A statistical approach to algebraic topology

Peter Bubenik

(joint work with Peter Kim)

peter.bubenik@epfl.ch

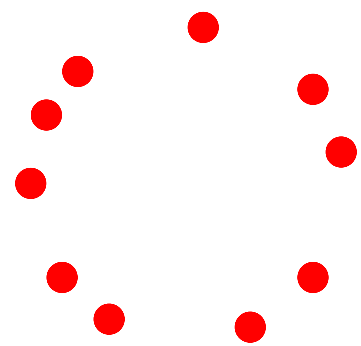
http://igat.epfl.ch/bubenik/talks/

Ecole Polytechnique Fédérale de Lausanne

June 10, 2005

1. Point Cloud Data from densities

Motivation: Given a set of points that "looks like a circle"



We would like to be able to say so mathematically.

Prior and current work

Background: Given a sample, techniques have been developed which detect topological features and give a rough measure of their 'size'.

Our goal: We aim to apply statistical techniques to provide some theoretical rigour to these largely experimental tools. We want to

- given a parameter, calculate expectations of the topological features detected from a sample
- apply bootstrapping to the study of these features

The setting

Given a manifold $\mathbb{M}$ and a (Radon) measure ν, a (probability) density is a function $f : \mathbb{M} \to [0, \infty]$ such that $\int_{\mathbb{M}} f \, d\nu = 1$.

We assume that we have a family of densities

$$\{f_\theta \mid \theta \in \Theta\}$$

where θ is called the parameter and Θ is the parameter space.

In our examples $(\mathbb{M}, \nu)$ is a sphere with spherical measure and Θ is a subset of a finite dimensional vector space.

Point Cloud Data

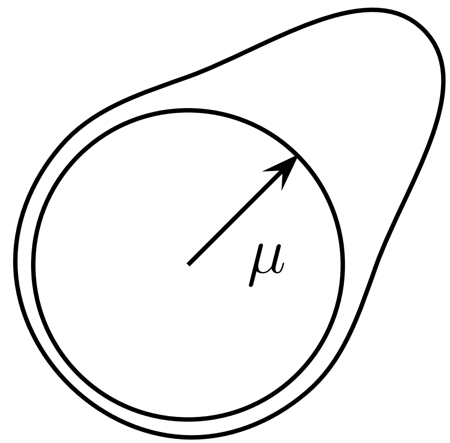
A sample $X_1, X_2, \ldots X_N$ is a sequence of independent and identically distributed random quantities on $\mathbb{M}$ drawn according to the density f_ϑ for some fixed but unknown $\vartheta \in \Theta$.

This sample is called Point Cloud Data (PCD).

We will show that from this PCD one can construct a Rips complex to which one can apply persistent homology to determine Betti barcodes.

Densities on S^1

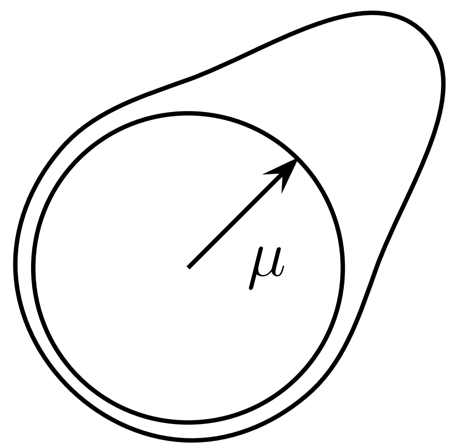
We will focus on unimodal densities on S^1, called von Mises distributions.



$$f_{\mu,\kappa}(x) = c(\kappa)e^{\kappa x^t \mu}.$$

Densities on S^1

We will focus on unimodal densities on S^1, called von Mises distributions.

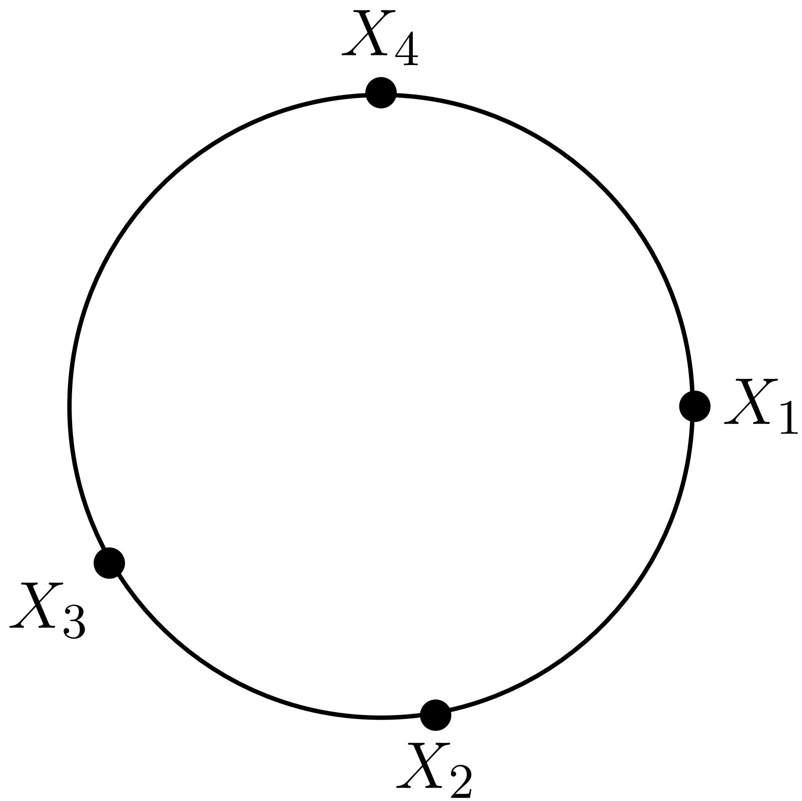


$$f_{\mu,\kappa}(x) = c(\kappa)e^{\kappa x^t \mu}.$$

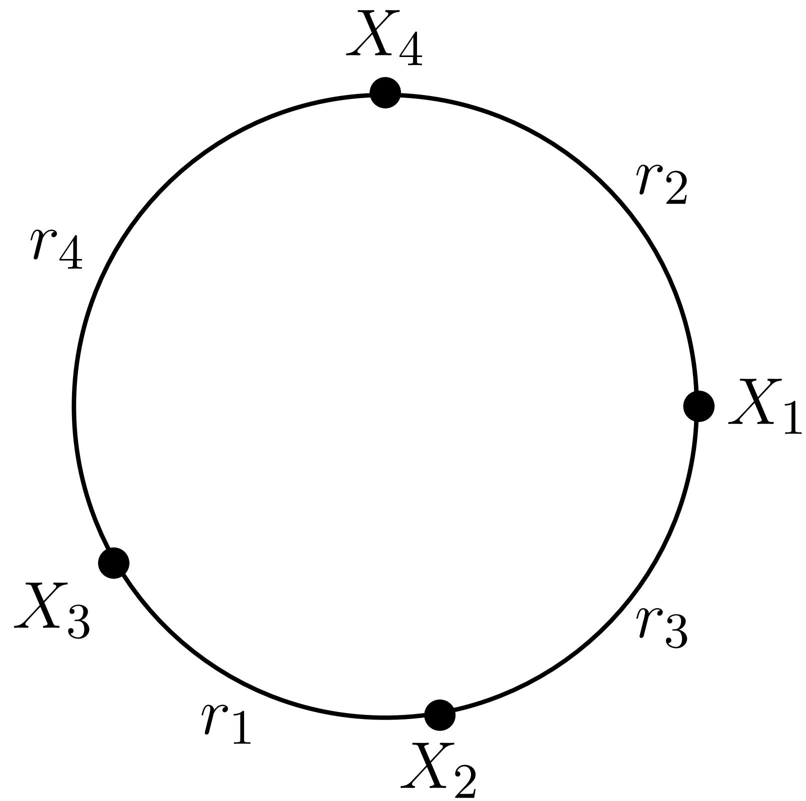
Plan: Sample $X = (X_1, \ldots, X_n) \subset S^1$ according to $f_{\mu,\kappa}$

