Statistical aspects of persistent homology

Matthew Arnold

University of Bristol

2nd July 2012

Problem specification

Larry Wasserman, June 2012, UCL:

Geometric problem: find a manifold $\hat{M}$ which is close to an unknown manifold M.

Topological problem: find a manifold $\hat{M}$ which has the same homology as an unknown manifold M.

NB: We are statisticians, not pure mathematicians. So we are approaching the topological problem from a statistical point of view, using persistent homology.

Problem specification

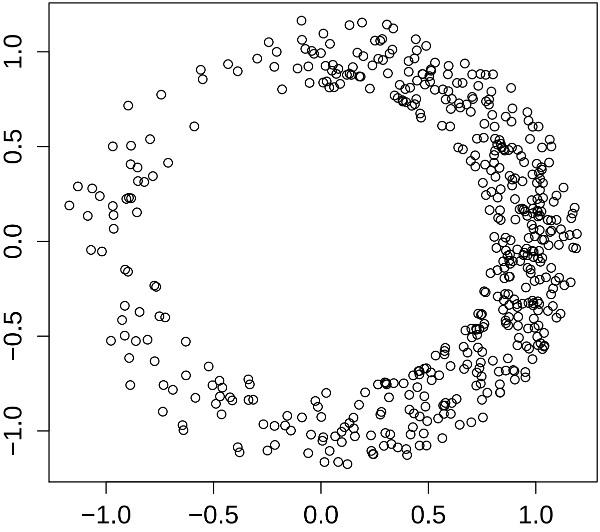
Larry Wasserman, June 2012, UCL:

Geometric problem: find a manifold $\hat{M}$ which is close to an unknown manifold M.

Topological problem: find a manifold $\hat{M}$ which has the same homology as an unknown manifold M.

NB: We are statisticians, not pure mathematicians. So we are approaching the topological problem from a statistical point of view, using persistent homology.

Example

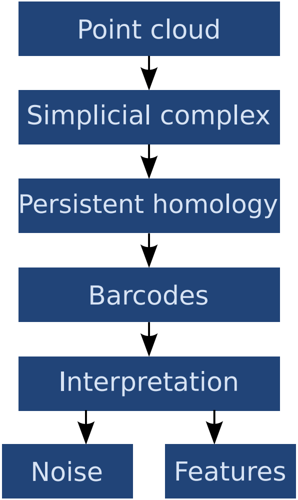


Problem

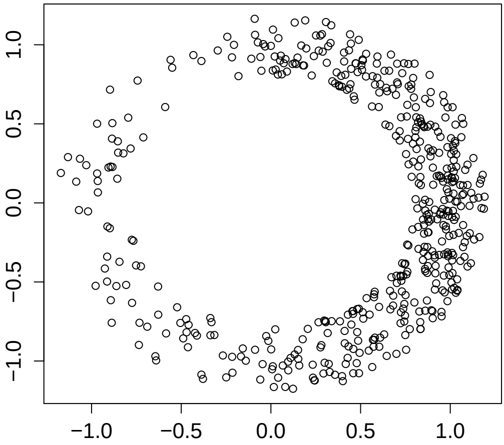
We sample points according to a distribution G on a manifold M, and call this sample X (point cloud).

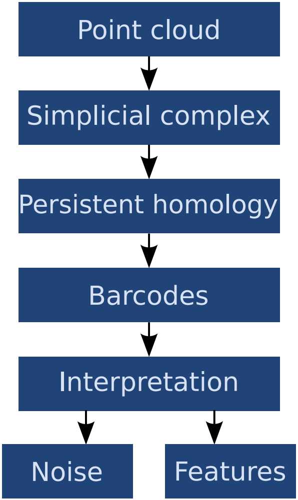
Once we have the barcodes, decide which persistence intervals are noise, and which represent true homology (let's call these features.)

Sometimes, this is easy. Just look for long persistence intervals.



