbb{N} \times \mathbb{R})$, $\|\lambda\|$ is a real random variable.

If $E\|\lambda\| < \infty$ then there exists $E(\lambda) \in L^p(\mathbb{N} \times \mathbb{R})$ such that $E(f(\lambda)) = f(E(\lambda))$ for all continuous linear functionals f.

Strong Law of Large Numbers (B)

$\overline{\lambda}^{(n)} \to E(\lambda)$ almost surely

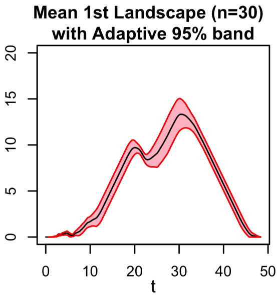
Central Limit Theorem (B)

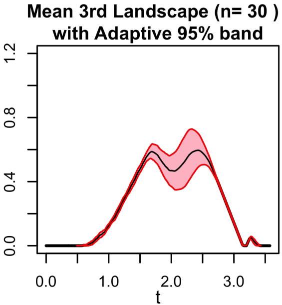
$\sqrt{n}[\overline{\lambda}^{(n)} - E(\lambda)]$ converges weakly to a Gaussian random variable

Understanding variance

Two approaches:

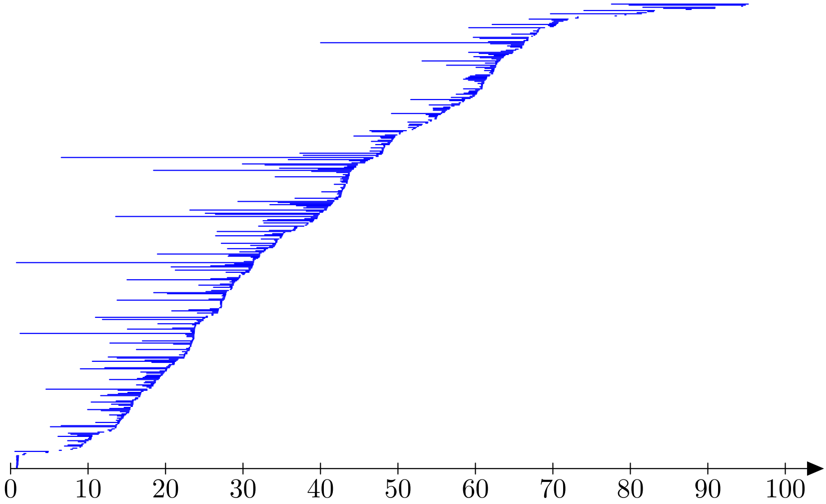
- Bootstrap and confidence intervals for persistence landscapes [Chazal, Fasy, Lecci, Rinaldo, Singh, Wasserman]