Construct Rips complex $\xrightarrow{\text{persistent homology}}$ Betti barcodes

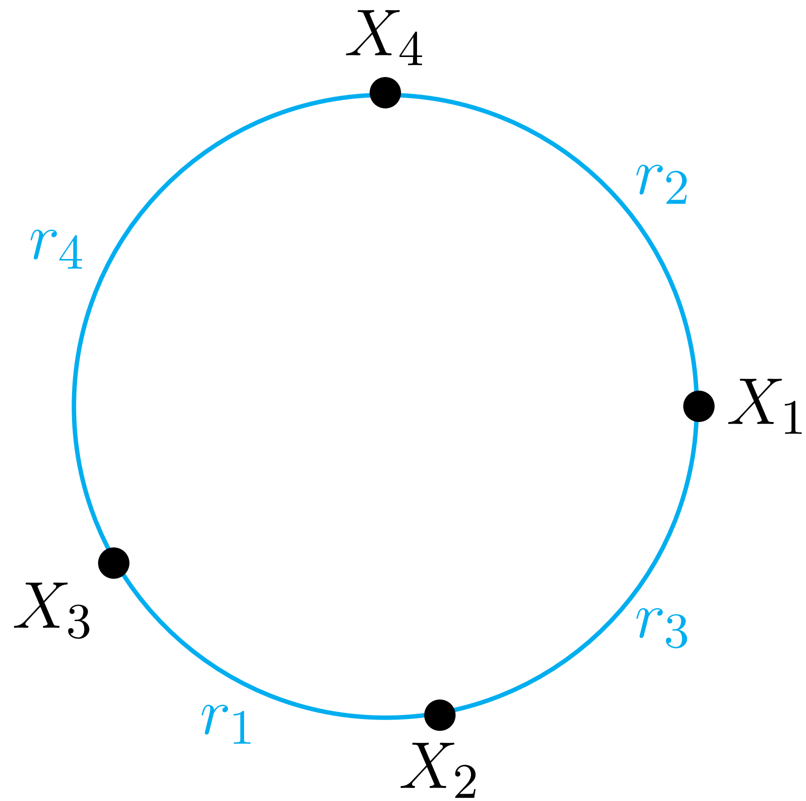
2. The Rips complex



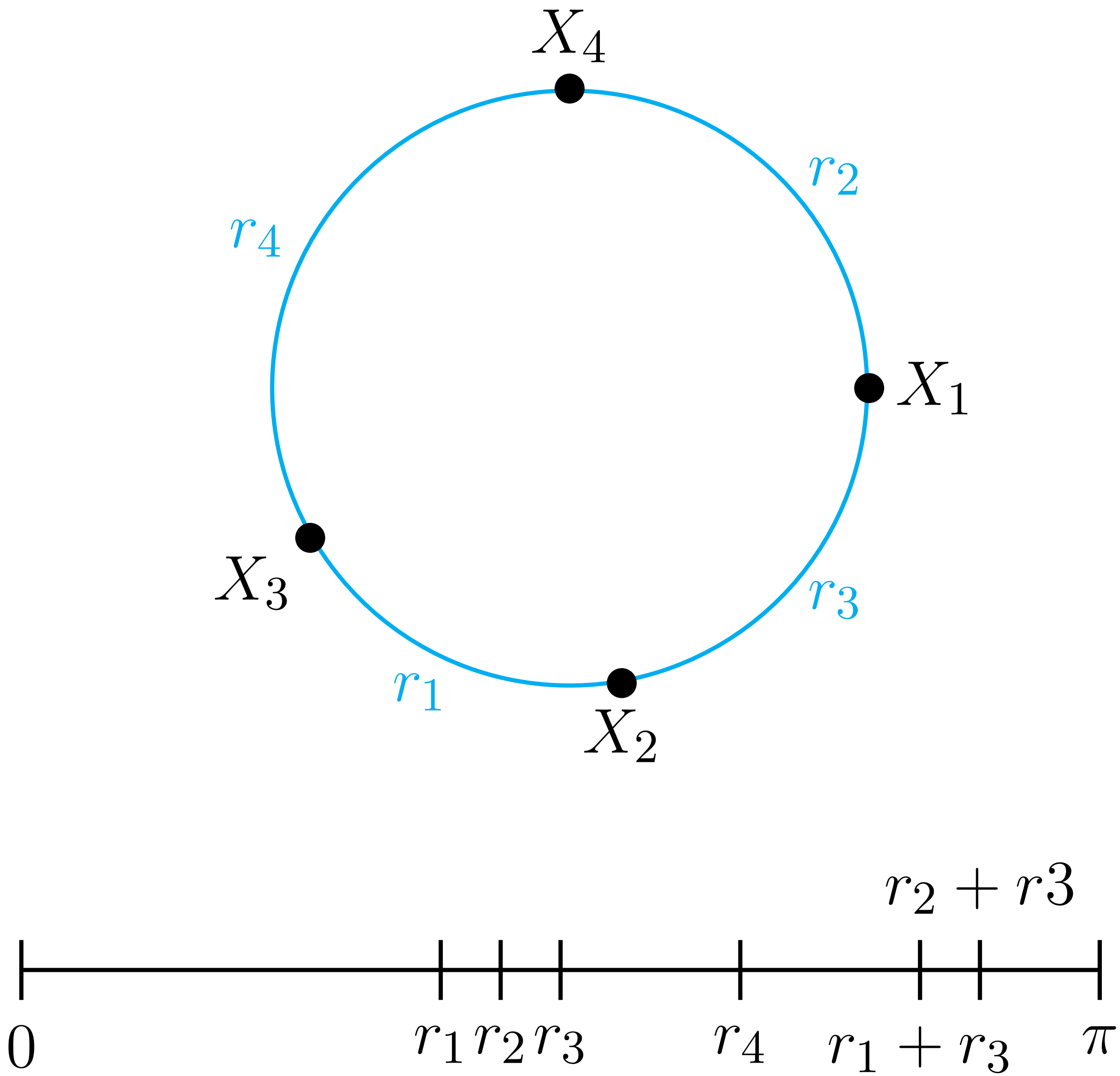
2. The Rips complex



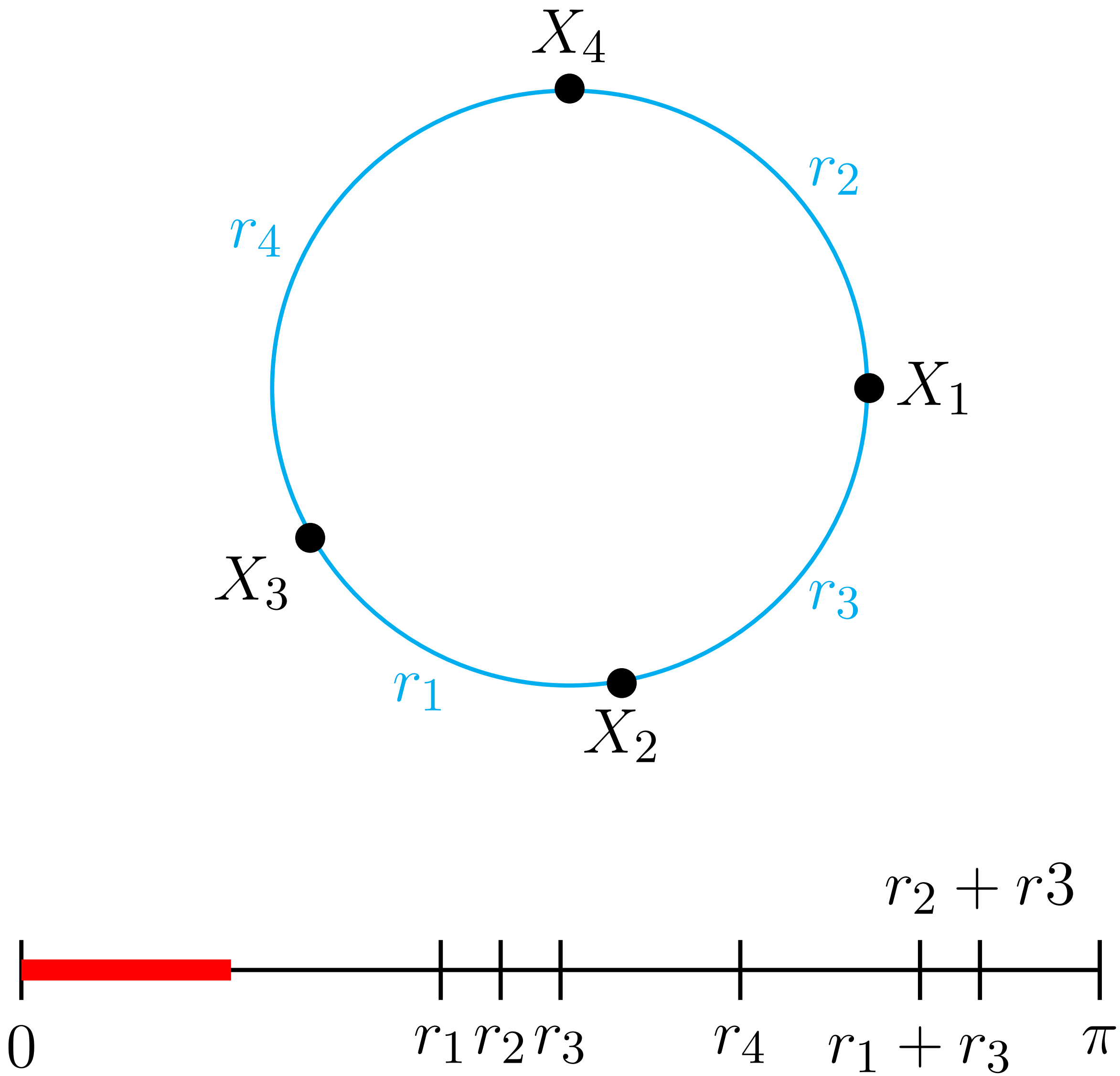
The Rips complex



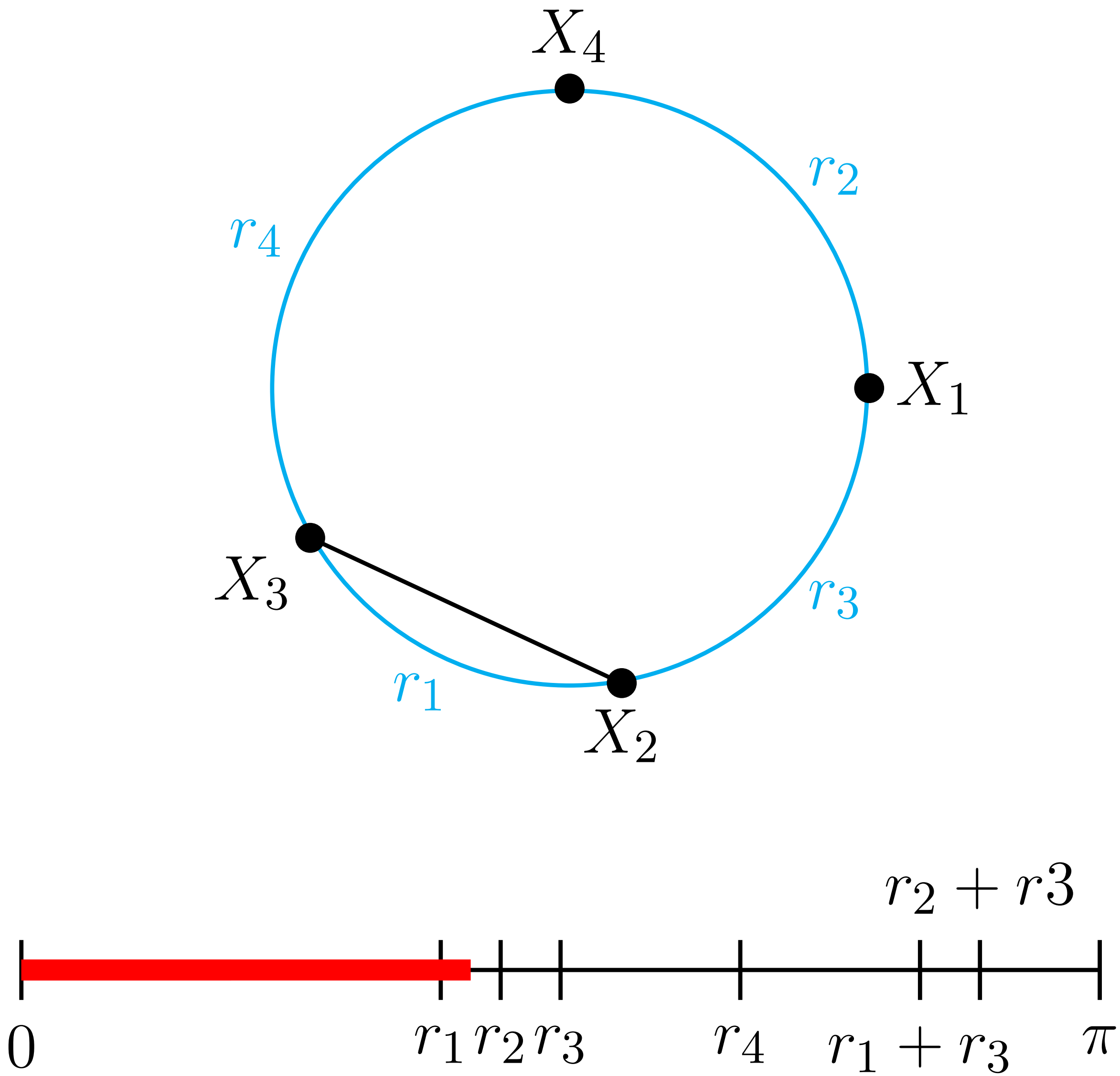
The Rips complex



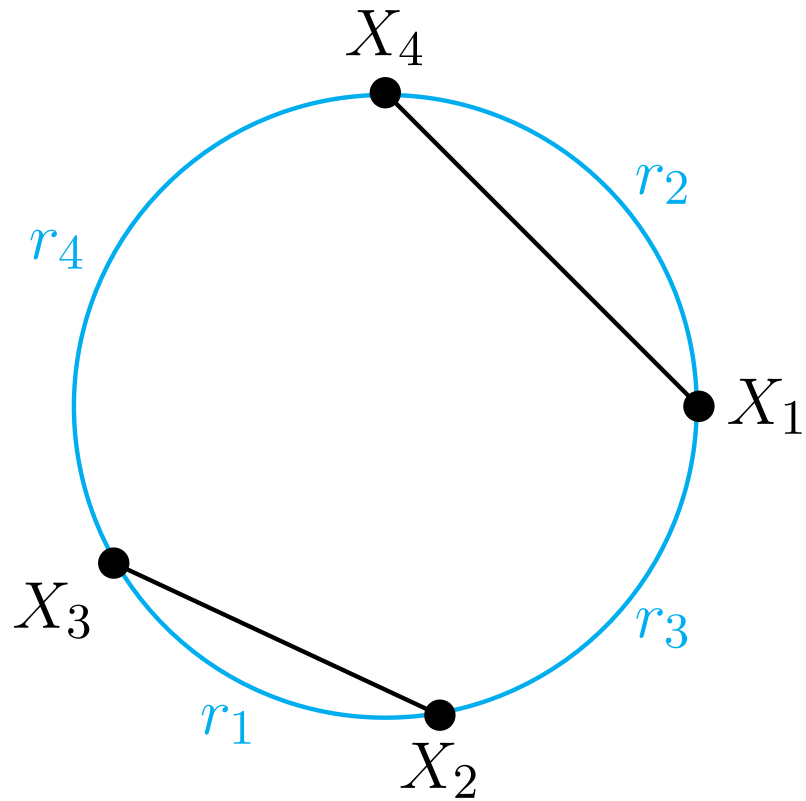
The Rips complex

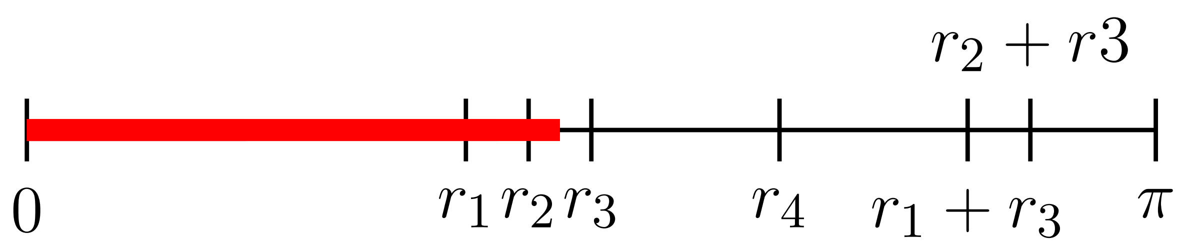


The Rips complex

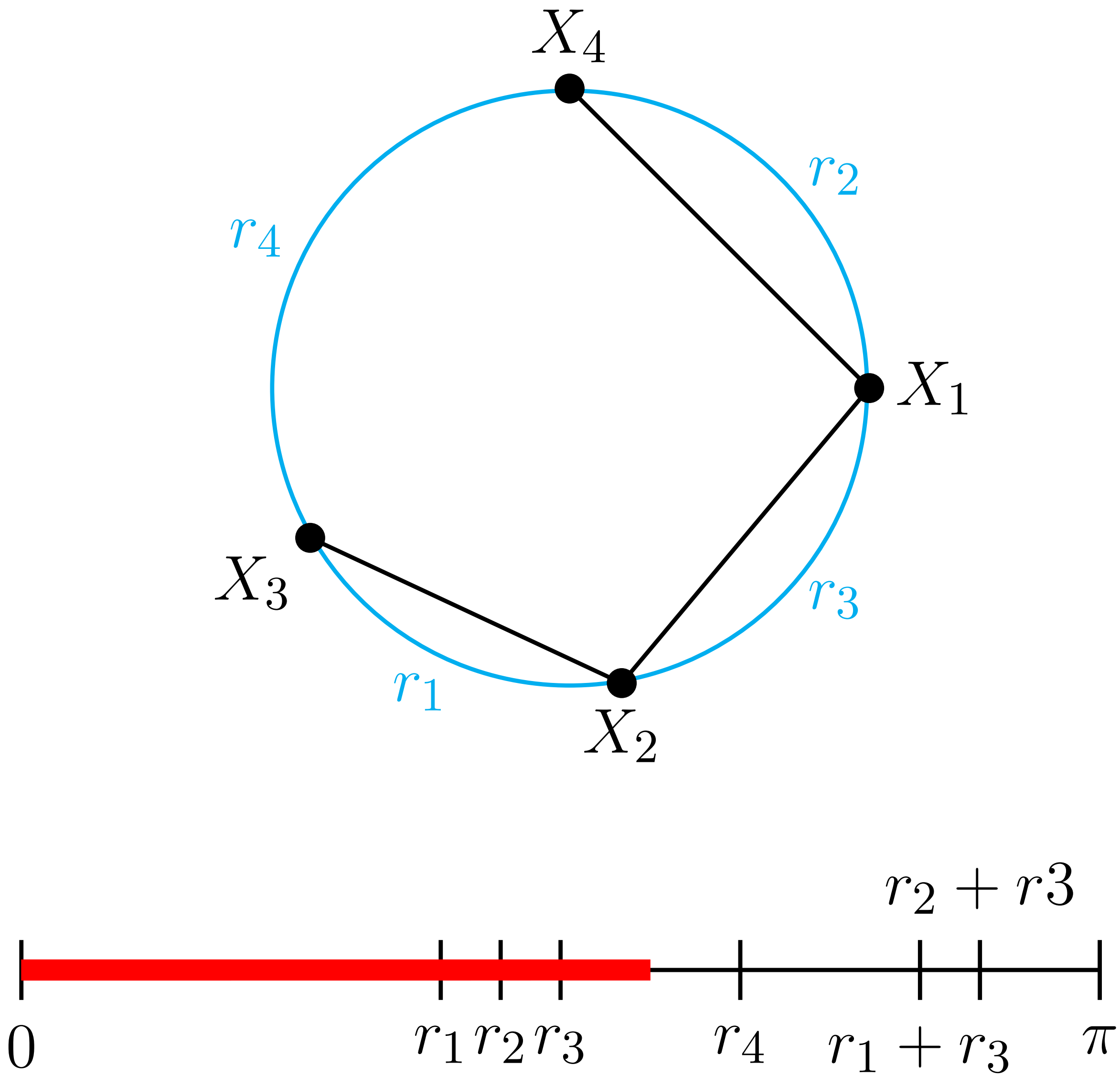


The Rips complex

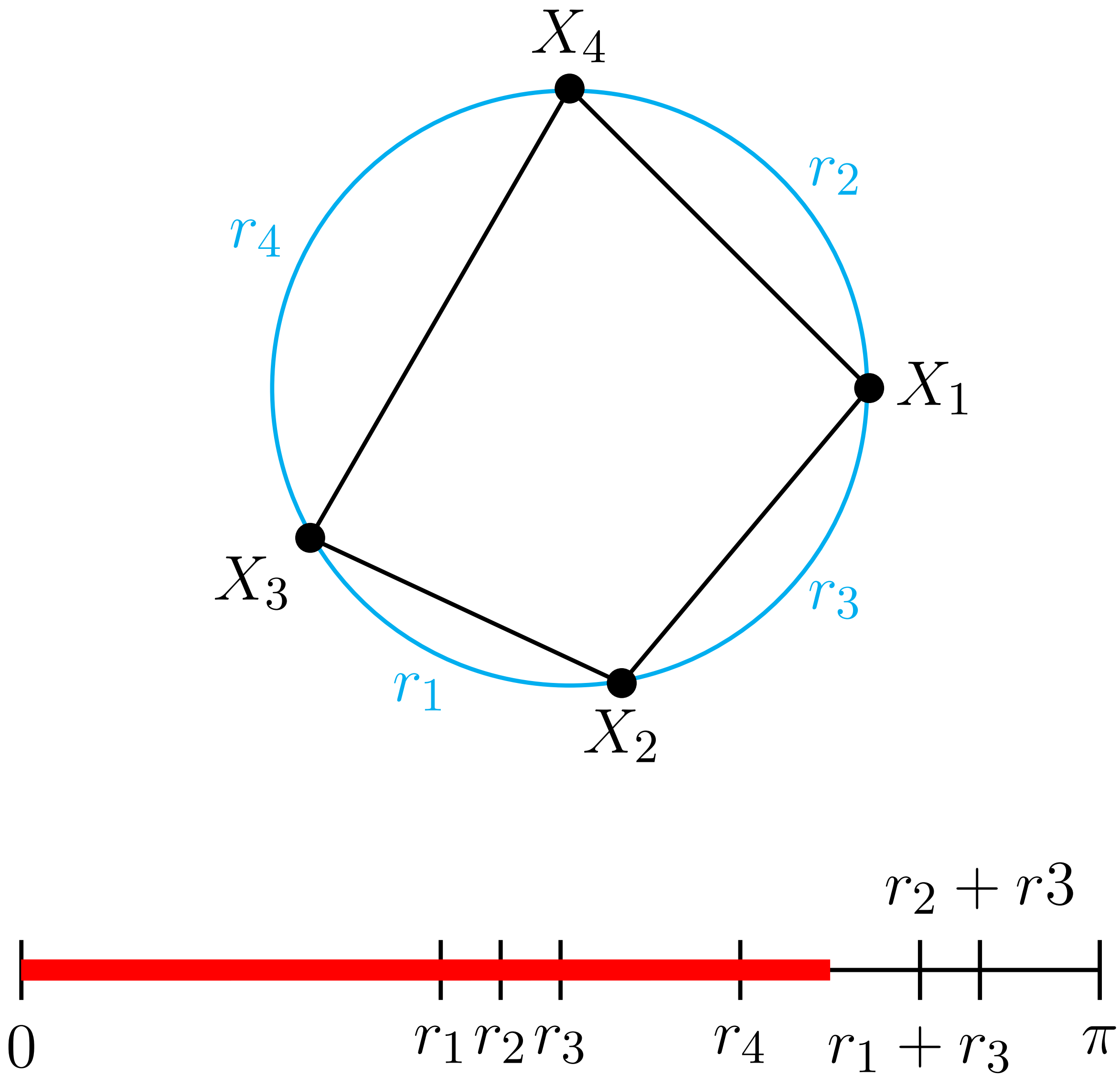




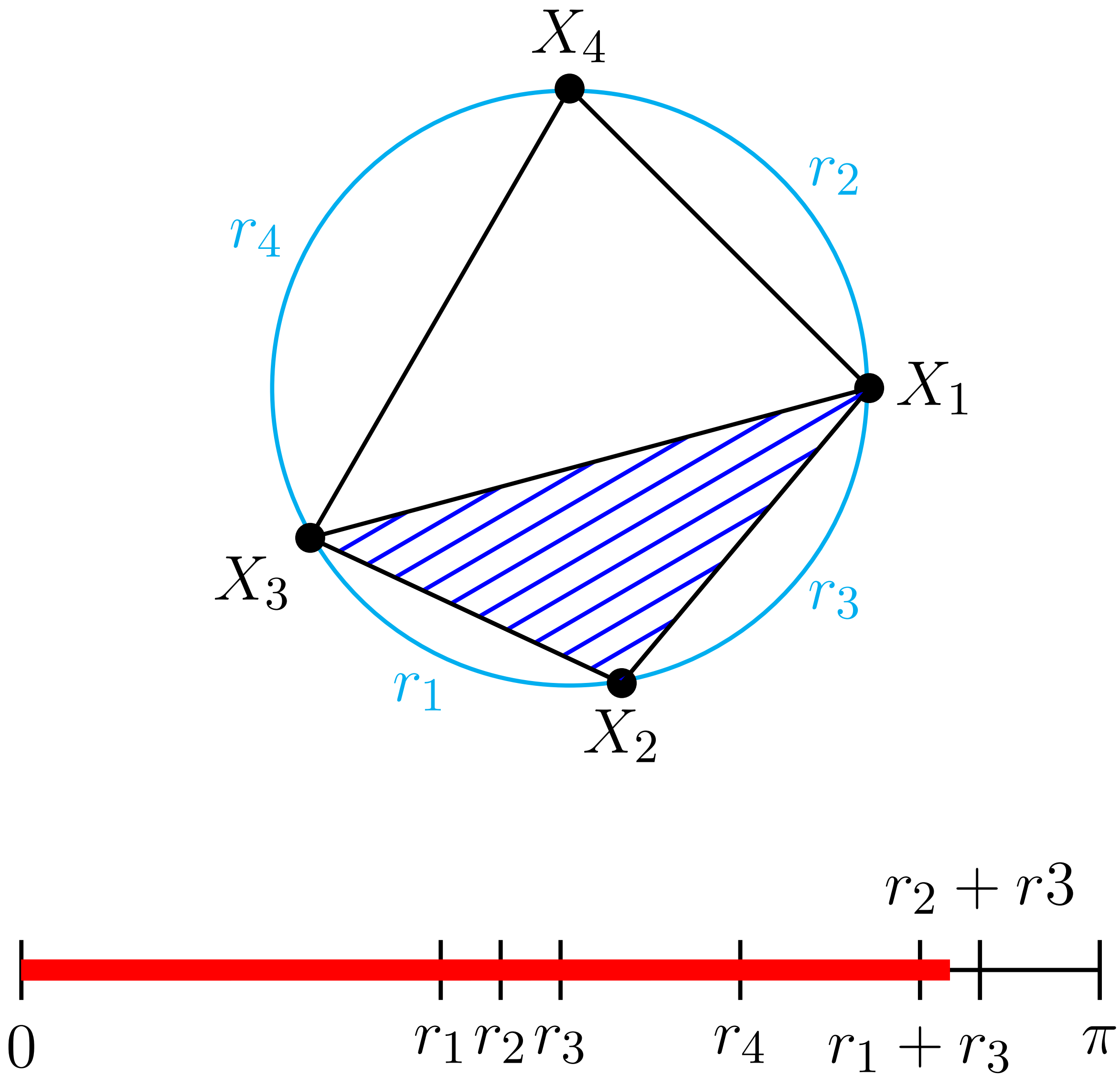
