

# A tutorial for topological data analysis 2

Otto van Koert

Seoul National University, South Korea

## From data to shape: invariants

Given a finite collection of data points  $\{p_i\}_{i=1}^N$  we want to understand the shape of the data. What does this mean?

- The data points are actually samples on a surface or other space: so we want to obtain this underlying space (which we want to think of as a manifold)
- The data points are approximated well by a “surface”

Here we describe some methods to reconstruct the shape and extract information.

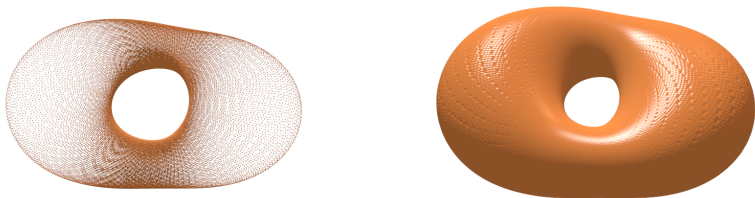


Figure: Points on a torus?

- We want to develop some tools to capture qualitative information of the shape of data
- We want this information to be robust against noise.

## Invariants of shapes

We want to associate invariants with shapes: these can be numbers, vector spaces, and even more general objects. This is the topic of algebraic topology.

### Example

Given a graph with vertices  $V$  and edges  $E$ , we have the following two invariants: the number of components and the number of loops: homology is one way to make this precise

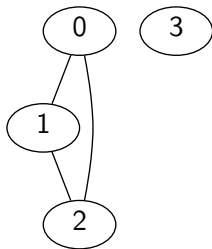


Figure: A graph with two components and a loop

## The idea of homology

Suppose that  $X$  is a topological space. Intuitively, the homology of  $X$  is a way to count “rooms” or “caves” that are in  $X$ .

- to make this a little bit more precise, recall the fundamental group. If  $[\gamma] \in \pi_1(X, x_0)$ , then  $[\gamma] = 0$  if and only if there is a disk  $D : D^2 \rightarrow X$  such  $\partial D = \gamma$ .
- in homology, we say that  $\gamma$  is null-homologous precisely when there is a surface  $\Sigma$  in  $X$  such that  $\partial \Sigma = \gamma$ .
- this makes homology more flexible
- the easiest way to include arbitrary surfaces, is to allow simplices: these can be combined to more complicated objects.

The easiest setting to make this precise (both mathematically and for the computer) is when  $X$  is a simplicial complex.

Formally, we fix a ring or field  $F$ . This is usually  $\mathbb{Z}_2$ ,  $\mathbb{Z}_p$  or  $\mathbb{Q}$ , but in an AT-course, it is the ring  $\mathbb{Z}$ . We also fix an abstract simplicial complex  $X$ , which we secretly identify with its geometric realization  $|X|$ .

- consider the  $F$ -vector space  $C_k(X; F)$  freely generated by the  $k$ -simplices of  $X$ , so

$$C_k(X; F) := \bigoplus_{\sigma \text{ } k\text{-simplex}} F\sigma.$$

We will call elements of  $C_k(X; F)$   $k$ -chains. These are just finite sums of  $k$ -simplices.

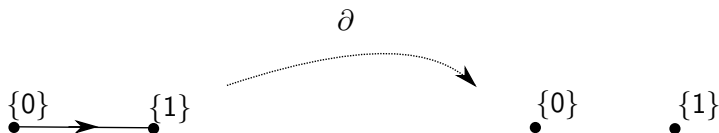
### Example

Consider  $X_0 = \{\{0\}, \{1\}, \{2\}\}$ ,  $X_1 = \{\{0, 1\}, \{1, 2\}\}$ . Then  $C_0 = \mathbb{R}\{0\} \oplus F\{1\} \oplus F\{2\} \cong F^3$ . Similarly,  $C_1 \cong F^2$

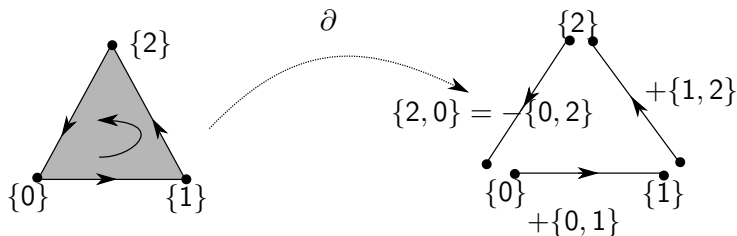
In other words, 0-chains are just combinations of points, 1-chains are combinations of edges, 2-chains are combinations of triangles and so on.

