Persistent topology of the reionisation bubble network. I and II

Articles by Willem Elbers and Rien van de Weygaert
Writeup and presentation by Elie Cueto
Niels Bohr Institute - Advanced Methods in applied statistics
(Dated: 8th March 2023)

Persistent Homology is a topic within topology, which deals with the formation and destruction of holes in different numbers of dimensions. From an arbitrary dataset, one can thus construct persistence diagrams and fields, to reflect the "shape" of the dataset, and how it compares to other data. This tool can be useful to illustrate high dimensional datasets in two dimensions, as well as illustrating and quantifying how much different datasets differ, or exploring how well different models comparatively fit your data.

These articles deal with reionisation of the intergalactic medium during the so called Epoch of Reionisation. Different models for galaxy formation and size distribution are explored and compared, both with each other and with numerical models for the epoch of reionisation, and it shows that different models can be differentiated, even with very noisy data.

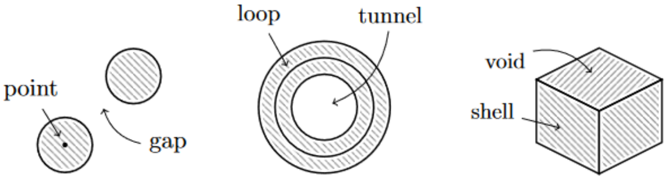

FIG. 1: Homology of shapes with holes in 0, 1, and 2 dimensions. Figure borrowed from [1]

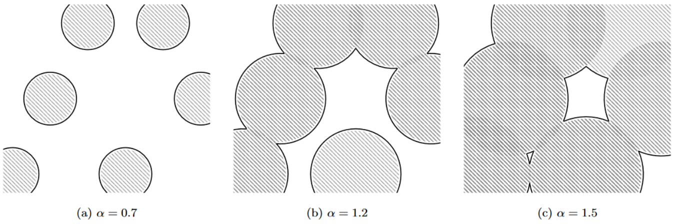
(a) $\alpha = 0.7$
(b) $\alpha = 1.2$
(c) $\alpha = 1.5$

FIG. 2: Illustration of how circles emanating from points merge to form different features. Figure borrowed from [1]

INTRODUCTION TO PERSISTENT HOMOLOGY

The study of persistent homology, in the context which will be relevant here, deals with characterising the number of different topological features in a dataset, at different scales.

By "topological features" we more precisely mean holes in different numbers of dimensions. The first three of these are illustrated on Figure 1 on page 1. A hole of n dimensions is one which can be enclosed by a loop of n dimensions. In 0, 1 and 2 dimensions these are points, loops and shells, respectively.

This gives that a zero dimensional hole is the boundary of a component, but may more intuitively be understood as the "gap" between two disjoint components. A one dimensional hole becomes what we usually refer to as a hole, a tunnel through a component, like the hole in a donut. A two dimensional hole becomes a hollow space enclosed on all sides by a component, like the hollow inside of a balloon. Higher dimensional holes can not easily be conceptualised, but can be computed.

To explain how this relates to the world of statistics, we can imagine having a number of data points in an arbitrary number of dimensions. For the sake of having to imagine it, 2 dimensions are recommended. Then, imagine a sphere - or circle in 2 dimensions - emanating from each point, becoming objects with a volume. As these circles grow in radius - which we will refer to as α, they will merge, forming different features, as illustrated on Figure 2 on page 1.

We can see that when α is small, the different points remain disjoint, keeping the number of $0-$d holes equal to the number of points, as seen on the left. As α grows larger, circles merge, "killing" some of the $0-$d holes seen in the middle panel. At some point the circles start enclosing empty spaces, thus "birthing" $1-$d holes, or tunnels, as seen on the right panel. When α grows yet larger, these spaces will be filled, killing the tunnels again.

This gives us the needed framework to start doing statistics.

PERSISTENCE DIAGRAMS

Given a set of data, we can find the number of holes in different dimensions as a function of α. By tracking when different features are born, and when they die, we can construct a persistence diagram. For 300 Monte Carlo simulated data for a Uniform and Normal distribution in 7 dimensions, such diagrams are shown on Figure 3 on page 2. The radius α at which a feature is born is noted on the x-axis, and the α at which it dies is noted on the y-axis. Note that all $0-$d holes are born at $\alpha = 0$, as all points enter the dataset simultaneously. The data would be very difficult to visualise in $7-d$, but we can clearly see the difference between the two distributions.

