Integrating Topological Data Analysis (TDA) with Statistical Learning Methods

Models, Inference, and Algorithms Seminar
Broad Institute

Emily T. Winn
Division of Applied Mathematics, Brown University

Website: www.emilytwinn.com
Twitter: @EmilyTWinn13

November 6, 2019

Table of Contents

Background

What is topology?

Figure: "A topologist cannot tell the difference between a coffee cup and a donut."

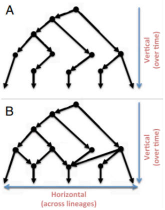
Frame from YouTube video (Sagerman, 2015)

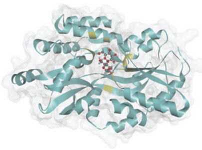
What is Topological Data Analysis (TDA)?

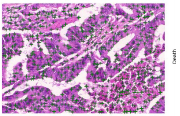
"TDA aims at providing well-founded mathematical, statistical and algorithmic methods to infer, analyze and exploit the complex topological and geometric structures underlying data that are often represented as point clouds in Euclidean or more general metric spaces." (Chazal and Michel, 2017)

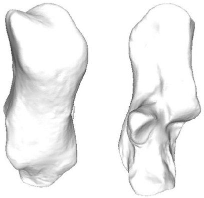
What is Topological Data Analysis (TDA)?

"TDA aims at providing well-founded mathematical, statistical and algorithmic methods to infer, analyze and exploit the complex topological and geometric structures underlying data that are often represented as point clouds in Euclidean or more general metric spaces." (Chazal and Michel, 2017)









Basic Outline of TDA Algorithm

Basic Outline of TDA Algorithm

1. Input: finite set of points with a notion of distance/similarity between them.

Basic Outline of TDA Algorithm

1. Input: finite set of points with a notion of distance/similarity between them.
2. A "continuous" shape is built on top of data to highlight underlying topology/geometry

Basic Outline of TDA Algorithm

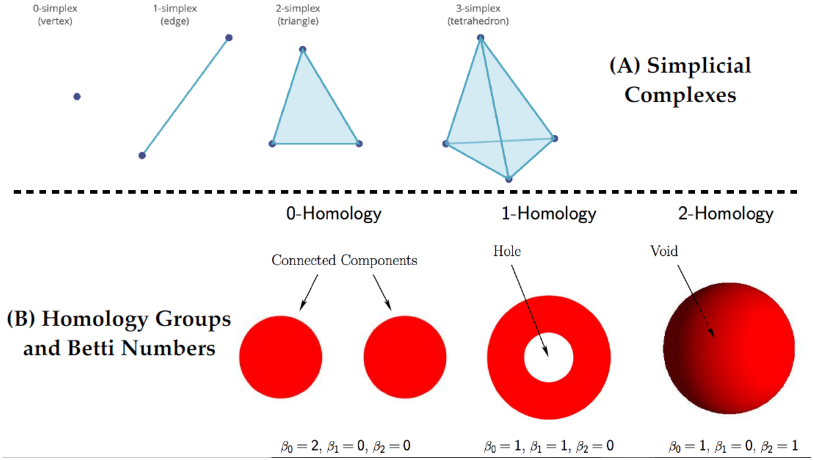
1. Input: finite set of points with a notion of distance/similarity between them.
2. A "continuous" shape is built on top of data to highlight underlying topology/geometry
3. Topological or geometric information is extracted from this structure built on top of the data.

Basic Outline of TDA Algorithm

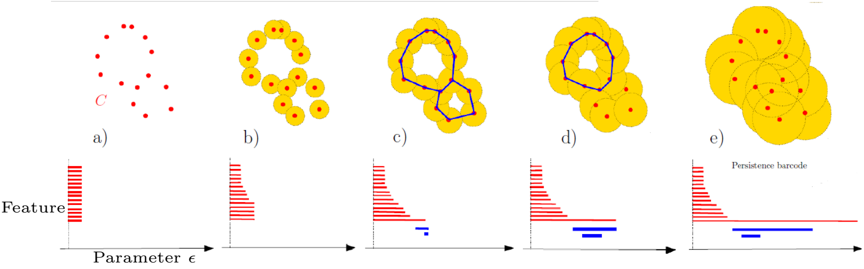
1. Input: finite set of points with a notion of distance/similarity between them.
2. A "continuous" shape is built on top of data to highlight underlying topology/geometry
3. Topological or geometric information is extracted from this structure built on top of the data.
4. Extracted features give new families of features/descriptors of the data.

(Chazal and Michel, 2017)

Persistence Homology



Persistence Diagrams and Barcodes

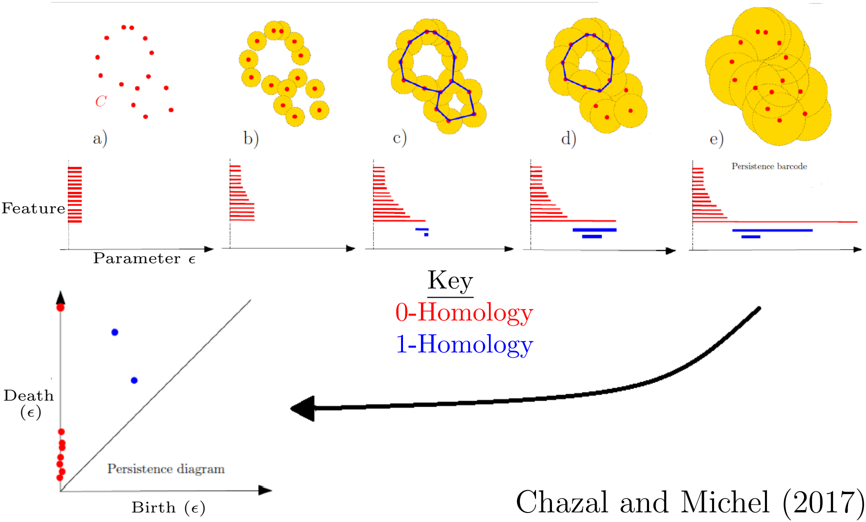


Key

0-Homology

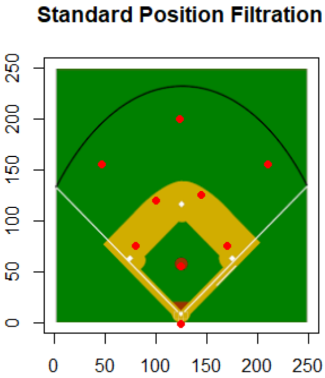
1-Homology

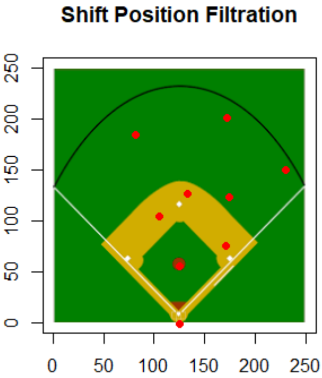
Persistence Diagrams and Barcodes



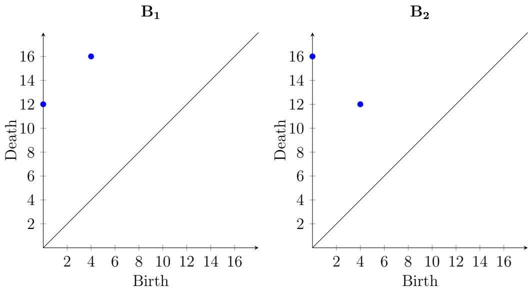
Chazal and Michel (2017)