## Boundary operator

Given a  $k$ -simplex, we can look at its boundary



In a formula, we have  $\partial\{0, 1\} = +\{1\} - \{0\}$ . Here is another example.



In the last case, the formula is

$$\partial\{0, 1, 2\} = \{0, 1\} + \{1, 2\} + \{2, 0\} = +\{1, 2\} - \{0, 2\} + \{0, 1\}$$

In general, if  $\sigma$  is a  $k$ -simplex, then we have

$$\partial\sigma = \sum_{j=0}^k (-1)^j \hat{\sigma}(j).$$

Here  $\hat{\sigma}(j)$  is the  $k - 1$ -simplex obtained from  $\sigma$  by removing the  $j$ -th vertex.

### Remark

*Since the  $k$ -simplices form a basis of  $C_k$ , we can extend  $\partial$  uniquely to a linear map  $\partial_k : C_k \rightarrow C_{k-1}$ . This is the **boundary operator**.*

The collection of pairs of vector spaces and boundary operators  $\bigoplus_k (C_k, \partial_k)$  is called a **chain complex**.

## An example and the boundary of a boundary vanishes

### Example

Consider  $X_0 = \{\{0\}, \{1\}, \{2\}\}$ ,  $X_1 = \{\{0, 1\}, \{0, 2\}, \{1, 2\}\}$   
and  $X_2 = \{\{0, 1, 2\}\}$ . Then  $C_2 \cong F$ , and  $C_1 \cong F^3$ . We find

$$\partial_2\{0, 1, 2\} = +\{1, 2\} - \{0, 2\} + \{0, 1\}$$

or in a matrix representation (with respect to the basis given  
above)

$$\partial_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

For  $\partial_1 : C_1 \cong F^3 \rightarrow C_0 \cong F^3$ , we do the same: work on the basis of simplices

$$\partial_1\{0, 1\} = +\{1\} - \{0\}, \quad \partial_1\{0, 2\} = +\{2\} - \{0\},$$

$$\partial_1\{1, 2\} = +\{2\} - \{1\}.$$

The matrix representation is hence

$$\partial_1 = \begin{pmatrix} -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{pmatrix}$$

We note that  $\partial_1 \circ \partial_2 = 0$ .

This is no coincidence:

### Theorem

*If  $(\bigoplus_k C_k, \partial)$  is a simplicial complex, then  $\partial_{k-1} \circ \partial_k = 0$ .*

## Proof.

We verify this on a basis. Take  $\sigma \in X_k$

$$\begin{aligned}
 \partial_{k-1} \circ \partial_k \sigma &= \partial_{k-1} \left( \sum_j (-1)^j \hat{\sigma}(j) \right) = \sum_j (-1)^j \partial_{k-1} \hat{\sigma}(j) \\
 &= \sum_j (-1)^j \sum_{i < j} (-1)^i \hat{\sigma}(j, i) + \\
 &\quad \sum_j (-1)^j \sum_{i \geq j, i < k} (-1)^i \hat{\sigma}(j, \substack{i+1 \\ \text{as } j \text{ is already removed}}) \\
 &= \sum_j (-1)^j \sum_{i < j} (-1)^i \hat{\sigma}(j, i) + \\
 &\quad \sum_j (-1)^j \sum_{i' > j} (-1)^{i'-1} \hat{\sigma}(j, i') = 0
 \end{aligned}$$

Last step holds since both the first and second sum remove two distinct vertices from  $\sigma$ , but with opposite signs. □

## Cycles and boundaries

We need a couple more definitions. As before, we consider a simplicial complex  $(C_*, \partial)$

### Definition

By a  **$k$ -cycle** we mean a  $k$ -chain  $c$  such  $\partial_k c = 0$ . By a **boundary of dimension  $k$**  we mean a  $k$ -chain  $b$  such that there is  $c'$  with  $b = \partial_{k+1} c'$ .