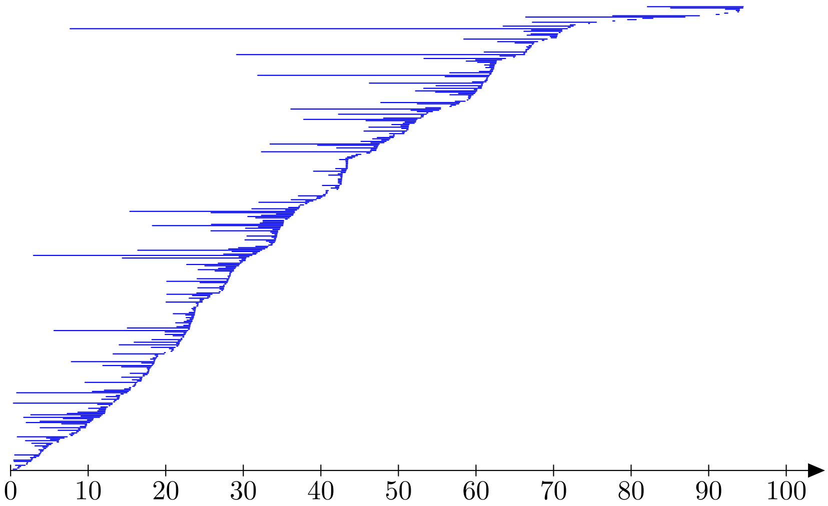


- Principal component analysis

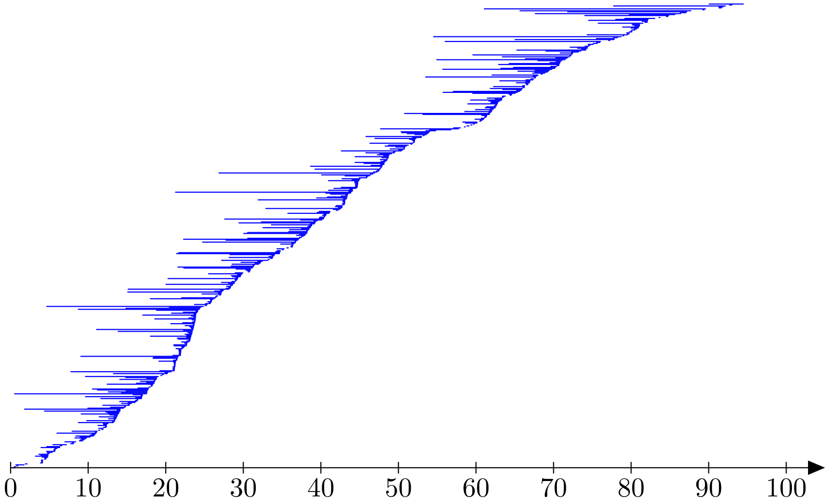
Brain artery barcodes



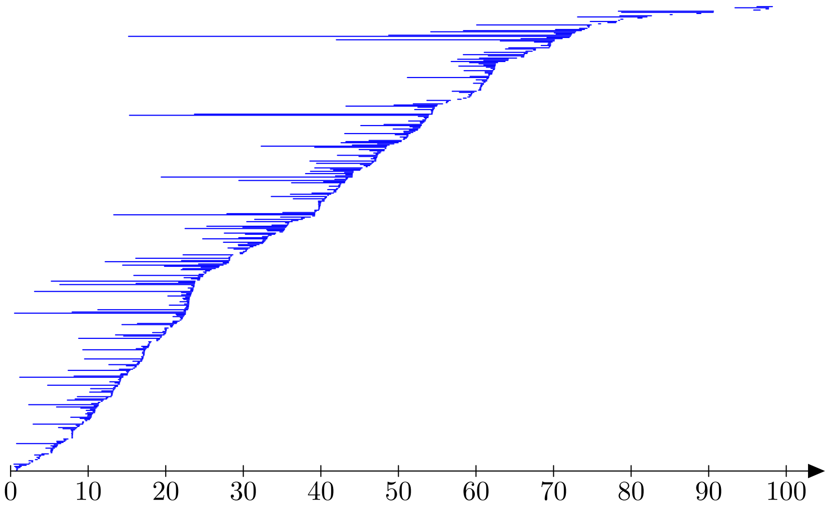
Brain artery barcodes



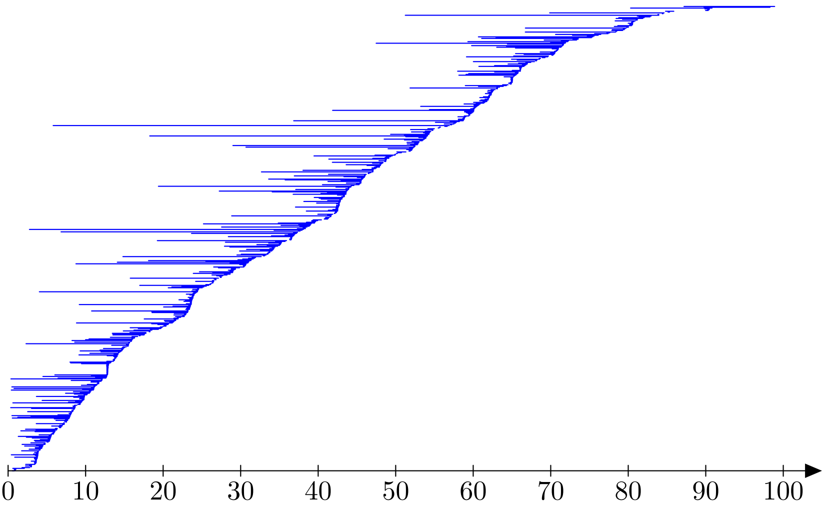
Brain artery barcodes



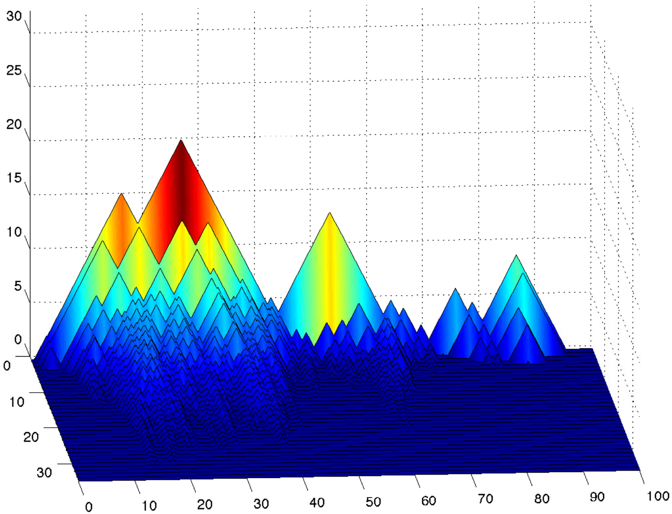
Brain artery barcodes



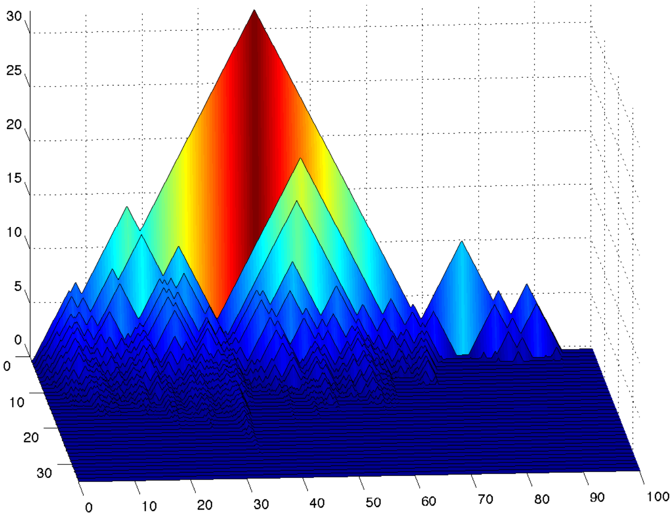
Brain artery barcodes



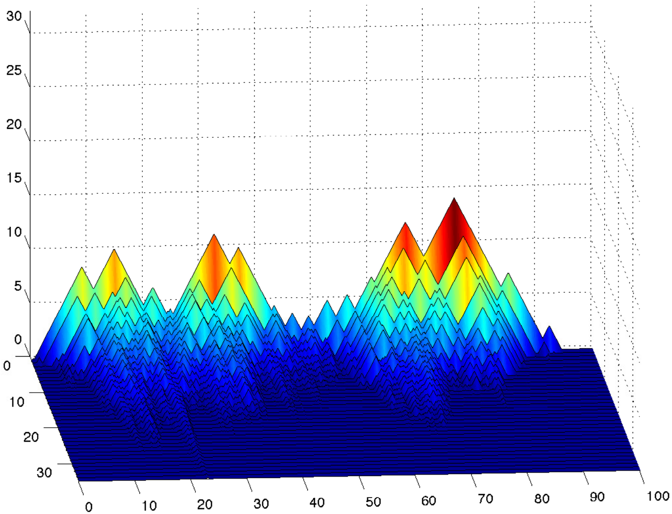
Brain artery persistence landscapes



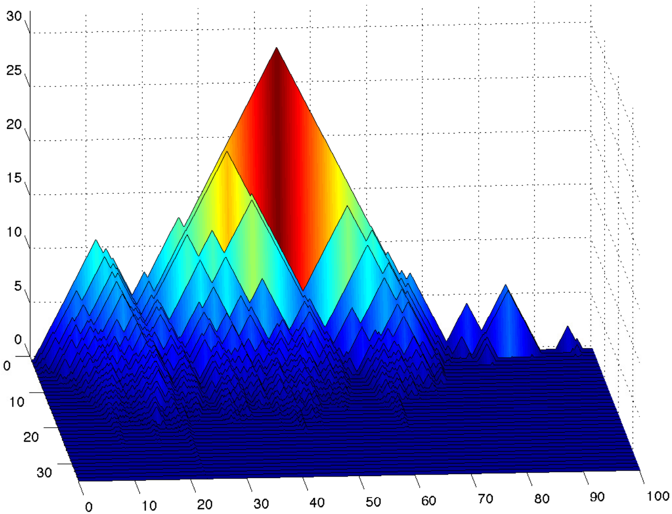
Brain artery persistence landscapes



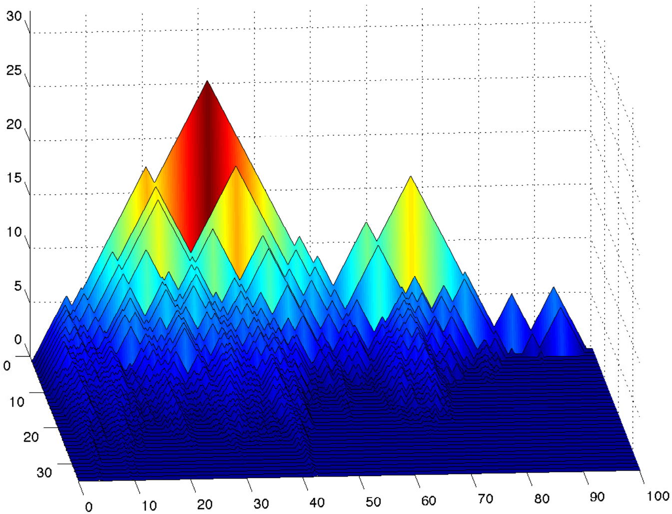
Brain artery persistence landscapes



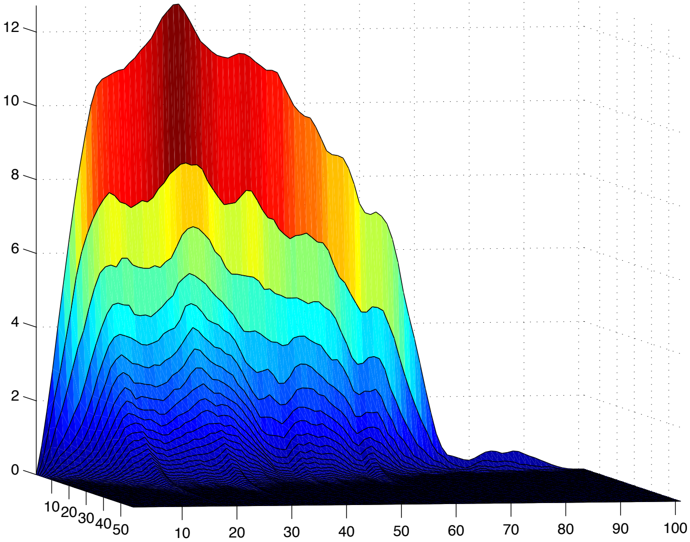
Brain artery persistence landscapes