Example: Baseball Fielding





Comparing Persistence Diagrams

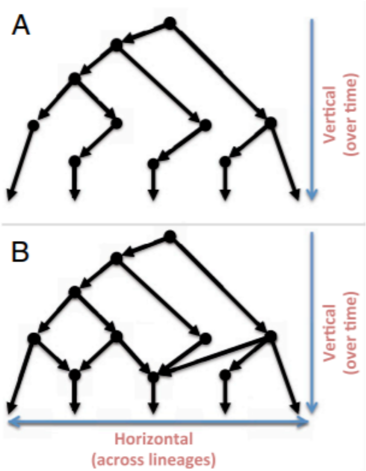


$$W_p(B_1, B_2) = \inf_{\gamma : B_1 \to B_2} \left(\sum_{u \in B_1} ||u - \gamma(u)||_\infty^p \right)^{1/p} \quad (1 \leq p < \infty)$$

$$W_\infty(B_1, B_2) = \inf_{\gamma : B_1 \to B_2} \sup_{u \in B_1} ||u - \gamma(u)||_\infty$$

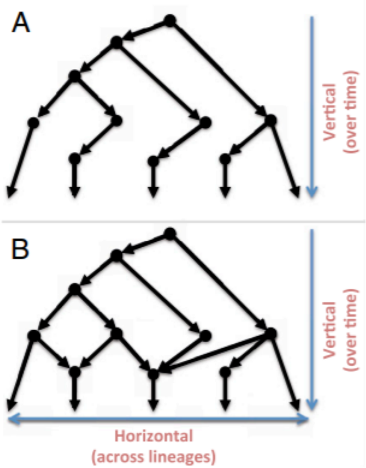
(Bubenik, 2015)

Persistence Diagrams Application: Viral Evolution



(Chan et al., 2013)

Persistence Diagrams Application: Viral Evolution



- Each genetic code is a point, visualize with Principal Coordinate Analysis
- Use genetic distance as the parameter ϵ
- Goal: Capture complex exchanges with more than two organisms, statistical patterns of cosegregation

(Chan et al., 2013)

Persistence Diagrams Application: Viral Evolution

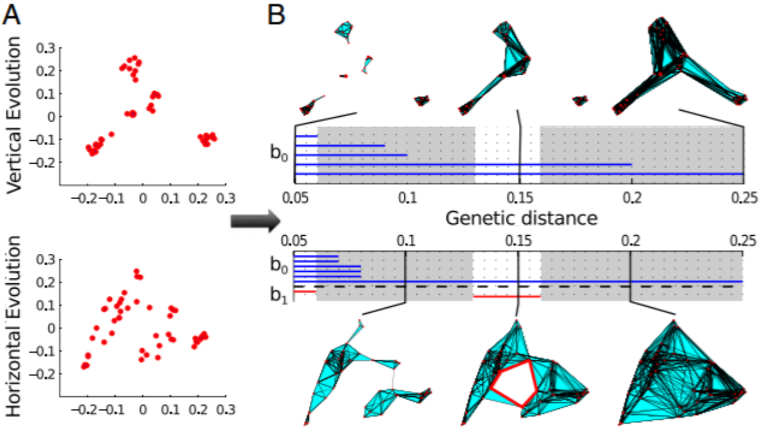


Figure: Simulated viral evolution, with and without reassortment. (Chan et al., 2013)

Persistence Diagrams Application: Viral Evolution

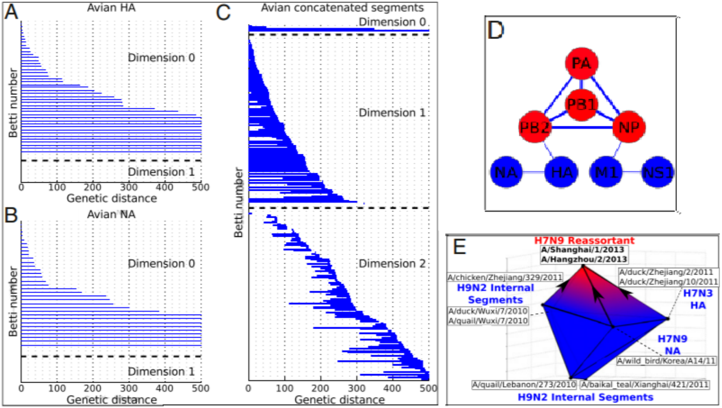
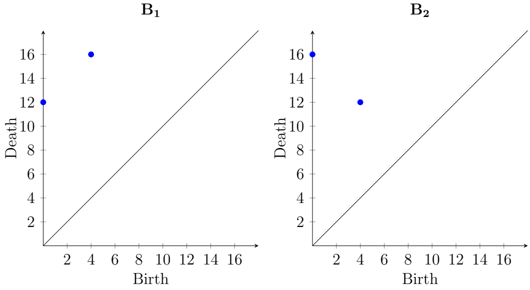


Figure: Persistent homology detects horizontal evolution (dimension 1) and complex reticulate evolution (dimension 2) in avian influenza. (Chan et al., 2013)

Comparing Persistence Diagrams

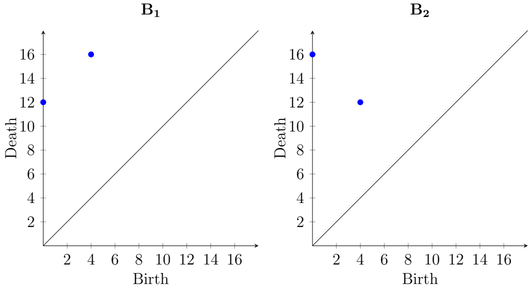


$$W_p(B_1, B_2) = \inf_{\gamma: B_1 \to B_2} \Big(\sum_{u \in B_1} ||u - \gamma(u)||_\infty^p \Big)^p \quad (1 \le p < \infty)$$

$$W_\infty(B_1, B_2) = \inf_{\gamma: B_1 \to B_2} \sup_{u \in B_1} ||u - \gamma(u)||_\infty$$

[(Bubenik, 2015), (Dey and Xin, 2019)]

Comparing Persistence Diagrams



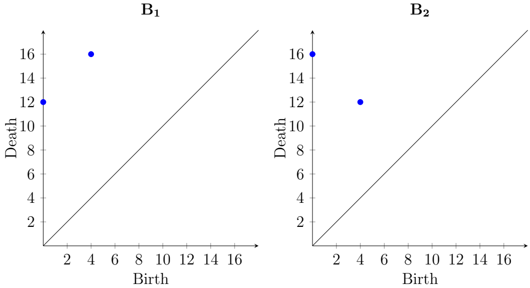
$$W_p(B_1, B_2) = \inf_{\gamma : B_1 \to B_2} \left(\sum_{u \in B_1} ||u - \gamma(u)||_\infty^p \right)^p \quad (1 \leq p < \infty)$$

$$W_\infty(B_1, B_2) = \inf_{\gamma : B_1 \to B_2} \sup_{u \in B_1} ||u - \gamma(u)||_\infty$$