In other words, the space of  $k$ -cycles  $Z_k$  is hence the subspace

$$Z_k = \text{Null}(\partial_k)$$

and the space of  $k$ -boundaries  $B_k$  is the subspace

$$B_k = \text{Range}(\partial_{k+1}).$$

According to the theorem, a boundary is always a cycle, so  $B_k$  is a subspace in  $Z_k$ .

# Homology

## Definition

The  $k$ -th homology vector space associated with the abstract simplicial complex  $X$  is the quotient space

$$H_k(X) := H_k := Z_k/B_k$$

We see that  $\dim H_k = \dim Z_k - \dim B_k$ , so we need only to compute the nullspace and ranges of the boundary operators.

## Remark

*From the definition we see that homology measures those cycles that cannot be filled in by boundaries.*

## Example

We consider the abstract simplicial complex

$X_0 = \{\{0\}, \{1\}, \{2\}\}$ ,  $X_1 = \{\{0, 1\}, \{0, 2\}, \{1, 2\}\}$  and  
 $X_2 = X_3 = \dots = \emptyset$ . We know  $\partial_0 = 0$ , so  $Z_0 = C_0 \cong F^3$

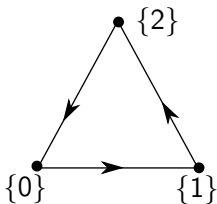
Furthermore  $\text{rk } \partial_1 = 2$ , and  $\dim \text{Null } \partial_1 = 1$ , so  $Z_1 \cong F$ . So we  
find

$$\dim H_0 = \dim Z_0 - \text{rk } \partial_1 = 3 - 2 = 1.$$

and

$$\dim H_1 = \dim Z_1 - \text{rk } \partial_2 = 1 - 0 = 1,$$

as  $\partial_2 = 0$  ( $C_2$  is 0-dimensional).



## Interpretation of homology

We have the following intuitive interpretation:

- the number of components ( $= \dim H_0$ )
- the number of independent loops ( $= \dim H_1$ ) that cannot be filled in
- more generally, the number of holes or rooms in the simplicial complex ( $= \dim H_k$  for  $k$  – *dimensional* rooms).

These dimensions are called **Betti numbers**, i.e. the  $k$ -th Betti number is defined as

$$b_k := \dim H_k.$$

### Remark

*Alternatively, think of  $b_k$  as the number of caves with  $k$ -dimensional walls...*

## More complicated homology classes

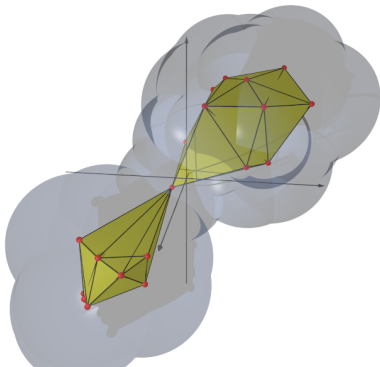


Figure: Rooms with 2-dimensional walls (elements in  $H_2$ )

## Data and simplicial complexes

A finite collection of data points has no intrinsic shape.  
To deal with this,

- measure distance (or dissimilarity) between data points.  
This can be very general, eg.  $d(\text{yes}, \text{yes}) = 0$ ,  
 $d(\text{yes}, \text{no}) = 1$ : we get a finite metric space.
- given a scale parameter  $\varepsilon$ , add a  $k$ -simplex  $\sigma$  if each of the vertices of  $\sigma$  lies within distance  $\varepsilon$ .
- depending on the scale parameter, we get different simplicial complexes.
- compute the homology of the different complexes. How does it vary with the scale parameter?

## Rips complex

Formally, we have for any metric space  $(M, d)$ :

### Definition

The Vietoris-Rips complex of  $(M, d)$  at threshold  $\varepsilon$  is the simplicial complex  $VR(M, \varepsilon)$  whose

- vertices are points in  $M$
- a  $k + 1$ -tuple of distinct points  $\{x_0, \dots, x_k\}$  spans a  $k$ -simplex precisely when  $d(x_i, x_j) < \varepsilon$  for  $i, j = 0, \dots, k$ .

Note that the Rips-complex is huge if  $M$  has many points.

