Persistent Homology: Distance Between Data Sets

Kiley Junker

Creighton University: Honors Day 2022

Department of Mathematics: Dr. Nathan Pennington

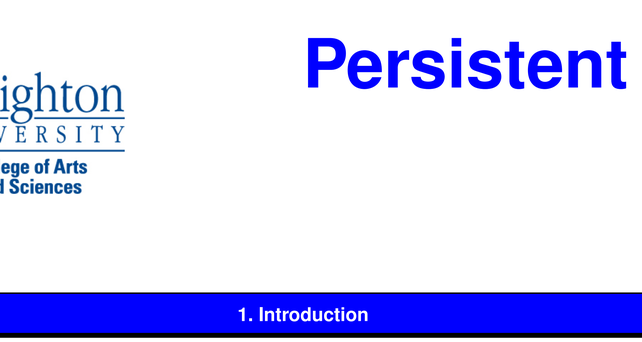
1. Introduction

- Tools to compare data sets are essential in this world that is increasingly dependent on data. Persistent homology is a topological tool used to examine the underlying structure of large data sets, filtering out noise. Practically, persistent homology takes in a data set and returns a persistence diagram, which is a collection of ordered pairs that represent the homological structure of the data across sequential filtration levels.
- Inspired by the work in [2], the goal of this research project is to define a notion of distance between different persistence diagrams, as this then allows us to quantify the distance between the original data sets. Specifically, we will construct a variety of candidate "distances" and then test them to determine which is optimal.

2. Introduction to Persistent Homology

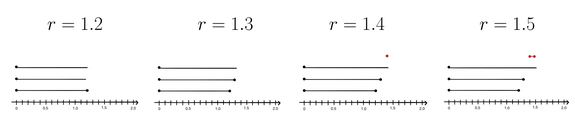
- Persistent homology is an analytical tool orginating from the field of topology that allows for the comparison of many variables. The process is as follows:

1. All observations from a data set are plotted in the appropriate Euclidean space.
2. For increasing values of $r \in \mathbb{R}_{>0}$, we draw a ball around each point.
3. When two balls overlap, the points they are centered around are connected by a line. This is called the birth of a component. Removing the balls yields a mathematical graph. The homology of that graph is called the filtration level of the data set for that value of r.



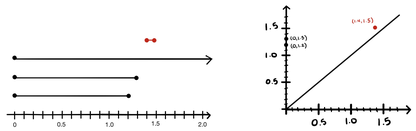
4. This process is repeated until there are no changes in the homology.

This process may be represented by a collection of barcodes, which record the birth and death of homological structures at increasing values of r.



3. Persistence Diagram

- These resulting barcodes can be represented as points in $\mathbb{R}^2$, using the birth and death of each component as the x and y values respectively:



- The black bars in the barcode represent the number of connected components in the graph, which is the zeroth homological level since they can be shrunk to a point. The red segment represents the only first-level homological structure present, which is the existence of a non-contractible loop [1]. This structure cannot be shrunk to a point, making it of more interest than those of the zeroth homological level.
- We will define a notion of "distance" between data sets by constructing a notion of "distance" between these persistence diagrams.

4. Inner Layer of Distance

- We define the length between two points in $\mathbb{R}^2$ by

$$l(x, y) := \max(|x_1 - y_1|, |x_2 - y_2|). \tag{1}$$

- Given two persistence diagrams P_1 and P_2, a *matching* μ is any assignment of each point in P_1 to a point in P_2 or the diagonal.
- We will not allow multiple points in P_1 to match with the same point in P_2, but multiple points in P_1 may be matched to the diagonal.

5. Best Match

- In order to find the best match for each point in P_1, we first define a function that finds the closest match between a point in P_1 and all points in P_2 or the diagonal. For $x \in P_1$, this is denoted $\mu(x)$.
- Let x^* be the point at which the set $\{l(x, \mu(x)) : x \in P_1\}$ reaches its minimum. Then we define $\eta(x^*) = \mu(x^*)$ and remove $(x^*, \mu(x^*))$ from $P_1 \times P_2$.
- This process repeats until all points in P_1 have been matched, which defines the minimal matching $\eta(x)$ for each $x \in P_1$.
- Finally, we may define the distance from P_1 to P_2 by

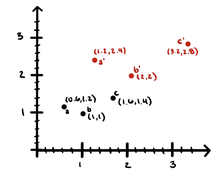
$$d(P_1, P_2) := \text{avg}_{x \in P_1}\, l(x, \eta(x)). \tag{2}$$

- Since this definition of distance is not commutative, the distance from P_1 to P_2 may be different than the distance from P_2 to P_1. So, we define the distance between P_1 and P_2 to be

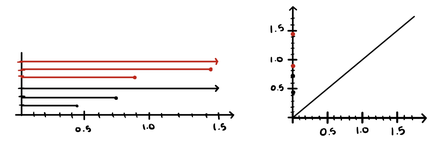
$$D(P_1, P_2) := \frac{d(P_1, P_2) + d(P_2, P_1)}{2}.$$

6. Comparing Adjusted Data Sets

- Once we defined the distance between two persistence diagrams, we investigated the distance between an adjusted data set and the original. First, we translated the data and compared it to the original set. As expected, the distance equation yielded zero.
- Next, we investigated the distance between a data set X and a scaled version of that data set X'. For example, if we scale the original data by two, we get



- We observe again that the distance between the points in X is different than the distance between the points in X'. Specifically, the euclidean distance between points $a', b' \in X'$ has also been scaled by the same value as the data points. This means that the barcode and persistence diagram will also be scaled by the same value:



- This seems to imply that there should be some sort of relationship between the scale and the calculated distance between the original and adjusted data. To investigate this, we ran trials where we scaled different data sets by the same amount and found the average distance between each data set and its scaled partner:

```
def scaleTest2(amount):
    ranScaleList = []
    for i in range(10000):
        x = np.random.random((10,2))
        d1 = scale(x,amount)
        ranScaleList.append(testFunction(x,d1))
    return summarizeList(ranScaleList)
scaleTest2(10)
```

```
min =  0.0
max =  0.9032038301229477
avg =  0.08134706804820649
```

- We conjectured that scaling by r would have a predicable effect on the distance equation. By running many trials, we found that, for a scale r,

$$D(P_1, rP_1) \approx \frac{0.04r}{5}.$$

7. Future Work

- We may investigate the effect of modifying equation $\{1\}$ to instead use the minimum or average, as opposed to the maximum.
- We may investigate what happens if we modify equation $\{2\}$ to instead take the maximum or minimum of $l(x, \eta(x))$, as opposed to the average.
- We may also wish to investigate whether or not to include outliers in the analysis of these data sets.

References

[1] M. A. Armstrong, *Basic topology*, Springer-Verlag Inc., New York, 1983.

[2] N. Yadav X. Zheng S.M. Han, T. Okonek, *Distributions of Matching Distances in Topological Data Analysis*, SIAM Undergraduate Research Online **13** (2020).