Statistical Shape Analysis using Topological Data Analysis

Part II: Representation and Modeling

Chul Moon

Southern Methodist University
chulm@smu.edu
https://github.com/chulmoon/TDA-lecture

ECSSC 2021
25 July 2021

Outline

Representations of Persistence Diagrams
- Representation in Euclidean Space
- Representation in Functional Space
- Representation in Reproducing Hilbert Kernel Space

Statistical Significance

Hypothesis Testing

Various Representations of Persistence Diagrams

Although persistence diagrams include topological persistence signal information, it cannot be directly used as input in data analysis

- Euclidean space
  - Summary function (Adcock et al., 2016)
  - Binning (Bendich et al., 2016)
  - Persistence image (Adams et al., 2017)
- L^2-space
  - Persistence landscape (Bubenik, 2015)
  - Persistence intensity function (Chen et al., 2015)
- Reproducing kernel Hilbert space
  - Persistence scale-space kernel (Reininghaus et al., 2015)
  - Persistence weighted Gaussian kernel (Kusano et al., 2016)

Representation using Summary Functions

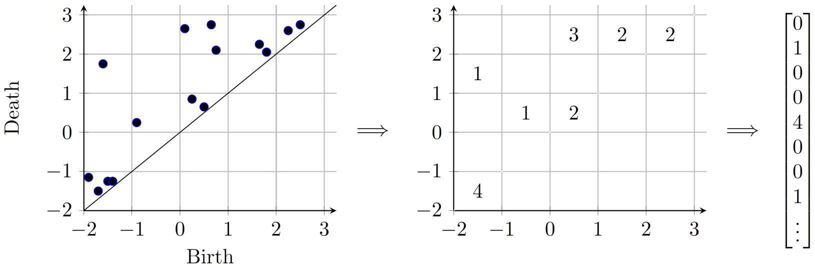
- Proposed by Adcock et al. (2016)
- Summarize birth, death, and persistence (death−birth)
  - Mean, median
  - Min, max
  - The n largest values
- Polynomials

$$
\begin{aligned}
&\sum_i x_i(y_i - x_i), \\
&\sum_i (y_{max} - y_i)(y_i - x_i), \\
&\sum_i x_i^2(y_i - x_i)^4, \\
&\sum_i (y_{max} - y_i)^2(y_i - x_i)^4,
\end{aligned}
$$

- Easy to compute
- Difficult to interpret

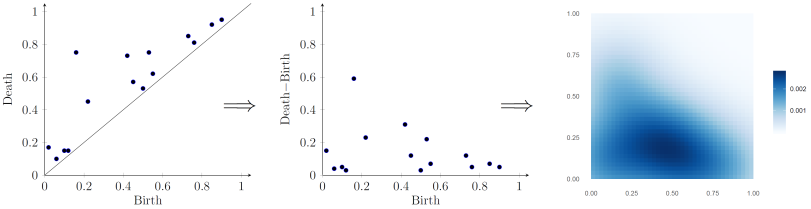
Vectorization by Binning

- Proposed by Bendich et al. (2016)
- Bin the persistence diagrams and count the number of points in each bin
- The representation by binning may not be stable



Persistence Image

- Proposed by Adams et al. (2017)
- Steps for generating persistence images
  1. Assigning weights to points
  2. Smoothing
  3. Converting to vector



Persistence Image

- Persistence diagram P is represented as persistence surface ρ_P
$$\rho_P(x, y) = \sum_{(b,d)\in P} g_{(b,d)}(x, y) \cdot w(b, d),$$
where x and y are the (x, y)-coordinates of the persistence function, $g_{(b,d)}$ is a smoothing function for $(b, d) \in P$, and $w(b, d) \geq 0$ is a non-negative weight function
- The surface is discretized to a vector

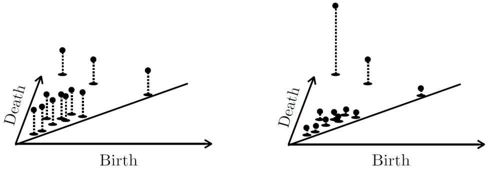
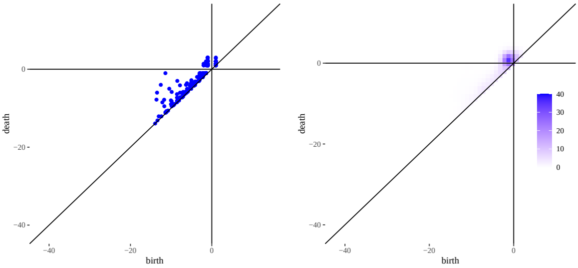


Figure: Unweighted vs. weighted persistence diagrams. Figure of Kusano et al. (2017).