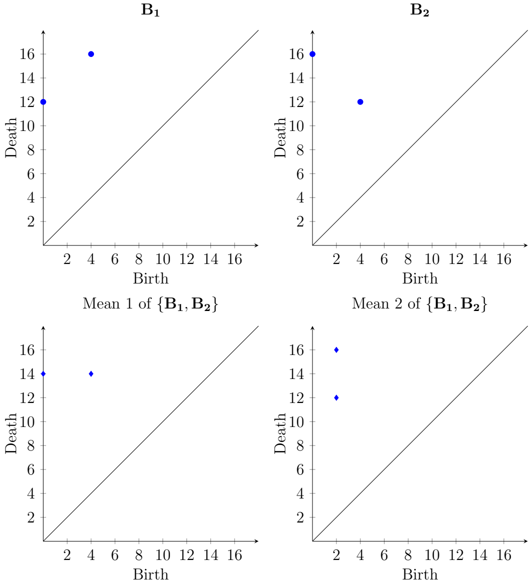
$O(m^{5/2} \log(m))$

[(Bubenik, 2015), (Dey and Xin, 2019)]

Average of Two Persistence Diagrams



Average of Two Persistence Diagrams



Recap: Persistence Diagrams

Pros:

Recap: Persistence Diagrams

Pros:

- Can look at underlying manifold, which contains information not available from data alone

Recap: Persistence Diagrams

Pros:

- Can look at underlying manifold, which contains information not available from data alone
- Descriptor in a metric space

Recap: Persistence Diagrams

Pros:

- Can look at underlying manifold, which contains information not available from data alone
- Descriptor in a metric space
- Stable against outliers, perturbations

Cons:

Recap: Persistence Diagrams

Pros:

- Can look at underlying manifold, which contains information not available from data alone
- Descriptor in a metric space
- Stable against outliers, perturbations

Cons:

- Difficult to integrate with statistics/machine learning tools we already have

Recap: Persistence Diagrams

Pros:

- Can look at underlying manifold, which contains information not available from data alone
- Descriptor in a metric space
- Stable against outliers, perturbations

Cons:

- Difficult to integrate with statistics/machine learning tools we already have
- Metric difficult to calculate

Recap: Persistence Diagrams

Pros:

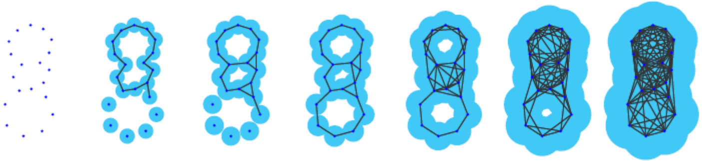
- Can look at underlying manifold, which contains information not available from data alone
- Descriptor in a metric space
- Stable against outliers, perturbations

Cons:

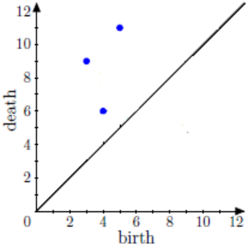
- Difficult to integrate with statistics/machine learning tools we already have
- Metric difficult to calculate
- No guarantee of a unique mean

Persistence and Statistics

Persistence Landscapes

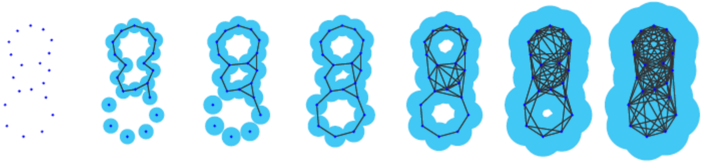


1-st Homology group (holes)

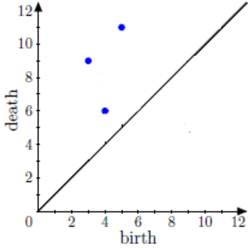


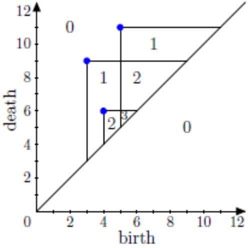
(Bubenik, 2015)

Persistence Landscapes



1-st Homology group (holes)

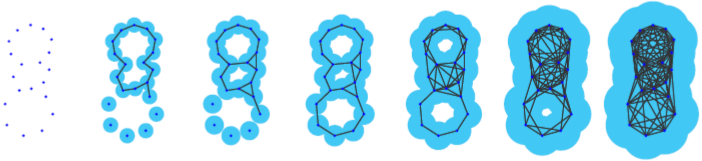




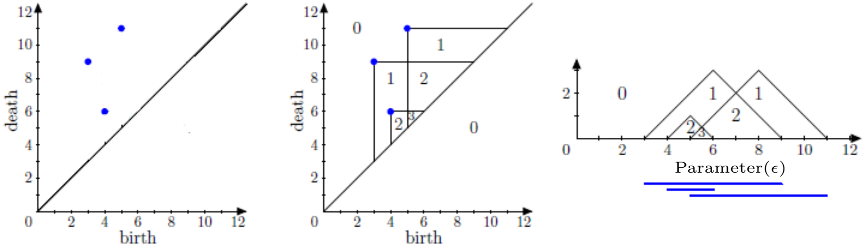
Label regions by Betti number β_1

(Bubenik, 2015)

Persistence Landscapes



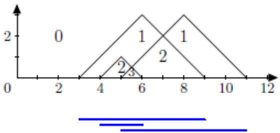
1-st Homology group (holes)

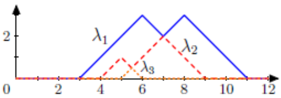


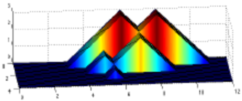
Label regions by Betti number β_1

(Bubenik, 2015)

Persistence Landscapes λ



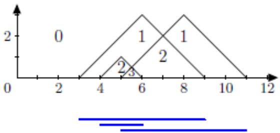


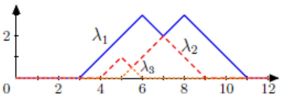


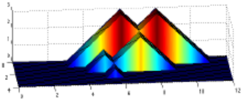
λ_1 bounds region where $\beta_1 \geq 1$

λ_2 bounds region where $\beta_1 \geq 2$

Persistence Landscapes λ







λ_1 bounds region where $\beta_1 \geq 1$
λ_2 bounds region where $\beta_1 \geq 2$

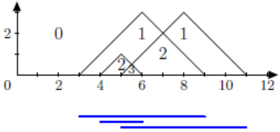
Definition

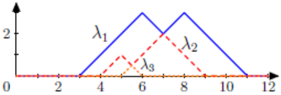
Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots), \boldsymbol{\lambda'} = (\lambda'_1, \lambda'_2, \ldots)$ be persistence landscapes corresponding to persistence diagrams B_1, B_2. The *p-landscape distance* $(1 \leq p < \infty)$ is given by

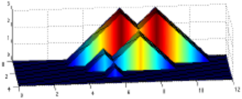
$$\Lambda_p(B_1, B_2) = ||\boldsymbol{\lambda} - \boldsymbol{\lambda'}||_p = \left[\sum_k \int_{\mathbb{R}} |\lambda_k(t) - \lambda'_k(t)|^p dt \right]^{1/p}$$