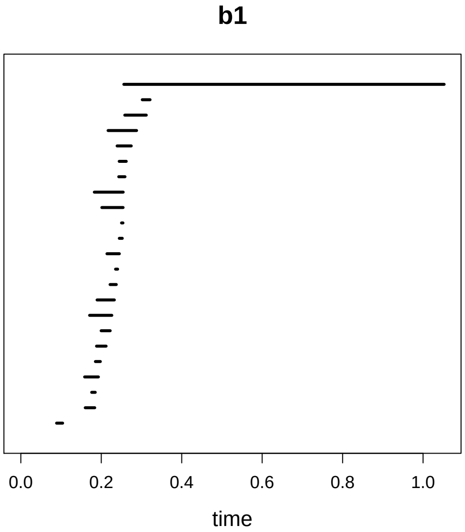
Problem

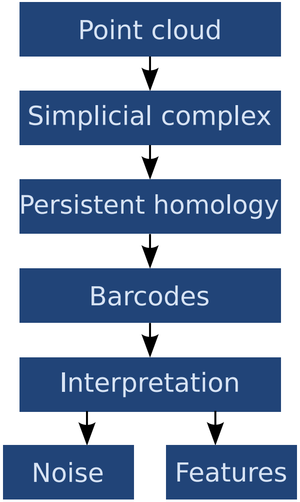




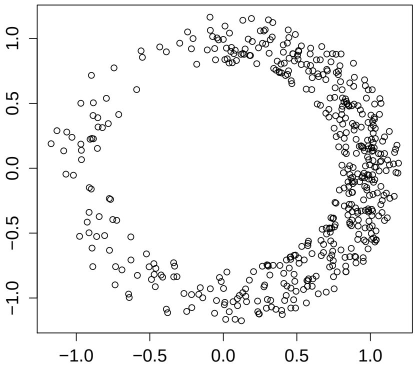
α-complex (Edelsbrunner and Harer, 2010)

Problem

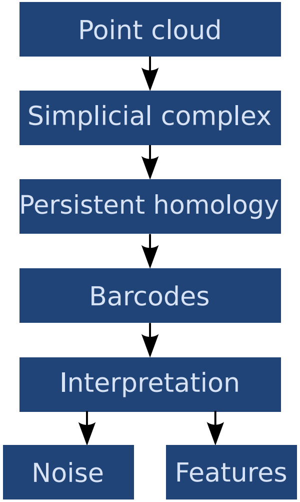

b1
0.0
0.2
0.4
0.6
0.8
1.0
time


Point cloud
Simplicial complex
Persistent homology
Barcodes
Interpretation
Noise
Features

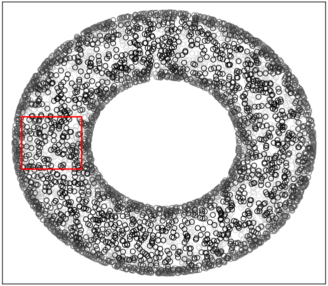
Outline

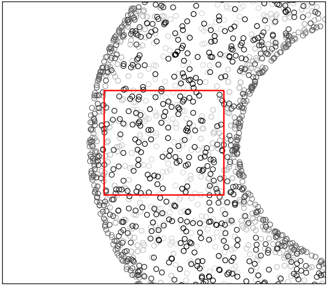


- Identify noise in barcodes.
- Identify noise in barcodes irrespective of G.



Discretisation noise





Discretisation noise $\perp$ topological features

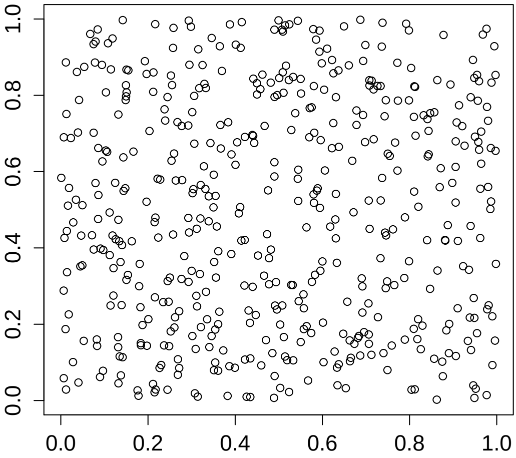
Look at discretisation noise in the absence of interesting topology, with $G = \text{Uniform}$.

All intervals in unit square barcodes as a result of discretisation noise.