Persistence Image Example

- Persistence image representation using the Gaussian smoothing function $g_{(b,d)}(x,y) = \frac{\exp\left[-\left((x-b)^2+(y-d)^2\right)\right]}{\sigma^2}$ with $\sigma = 1$ and the linear weight $w(b,d) = d - b$



Properties of Persistence Image

- Persistence image is a stable representation of persistence diagram with respect to 1-Wasserstein distance (Adams et al., 2017)

$$\|\rho_B - \rho_{B'}\|_\infty \leq \sqrt{10}\big(\|f\|_\infty |\nabla \phi| + \|\phi\|_\infty |\nabla f|\big) W_1(B, B').$$

- The persistence image can be used as an input of statistical models or ML algorithms
- Due to the high-dimensionality of persistence image, dimension reduction or feature selection is often required

Outline

Representations of Persistence Diagrams

Representation in Euclidean Space

Representation in Functional Space

Representation in Reproducing Hilbert Kernel Space

Statistical Significance

Hypothesis Testing

Idea of Persistence Landscape

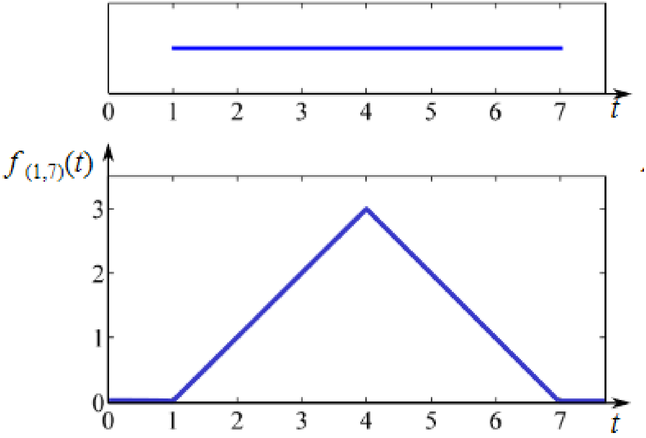


Figure: Figure from Kovacev-Nikolic et al. (2016)

Persistence Diagram

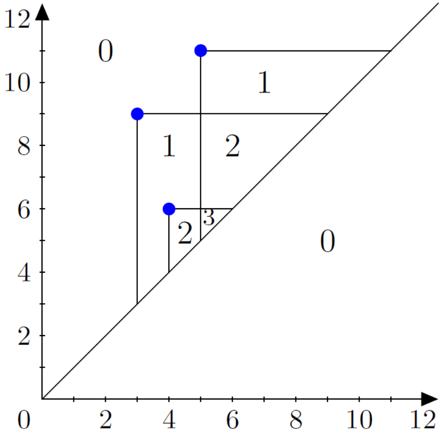


Figure: Figure from Bubenik (2015)

Persistence Landscape

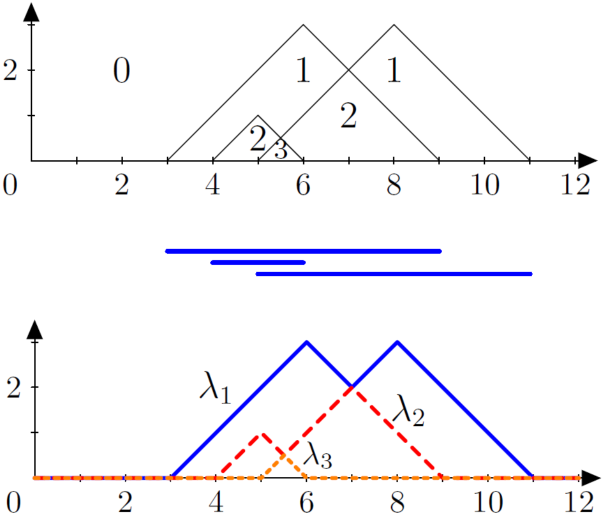


Figure: Figure from Bubenik (2015)

Persistence Landscape

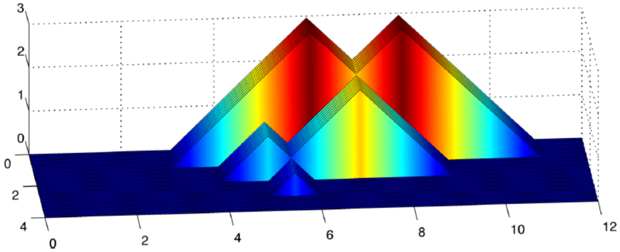


Figure: Figure from Bubenik (2015)