[(Bubenik, 2015),(Kovacev-Nikolic et al., 2016), (Bubenik and Dłotko, 2014)]

Persistence Landscapes λ







λ_1 bounds region where $\beta_1 \geq 1$

λ_2 bounds region where $\beta_1 \geq 2$

Definition

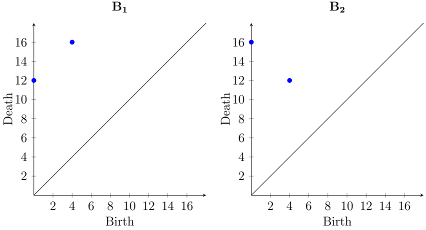
Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ...), \boldsymbol{\lambda'} = (\lambda'_1, \lambda'_2, ...)$ be persistence landscapes corresponding to persistence diagrams B_1, B_2. The *p-landscape distance* $(1 \leq p < \infty)$ is given by

$$\Lambda_p(B_1, B_2) = ||\boldsymbol{\lambda} - \boldsymbol{\lambda'}||_p = \left[\sum_k \int_{\mathbb{R}} |\lambda_k(t) - \lambda'_k(t)|^p dt \right]^{1/p}$$

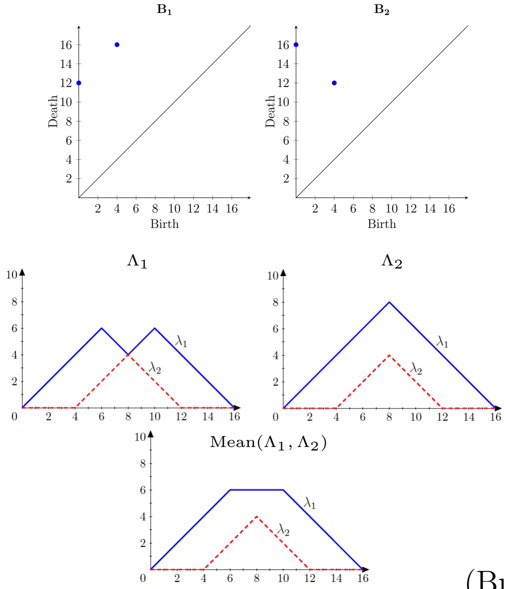
$O(m^2)$

[(Bubenik, 2015),(Kovacev-Nikolic et al., 2016), Bubenik and Dłotko (2014)]

Persistence Landscape Advantage: Unique Means!



Persistence Landscape Advantage: Unique Means!



(Bubenik, 2015)

Persistent Landscape Advantage: It's a random variable!

$$X = f(\lambda_k(t)) = \sum_k \int_{\mathbb{R}} t\lambda_k(t)dt$$

Persistent Landscape Advantage: It's a random variable!

$$X = f(\lambda_k(t)) = \sum_k \int_{\mathbb{R}} t\lambda_k(t)dt$$

- Persistent landscapes are in a separable, Banach space. ($\mathbb{L}^p(\mathcal{S})$, where $\mathcal{S} = \mathbb{N} \times \mathbb{R}$ or $\mathbb{R}^2$)

Persistent Landscape Advantage: It's a random variable!

$$X = f(\lambda_k(t)) = \sum_k \int_{\mathbb{R}} t\lambda_k(t)dt$$

- Persistent landscapes are in a separable, Banach space. ($\mathbb{L}^p(\mathcal{S})$, where $\mathcal{S} = \mathbb{N} \times \mathbb{R}$ or $\mathbb{R}^2$)
- Translation: we can use the **Strong Law of Large Numbers** and the **Central Limit Theorem** (with enough samples, we can assume a Gaussian distribution).

Persistent Landscape Advantage: It's a random variable!

$$X = f(\lambda_k(t)) = \sum_k \int_{\mathbb{R}} t\lambda_k(t)dt$$

- Persistent landscapes are in a separable, Banach space. ($\mathbb{L}^p(\mathcal{S})$, where $\mathcal{S} = \mathbb{N} \times \mathbb{R}$ or $\mathbb{R}^2$)
- Translation: we can use the **Strong Law of Large Numbers** and the **Central Limit Theorem** (with enough samples, we can assume a Gaussian distribution).
- When $p = 2$, this space is also Hilbert.

Persistent Landscape Advantage: It's a random variable!

$$X = f(\lambda_k(t)) = \sum_k \int_{\mathbb{R}} t\lambda_k(t)dt$$

- Persistent landscapes are in a separable, Banach space. ($\mathbb{L}^p(\mathcal{S})$, where $\mathcal{S} = \mathbb{N} \times \mathbb{R}$ or $\mathbb{R}^2$)
- Translation: we can use the **Strong Law of Large Numbers** and the **Central Limit Theorem** (with enough samples, we can assume a Gaussian distribution).
- When $p = 2$, this space is also Hilbert. ...which gives us a positive definite kernel!

(Kovacev-Nikolic et al., 2016)

Persistence Landscape Application: Conformations of Maltose-Binding Protein (MBP)

- MBP can have an open or closed conformation

Persistence Landscape Application: Conformations of Maltose-Binding Protein (MBP)

- MBP can have an open or closed conformation
- Setup: shape of one MBP can be represented as 370 points in $\mathbb{R}^3$.

Persistence Landscape Application: Conformations of Maltose-Binding Protein (MBP)

- MBP can have an open or closed conformation
- Setup: shape of one MBP can be represented as 370 points in $\mathbb{R}^3$.
- Dynamic cross correlation on 370×370 matrix for 7 closed, 7 open MBPs

Persistence Landscape Application: Conformations of Maltose-Binding Protein (MBP)

- MBP can have an open or closed conformation
- Setup: shape of one MBP can be represented as 370 points in $\mathbb{R}^3$.
- Dynamic cross correlation on 370×370 matrix for 7 closed, 7 open MBPs
- Underlying distribution: Two-sample permutation t-test

Persistence Landscape Application: Conformations of Maltose-Binding Protein (MBP)

- MBP can have an open or closed conformation
- Setup: shape of one MBP can be represented as 370 points in $\mathbb{R}^3$.
- Dynamic cross correlation on 370×370 matrix for 7 closed, 7 open MBPs
- Underlying distribution: Two-sample permutation t-test
- $H_0 : \mu_C = \mu_O$, $H_a : \mu_C \neq \mu_O$

Persistence Landscape Application: Conformations of Maltose-Binding Protein (MBP)

- MBP can have an open or closed conformation
- Setup: shape of one MBP can be represented as 370 points in $\mathbb{R}^3$.
- Dynamic cross correlation on 370×370 matrix for 7 closed, 7 open MBPs
- Underlying distribution: Two-sample permutation t-test
- $H_0 : \mu_C = \mu_O$, $H_a : \mu_C \neq \mu_O$
- Classified via SVM (using 50 points from the persistence landscapes)

(Kovacev-Nikolic et al., 2016)