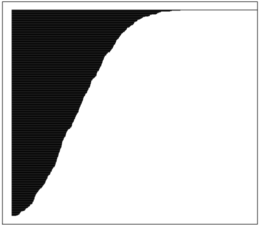
Method

Sample independently and uniformly at random from the unit square.

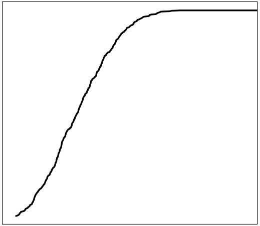
What do the barcodes look like from samples like this?



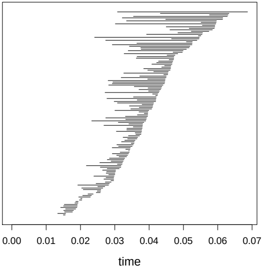
Discretisation noise

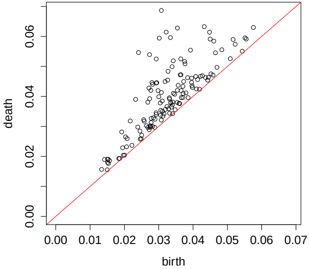


time



Betti-0 barcode becomes Betti-0 trace





Betti-1 barcode becomes persistence diagram (Edelsbrunner and Harer).

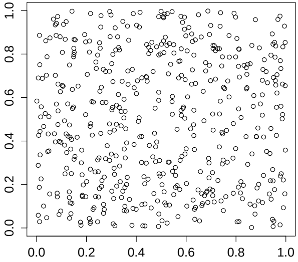
Discretisation noise

Those two barcodes were for a single finite uniform sample of 500 points from the unit square.

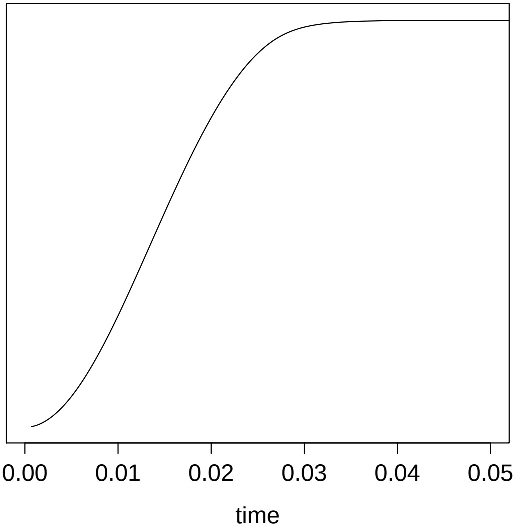
We can use Monte Carlo methods to estimate the expected trace and diagram for the unit square.

We use 10000 independent finite uniform samples of 500 points from the unit square, and look at the expected Betti-0 trace and Betti-1 diagram.

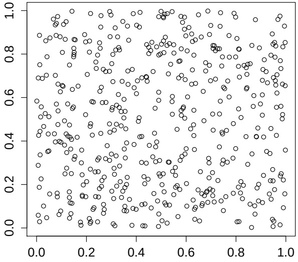
Monte Carlo barcodes



Expected Betti-0 trace.

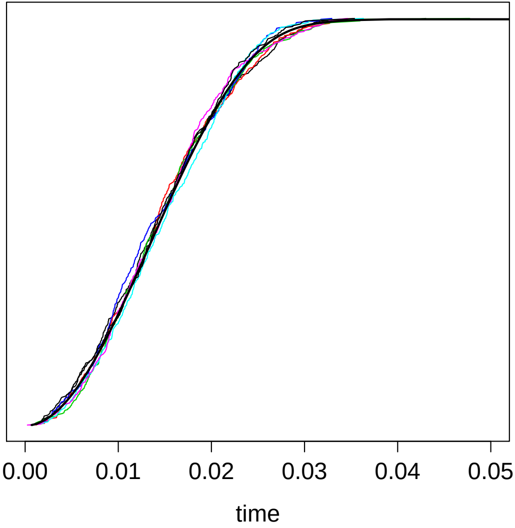


Monte Carlo barcodes

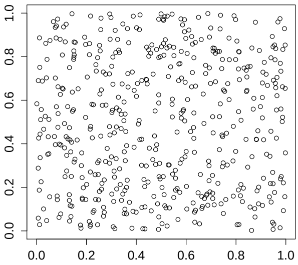


Expected Betti-0 trace.