Persistence Landscape Example

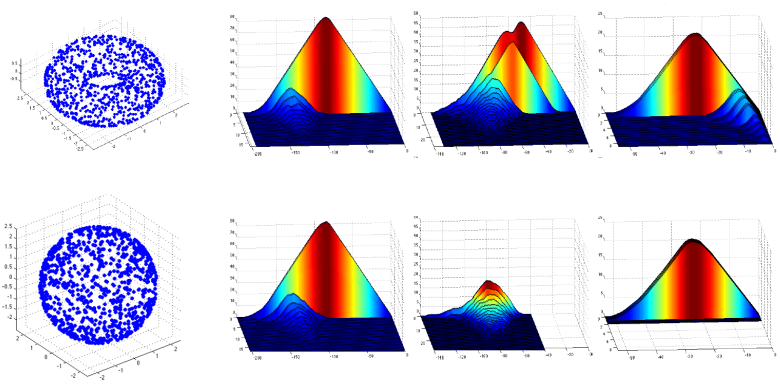


Figure: Figure from Bubenik (2015)

Properties of Persistence Landscape

- Persistence landscape is stable

Theorem 5.1 (∞-Landscape Stability Theorem). *Let $f, g : X \to \mathbb{R}$. Then*

$$\Lambda_\infty(M(f), M(g)) \leq \|f - g\|_\infty.$$

Theorem 5.5 (p-Landscape stability theorem). *Let X be a triangulable, compact metric space that implies bounded degree-k total persistence for some real number $k \geq 1$, and let f and g be two tame Lipschitz functions. Then*

$$\Lambda_p(D(f), D(g))^p \leq C\|f - g\|_\infty^{p-k},$$

for all $p \geq k$, where $C = C_X \max\{\mathrm{Lip}(f)^k, \mathrm{Lip}(g)^k, \mathrm{Lip}(f)^{k+1}, \mathrm{Lip}(g)^{k+1}\}(W_\infty(D, \emptyset) + \frac{1}{p+1})$.

Properties of Persistence Landscape

- Mean persistence landscape exists: pointwise average
- SLLN and CLT holds for persistence landscape

mean landscape $\overline{\lambda(X)}_n$ is given by the pointwise mean.

$$\overline{\lambda(X)}_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \lambda(X_i)(x, y)$$

Theorem 3.4 (Strong Law of Large Numbers for persistence landscapes). $\overline{\lambda(X)}_n \to E(\lambda(X))$ *almost surely if and only if* $E\|\lambda(X)\| < \infty$.

Theorem 3.5 (Central Limit Theorem for peristence landscapes). *Assume* $\lambda(X) \in L^p(\mathcal{S})$ *with* $2 \le p < \infty$. *If* $E\|\lambda(X)\| < \infty$ *and* $E(\|\lambda(X)\|^2) < \infty$ *then* $\sqrt{n}[\overline{\lambda(X)}_n - E(\lambda(X))]$ *converges weakly to a Gaussian random variable with the same covariance structure as* $\lambda(X)$.

Outline

Representations of Persistence Diagrams
Representation in Euclidean Space
Representation in Functional Space
Representation in Reproducing Hilbert Kernel Space

Statistical Significance

Hypothesis Testing

Representation in Reproducing Kernel Hilbert Space

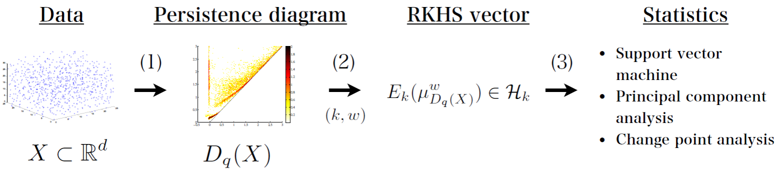


Figure: Figure from Kusano (2018)

- Persistence scale-space kernel (Reininghaus et al., 2015)
- Persistence weighted Gaussian kernel (PWGK) (Kusano, 2018)