The Rips complex



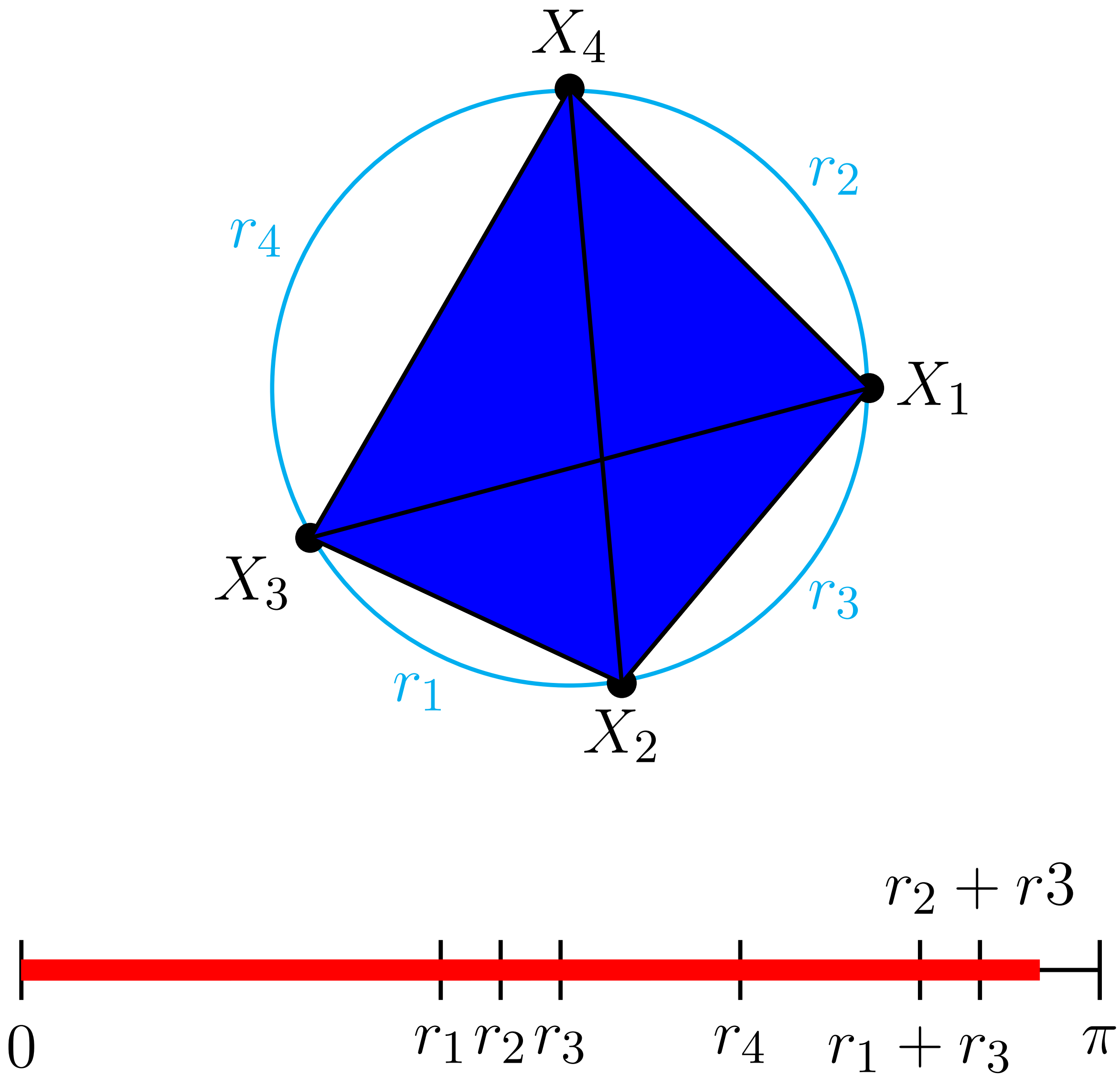
The Rips complex



The Rips complex

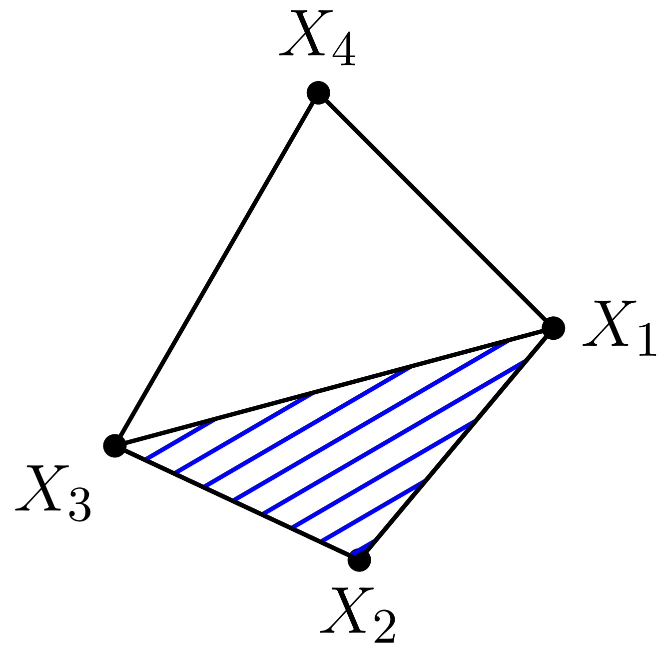


The Rips complex



3. Homology of Simplices

Consider the following simplicial complex Δ:

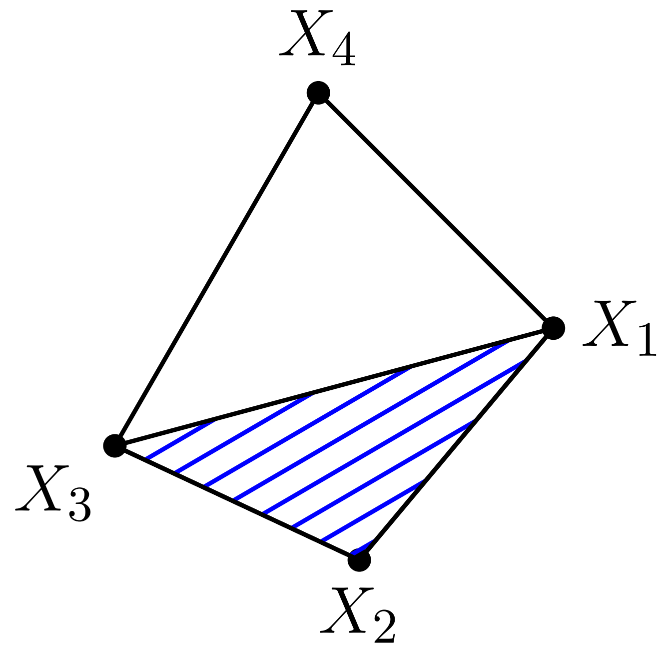


It consists of

- four 0-simplices: X_1, X_2, X_3, X_4,
- five 1-simplicies: $\{X_1, X_2\}$, $\{X_1, X_3\}$, $\{X_2, X_3\}$, $\{X_1, X_4\}$, $\{X_3, X_4\}$,
- and one 2-simplex: $\{X_1, X_2, X_3\}$.

Homology of Simplicies

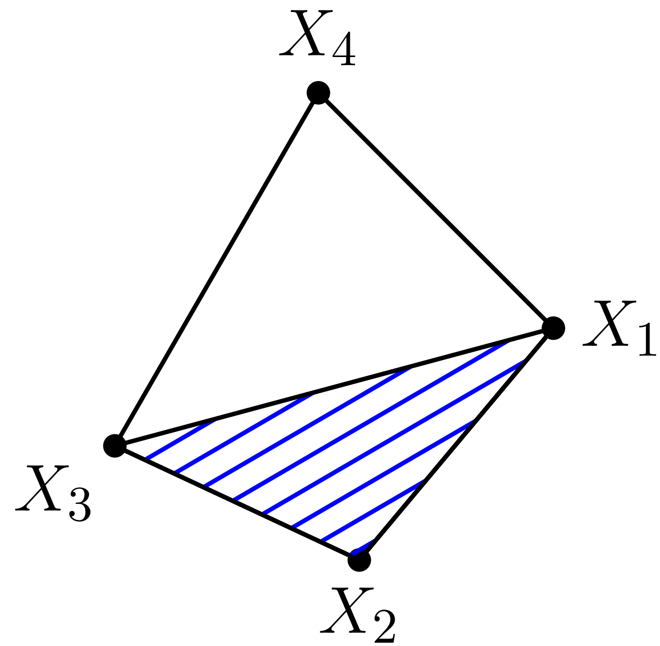
Sums of n-simplicies are called n-chains.