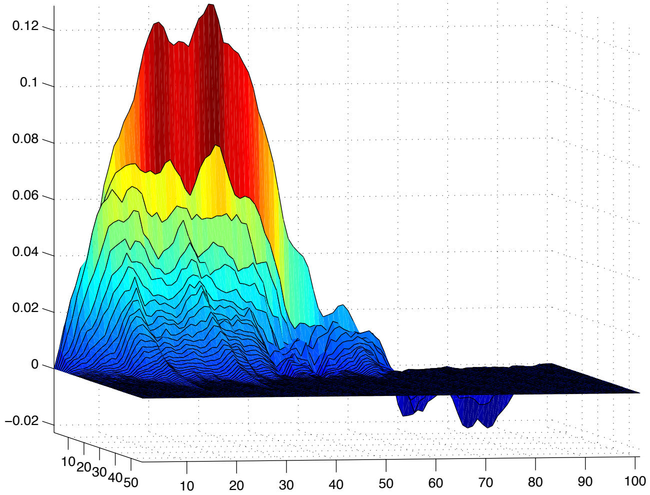
Brain artery persistence landscapes



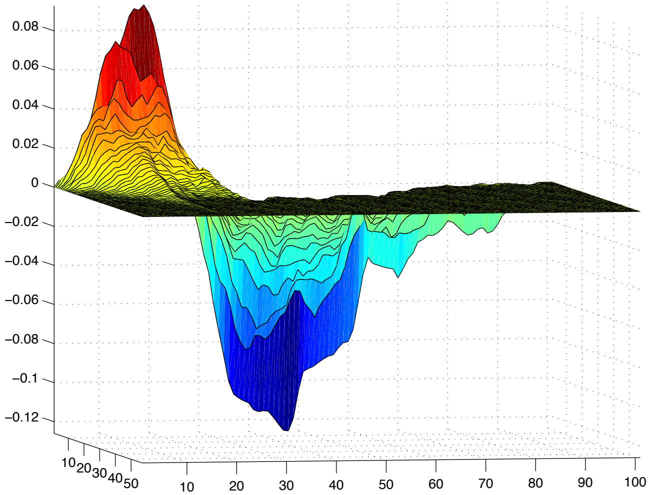
Average landscape for brain arteries



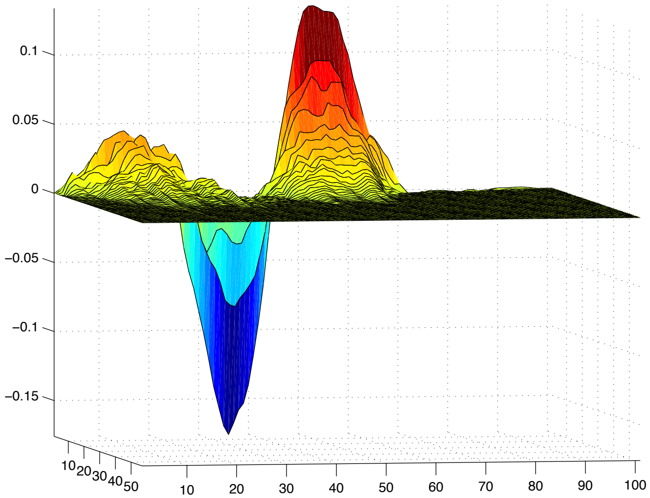
Principal Component Analysis



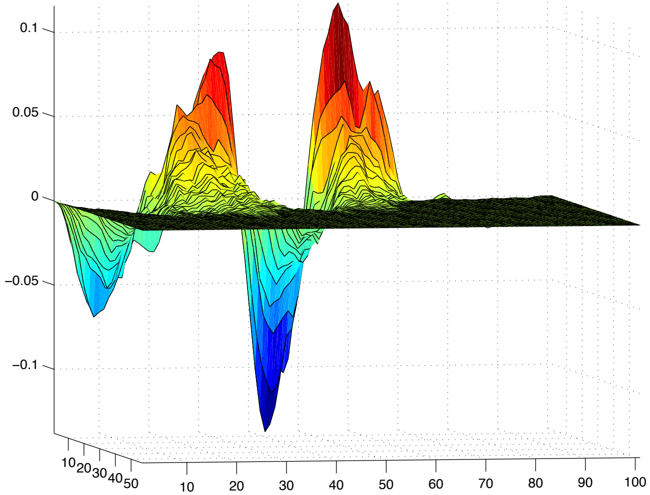
Principal Component Analysis



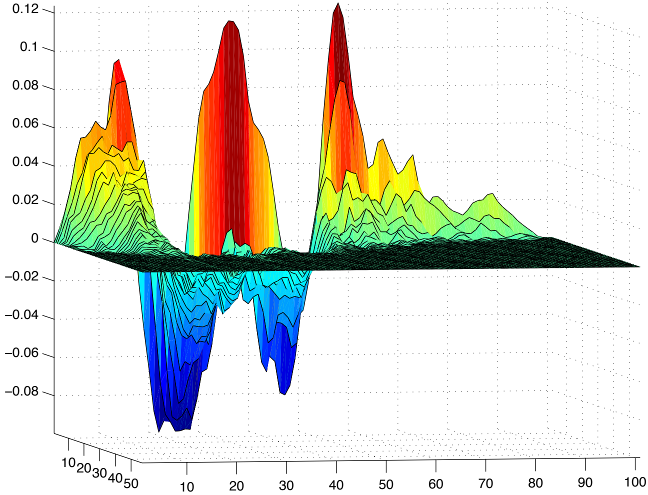
Principal Component Analysis



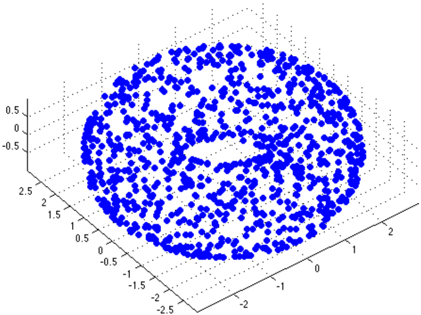
Principal Component Analysis

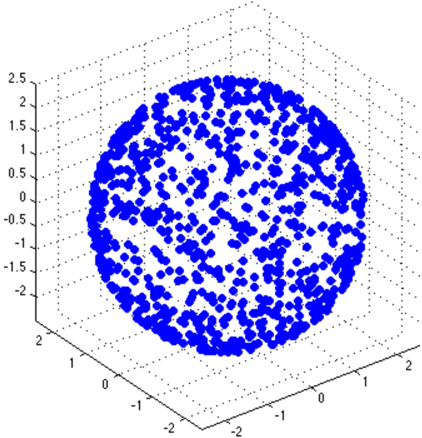


Principal Component Analysis



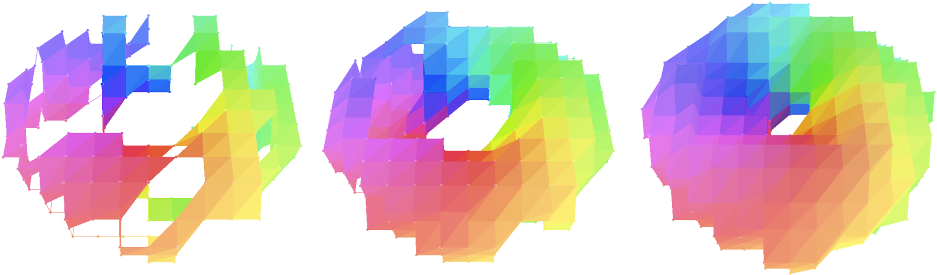
Topological hypothesis testing



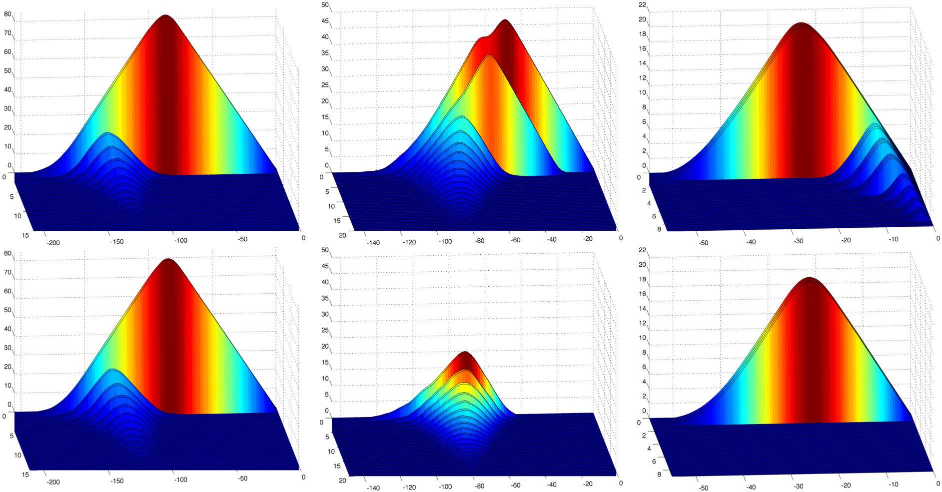


Topological hypothesis testing