Kernel Trick

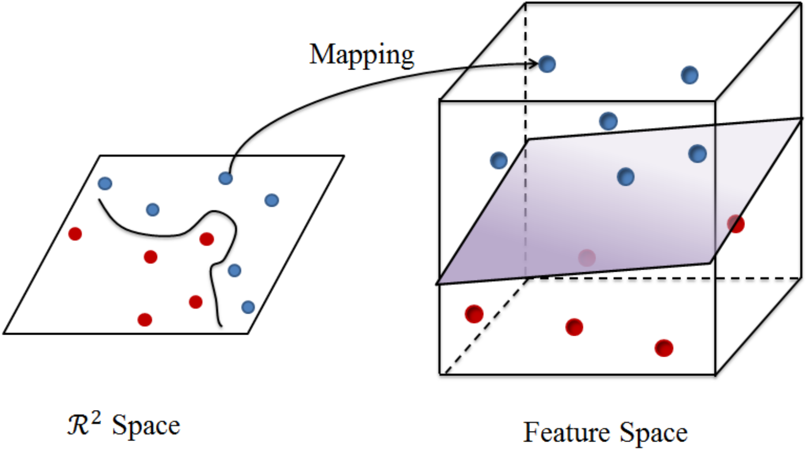


Figure: Figure from
`http://songcy.net/posts/story-of-basis-and-kernel-part-2/`

Kernel Trick

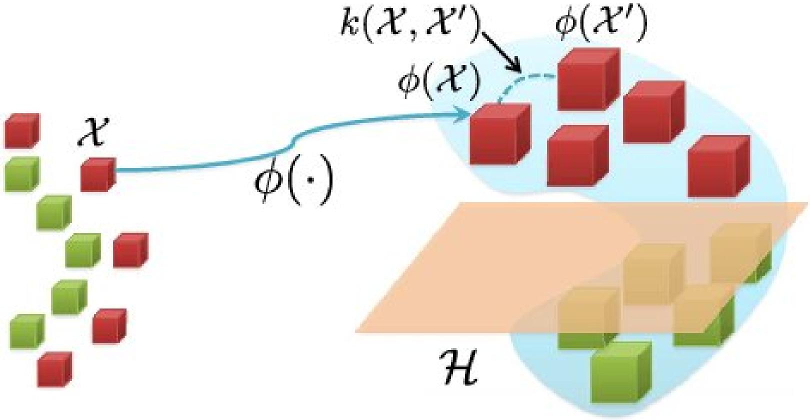


Figure: Figure from Zhao et al. (2013)

Kernel Methods

- Support vector machine
- Kernel ridge regression
- Kernel principal component analysis
- Gaussian process

Outline

Representations of Persistence Diagrams
Representation in Euclidean Space
Representation in Functional Space
Representation in Reproducing Hilbert Kernel Space

Statistical Significance

Hypothesis Testing

Motivation

- How can we identify a signal from persistence homology results?

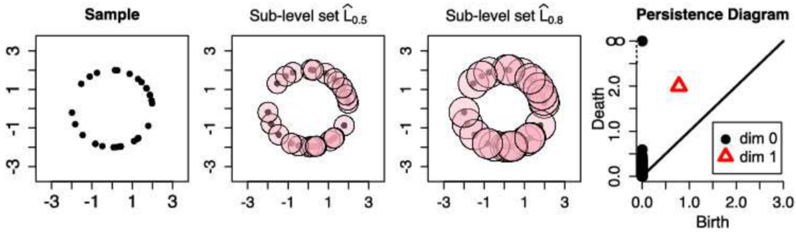


Figure: Figure from Fasy et al. (2014)

Confidence Sets for Persistence Diagram

- Proposed by Fasy et al. (2014)

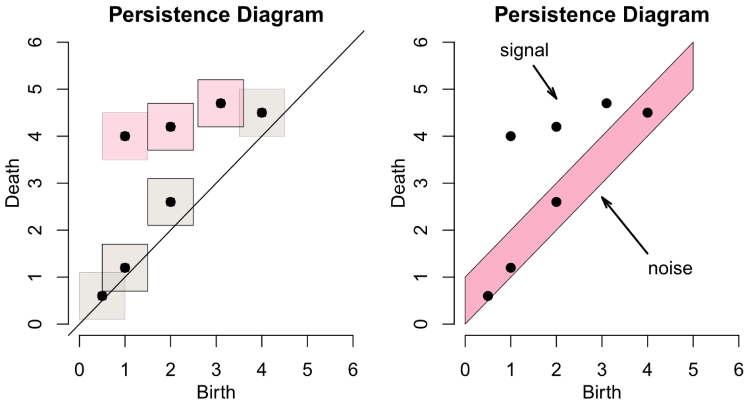


Figure: Figure from Fasy et al. (2014)

