MATHEMATICS OF COMPUTATION
Volume 00, Number 0, Pages 000–000
S 0025-5718(XX)0000-0

PERSISTENT HOMOLOGY: HOW IT IS USED TO ANALYSE POINT CLOUD DATA SETS

MATT O'REILLY AND PAUL ARMSTRONG

Abstract. Topological data analysis (TDA) is a way to analyse datasets using topological techniques. Extracting information from datasets that are high-dimensional and noisy can be challenging. The main tool that is used in TDA is Persistent Homology, It is a method for computing topological features of a space at differential spatial resolutions. Simply put Persistent homology is an algebraic tool for measuring topological features of shapes and functions.

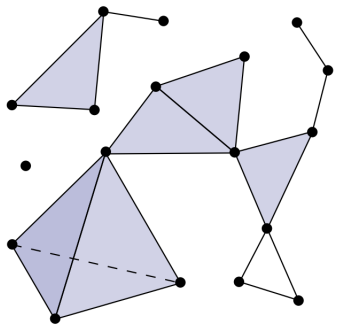
1. What is Homology?

Homology in general is a way to associate other parts of abstract algebra with topological spaces,this is typically done by using algebraic concepts such as open balls to describe the simplicial complexes (which we will define below) of a topological space. Homology is a sub section of Topology which slowly developed with the subject. Homology has been used as a means of categorising n dimensional manifolds alongside the euler characteristic. After Euler, Reimann defined genus and n fold connectedness as invariants of a topological space. Later Mark Goresky and Robert MacPherson would define persistent intersection homology (the basis of this paper) and prove it to be another topological invariant. [2]

2. Background

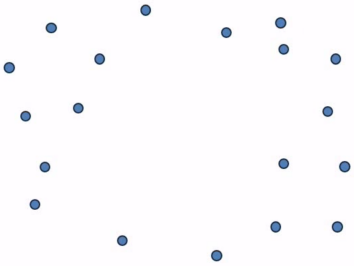
Persistent homology is an algebraic method for measuring topological features of shapes and functions. Persistent features of data are deemed more likely to represent true features of the underlying space than noise. To begin talking about topological features of a given data set we must first introduce the concept of a Simplicial Complex. A **simplicial complex** is the combination of a number of simplices.

2010 *Mathematics Subject Classification.* 55N33.



3. Topological Features

Example: What topological features does the following data exhibit?



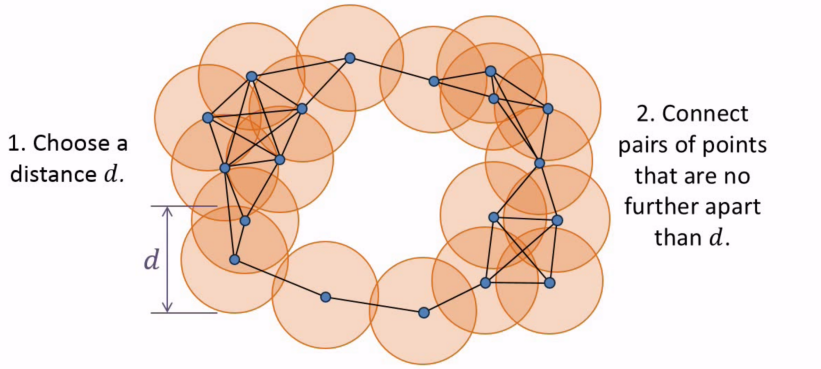
The data appears to be represented by a ring. However, this is just intuition. How can we verify that is a topological feature? Let X represent the set of data points.

3.1. **How do we turn our data points into shape?** Fix a real-valued distance, $d > 0$. Take the open ball of radius d around every point (x,y) in the data-set.

$$\forall\, x, y \in X,\ \exists\, d > 0\ B_d(x, y) = \{(a, b) : |(x, y) - (a, b)| < d\}$$

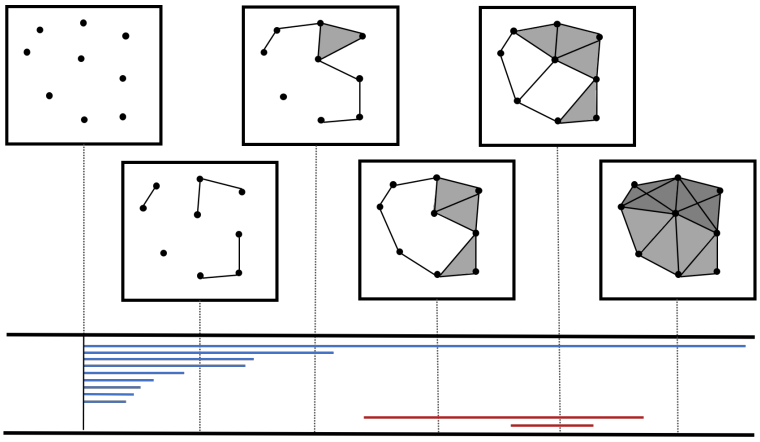
If n+1 circles intersect, draw a n-simplex between the n+1 points at the centers of the circles.