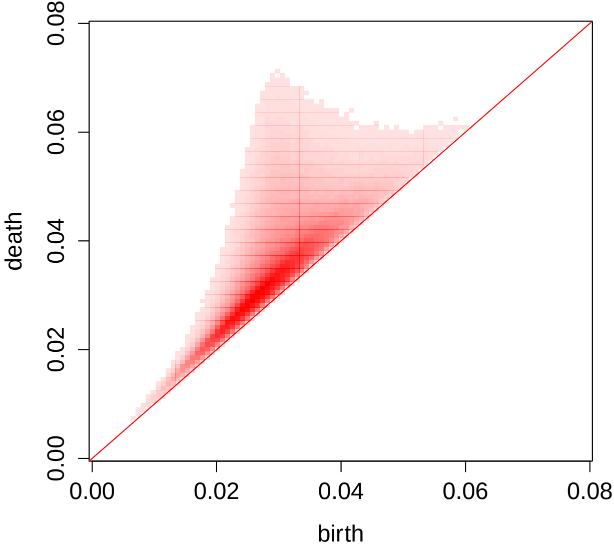
Highlighted: Betti-0 traces for samples.



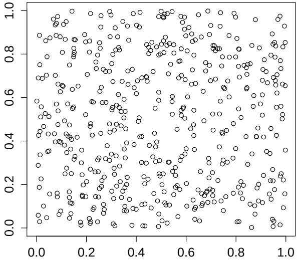
Monte Carlo barcodes



Expected Betti-1 persistence diagram.

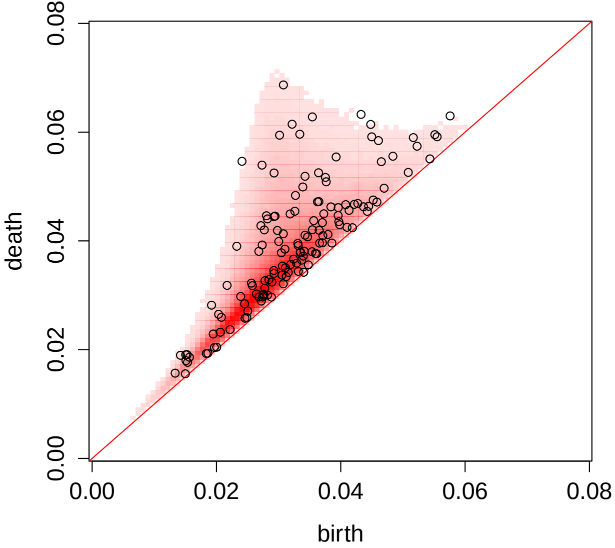


Monte Carlo barcodes



Expected Betti-1 persistence diagram.

Actual Betti-1 persistence diagram for sample.



Normalisation

The length of the persistence intervals depends on the sample size n and the area A of the manifold M (which we may have to estimate).

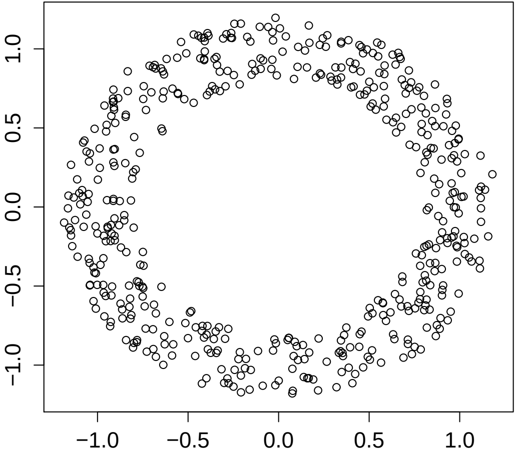
Therefore, we multiply the Monte Carlo noise estimates by a constant $\frac{\sqrt{n}}{\sqrt{A}}$ to give a estimate of the noise for any 2-d manifold for any sample size.

Example

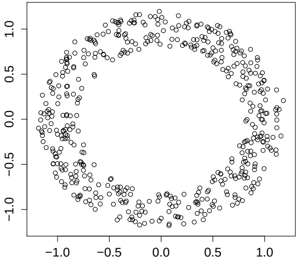
Now let's add some interesting topology.