Persistence Landscape Application: Conformations of Maltose-Binding Protein

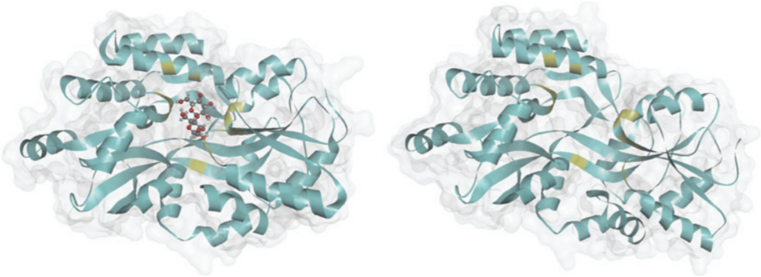


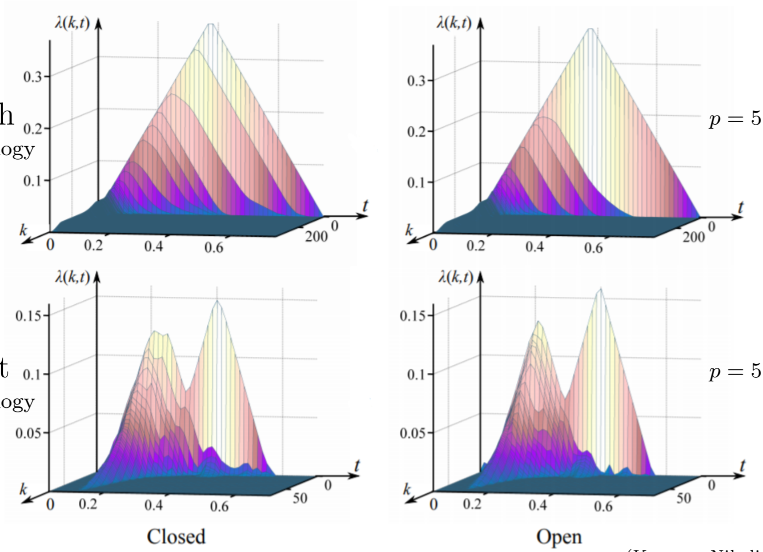
Figure: Left: closed conformal structure with ligand, Right: open conformal strcture (Kovacev-Nikolic et al., 2016)

Conformations of Maltose-Binding Protein (MBP)

Mean Landscapes | p-values

0-th Homology

1-st Homology



0-th Homology: $p = 5.83 \times 10^{-4}$

1-st Homology: $p = 5.83 \times 10^{-4}$

(Kovacev-Nikolic et al., 2016)

Recap: Persistence Landscape

Pros:

Recap: Persistence Landscape

Pros:

- Can treat persistence landscapes as random variables

Recap: Persistence Landscape

Pros:

- Can treat persistence landscapes as random variables
- Distance easier to calculate

Recap: Persistence Landscape

Pros:

- Can treat persistence landscapes as random variables
- Distance easier to calculate and gives a lower bound for the p-Wasserstein distance/bottleneck distance.

Recap: Persistence Landscape

Pros:

- Can treat persistence landscapes as random variables
- Distance easier to calculate and gives a lower bound for the p-Wasserstein distance/bottleneck distance.
- Set up to apply hypothesis testing and machine learning methods.

Cons:

Recap: Persistence Landscape

Pros:

- Can treat persistence landscapes as random variables
- Distance easier to calculate and gives a lower bound for the p-Wasserstein distance/bottleneck distance.
- Set up to apply hypothesis testing and machine learning methods.

Cons:

- Vector form takes extra processing

Recap: Persistence Landscape

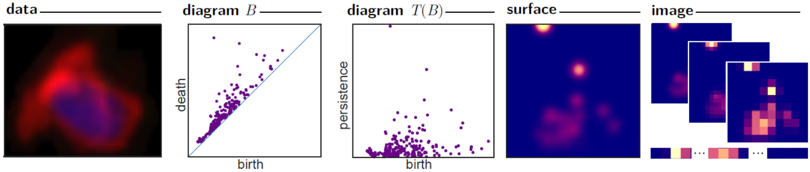
Pros:

- Can treat persistence landscapes as random variables
- Distance easier to calculate and gives a lower bound for the p-Wasserstein distance/bottleneck distance.
- Set up to apply hypothesis testing and machine learning methods.

Cons:

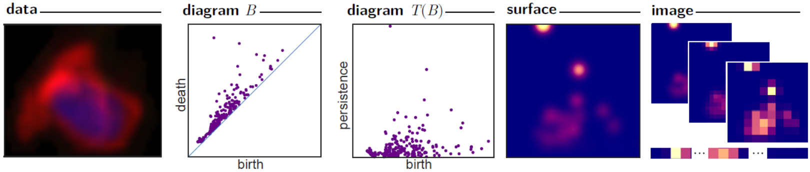
- Vector form takes extra processing
- Limited in which machine learning methods can be used

Persistence Images



Persistence Images

Steps: 0 1 2

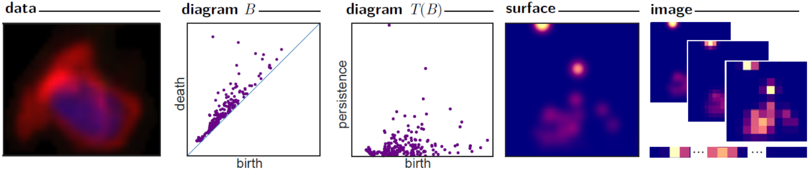


0. Calculate persistence diagram from data
1. Define $T : \mathbb{R}^2 \to \mathbb{R}^2$ by $T(x, y) = T(x, y - x)$. Then $T(B)$ is transformation of persistence diagram.
2. Choose f weighting function (depends on the application)

(Adams et al., 2017)

Persistence Images: Algorithm

Steps: 0 1 2-4

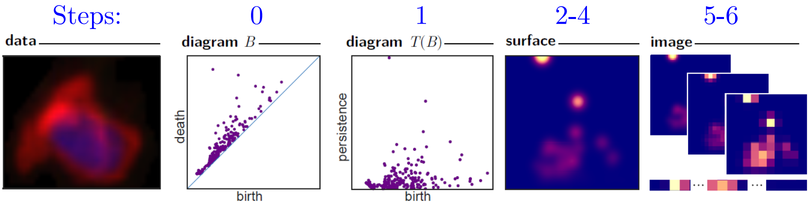


3. Choose ϕ probability function over $\mathbb{R}^2_+$ (Adams et al used joint Gaussian with mean μ and parameter σ^2).
4. Calculate the *persistence surface*, given by

$$\rho(B) = \sum_{u \in B} f(u)\phi(u)$$

(Adams et al., 2017)

Persistence Images Algorithm



5. Divide the surface into a grid (can be as coarse or fine as user decides)
6. The *persistence image* of PD B is the collection of pixels given by

$$I(\rho_B))_p = \int \int_p \rho_B dydx$$

(Adams et al., 2017)

Persistence Image Application: Histology Image

- Goal: Characterize the glandular architecture of histology images and use for classification

Persistence Image Application: Histology Image

- Goal: Characterize the glandular architecture of histology images and use for classification
- Data: MICCAI 2015 Gland Segmentation Challenge Contest data set (165 images, 85 training, 80 test)

Persistence Image Application: Histology Image

- Goal: Characterize the glandular architecture of histology images and use for classification
- Data: MICCAI 2015 Gland Segmentation Challenge Contest data set (165 images, 85 training, 80 test)
- Marked nucleoids in the images and used those as their "point cloud")

(Chittajallu et al., 2018)

Persistence Image Application: Histology Image

- Weighting function:

$$f(b, p; c) = \begin{cases} 0 & \text{if } p \leq 0 \\ p/c & \text{if } p \leq c \\ 1 & \text{otherwise} \end{cases}$$

  where b is the birth, p is the persistence, and c is the maxixum persistence over all features.