This gives a simplicial complex which roughly describes the shape of the data



3.2. **Choosing the right distance.** Our final hurdle is choosing the right d, If we choose a d too small then we might detect multiple connected components. This we can call noise. If we choose d too large then we get a giant simplex of all of the connected points. This has the trivial topology. For different values of d the topological structure will change.

So, What distance d should we choose? A solution is to consider all values of d. We let d vary over a range of values and count the components at each value of d, We record a barcode of the data which is seen below. [3]



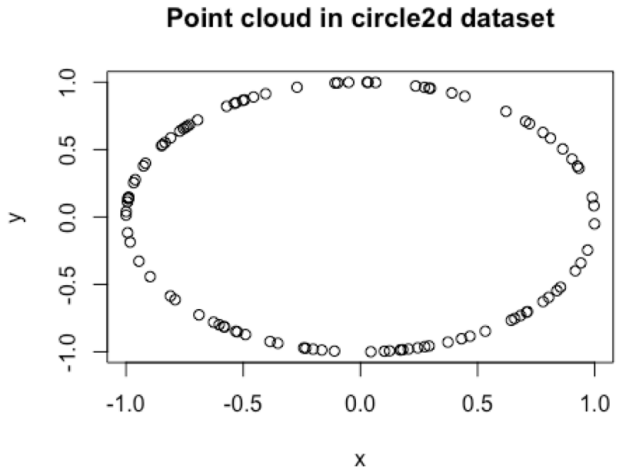
The small holes which we record are due to irregularities and noise, These are represented by the short bars in the barcode. The large bars represent significant features of the data

4. Topological Data Analysis using R Studio

```
library("TDAstats")

head(circle2d)
##              [,1]       [,2]
## [1,] -0.09728967  0.9952561
## [2,] -0.69421384  0.7197688
## [3,] -0.89704859 -0.4419319
## [4,]  0.83824069 -0.5453004
## [5,]  0.29894927  0.9542690
## [6,]  0.80302853 -0.5959406

plot(circle2d, xlab = "x", ylab = "y")
```



Calculating persistent homology

```
circle.phom <- calculate_homology(circle2d)
head(circle.phom)
##      dimension  birth        death
## [1,]         0      0  0.0007823978
## [2,]         0      0  0.0024476588
## [3,]         0      0  0.0043431663
## [4,]         0      0  0.0050472712
## [5,]         0      0  0.0055217817
## [6,]         0      0  0.0058158724
tail(circle.phom)
##       dimension      birth      death
##  [95,]         0  0.0000000  0.1655777
##  [96,]         0  0.0000000  0.1676424
##  [97,]         0  0.0000000  0.1724907
##  [98,]         0  0.0000000  0.1962725
##  [99,]         0  0.0000000  0.2085510
## [100,]         1  0.2232352  1.7339087
```

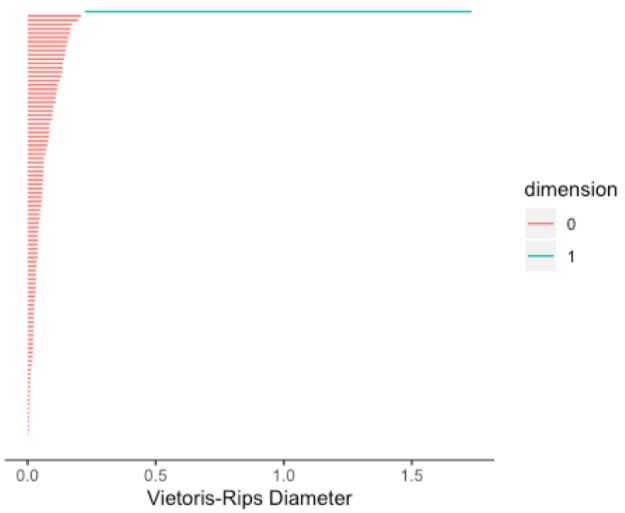
Each row in the homology matrix returned by the calculate homology function (variable named circle.phom) represents a single feature (cycle). The homology matrix has 3 columns in the following order:

1. Dimension: if 0, represents a 0-cycle; if 1, represents a 1-cycle; and so on.
2. Birth: the radius of the Vietoris-Rips complex at which this feature was first detected.
3. Death: the radius of the Vietoris-Rips complex at which this feature was last detected.

Persistence of a feature is generally defined as the length of the interval of the radius within which the feature exists. [1]

the vietoris rips complex is a special simplicial complex which for a defined metric and a value for distance e connects all vertices of distance less than e with an edge.

```
plot_barcode(circle.phom)
```



We can see a number of 0-cycles (The Red Bars) which represent noise in the data. The single 1-cycle at the top of the barcode (The Blue bar) is what is of interest to us. We can see that it clearly is a significant feature.

References

[1] Gunnar Carlsson. "Topology and Data". In: *AMS* (2009).
[2] Siddharth Venkatesh. "IntersectionHomology". In: *AMS* (2015).
[3] Matthew Wright. "Interactive Visualization of 2-D Persistence Modules". In: *AMS* (2015).