Sample uniformly at random from this ring.

If intervals in the barcode look like the unit square noise, classify them as noise.

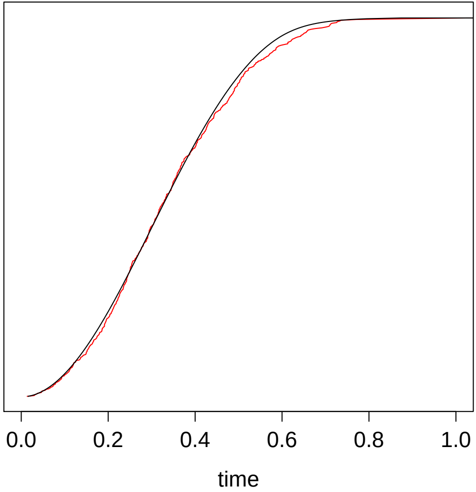


Example

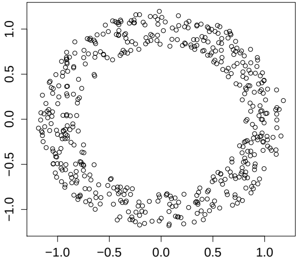


Expected (rescaled) Betti-0 trace for unit square.

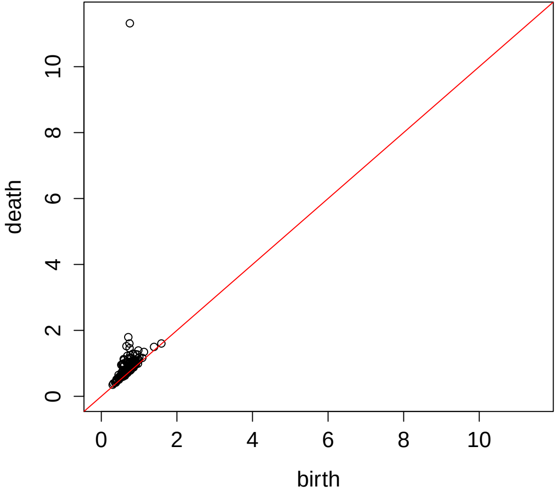
Actual (rescaled) trace for ring



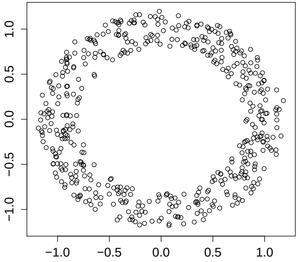
Example



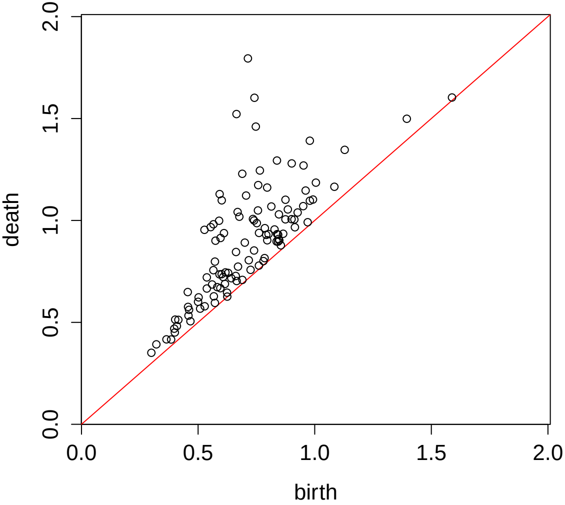
Actual (rescaled) Betti-1 persistence diagram for ring.