Persistence Image Application: Histology Image

- Weighting function:

$$f(b, p; c) = \begin{cases} 0 & \text{if } p \leq 0 \\ p/c & \text{if } p \leq c \\ 1 & \text{otherwise} \end{cases}$$

  where b is the birth, p is the persistence, and c is the maxixum persistence over all features.
- Probability distribution: Gaussian

Persistence Image Application: Histology Image

- Weighting function:

$$f(b, p; c) = \begin{cases} 0 & \text{if } p \leq 0 \\ p/c & \text{if } p \leq c \\ 1 & \text{otherwise} \end{cases}$$

  where b is the birth, p is the persistence, and c is the maxixum persistence over all features.
- Probability distribution: Gaussian
- Persistence Surface ($u = (u_b, u_p)$):

$$\rho(B) = \sum_{u \in T(B)} f(u_b, u_p; c) \mathcal{N}(u, \sigma^2 I)$$

(Chittajallu et al., 2018)

Persistence Image Application: Histology Images

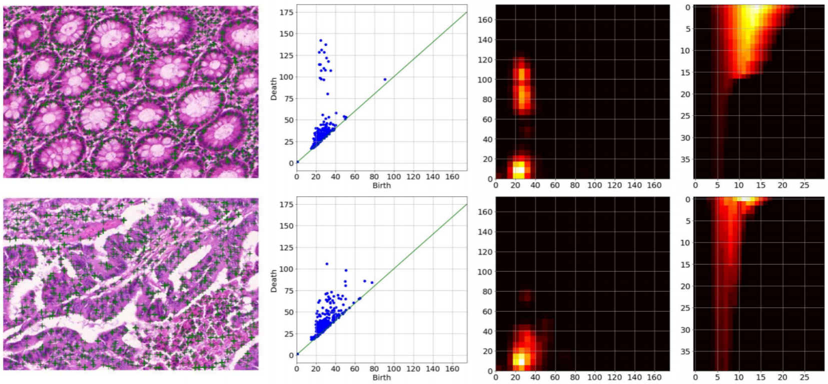


Figure: Top row: benign tissue. Bottom Row: malignant tissue. (Chittajallu et al., 2018)

Recap: Persistence Images

Pros:

Recap: Persistence Images

Pros:

- Takes just as much computational power as Persistence Landscapes, but far better in classification tasks

Recap: Persistence Images

Pros:

- Takes just as much computational power as Persistence Landscapes, but far better in classification tasks
- Once calculate, have vector, so can use for almost all machine learning tasks

Recap: Persistence Images

Pros:

- Takes just as much computational power as Persistence Landscapes, but far better in classification tasks
- Once calculate, have vector, so can use for almost all machine learning tasks
- Computational efficiency in distance calculations

Recap: Persistence Images

Pros:

- Takes just as much computational power as Persistence Landscapes, but far better in classification tasks
- Once calculate, have vector, so can use for almost all machine learning tasks
- Computational efficiency in distance calculations
- Flexible in applications, parameters can be tailored

Recap: Persistence Images

Pros:

- Takes just as much computational power as Persistence Landscapes, but far better in classification tasks
- Once calculate, have vector, so can use for almost all machine learning tasks
- Computational efficiency in distance calculations
- Flexible in applications, parameters can be tailored

Cons:

- Difficult to recover persistence diagram from persistence image

Recap: Persistence Images

Pros:

- Takes just as much computational power as Persistence Landscapes, but far better in classification tasks
- Once calculate, have vector, so can use for almost all machine learning tasks
- Computational efficiency in distance calculations
- Flexible in applications, parameters can be tailored

Cons:

- Difficult to recover persistence diagram from persistence image
- Computational efficiency for preprocessing into vector form can be improved

(Adams et al., 2017)

Topological Modeling of Surfaces

Topological Modeling of 3D Shapes

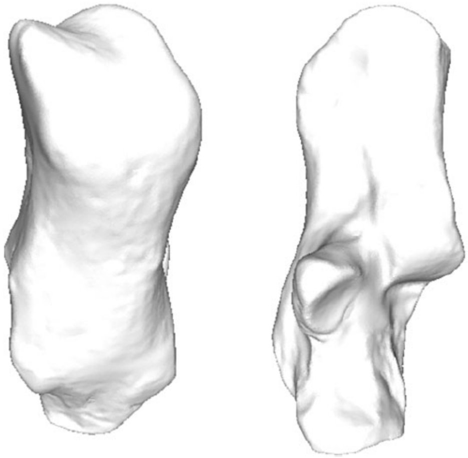


Figure: "Images of a calcaneous [heel bone] from two different angles" Turner et al. (2014)

Persistence Homology Transform (PHT)

Let M be a shape of $\mathbb{R}^d$ that can be written as a finite simplicial complex K.

Persistence Homology Transform (PHT)

Let M be a shape of $\mathbb{R}^d$ that can be written as a finite simplicial complex K.

And let $v \in S^d$ be any unit vector over the unit sphere.

Persistence Homology Transform (PHT)

Let M be a shape of $\mathbb{R}^d$ that can be written as a finite simplicial complex K.

And let $v \in S^d$ be any unit vector over the unit sphere.

We define a *filtration* $K(\nu)$ of K parameterized by a height function r as

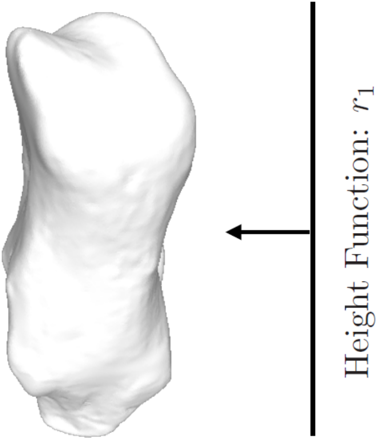
$$K(\nu)_r = \{x \in K | x \cdot \nu \leq r\}$$

The k-th dimensional persistence diagram $X_k(K, \nu)$ summarizes how topology of the filtration $K(\nu)$ changes over the height parameter r.