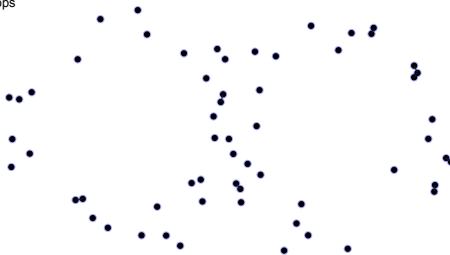
### Remark

*We will now apply the Rips-construction to finite metric spaces. By looking at the Rips-complex at varying scales (different  $\varepsilon$ ) we can understand the shape of data.*

## Visualizing the rips complex

To minimize clutter, we draw balls of radius  $\varepsilon/2$  rather than  $\varepsilon$ .

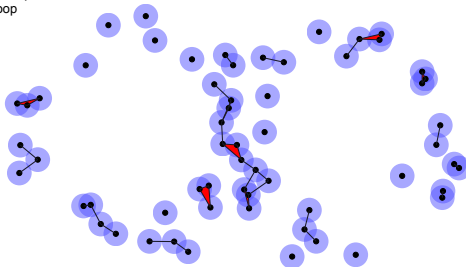
52 components  
0 loops



## Visualizing therips complex

To minimize clutter, we draw balls of radius  $\varepsilon/2$  rather than  $\varepsilon$ .

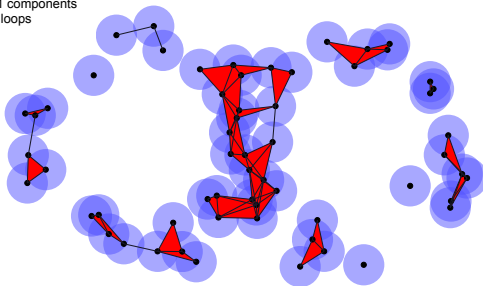
26 components  
1 loop



## Visualizing the rips complex

To minimize clutter, we draw balls of radius  $\varepsilon/2$  rather than  $\varepsilon$ .

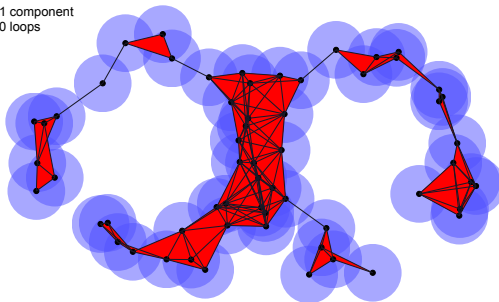
11 components  
2 loops



## Visualizing the rips complex

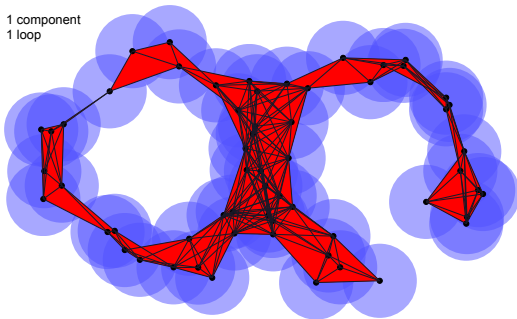
To minimize clutter, we draw balls of radius  $\varepsilon/2$  rather than  $\varepsilon$ .

1 component  
0 loops



## Visualizing the rips complex

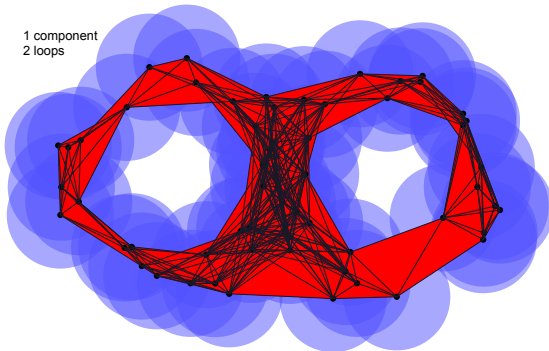
To minimize clutter, we draw balls of radius  $\varepsilon/2$  rather than  $\varepsilon$ .



## Visualizing the rips complex

To minimize clutter, we draw balls of radius  $\varepsilon/2$  rather than  $\varepsilon$ .