Confidence Set Example

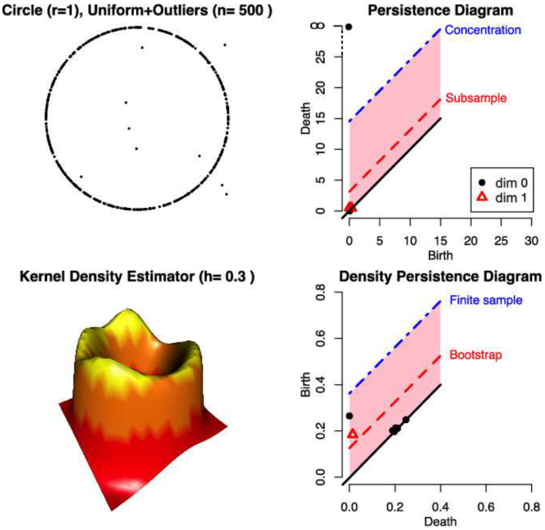


Figure: Figure from Fasy et al. (2014)

Bootstrap Band for Persistence Landscape

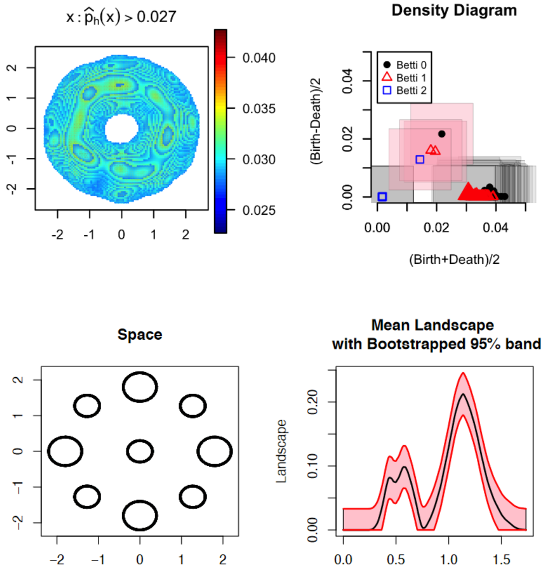
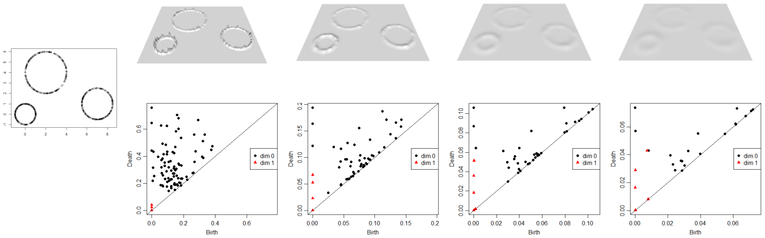


Figure: Figure from Chazal et al. (2013)

Significance for Point Cloud Data?

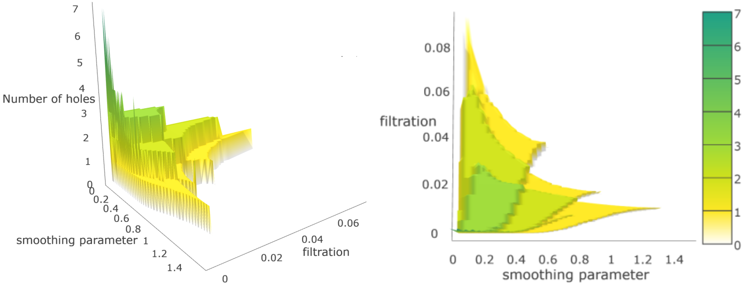
	Direct Estimation	Density Estimator
Pros	Fast and simple	Robust
Cons	Sensitive	Smoothing parameter selection
Significance	Size	Point density

- Use a range of smoothing parameters instead of a single smoothing parameter



Persistence Terrace

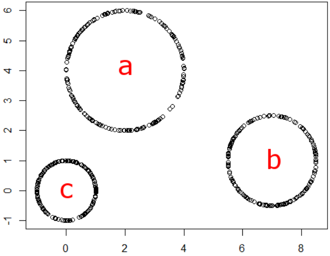
- Persistence terrace is a 3D summary plot where features are represented as terrace layers (Moon et al., 2018)