- the boundary of $\{X_1, X_2, X_3\}$ is $\{X_1, X_2\} + \{X_2, X_3\} + \{X_3, X_1\}$
- the boundary of $\{X_i, X_j\}$ is $X_j - X_i$,
- the boundary of $\{X_i\}$ is 0

Homology of Simplicies

By linearity this defines the boundary on all n-chains.



Cycles are n-chains with boundary equal to zero.

For example $\{X_1, X_2\} + \{X_2, X_3\} + \{X_3, X_1\}$ is a cycle and so is X_1.

One can check that boundaries are always cycles.

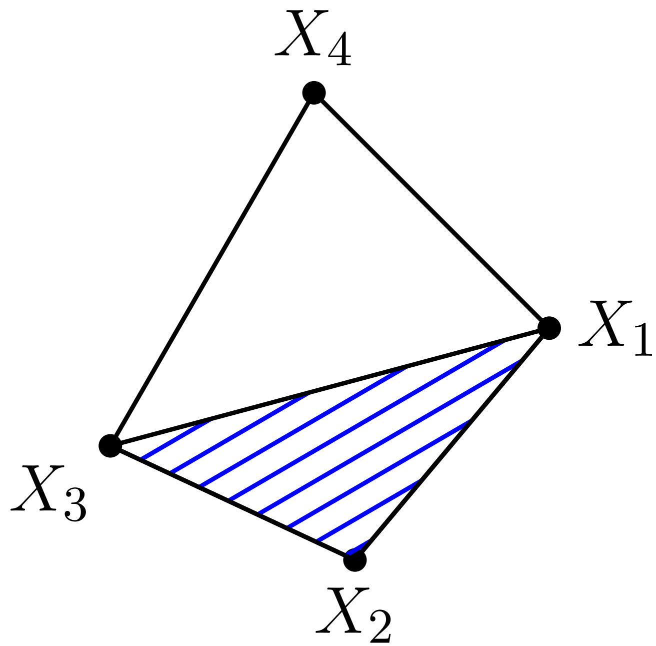
Homology of Simplicies

The homology $H_n(\Delta)$ is the quotient of the cycles modulo the boundaries.

The Betti number $B_n(\Delta)$ is the dimension of $H_n(\Delta)$.

$B_0(\Delta)$ is the number of connected components of Δ.
$B_1(\Delta)$ is the number of holes in Δ.
$B_2(\Delta)$ is the number of voids in Δ.

Homology of Simplicies



In our example $B_0(\Delta) = 1$, $B_1(\Delta) = 1$, and all higher Betti numbers are zero.

4. Persistent homology

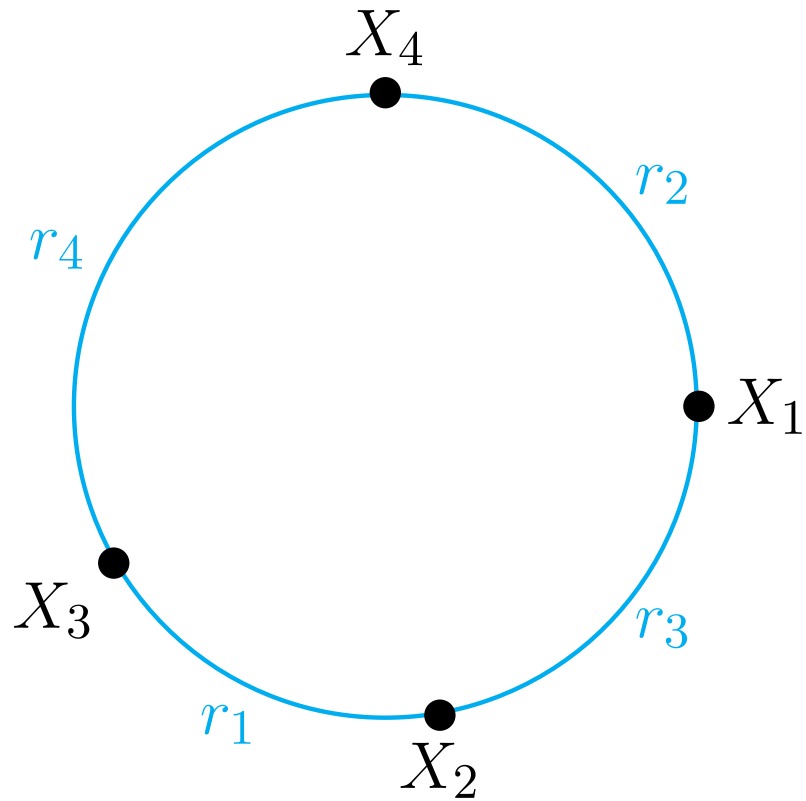
Assume we have a simplicial complex that changes as we vary some parameter r.

The homology that persists as r changes is called persistent homology.

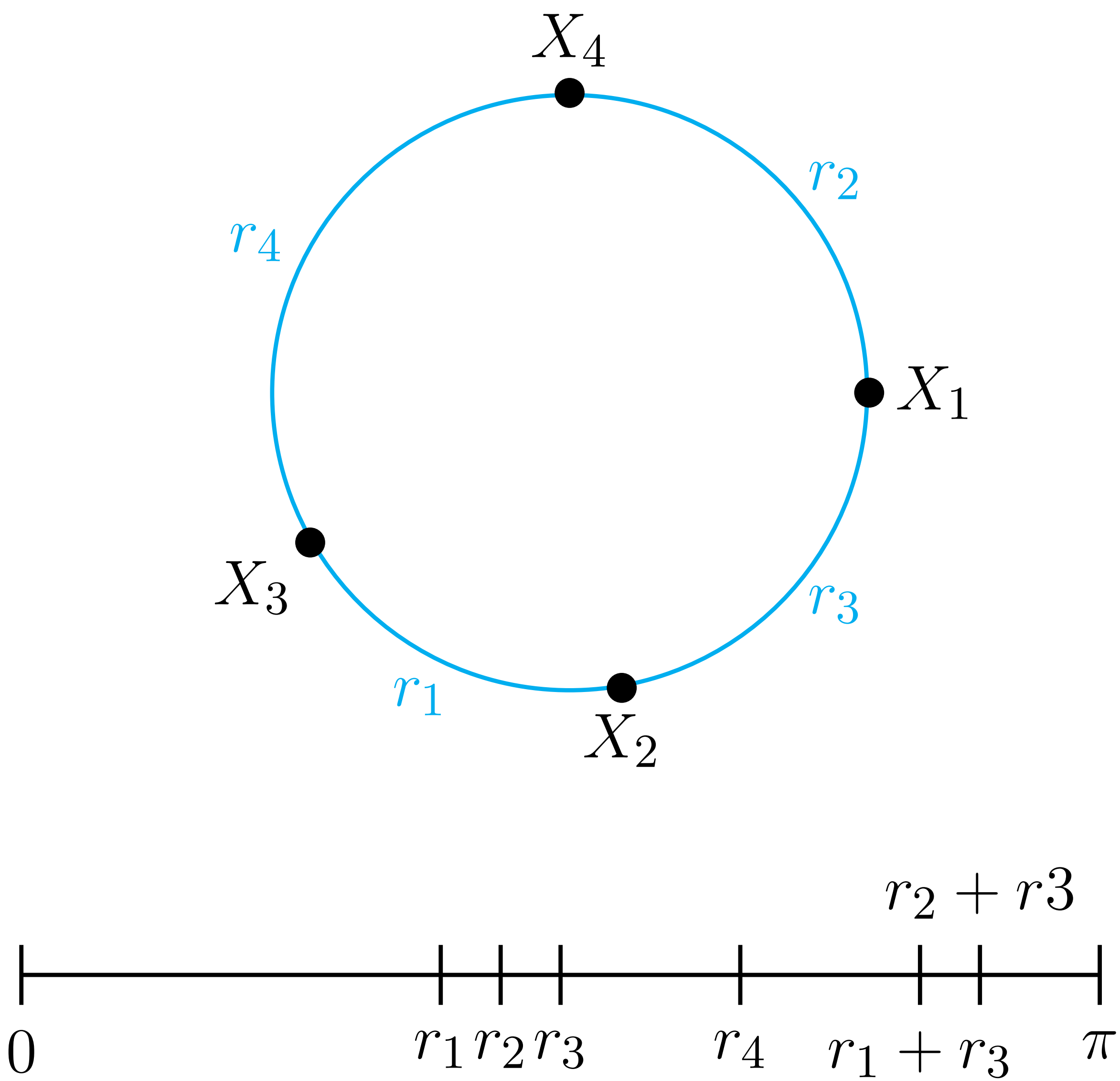
We can record how the Betti numbers change as r changes using Betti barcodes.

We illustrate this using the Rips complex on our earlier example.