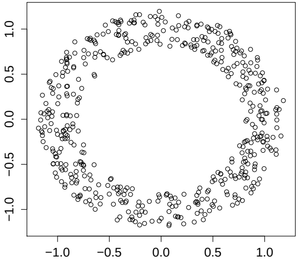
Example

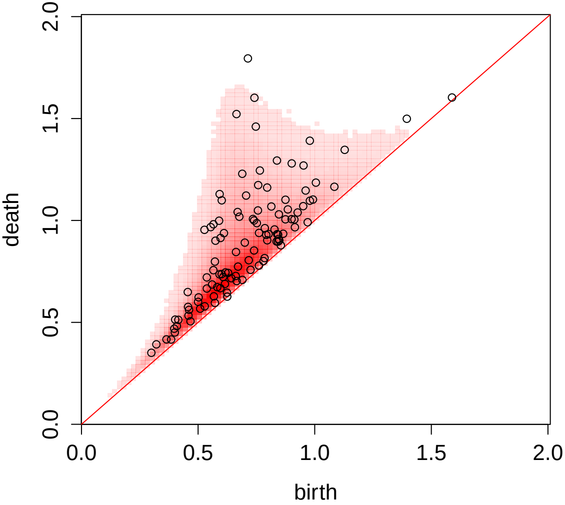


Discretisation noise in ring persistence diagram.



Example





Discretisation noise in ring persistence diagram.

Expected discretisation noise.

Non-uniform sampling

To understand discretisation noise, we assumed X is a finite sample from the Uniform density supported on M.

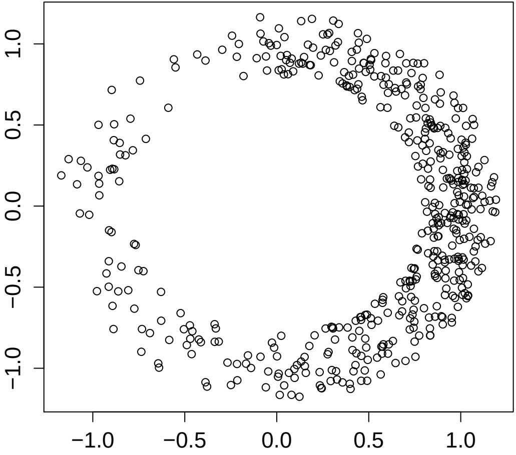
What happens if G is some arbitrary distribution? (but still supported on M)?

What is our model for sampling noise in the barcodes for this case?

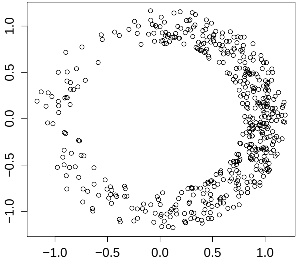
Example

Sample non-uniformly from the same annulus as earlier.

Refer to as 'non-uniform ring'.

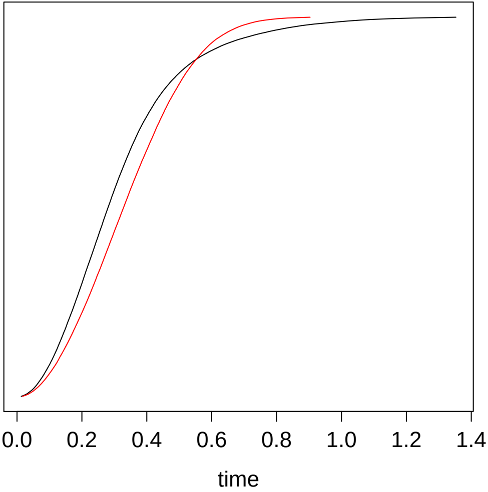


Example

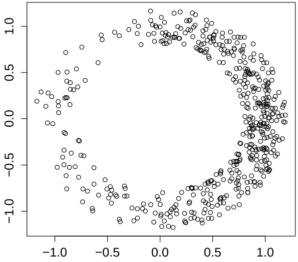


Expected Betti-0 trace for:

Non-uniform ring
Uniform ring

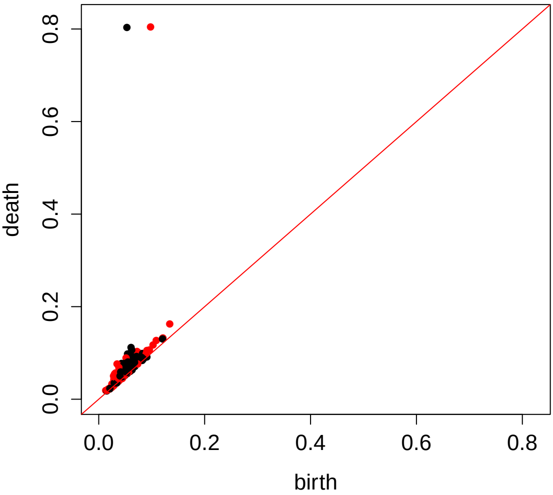


Example

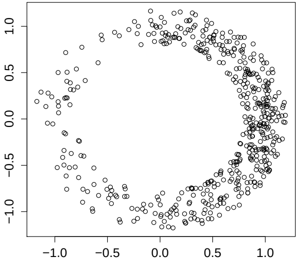


Example Betti-1 diagrams for:

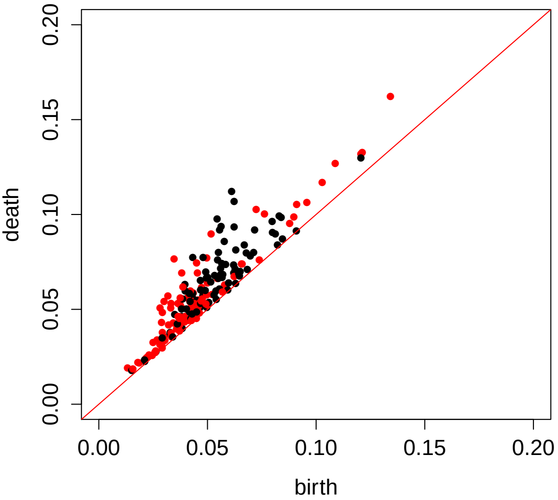
Uniform ring
Non-uniform ring



Example



Discretisation noise:



Uniform ring
Non-uniform ring

Persistent homology recap

Persistent homology works by measuring the size of simplices, and then adding them to the complex in order of smallest to biggest.

The function f that measures the simplices is called the filtration function. In standard persistent homology, this is the miniball of a simplex σ, i.e. $f(\sigma) = \text{miniball}(\sigma)$.

We modify the filtration function to take account of the local sampling density, via a (kernel) density estimate h.

Kernel density estimation

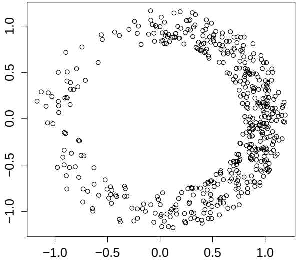
Our sample X is a set of points $X_1, X_2, \ldots, X_n$. A simple kernel density estimate of X evaluated at some point x is

$$h(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right),$$

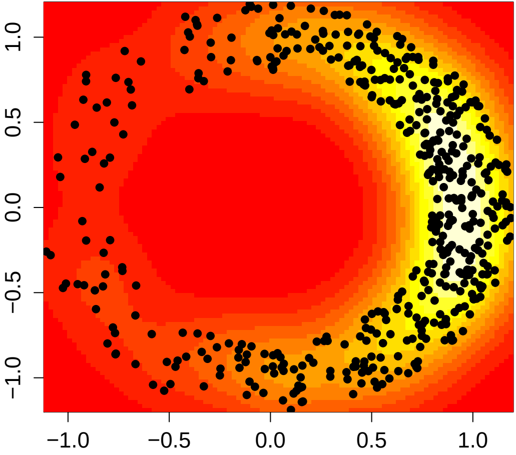
for a kernel K and a bandwidth b.

In what follows, we use the 2-dimensional multivariate normal distribution as our kernel. The bandwidth is selected automatically.

Example



Kernel density estimate



Density corrected persistent homology

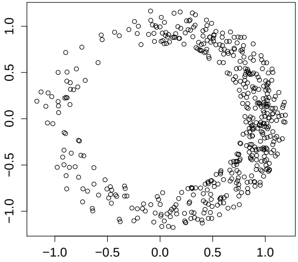
In density corrected persistent homology, we measure simplices by the filtration function