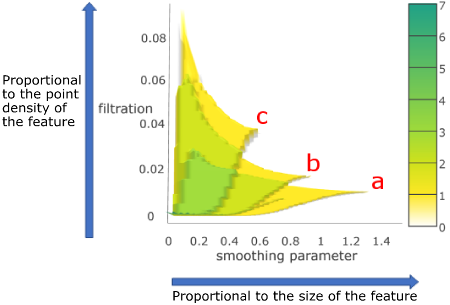


(a) Overall view (b) Satellite view

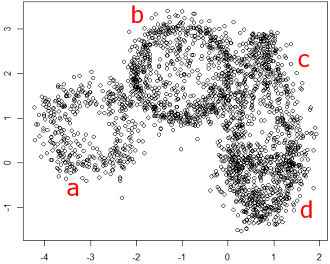
Interpretation of Persistence Terrace

- Robust inference of topological features while capturing both size point density information

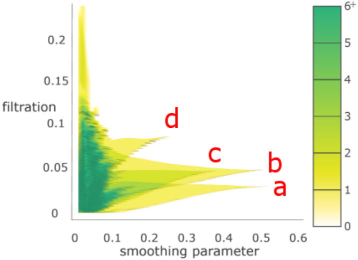




Persistence Terrace Example



(a) Scatterplot of noisy loops



(b) Dimension 1 persistence terrace

- Persistence terrace identifies four noisy loops
- The two large loops (square a and circle b) with different point density
- One slightly smaller loop (triangle c) and one smaller and denser loop (triangle d)

Additional Notes

- These approaches are based on the idea that the longer-surviving topological features are a signal and shorter-surviving topological features are noise
- However, short-surviving topological features may contain valuable information

Outline

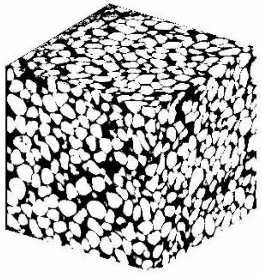
Representations of Persistence Diagrams
Representation in Euclidean Space
Representation in Functional Space
Representation in Reproducing Hilbert Kernel Space

Statistical Significance

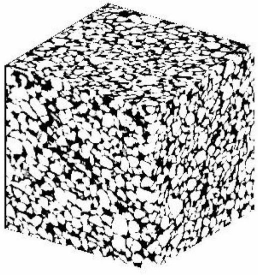
Hypothesis Testing

Motivating Example: Rock Comparison

- Data: three-dimensional rock images
- Scanned by a focus ion beam scanning electron microscopes



(a) Unground silica

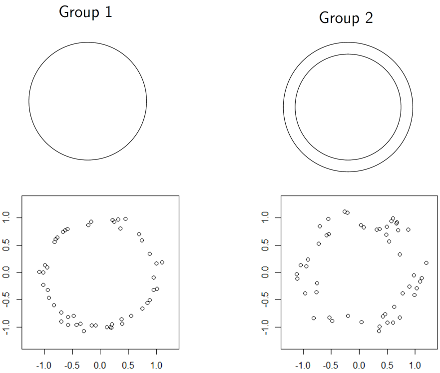


(b) Leavenseat sand

Figure: Figures of Talabi et al. (2009)

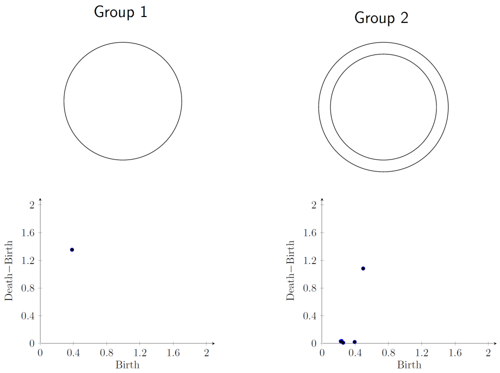
- We can summarize structure/connectivity information of rocks using topological data analysis
  - How can we compare based on topological characteristics?
  - Which topological features differ the most?

Toy Example



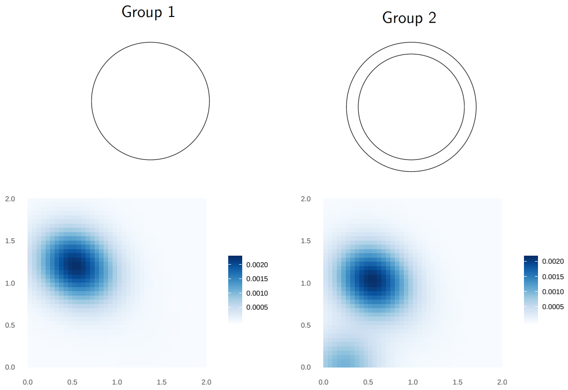
- How can we compare based on topological characteristics?
- Which topological features differ the most?