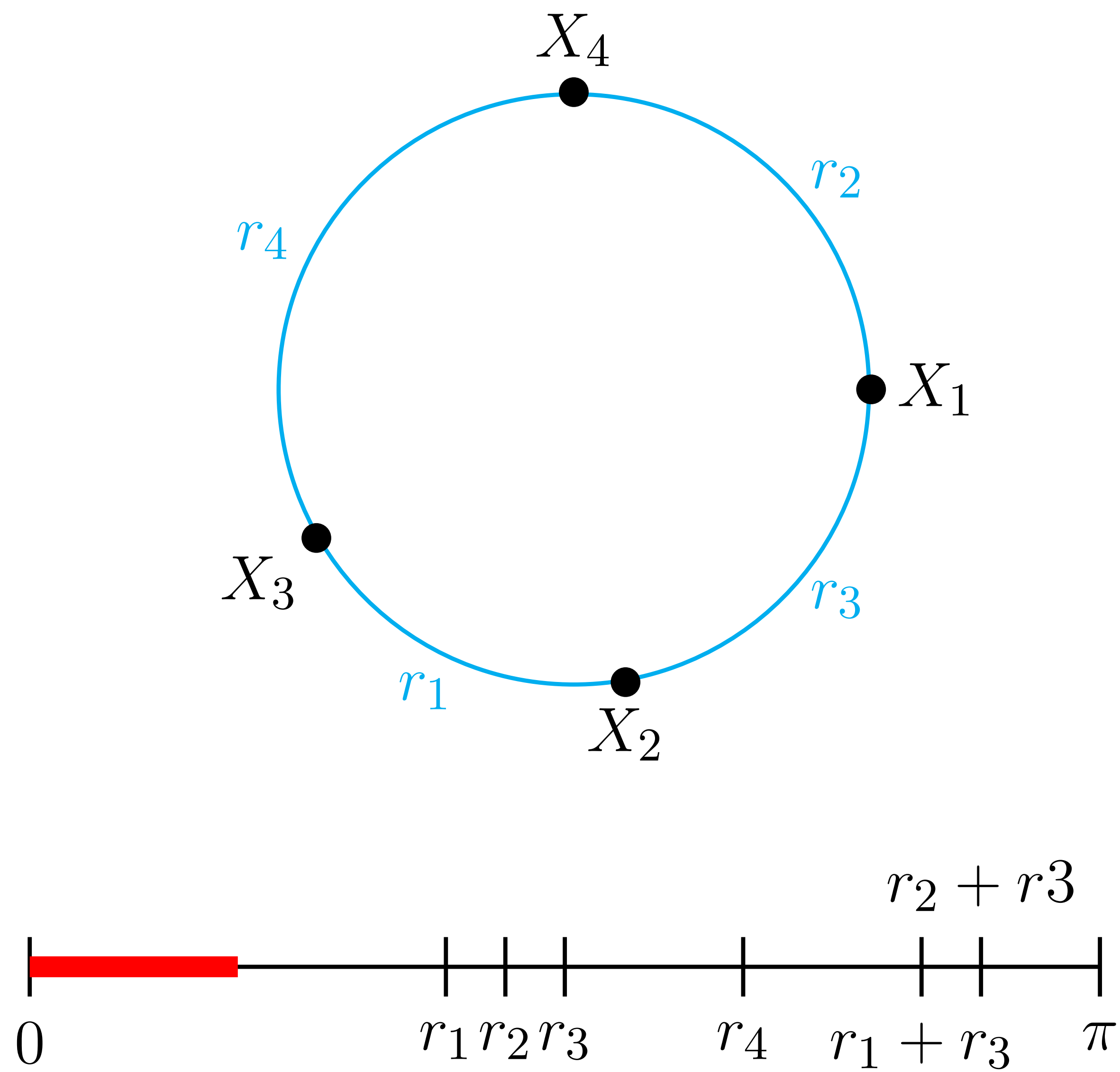
The Rips complex



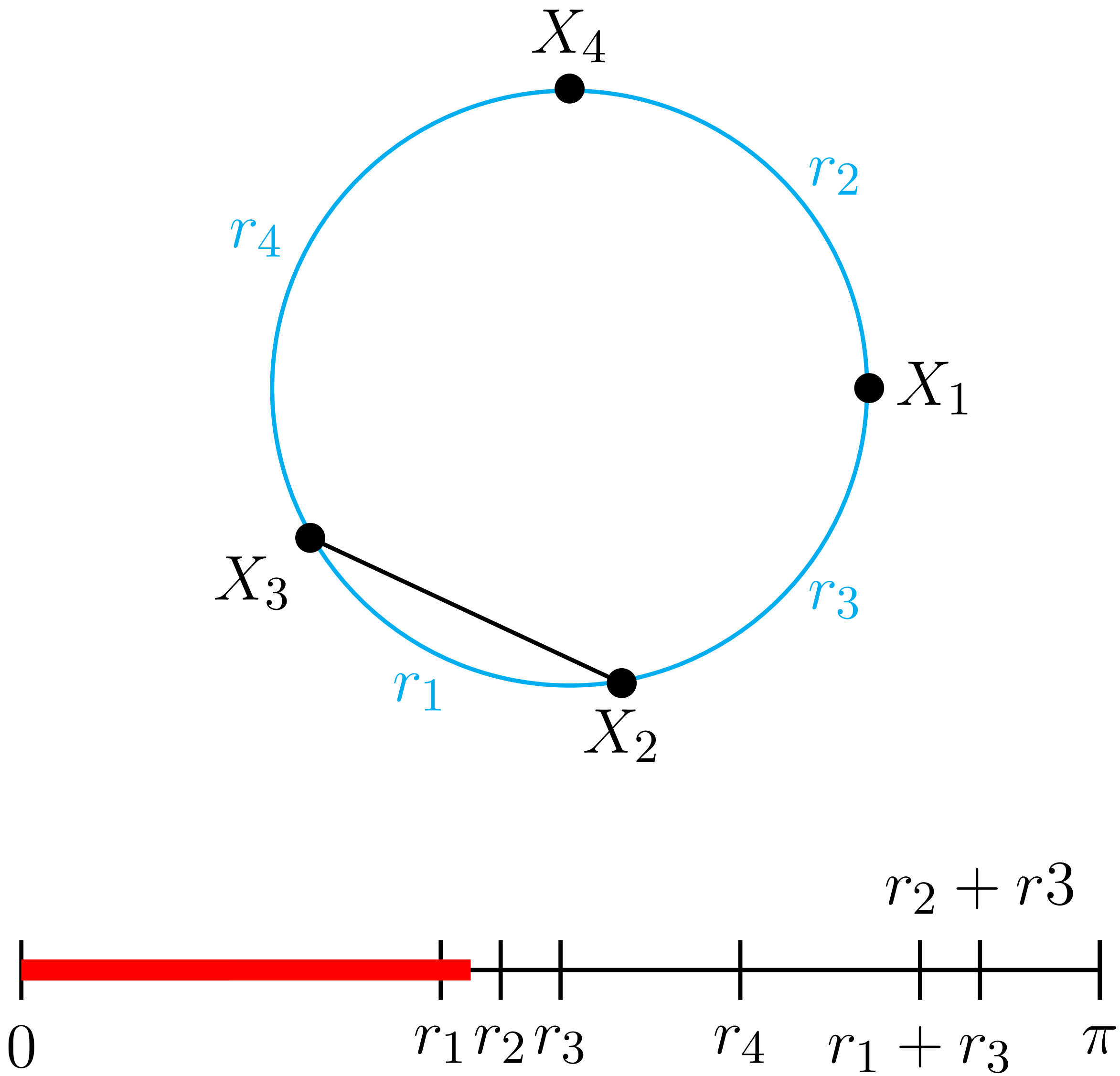
The Rips complex



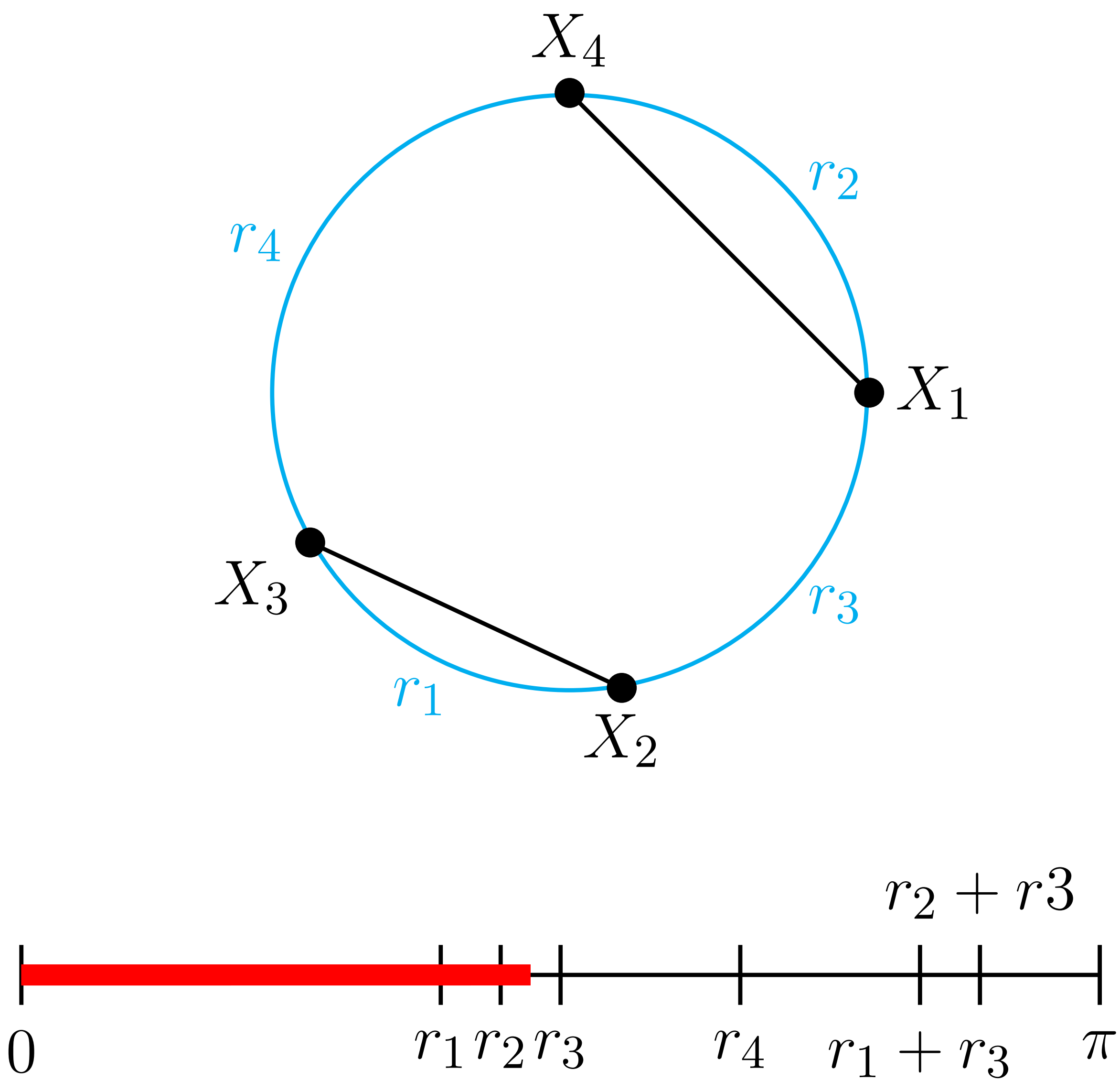
The Rips complex



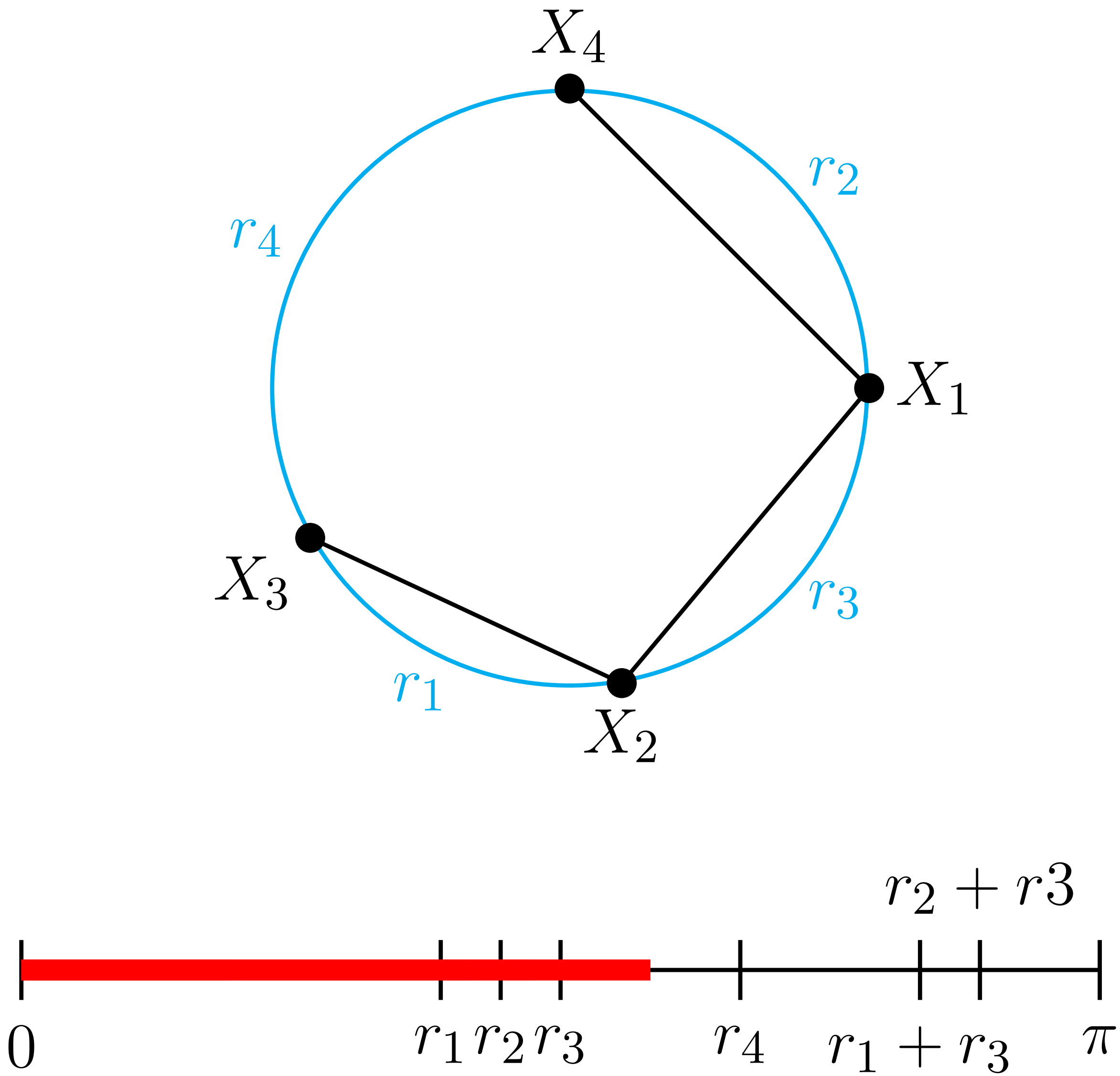
The Rips complex



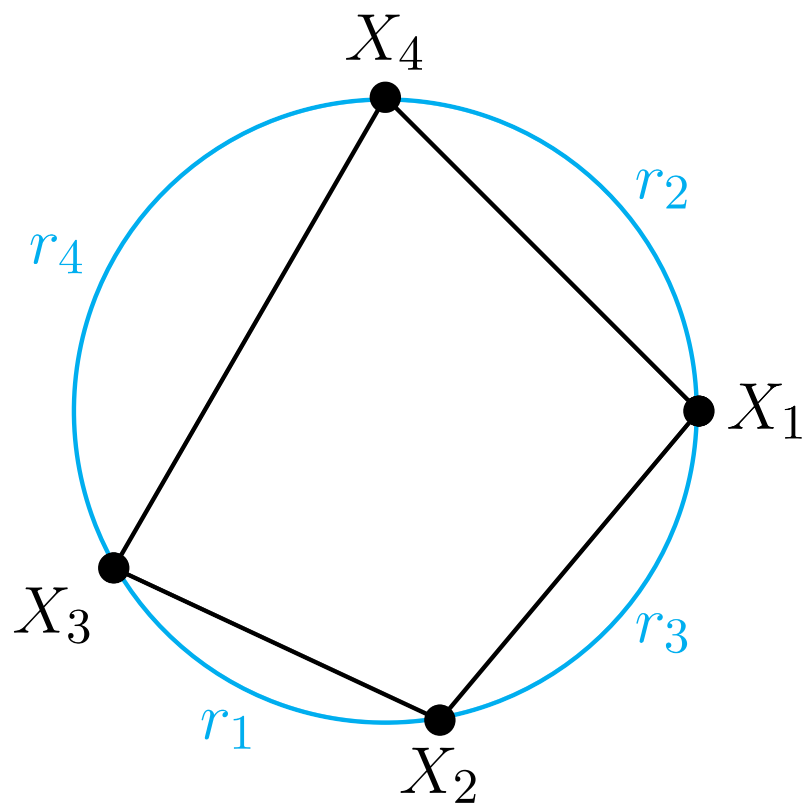
The Rips complex

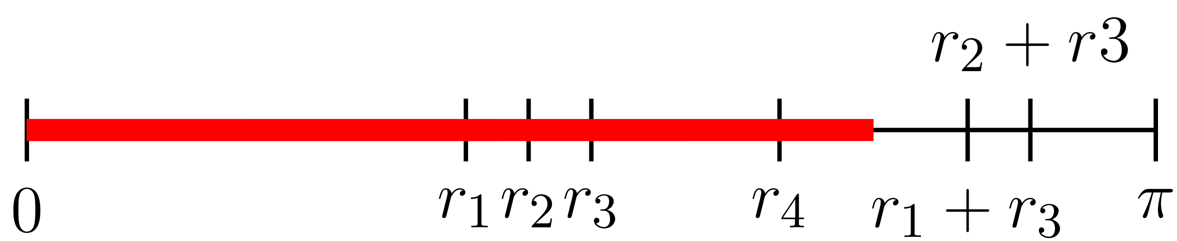


The Rips complex

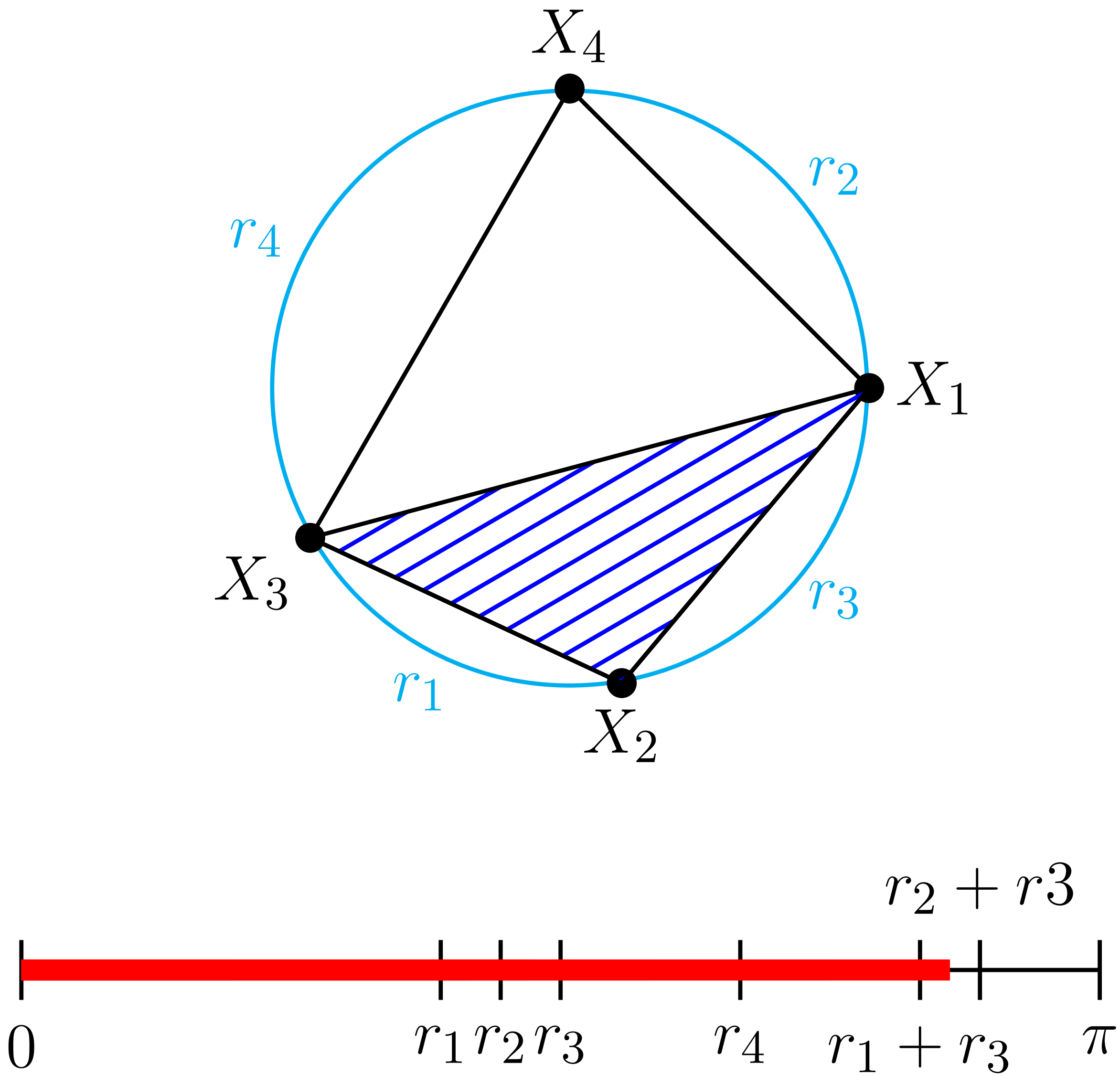


The Rips complex

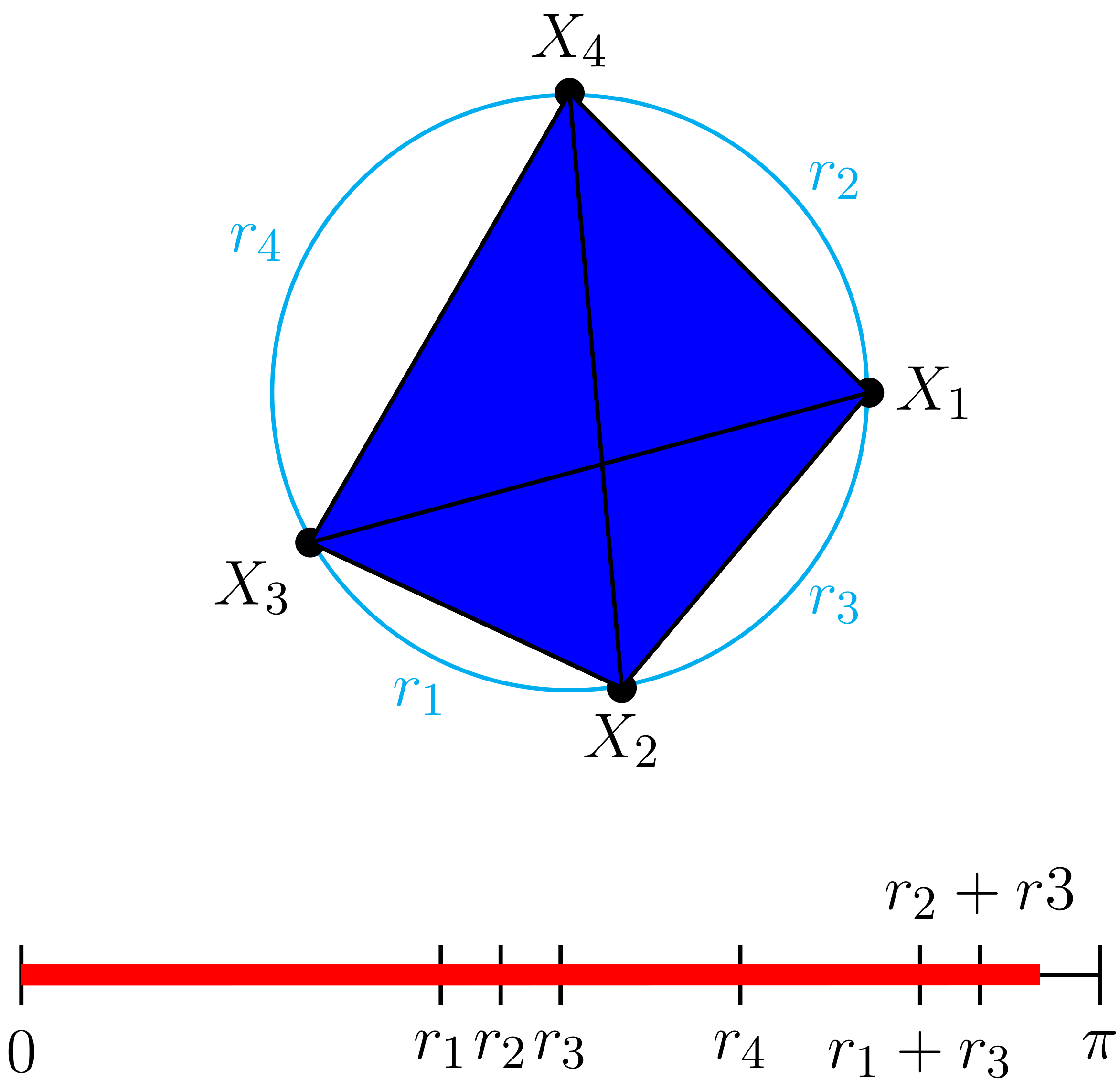




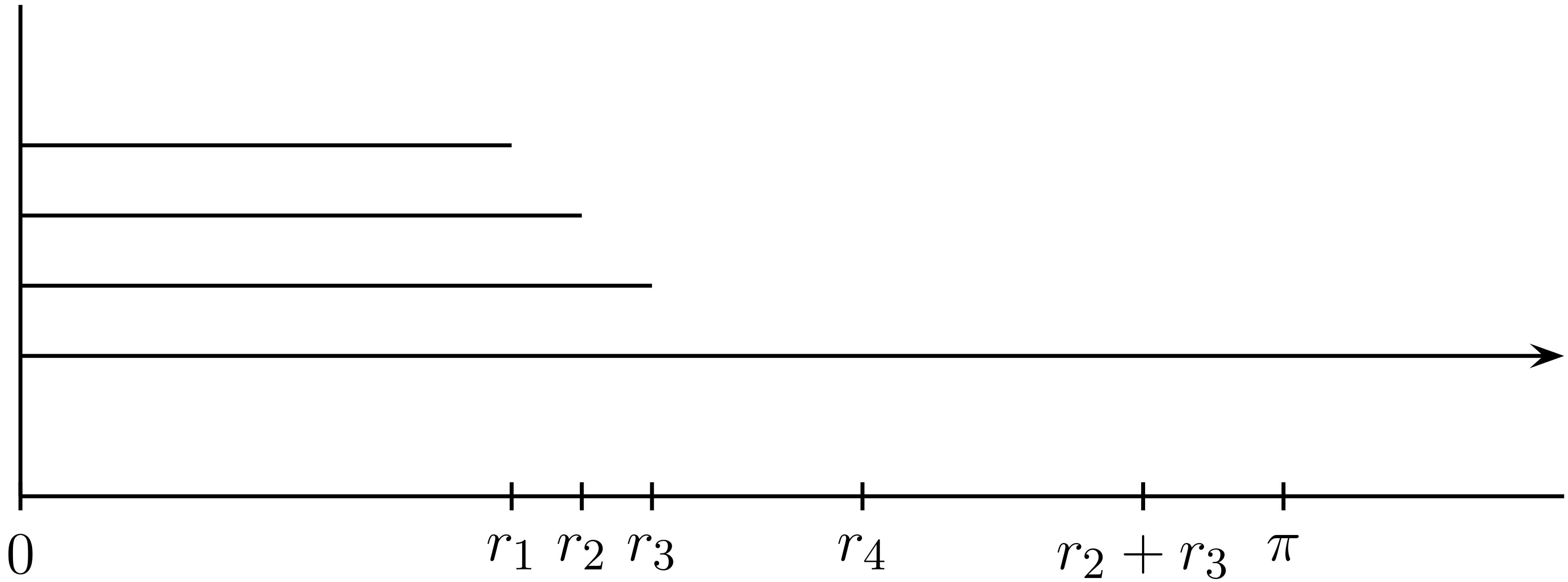
The Rips complex



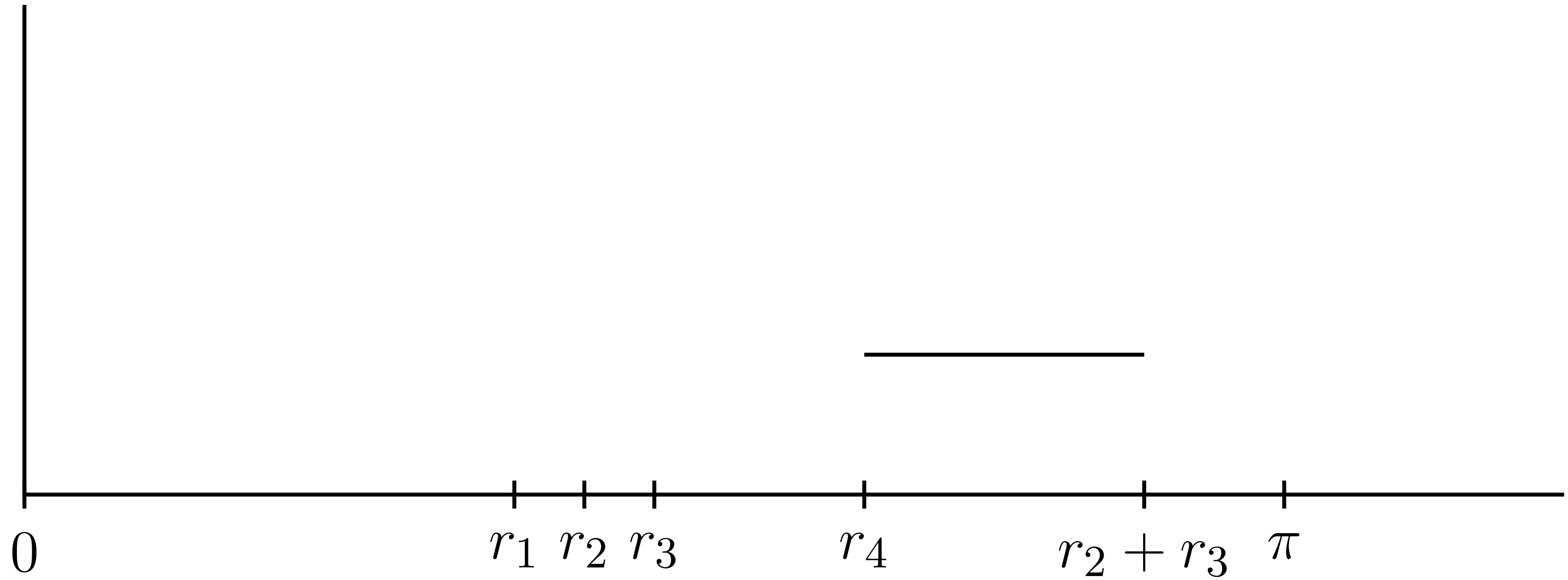
The Rips complex



Betti 0-barcode



Betti 1-barcode



V

5. Spacings

Let $X = (X_1, \ldots X_n) \subset S^1$ be a sample.

5. Spacings

Let $X = (X_1, \ldots X_n) \subset S^1$ be a sample.
Choose α such that $X_1 = e^{2\pi i \alpha}$.
For $2 \leq k \leq n$, choose $U_k \in [0, 1]$ such that

$$X_k = e^{2\pi i (\alpha + U_k)}.$$

5. Spacings

Let $X = (X_1, \ldots X_n) \subset S^1$ be a sample.
Choose α such that $X_1 = e^{2\pi i \alpha}$.
For $2 \leq k \leq n$, choose $U_k \in [0, 1]$ such that

$$X_k = e^{2\pi i (\alpha + U_k)}.$$

Order $\{U_2, \ldots U_n\}$ to obtain

$$0 < U_{n:1} < U_{n:2} < \ldots < U_{n:n-1} < 1.$$

Let $U_{n:0} = 0$ and $U_{n:n} = 1$.
For $1 \leq k \leq n$, let

$$S_k = U_{n:k} - U_{n:k-1}.$$

These are called the spacings.

Ordered Spacings

Order S:

$$0 < S_{n:1} < S_{n:2} < \ldots < S_{n:n} < 1$$

Ordered Spacings

Order S:

$$0 < S_{n:1} < S_{n:2} < \ldots < S_{n:n} < 1$$

Let $B_0^n(k)$ equal the k-th shortest interval in the Betti-0 barcode.

Ordered Spacings

Order S:

$$0 < S_{n:1} < S_{n:2} < \ldots < S_{n:n} < 1$$

Let $B_0^n(k)$ equal the k-th shortest interval in the Betti-0 barcode.

Therefore,

$$B_0^n(k) = 2\pi S_{n:k}.$$

Uniform spacings

If X is sampled uniformly from S^1 then each U_k is uniform on $[0,1]$ and $S = (S_1, \ldots, S_n)$ are called uniform spacings.

Assume X is sampled uniformly.

Uniform spacings

If X is sampled uniformly from S^1 then each U_k is uniform on $[0,1]$ and $S = (S_1, \ldots, S_n)$ are called uniform spacings.

Assume X is sampled uniformly.

Lemma: $S = (S_1, \ldots, S_n)$ has a constant density on the standard $(n-1)$-simplex Δ^{n-1}

$$S_k \geq 0, \quad \sum_{k=1}^{n} S_i = 1.$$

Expected spacings

For $1 \leq k \leq n$, $ES_{n:k}$ can be found be integrating over Δ^{n-1}.

Expected spacings

For $1 \le k \le n$, $ES_{n:k}$ can be found be integrating over Δ^{n-1}.

Proposition:

$$ES_{n:k} = \frac{1}{n} \sum_{j=n+1-k}^{n} \frac{1}{j}$$

Corollary:

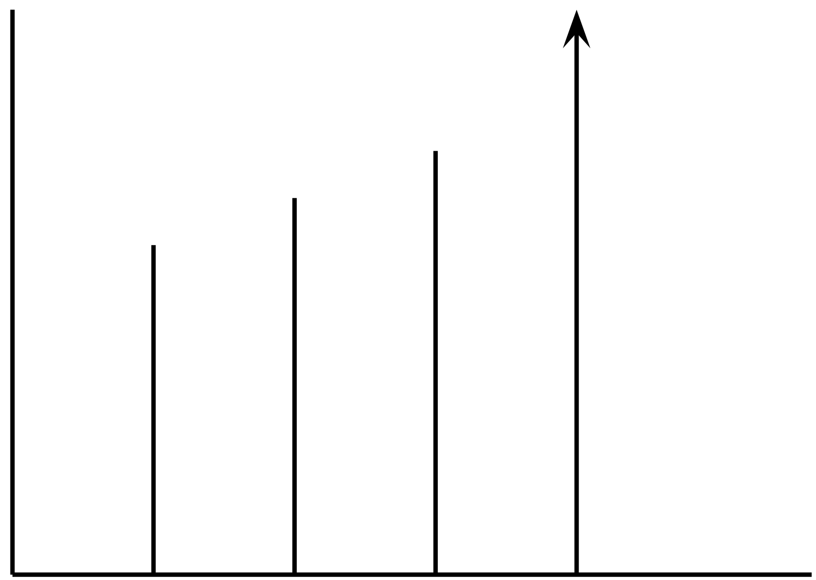
$$EB_0^n(k) = \frac{2\pi}{n} \sum_{j=n+1-k}^{n} \frac{1}{j}$$

6. Asymptotic behavior

We would like to understand what happens to the Betti barcodes as the number of samples n increases.