$$f(\sigma) = \mathrm{miniball}(\sigma) h(\sigma).$$

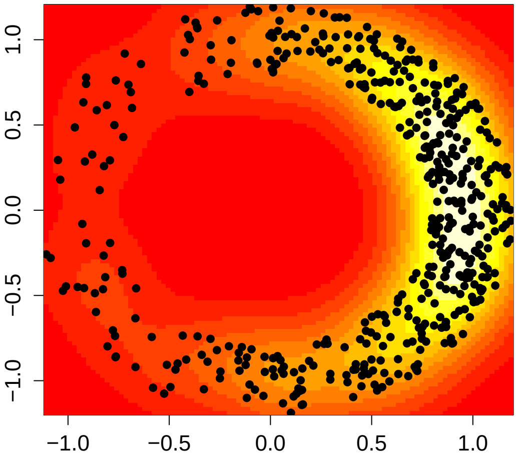
This has the effect of enlarging simplices in high density areas, and shrinking simplices in low density areas. The barcodes we obtain are as if we had sampled from a Uniform distribution on M, instead of from G.

Therefore, we can use our classification method from earlier.

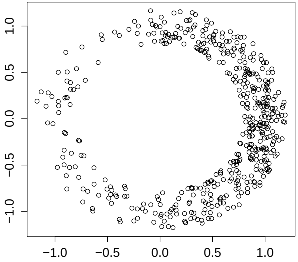
Example



We enforce a minimum value of h to stop the simplices shrinking too small.

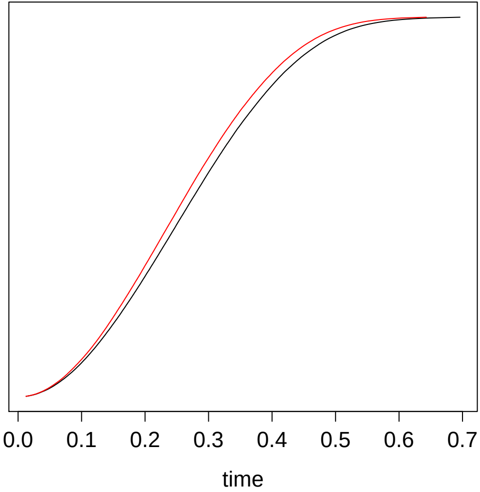


Example

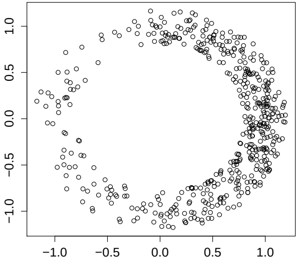


Expected Betti-0 traces with DCPH:

Non-uniform ring
Uniform ring

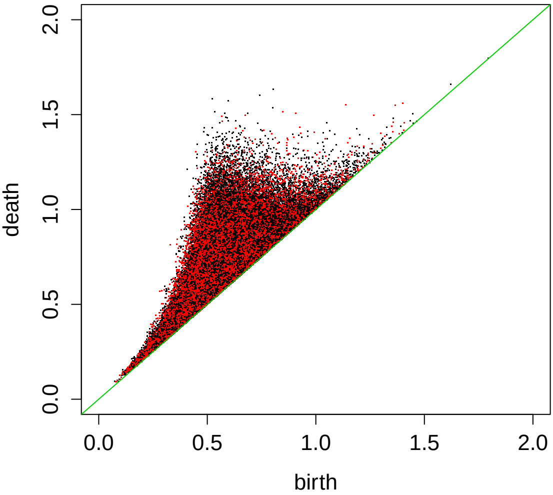


Example

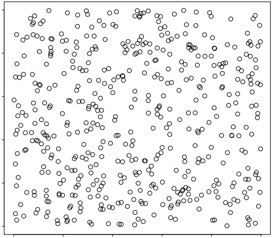


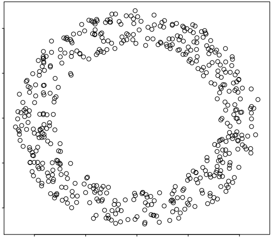
Discretisation noise with DCPH:

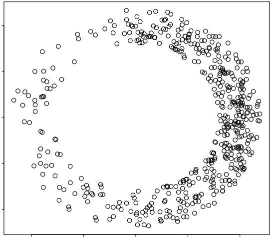
Non-uniform ring
Uniform ring



Summary







Discretisation noise in the unit square

Correct for surface area and identify topological features

Remove restriction on uniform sampling

Ongoing work

- Models for ‘off-manifold’ noise
- Statistical significance
- Machine implementable classification for persistence diagrams
- Applications