Points $\to$ kernel density estimator $\to$ filtered simplicial complex



Topological hypothesis testing



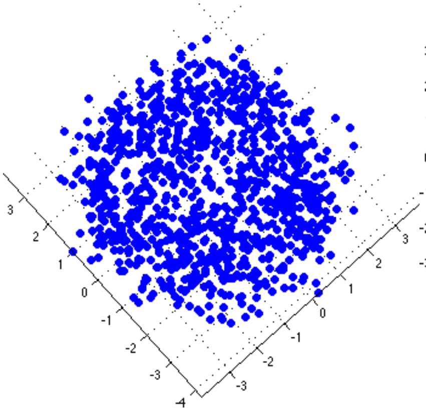
Topological hypothesis testing

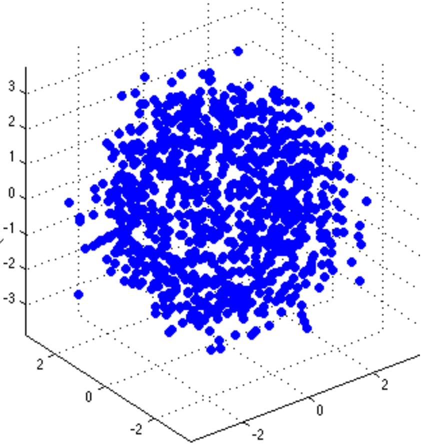
Null hypothesis: $\|\overline{\lambda_S}\|_1 = \|\overline{\lambda_T}\|_1$.

t-test:

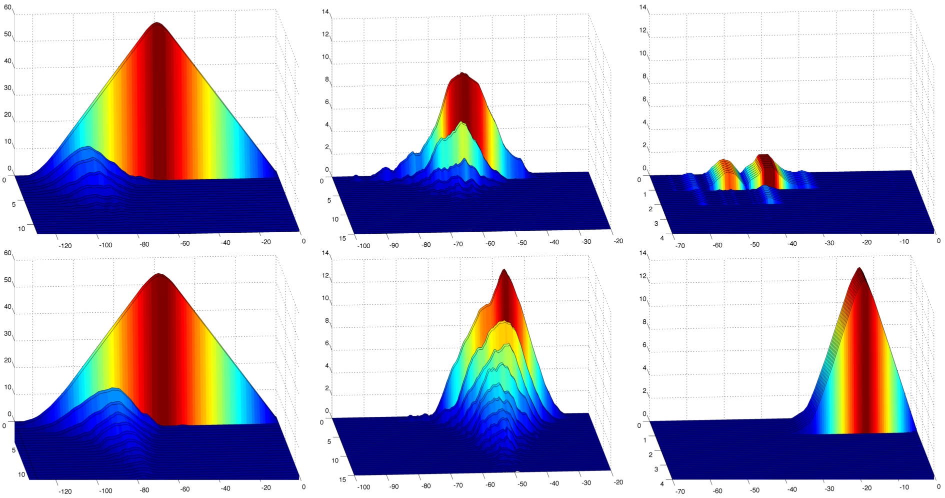
dim	decision	p-value
0	cannot reject	
1	reject	3×10^{-6}
2	cannot reject	

Topological hypothesis testing, noisy





Topological hypothesis testing, noisy



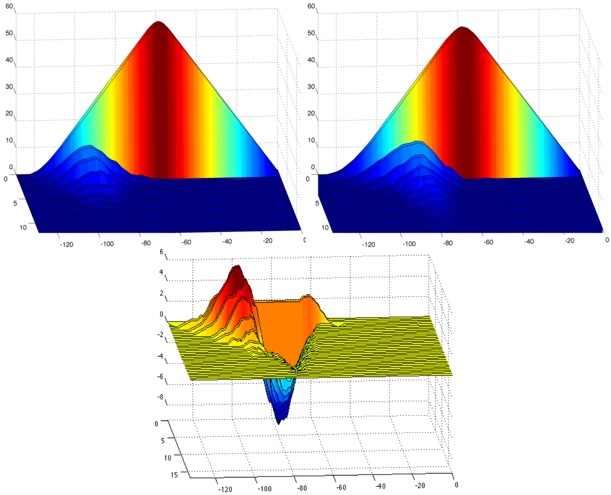
Topological hypothesis testing, noisy

Null hypothesis: $\|\overline{\lambda_S} - \overline{\lambda_T}\|_2 = 0$.

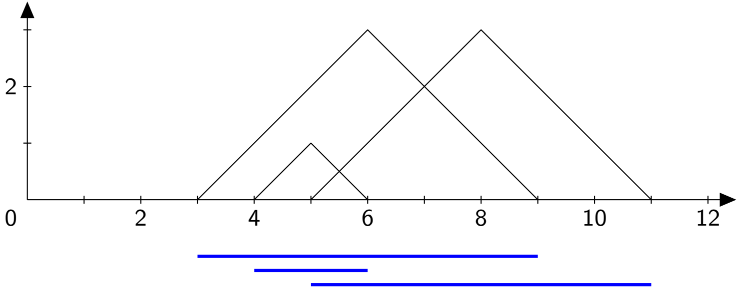
Permutation test:

dim	decision	p-value
0	reject	0.0111
1	reject	0.0000
2	reject	0.0000

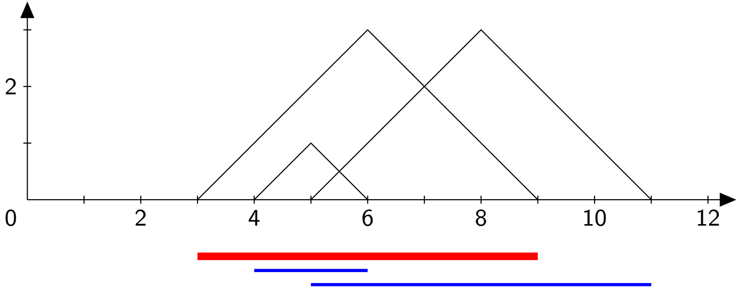
Topological hypothesis testing, noisy



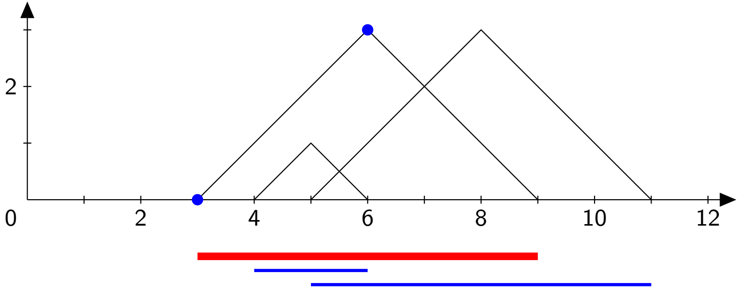
Constructing the Persistence Landscape



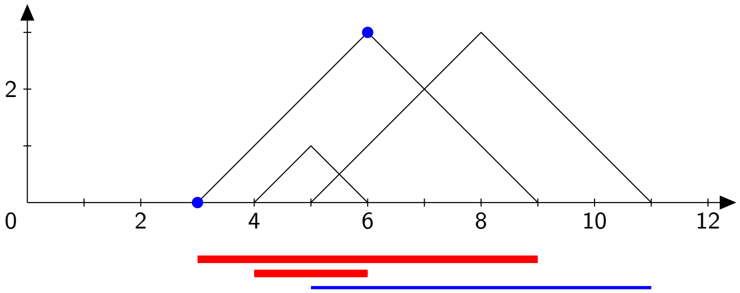
Constructing the Persistence Landscape



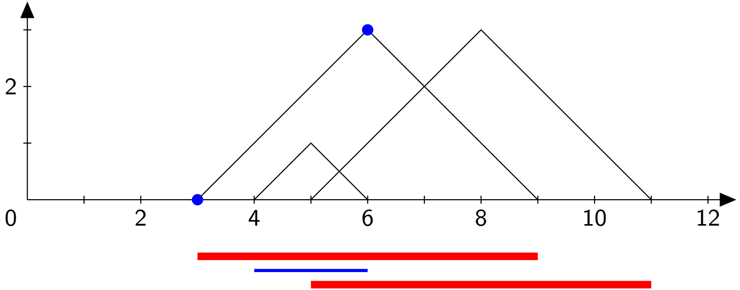
Constructing the Persistence Landscape



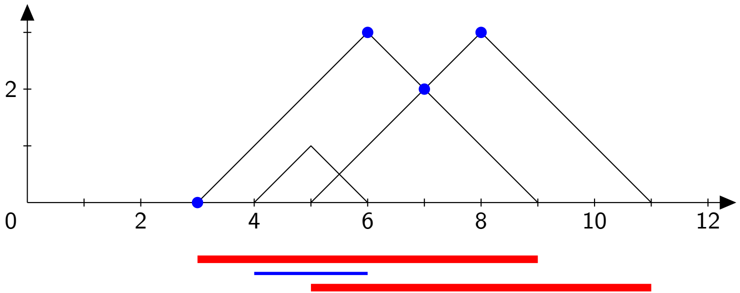
Constructing the Persistence Landscape



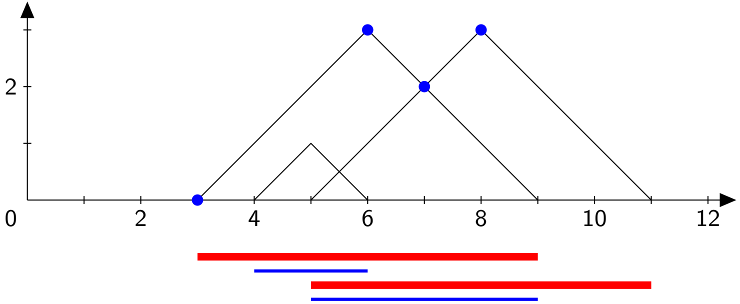
Constructing the Persistence Landscape



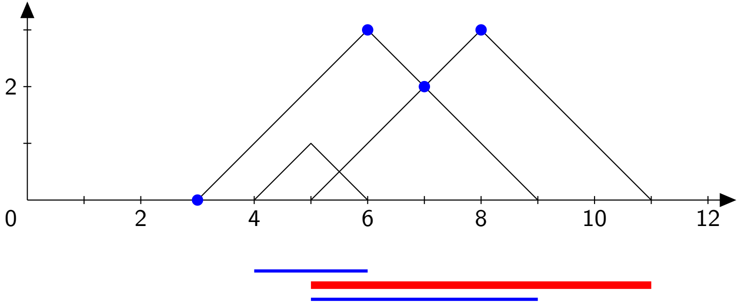
Constructing the Persistence Landscape



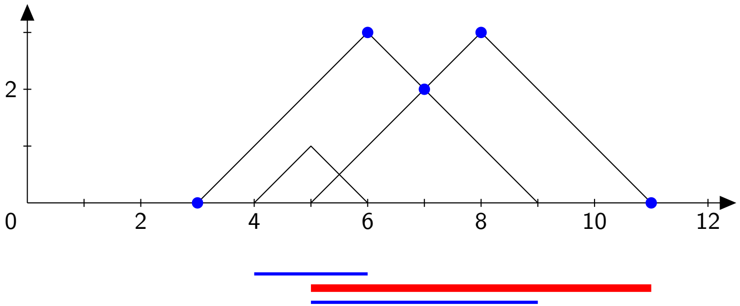
Constructing the Persistence Landscape



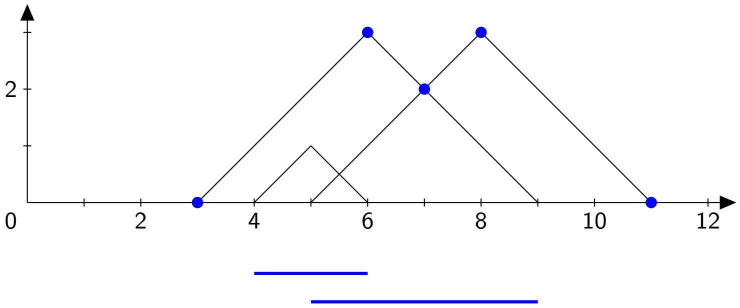
Constructing the Persistence Landscape



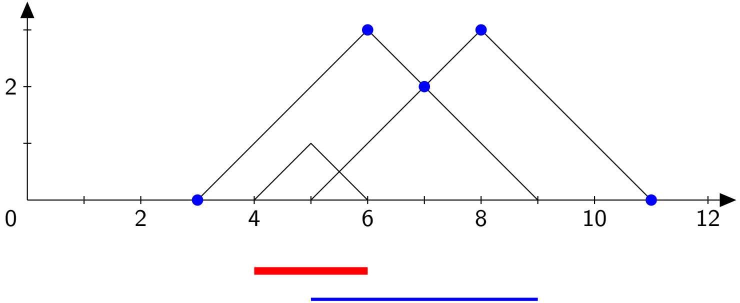
Constructing the Persistence Landscape



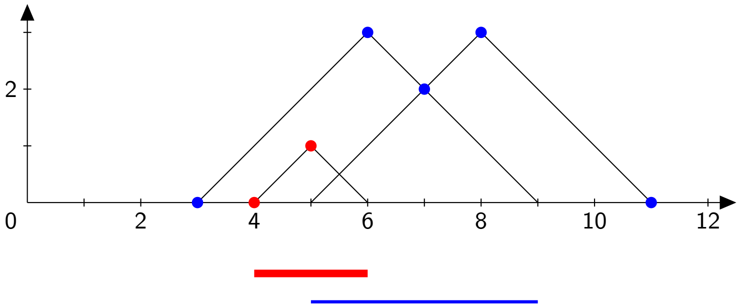
Constructing the Persistence Landscape



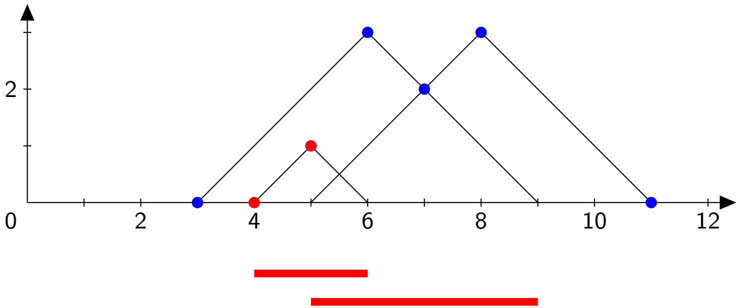