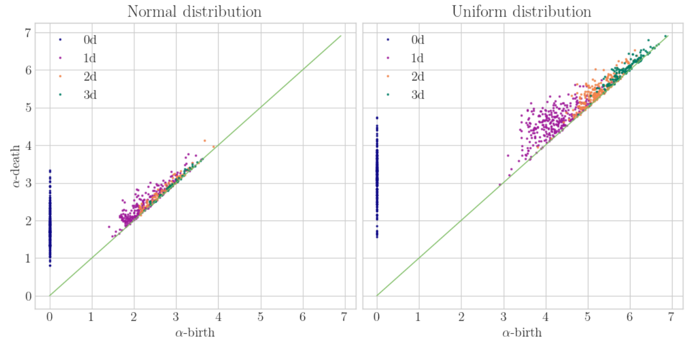

FIG. 3: Persistence diagrams constructed from data drawn from a Normal and Uniform distribution in 7 dimensions respectively.

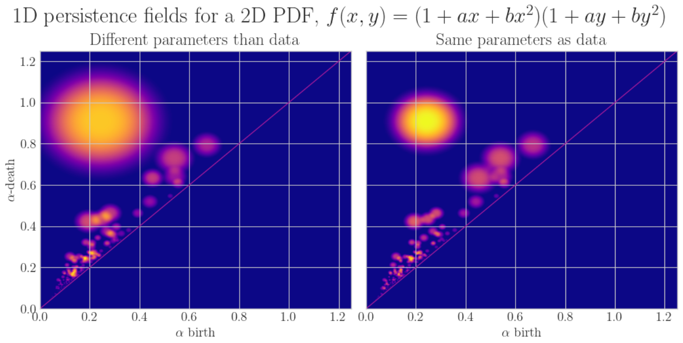

FIG. 4: Persistence fields. Initial dataset is generated from a known PDF, and two different test models are used

An important feature of these diagrams is the "psersistence" of different features. That is, how long does a particular feature last after it is born, which correspond to the vertical distance of a feature above the $x = y$ diagonal also drawn on the figures. This is the term that gives "persistent homology" its name, and can contain important information about the shape of ones data, though this will depend heavily on the data, and on what is being studied.

STATISTICAL METHODS

If we have a set of data, and we believe to have found a new model, which better describes this data than an existing model, we can compare how well the two different models fit the data by utilising persistence diagrams. First we construct persistence diagrams based on Monte Carlo simulations of the two different models, as well as our data. Then we calculate the "distance" between the diagrams. This distance can be found using the L^2-Wasserstein metric, in which, for two diagrams X and Y

$$d(X, Y) = \left[\inf_{\phi : X \to Y} \sum_{x \in X} ||x - \phi(x)||^2 \right]^{1/2} \tag{1.1}$$

where ϕ pairs a point in X to a point in Y, and the diagonal line $x = y$ is included as many times as is needed, such that a different number of points in X and Y is not an issue.

A smaller distance will mean that the two datasets are more similar. By doing a large number of monte carlo simulations for each model, we can construct the Fréchet Average as

$$F(Y) = \frac{1}{n} \sum_{i=1}^{n} d(Y, X_i)^2 \tag{1.2}$$

where Y is our original data, and $\{X_i\}$ is the set of Monte Carlo datasets for each model. A smaller Fréchet Average, will mean that a model is more likely to be consistent with a dataset. The "variance" on each point $y \in Y$ is described as

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{n} ||y - \phi_i(y)||^2 \tag{1.3}$$

where ϕ_i is an optimal matching with X_i. This allows for the construction of persistence fields, as shown on Figure 4 on page 2. The radius of a feature is proportional to its variance, and its "brightness" is proportional to the square root of its persistence. More persistent features reflect more significant, but less probable features in the shape of the data, and these are thus brighter.

CONCLUSIONS: USE IN THE STUDY OF REIONISATION

By looking at a semi-numerical model of galaxy formation and evolution, the topology of the cosmos during the era of reionisation can be studied. Simulating a "Bright Galaxy" and a "faint galaxy" model and constructing persistence diagrams, with ionised regions of space replacing the shperes discussed above, and not constaining all galaxies to form simultaneously, as is done above, Persistence fields consistent with each model is constructed [2]. It is shown that even when artificially introducing noise consistent with what might be expected for actual observational data, the two models can be effectively differentiated, suggesting a promising way to learn about the epoch of reionisation, though the lens of persistent homology.

[1] Willem Elbers, Rien van de Weygaert (2018). *Persistent topology of the reionisation bubble network. I: Formalism and Phenomenology.*

[2] Willem Elbers, Rien van de Weygaert (2022). *Persistent topology of the reionisation bubble network. I: Evolution and classification.*