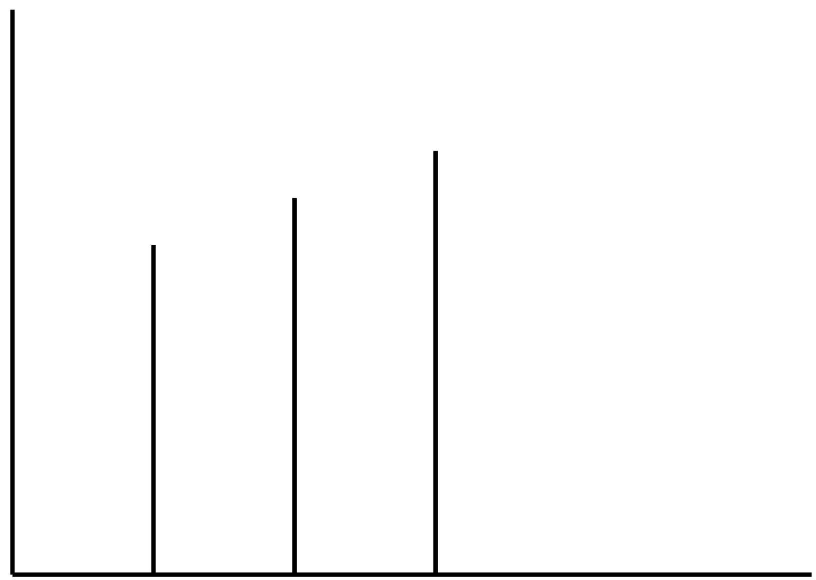
A different view

Consider the following view of the Betti-0 barcode.



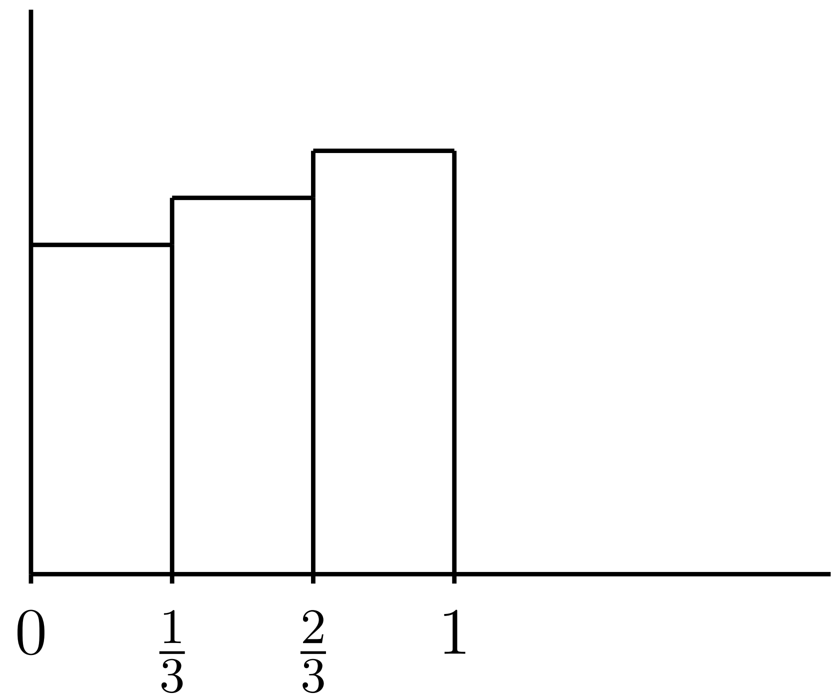
A different view

Consider the following view of the Betti-0 barcode.



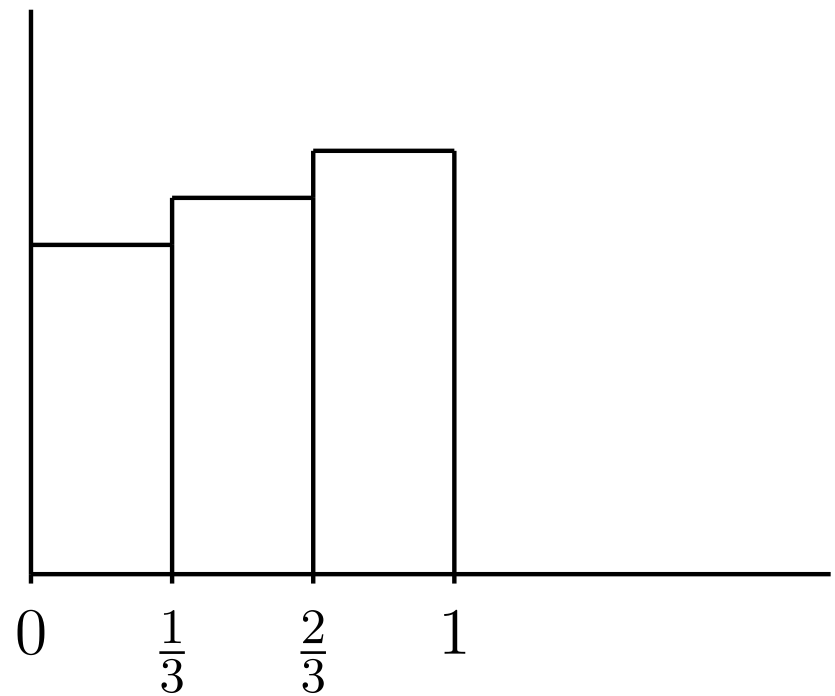
A different view

Consider the following view of the Betti-0 barcode.



A different view

Consider the following view of the Betti-0 barcode.



This gives us a function $B_0^n : [0,1] \longrightarrow \mathbf{R}_{\geq 0}$, where $B_0^n(x) = B_0^n(k)$, where $k = \lceil (n-1)x \rceil$.

Normalization

To see the effect of κ as $n \to \infty$, we normalize the above function.

$$\bar{B}_0^n(x) = \frac{B_0^n(x)}{\int_0^1 \; B_0^n(t)dt}$$

We call this the Betti-0 function.

Expected Betti-0 function

Recall that $B_0^n(x) = 2\pi S_{n:\lceil (n-1)x \rceil}$.

Thus,

$$\begin{aligned} E(B_0^n)(x) &= 2\pi E S_{n:\lceil (n-1)x \rceil} \\ &= \frac{2\pi}{n} \left(\sum_{j=n+1-\lceil (n-1)x \rceil}^{n} \frac{1}{j} \right). \end{aligned}$$

Expected Betti-0 function

Recall that $B_0^n(x) = 2\pi S_{n:\lceil (n-1)x \rceil}$.

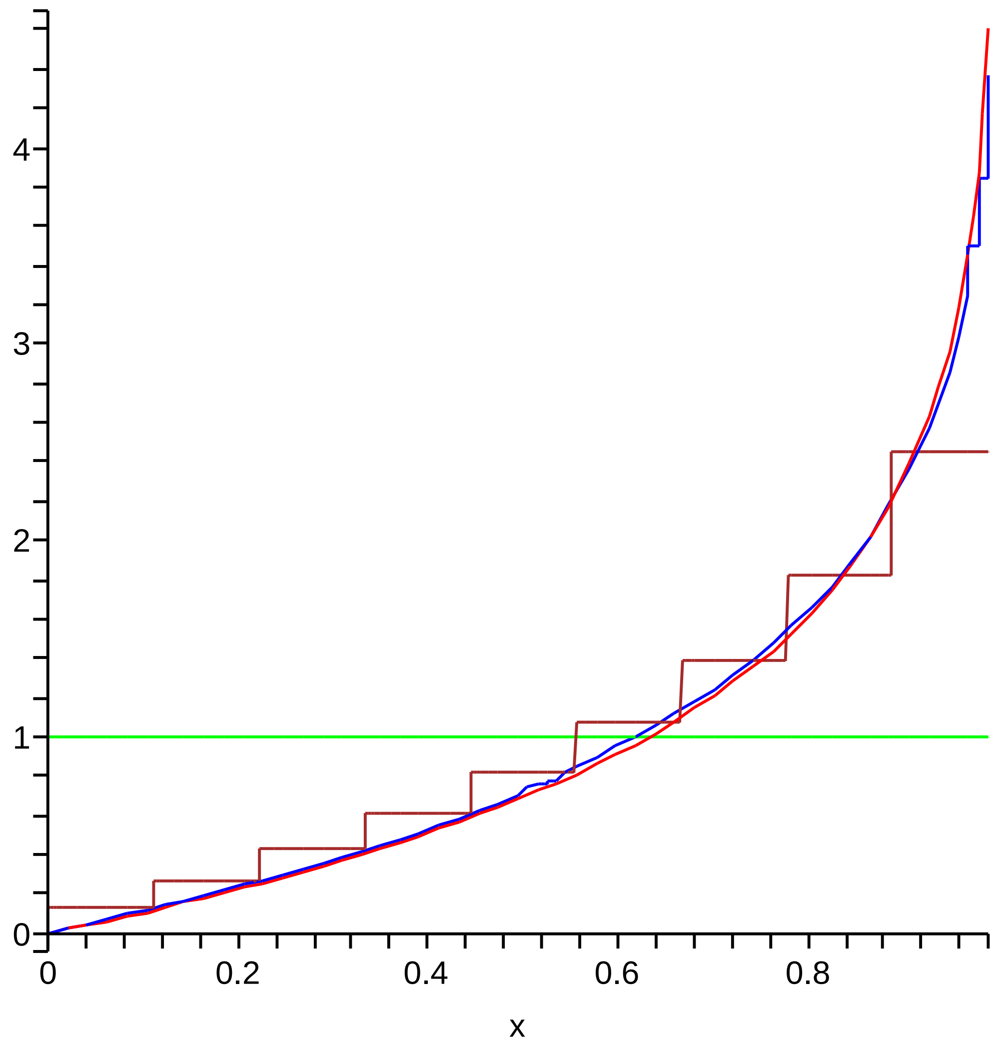
Thus,

$$\begin{aligned} E(B_0^n)(x) &= 2\pi E S_{n:\lceil (n-1)x \rceil} \\ &= \frac{2\pi}{n} \left(\sum_{j=n+1-\lceil (n-1)x \rceil}^{n} \frac{1}{j} \right). \end{aligned}$$

Proposition[B-K]:

$E(\bar{B}_0^n)(x) \to -\log(1-x)$ as $n \to \infty$.

Betti-0 Graphs



7. Topology of densities

We have also adapted persistent homology to study densities directly.

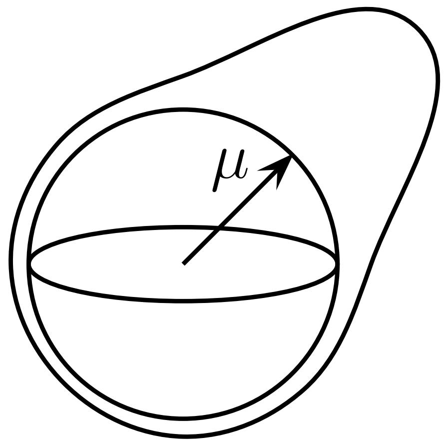
Recall that to use persistent homology one needs a parameter r.

We consider the portion of a manifold whose density is less that r, or alternately the portion whose density is greater than $\frac{1}{r}$.

Again we get Betti barcodes and Betti-0 functions.

The von Mises-Fischer distribution

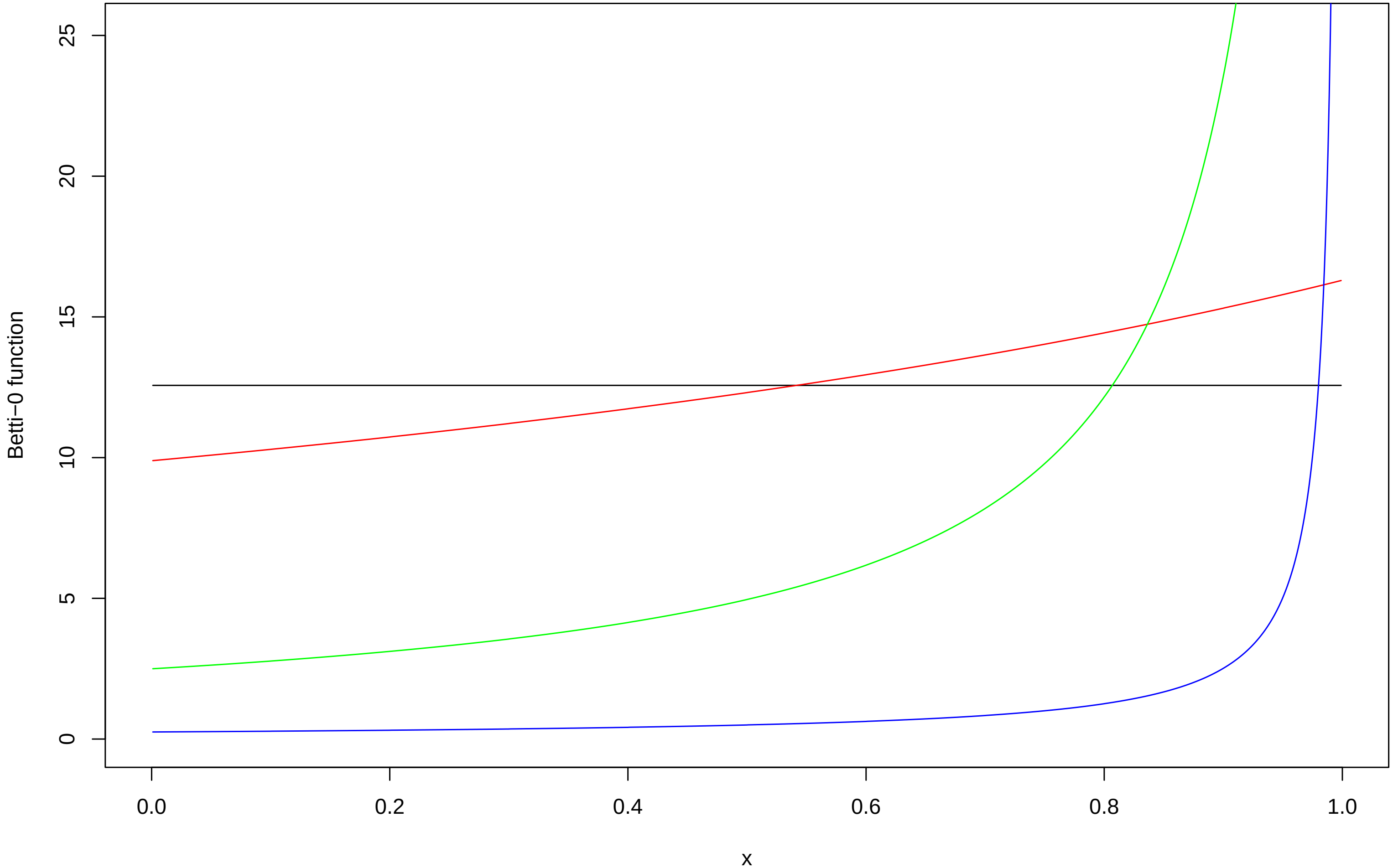
The von Mises-Fischer distribution gives unimodal density on S^{p-1}.



$$f_{\mu,\kappa}(x) = c(\kappa)e^{\kappa x^t \mu}.$$

For different values of κ we get different Betti barcodes and Betti-0 functions.