Constructing the Persistence Landscape



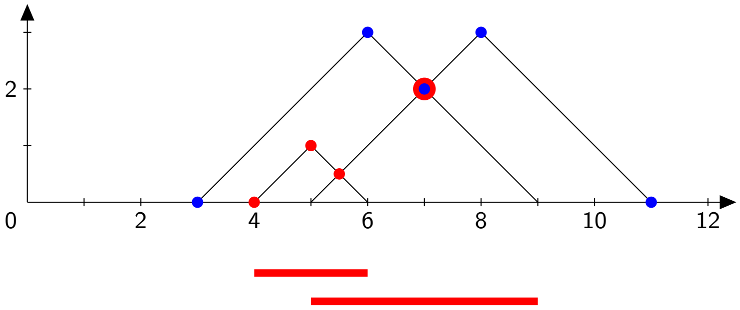
Constructing the Persistence Landscape



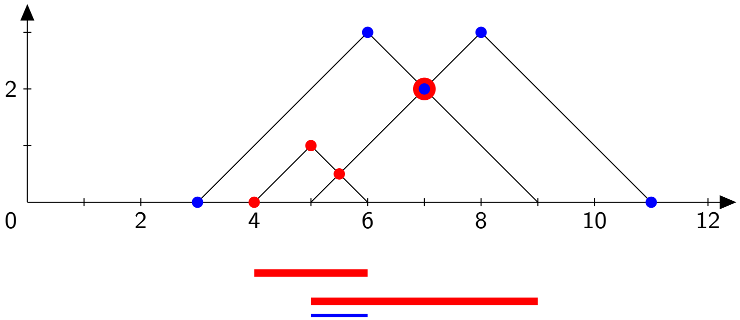
Constructing the Persistence Landscape



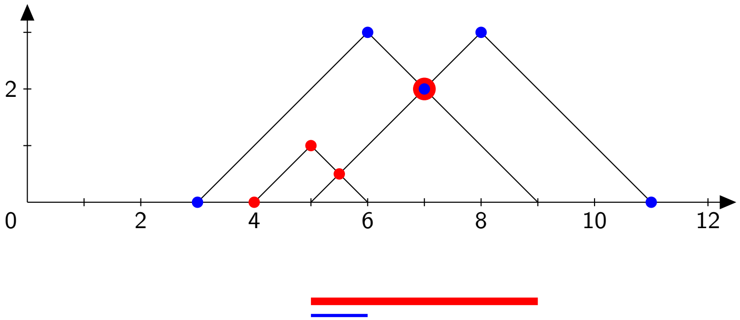
Constructing the Persistence Landscape



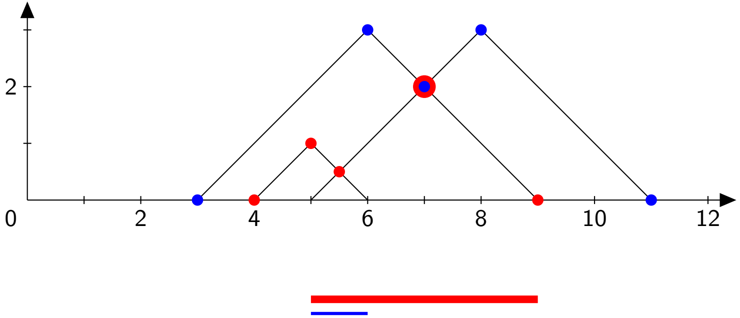
Constructing the Persistence Landscape



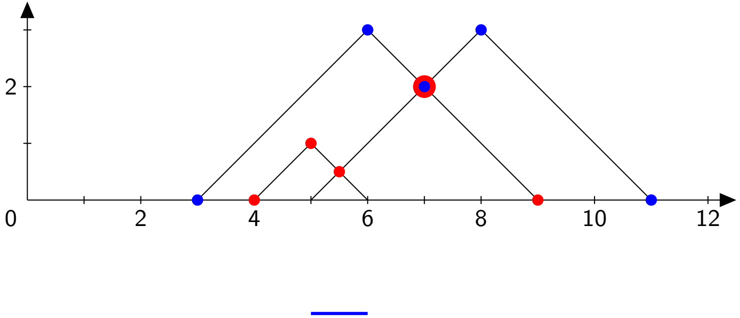
Constructing the Persistence Landscape



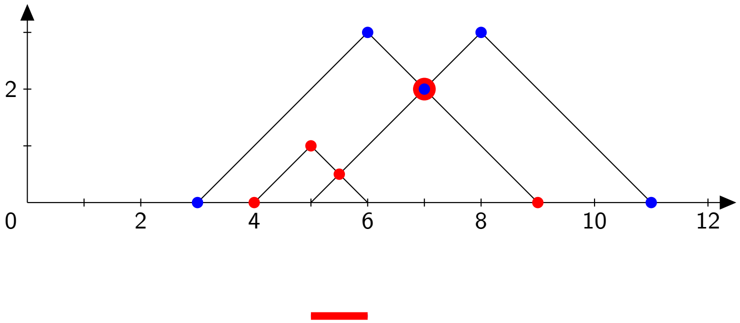
Constructing the Persistence Landscape



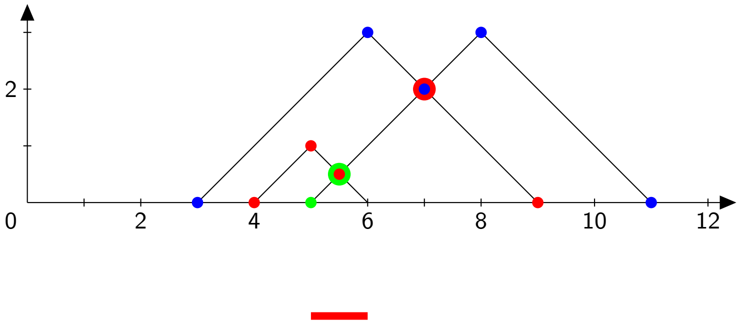
Constructing the Persistence Landscape



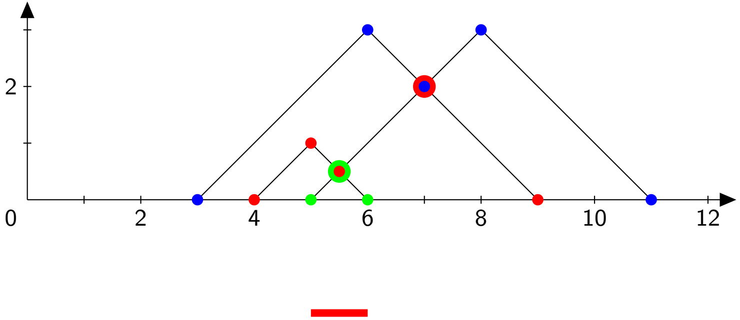
Constructing the Persistence Landscape



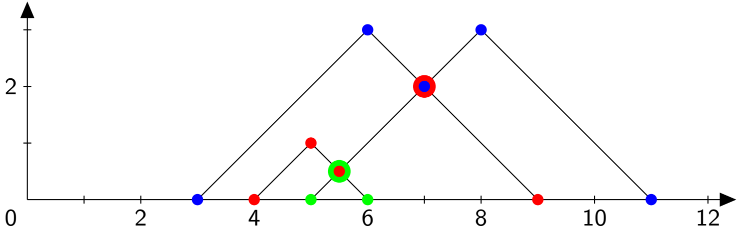
Constructing the Persistence Landscape



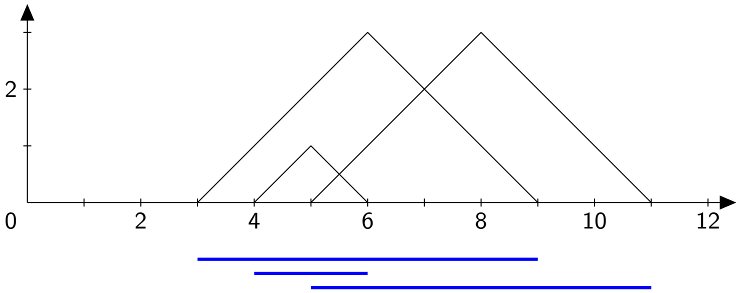
Constructing the Persistence Landscape



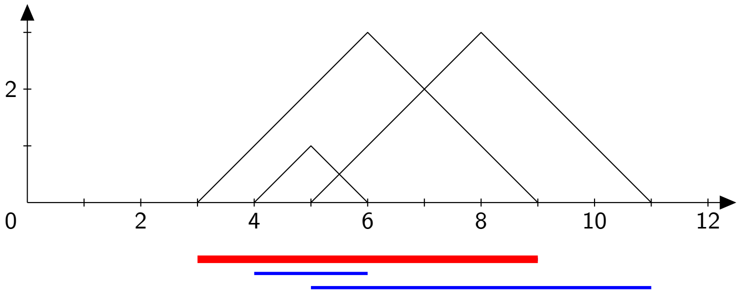
Constructing the Persistence Landscape



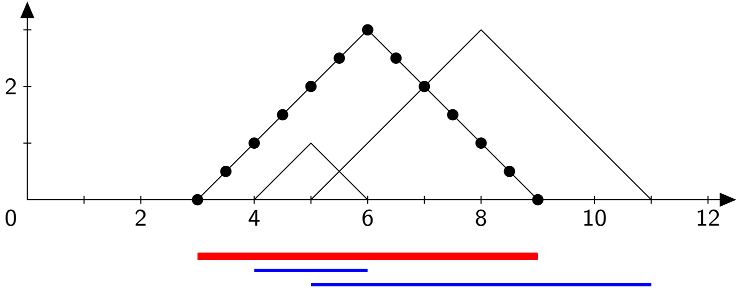
Constructing the Persistence Landscape on a grid