Toy Example



- How can we compare based on topological characteristics?
- Which topological features differ the most?

Toy Example



- How can we compare based on topological characteristics?
- Which topological features differ the most?

Permutation-based Hypothesis Test

Permutation test using Bottleneck/Wasserstein distances

- Two labels by Robinson and Turner (2017)
- Multiple labels using ANOVA by Cericola et al. (2018)
- Implementing multiple testing by Vejdemo-Johansson and Mukherjee (2018)

Algorithm S1 Permutation test (Robinson and Turner, 2017)

Input: Persistence diagrams, given group label $G_{\text{unshuffled}}$, number of repetitions N, joint loss function L

Output: Permutation p-value Z

Compute $L(G_{\text{unshuffled}})$

Create an empty vector L of size N

for $i = 1 \to N$ **do**

  Randomly generate shuffled group label G^i_{shuffled}

  $L[i] \leftarrow L(G^i_{\text{shuffled}})$

end for

$Z \leftarrow \text{sum}(L < L(G_{\text{unshuffled}}))/N$

Kernel Test

Kernel test using maximum mean discrepancy

- Kernel two-sample test of Gretton et al. (2006) is applied to TDA literature by Kusano (2018)

Definition 2 *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ and let p, q, x, y, X, Y be defined as above. We define the maximum mean discrepancy (MMD) as*

$$\mathrm{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} \left(\mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)]\right). \tag{1}$$

In the statistics literature, this is known as an integral probability metric (Müller, 1997). A biased[2] empirical estimate of the MMD is obtained by replacing the population expectations with empirical expectations computed on the samples X and Y,

$$\mathrm{MMD}_b[\mathcal{F}, X, Y] := \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(y_i) \right). \tag{2}$$

Some Limitations

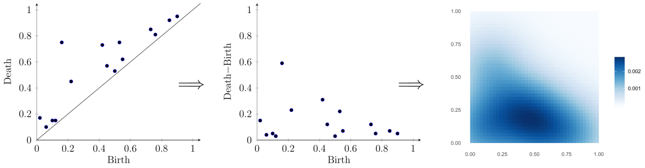
Permutation test using Bottleneck/Wasserstein distances

- High computation cost
- No information on which topological features contribute how much to the differences

Kernel test using maximum mean discrepancy

- No information on which topological features contribute how much to the differences

Two-stage Test using Persistence Image



Two-stage hypothesis test by Moon and Lazar (2020)

1. Filtering
   - Filter statistic: T^I
   - Filter pixels in the sparse region
   - Remove the pixel i if $\bar{X}^i < C^{th}$ percentile of $\bar{X}'s$
2. Testing
   - Test statistics: T^{II}
   - Control the false discovery rate using the q-value procedure by Storey (2002)

Hypothesis Test Example

- Generate two-dimensional pseudo-rock images using algorithm of Obayashi et al. (2018) using three sets of parameters: (M=180, S=80), (M=190, S=75), (M=200, S=70)
- 50 images for each parameter set

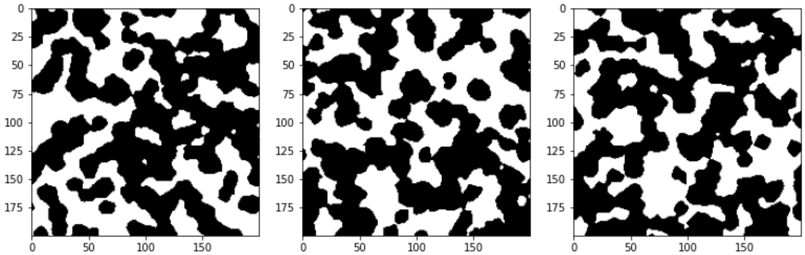


Figure: Examples of two-dimensional pseudo-rock images with parameters $(M = 180, S = 80)$ (left), $(M = 190, S = 75)$ (center), and $(M = 200, S = 70)$ (right).