(Turner et al., 2014)

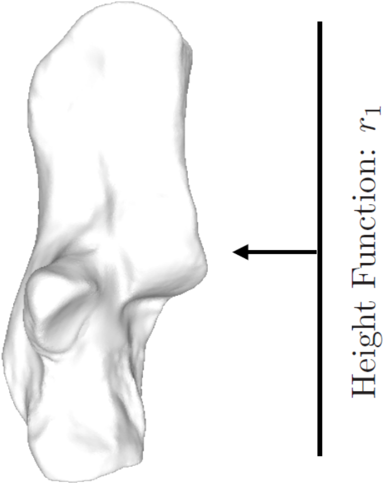
Persistent Homology Transform: Illustration

For direction ν_1:



Persistent Homology Transform: Illustration

For direction ν_2:



Persistence Homology Transform: Shape Anlaysis

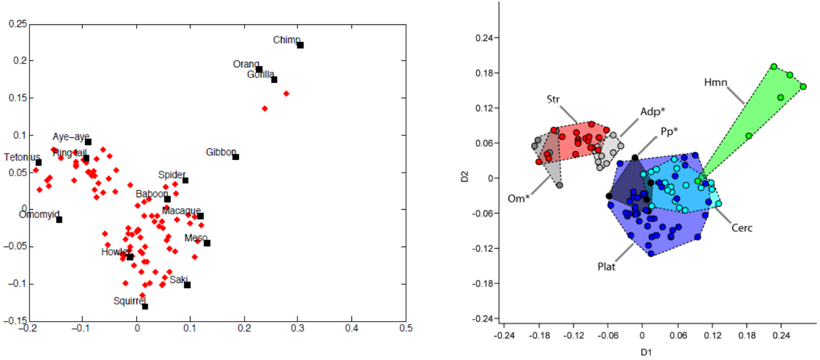
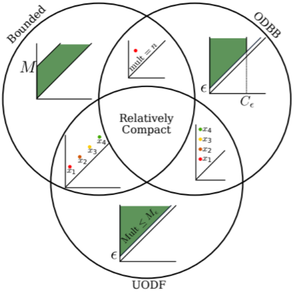


Figure: Phylogenetic groups for primate calcanei with 67 genera (Turner et al., 2014)

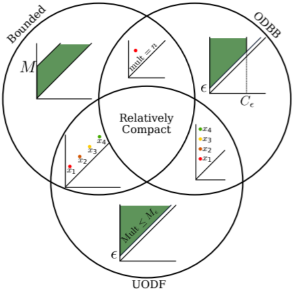
Future of TDA and Statistics

Future Directions: Theoretical Framework



(Perea et al., 2019)

Future Directions: Theoretical Framework

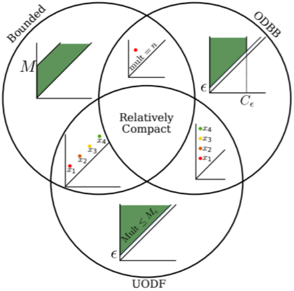


(Perea et al., 2019)

$$
\begin{array}{llllll}
\mathsf{H}_*(\mathbb{X}): & & \mathsf{H}_*(X_1) & \to \cdots \to \mathsf{H}_*(X_{n-1}) & \to \mathsf{H}_*(X_n) \\
\mathsf{H}^*(\mathbb{X}): & & \mathsf{H}^*(X_1) & \leftarrow \cdots \leftarrow \mathsf{H}^*(X_{n-1}) & \leftarrow \mathsf{H}^*(X_n) \\
\mathsf{H}_*(X_\infty, \mathbb{X}): & \mathsf{H}_*(X_n) \to & \mathsf{H}_*(X_n, X_1) \to & \cdots \to \mathsf{H}_*(X_n, X_{n-1}) & \\
\mathsf{H}^*(X_\infty, \mathbb{X}): & \mathsf{H}^*(X_n) \leftarrow & \mathsf{H}^*(X_n, X_1) \leftarrow & \cdots \leftarrow \mathsf{H}^*(X_n, X_{n-1}). &
\end{array}
$$

(Silva et al., 2011)

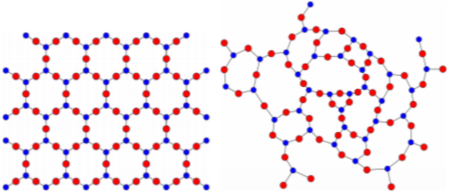
Future Directions: Theoretical Framework



(Perea et al., 2019)

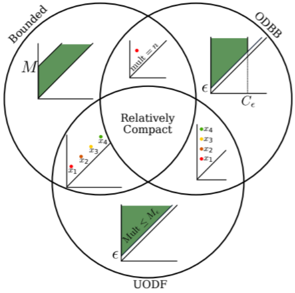
$$
\begin{array}{lllll}
\mathsf{H}_*(\mathbb{X}): & \mathsf{H}_*(X_1) & \to \cdots \to \mathsf{H}_*(X_{n-1}) & \to \mathsf{H}_*(X_n) \\
\mathsf{H}^*(\mathbb{X}): & \mathsf{H}^*(X_1) & \leftarrow \cdots \leftarrow \mathsf{H}^*(X_{n-1}) & \leftarrow \mathsf{H}^*(X_n) \\
\mathsf{H}_*(X_\infty, \mathbb{X}): & \mathsf{H}_*(X_n) \to \mathsf{H}_*(X_n, X_1) \to & \cdots \to \mathsf{H}_*(X_n, X_{n-1}) & \\
\mathsf{H}^*(X_\infty, \mathbb{X}): & \mathsf{H}^*(X_n) \leftarrow \mathsf{H}^*(X_n, X_1) \leftarrow & \cdots \leftarrow \mathsf{H}^*(X_n, X_{n-1}). &
\end{array}
$$

(Silva et al., 2011)



(Schweinhart et al., 2019)

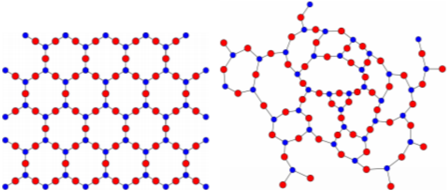
Future Directions: Theoretical Framework



(Perea et al., 2019)

$$
\begin{array}{ll}
\mathsf{H}_*(\mathbb{X}): & \quad \mathsf{H}_*(X_1) \to \cdots \to \mathsf{H}_*(X_{n-1}) \to \mathsf{H}_*(X_n) \\
\mathsf{H}^*(\mathbb{X}): & \quad \mathsf{H}^*(X_1) \leftarrow \cdots \leftarrow \mathsf{H}^*(X_{n-1}) \leftarrow \mathsf{H}^*(X_n) \\
\mathsf{H}_*(X_\infty, \mathbb{X}): & \mathsf{H}_*(X_n) \to \mathsf{H}_*(X_n, X_1) \to \cdots \to \mathsf{H}_*(X_n, X_{n-1}) \\
\mathsf{H}^*(X_\infty, \mathbb{X}): & \mathsf{H}^*(X_n) \leftarrow \mathsf{H}^*(X_n, X_1) \leftarrow \cdots \leftarrow \mathsf{H}^*(X_n, X_{n-1}).
\end{array}
$$

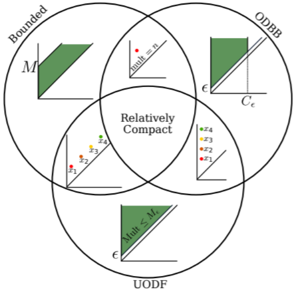
(Silva et al., 2011)



(Schweinhart et al., 2019)

(Moore and Vazquez, 2018)

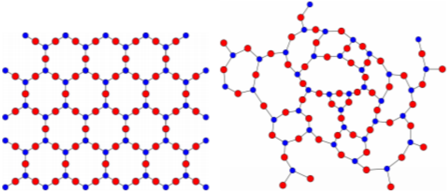
Future Directions: Theoretical Framework