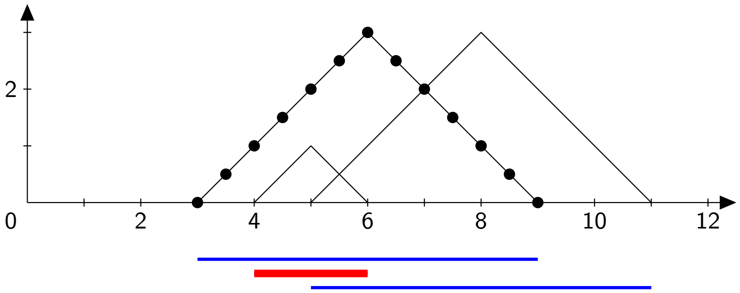
Constructing the Persistence Landscape on a grid



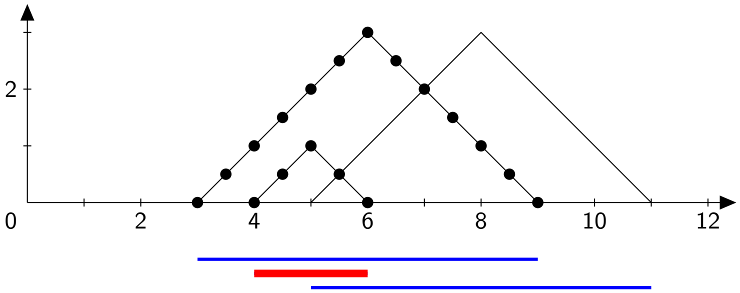
Constructing the Persistence Landscape on a grid



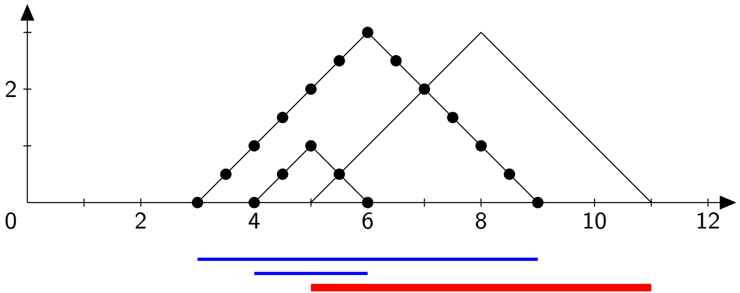
Constructing the Persistence Landscape on a grid



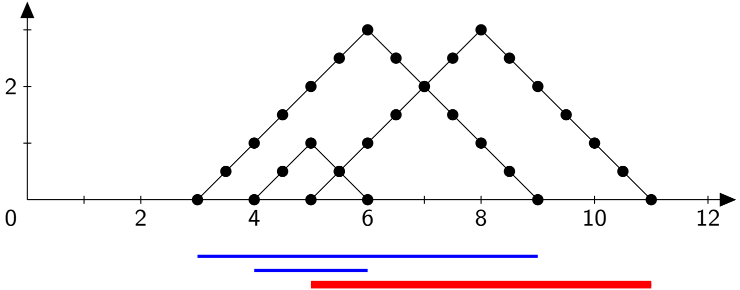
Constructing the Persistence Landscape on a grid



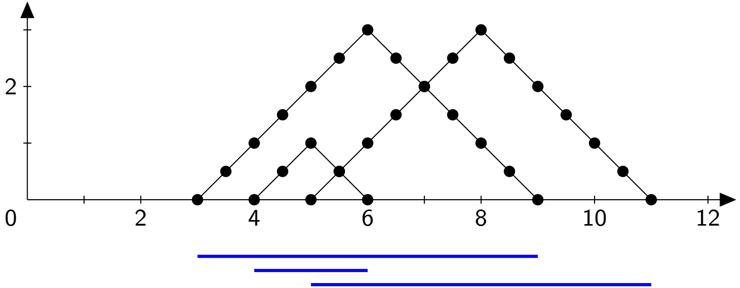
Constructing the Persistence Landscape on a grid



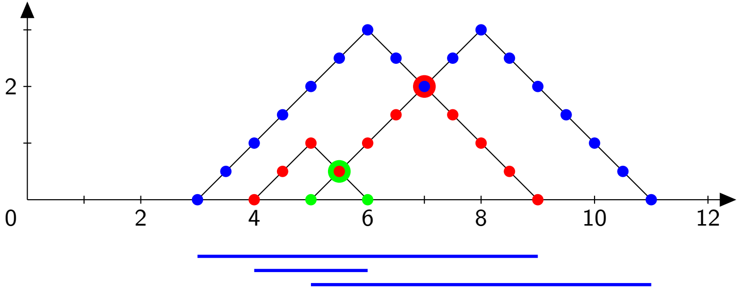
Constructing the Persistence Landscape on a grid



Constructing the Persistence Landscape on a grid



Constructing the Persistence Landscape on a grid



Computational complexity

Starting with n birth-death pairs

	no grid	grid
Construct persistence landscape	$O(n^2)$	$O(n \log n)$
Distance between two landscapes	$O(n^2)$	$O(n)$
Average of N landscapes	$O(n^2 N^2)$	$O(nN)$

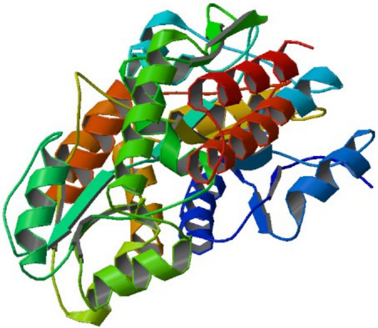
Joint work with Pawel Dlotko.

Computational packages

Code for computing the persistence landscape and associated constructions is available.

- The Persistence Landscape Toolbox, Pawel Dlotko
- the R package **TDA**, Brittany Fasy, Jisu Kim, Fabrizio Lecci and Clément Maria

Maltose Binding Protein, two 'conformations'



Joint work with Giseon Heo and Violeta Kovacev-Nikolic (Alberta) and Dragan Nikolic (Caltech–JPL)

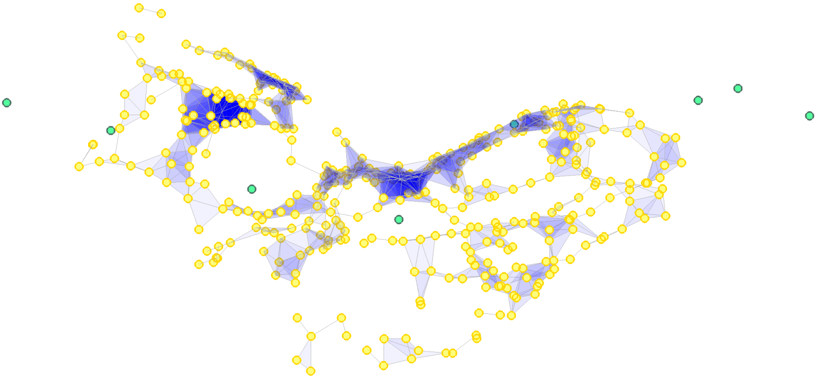
Maltose Binding Protein (MBP)

Fourteen MBP structures from the Protein Data Bank (PDB).