Hypothesis Test Example

- Scenario 1 examines two groups of $(M = 180, S = 80)$ images
- Scenario 2 compares $(M = 180, S = 80)$ and $(M = 190, S = 75)$ groups
- Scenario 3 tests $(M = 180, S = 80)$ and $(M = 200, S = 70)$ groups

	Scenario 1		Scenario 2		Scenario 3	
	Dim 0	Dim 1	Dim 0	Dim 1	Dim 0	Dim 1
Two-stage test	0.763	0.798	0.121	0.792	0.003	0.019
Permutation test	0.520	0.700	0.090	0.475	0.000	0.000

Table: Minimum q-values of two-stage tests and p-values of permutation tests.

References I

Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L. (2017), "Persistence Images: A Stable Vector Representation of Persistent Homology," *Journal of Machine Learning Research*, 18, 1–35.

Adcock, A., Carlsson, E., and Carlsson, G. (2016), "The ring of algebraic functions on persistence bar codes," *Homology, Homotopy and Applications*, 18, 381–402.

Bendich, P., Chin, S. P., Clark, J., Desena, J., Harer, J., Munch, E., Newman, A., Porter, D., Rouse, D., Strawn, N., and Watkins, A. (2016), "Topological and statistical behavior classifiers for tracking applications," *IEEE Transactions on Aerospace and Electronic Systems*, 52, 2644–2661.

Bubenik, P. (2015), "Statistical Topological Data Analysis Using Persistence Landscapes," *Journal of Machine Learning Research*, 16, 77–102.

Cericola, C., Johnson, I. J., Kiers, J., Krock, M., Purdy, J., and Torrence, J. (2018), "Extending hypothesis testing with persistent homology to three or more groups," *Involve: A Journal of Mathematic*, 11, 27–51.

References II

Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., Singh, A., and Wasserman, L. (2013), "On the bootstrap for persistence diagrams and landscapes," *arXiv preprint arXiv:1311.0376*.

Chen, Y.-C., Wang, D., Rinaldo, A., and Wasserman, L. (2015), "Statistical analysis of persistence intensity functions," *arXiv preprint arXiv:1510.02502*.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014), "Confidence Sets for Persistence Diagrams," *Annals of Statistics*, 42, 2301–2339.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. J. (2006), "A Kernel Method for the Two-sample-problem," in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA: MIT Press, NIPS'06, pp. 513–520.

Kovacev-Nikolic, V., Bubenik, P., Nikolić, D., and Heo, G. (2016), "Using persistent homology and dynamical distances to analyze protein binding," *Statistical applications in genetics and molecular biology*, 15, 19–38.

Kusano, G. (2018), "On the Expectation of a Persistence Diagram by the Persistence Weighted Kernel," *arXiv e-prints*, arXiv:1803.08269.

References III

Kusano, G., Fukumizu, K., and Hiraoka, Y. (2017), "Kernel method for persistence diagrams via kernel embedding and weight factor," *The Journal of Machine Learning Research*, 18, 6947–6987.

Kusano, G., Hiraoka, Y., and Fukumizu, K. (2016), "Persistence weighted Gaussian kernel for topological data analysis," in *ICML*.

Moon, C., Giansiracusa, N., and Lazar, N. A. (2018), "Persistence terrace for topological inference of point cloud data," *Journal of Computational and Graphical Statistics*, 27, 576–586.

Moon, C. and Lazar, N. A. (2020), "Hypothesis Testing for Shapes using Vectorized Persistence Diagrams," *arXiv preprint arXiv:2006.05466*.

Obayashi, I., Hiraoka, Y., and Kimura, M. (2018), "Persistence diagrams with linear machine learning models," *Journal of Applied and Computational Topology*, 1, 421–449.

Reininghaus, J., Huber, S. M., Bauer, U., and Kwitt, R. (2015), "A stable multi-scale kernel for topological machine learning," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4741–4748.

References IV

Robinson, A. and Turner, K. (2017), “Hypothesis testing for topological data analysis,” *Journal of Applied and Computational Topology*, 1, 241–261.

Storey, J. D. (2002), “A direct approach to false discovery rates,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 479–498.

Talabi, O., AlSayari, S., Iglauer, S., and Blunt, M. J. (2009), “Pore-scale simulation of NMR response,” *Journal of Petroleum Science and Engineering*, 67, 168 – 178.

Vejdemo-Johansson, M. and Mukherjee, S. (2018), “Multiple testing with persistent homology,” *arXiv e-prints*, arXiv:1803.08269.

Zhao, Q., Zhou, G., Adali, T., Zhang, L., and Cichocki, A. (2013), “Kernelization of tensor-based models for multiway data analysis: Processing of multidimensional structured data,” *IEEE Signal Processing Magazine*, 30, 137–148.