Betti-0 for various κ



The Betti-0 function for S^2 for $\kappa = 0, 0.25, 2.5, 25$

Maximum Likelihood Estimators

For PCD $x_1, \ldots, x_n$ on S^{p-1} sampled from the von Mises-Fischer distribution let $\bar{x}$ be the sample mean.

Maximum Likelihood Estimators

For PCD $x_1, \ldots, x_n$ on S^{p-1} sampled from the von Mises-Fischer distribution let $\bar{x}$ be the sample mean.

A statistical estimator for μ is

$$\frac{\bar{x}}{\|\bar{x}\|}.$$

A statistical estimator for κ is

$$\hat{\kappa} = A_p^{-1}(\|\bar{x}\|) \text{ where } A_p(\lambda) = \frac{I_{p/2}(\lambda)}{I_{p/2-1}(\lambda)}.$$

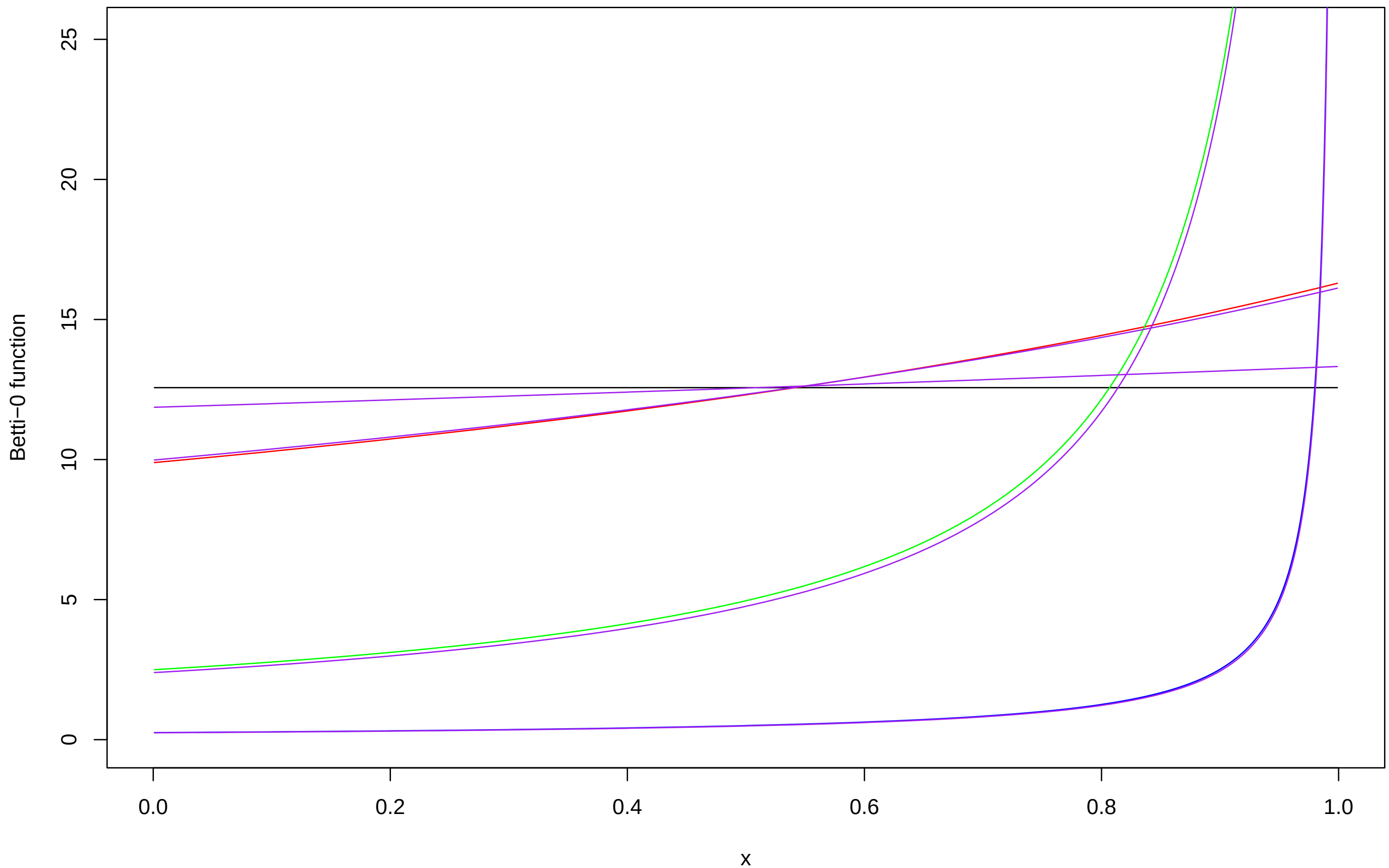
Example S^2

Simulating 1000 samples on S^2 from $\kappa = 0, 0.25, 2.5, 25$ we calculate the maximum likelihood estimators $\hat{\kappa}$.

We get $\hat{\kappa} = 0.058, 0.240, 2.611, 25.657$.

For these $\hat{\kappa}$, we can again calculate the Betti-0 functions.

Betti-0 for various $\kappa, \hat{\kappa}$



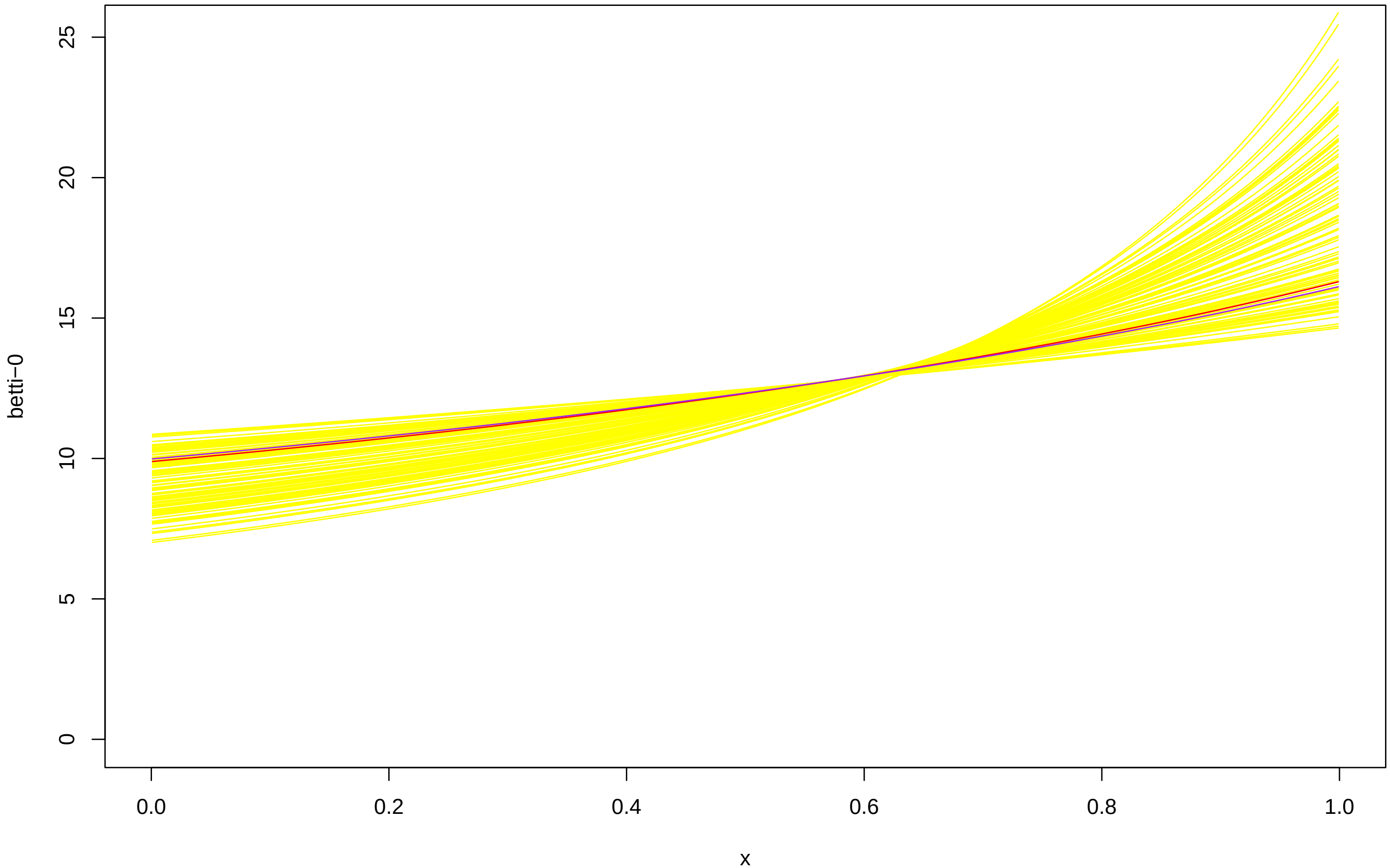
$$\kappa = 0, 0.25, 2.5, 25,\ \hat{\kappa} = 0.058, 0.240, 2.611, 25.657$$

Bootstrapping

Finally we sample from the maximum likelihood estimators, and again calculate the persistent homology.

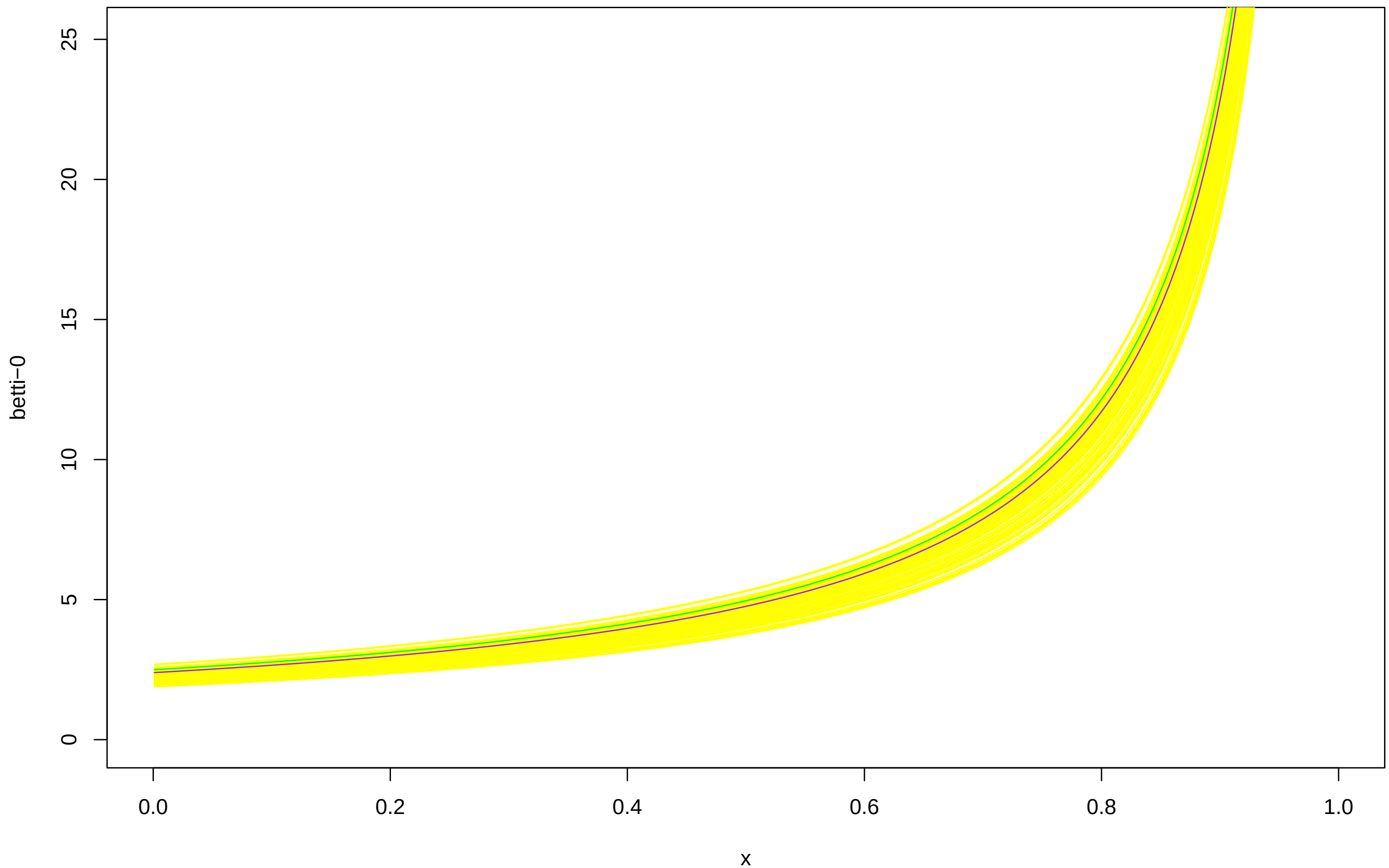
This provides confidence intervals on the Betti barcodes and Betti-0 functions.

Bootstrapping Betti-0



$\kappa = 0.25$, $\hat{\kappa} = 0.240$ and 95% bootstrap confidence interval

Bootstrapping Betti-0



$\kappa = 2.5$, $\hat{\kappa} = 2.611$ and 95% bootstrap confidence interval

Summary

- Persistent homology detects topological features of samples.
- Using the theory of spacings we were able to work out exact estimates of the Betti barcodes for samples on the uniform density on S^1.
- The asymptotic behaviour was $-\log(1-x)$.
- Applied directly to densities, persistent homology depends on the parameter.
- Using bootstrapping we can give confidence intervals on the Betti barcodes and the Betti-0 function of a sample.