(Perea et al., 2019)

$$
\begin{array}{llllll}
\mathsf{H}_*(\mathbb{X}): & & \mathsf{H}_*(X_1) & \to \cdots \to \mathsf{H}_*(X_{n-1}) & \to \mathsf{H}_*(X_n) \\
\mathsf{H}^*(\mathbb{X}): & & \mathsf{H}^*(X_1) & \leftarrow \cdots \leftarrow \mathsf{H}^*(X_{n-1}) & \leftarrow \mathsf{H}^*(X_n) \\
\mathsf{H}_*(X_\infty, \mathbb{X}): & \mathsf{H}_*(X_n) \to \mathsf{H}_*(X_n, X_1) \to & \cdots & \to \mathsf{H}_*(X_n, X_{n-1}) & \\
\mathsf{H}^*(X_\infty, \mathbb{X}): & \mathsf{H}^*(X_n) \leftarrow \mathsf{H}^*(X_n, X_1) \leftarrow & \cdots & \leftarrow \mathsf{H}^*(X_n, X_{n-1}). &
\end{array}
$$

(Silva et al., 2011)



(Schweinhart et al., 2019)

(Moore and Vazquez, 2018)

Applied Algebraic Topology Research Network

Future Directions: Statistics/Machine Learning

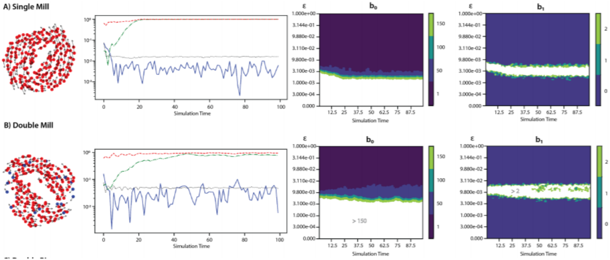
Learning Simplicial Complexes from Persistence Diagrams

Robin Lynne Belton* Brittany Terese Fasy*† Rostik Mertz† Samuel Micka† David L. Millman†
Daniel Salinas† Anna Schenfisch* Jordan Schupbach* Lucia Williams†

Future Directions: Statistics/Machine Learning

Learning Simplicial Complexes from Persistence Diagrams

Robin Lynne Belton[*] Brittany Terese Fasy[*,†] Rostik Mertz[†] Samuel Micka[†] David L. Millman[†]
Daniel Salinas[†] Anna Schenfisch[*] Jordan Schupbach[*] Lucia Williams[†]

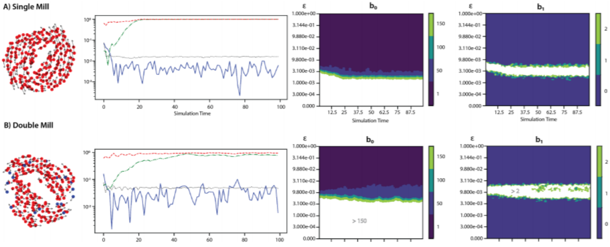


(Bhaskar et al., 2019)

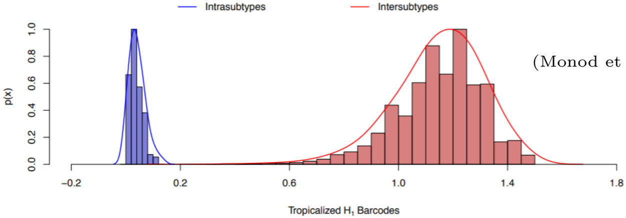
Future Directions: Statistics/Machine Learning

Learning Simplicial Complexes from Persistence Diagrams

Robin Lynne Belton* Brittany Terese Fasy*† Rostik Mertz† Samuel Micka† David L. Millman†
Daniel Salinas† Anna Schenfisch* Jordan Schupbach* Lucia Williams†



(Bhaskar et al., 2019)



(Monod et al., 2019)

Acknowledgements

Crawford Lab.

- Lorin Crawford, PhD. (PI)
- Pinar Dimetci
- Alan DenAdel
- Chibuikem (Chib) Nwizu
- Dana Udwin
- Gabrielle Ferra
- Isabella Ting

www.lcrawlab.com

Funding Support

- National Science Foundation Graduate Research Fellowship Program, Grant No. 1644760.
- Division of Applied Mathematics, Brown University

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1644760. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References I

H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *J. Mach. Learn. Res.*, 18(1):218–252, Jan. 2017. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=3122009.3122017`.

D. Bhaskar, A. Manhart, J. Milzman, J. T. Nardini, K. Storey, C. M. Topaz, and L. Ziegelmeier. Analyzing collective motion with machine learning and topology, 2019.

References II

P. Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(1):77–102, Jan 2015. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2789272.2789275`.

P. Bubenik and P. Dłotko. A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation*, 78, 12 2014. doi: 10.1016/j.jsc.2016.03.009.

J. M. Chan, G. Carlsson, and R. Rabadan. Topology of viral evolution. *Proceedings of the National Academy of Sciences*, 110:18566–18571, Nov 2013. doi: 10.1073/pnas.1313480110.

References III

F. Chazal and B. Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *ArXiv*, abs/1710.04019, 2017.

D. R. Chittajallu, N. Siekierski, S. Lee, S. Gerber, J. D. Beezley, D. Manthey, D. A. Gutman, and L. A. D. Cooper. Vectorized persistent homology representations for characterizing glandular architecture in histology images. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 232–235, 2018.

T. K. Dey and C. Xin. Computing bottleneck distance for multi-parameter interval decomposable persistence modules. *arXiv*, 2019. URL `https://arxiv.org/pdf/1803.02869.pdf`.

References IV

V. Kovacev-Nikolic, P. Bubenik, D. Nikolic, and G. Heo. Using persistent homology and dynamical distances to analyze protein binding. *Statistical applications in genetics and molecular biology*, 15:19–38, 03 2016. doi: 10.1515/sagmb-2015-0057.

A. Monod, S. Kalisnik Verovsek, J. Patiño Galindo, and L. Crawford. Tropical sufficient statistics for persistent homology. *SIAM Journal on Applied Algebra and Geometry*, 3:337–371, 01 2019. doi: 10.1137/17M1148037.

A. Moore and M. Vazquez. Recent advances on the non-coherent band surgery model for site-specific recombination. 10 2018. URL `https://arxiv.org/pdf/1810.08751.pdf`.

References V

J. A. Perea, E. Munch, and F. A. Khasawneh. Approximating continuous functions on persistence diagrams using template functions. *CoRR*, abs/1902.07190, 2019. URL `http://arxiv.org/abs/1902.07190`.

H. Sagerman. Topology joke, 2015. URL `https://www.youtube.com/watch?v=9NlqYr6-TpA`.

B. Schweinhart, D. Rodney, and J. Mason. Statistical topology of bond networks with applications to silica. *arxiv*, 10 2019. URL `https://people.math.osu.edu/schweinhart.2/TopologyBondNetworks.pdf`.

References VI

V. Silva, D. Morozov, and M. Vejdemo-Johansson. Dualities in persistent (co)homology. *Inverse Problems - INVERSE PROBL*, 27, 07 2011. doi: 10.1088/0266-5611/27/12/124003.

K. Turner, S. Mukherjee, and D. Boyer. Persistent homology transform for modeling shapes and surfaces. *Information and Inference*, 3:310–344, 01 2014. doi: 10.1093/imaiai/iau011.