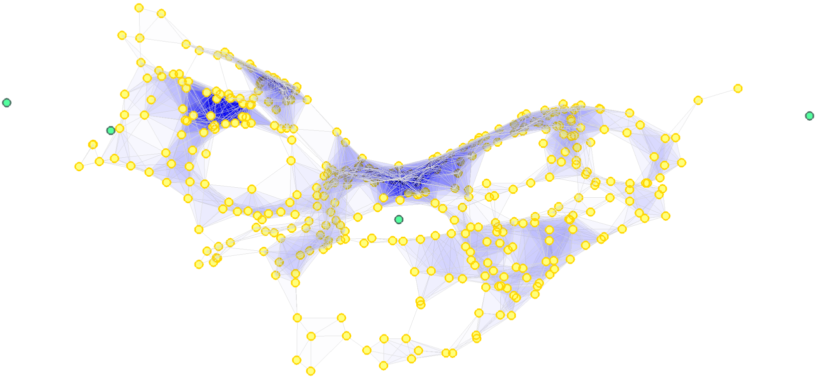
No.	PDB code	Ligand name	Protein structure
1	1ANF	maltose	closed-*holo*
2	1FQC	maltotriotol	closed-*holo*
3	1FQD	maltotetraitol	closed-*holo*
4	1MPD	maltose	closed-*holo*
5	3HPI	sucrose	closed-*holo*
6	3MBP	maltotriose	closed-*holo*
7	4MBP	maltotetraose	closed-*holo*
8	1EZ9	maltotetraitol	open-*holo*
9	1FQA	maltotetraitol	open-*holo*
10	1FQB	maltotetraitol	open-*holo*
11	1JW4	-	open-*apo*
12	1JW5	maltose	open-*holo*
13	1LLS	-	open-*apo*
14	1OMP	-	open-*apo*

MBP Vietoris-Rips complex

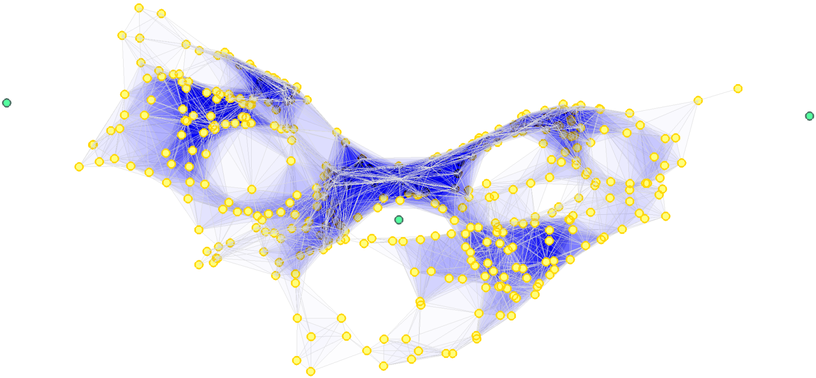
MBP Vietoris-Rips complex



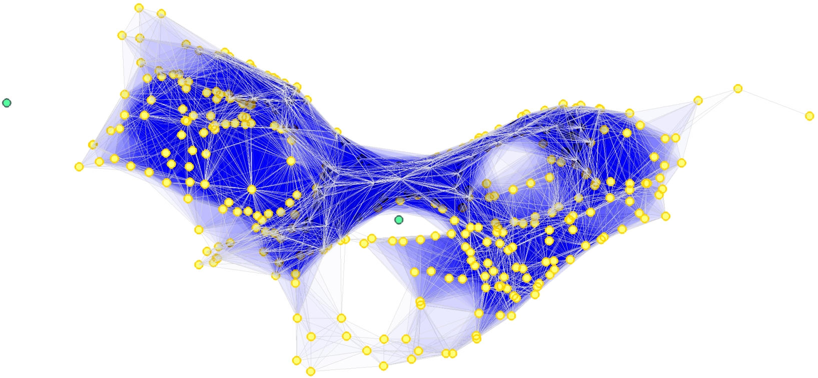
MBP Vietoris-Rips complex



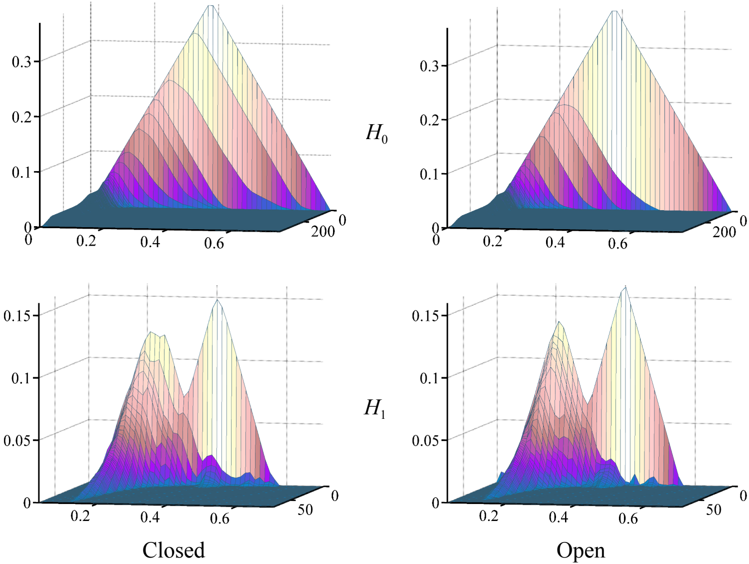
MBP Vietoris-Rips complex



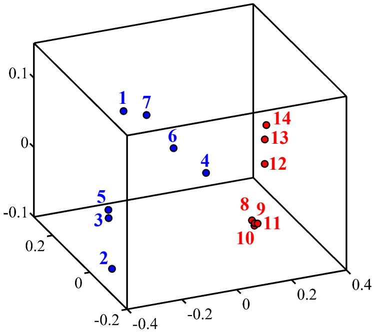
MBP Vietoris-Rips complex

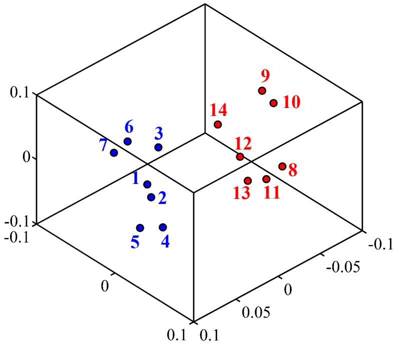


MBP average landscapes

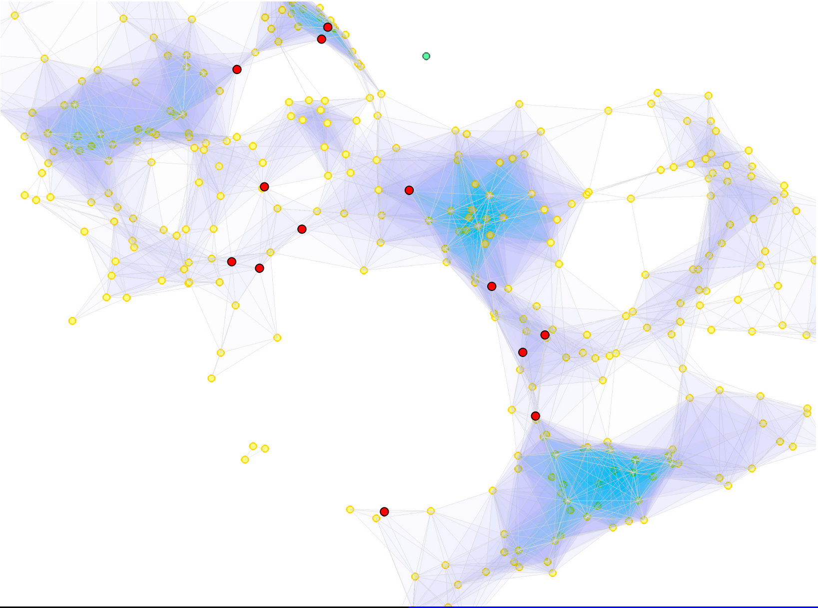


MBP landscape distance





Active sites and the most persistent cycle



Acknowledgments

Collaborators:

- Brain arteries – Paul Bendich and Ezra Miller (Duke), J.S. Marron and Sean Skwerer (UNC-CH)
- Protein data – Giseon Heo and Violeta Kovacev-Nikolic (Alberta) and Dragan Nikolic (Caltech–JPL)
- Algorithms and software – Pawel Dlotko (Penn)

Sponsor:

- Air Force Office of Scientific Research (AFOSR)