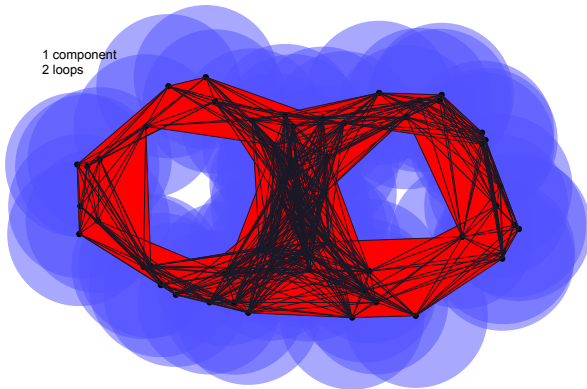
1 component  
2 loops



## Visualizing therips complex

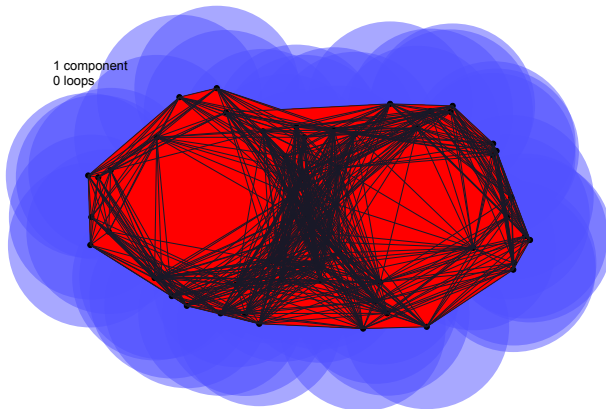
To minimize clutter, we draw balls of radius  $\varepsilon/2$  rather than  $\varepsilon$ .

1 component  
2 loops



## Visualizing the rips complex

To minimize clutter, we draw balls of radius  $\varepsilon/2$  rather than  $\varepsilon$ .

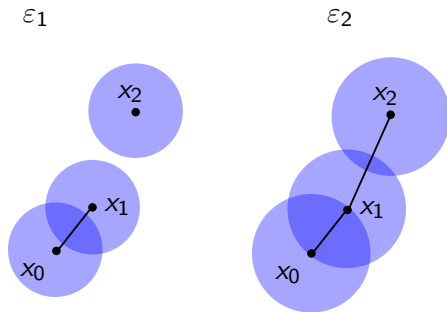




## Nice properties of inclusion maps

In order to see what shapes survive for a long time/parameter scale, we consider the inclusion map

$$f^{\varepsilon_1, \varepsilon_2} : VR(XM, \varepsilon_1) \rightarrow VR(M, \varepsilon_2)$$



This works because if  $\sigma$  is a simplex in  $VR(X, \varepsilon_1)$ , then it is also a simplex in  $VR(X, \varepsilon_2)$  with  $\varepsilon_2 \geq \varepsilon_1$ .

If we denote the Rips chain complex at filtration level  $\varepsilon$  (this is the scale parameter we mentioned earlier) by  $(C_k^\varepsilon, \partial_k^\varepsilon)$ , then we find

$$f_k^{\varepsilon_1, \varepsilon_2} : C_k^{\varepsilon_1} \longrightarrow C_k^{\varepsilon_2}$$

Again, these are linear maps determined by their effect on a basis.

So to get a matrix representation, choose a basis  $\sigma_1, \dots, \sigma_{n_k(\varepsilon_1)}$  of  $C_k^{\varepsilon_1}$ . Extend this to a basis of  $C_k^{\varepsilon_2}$ . Then

$$[f_k^{\varepsilon_1, \varepsilon_2}] = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

So  $[f_k^{\varepsilon_1, \varepsilon_2}]$  is

- an  $n_k(\varepsilon_2) \times n_k(\varepsilon_1)$ -matrix,
- whose upper  $n_k(\varepsilon_1) \times n_k(\varepsilon_1)$ -block is the identity, and its remaining entries are 0.

## Lemma

If  $\varepsilon_3 \geq \varepsilon_2 \geq \varepsilon_1$ , then  $f_k^{\varepsilon_2, \varepsilon_3} \circ f_k^{\varepsilon_1, \varepsilon_2} = f_k^{\varepsilon_1, \varepsilon_3}$ .

## Proof.

Use the observations we just made. □

The above leads to the notion of persistence vector space. To keep things concrete, we keep the following example in mind.

## Example

In  $F^2$ , consider the data points

$$p_0 = (0, 0), p_1 = (1, 0), p_2 = (1, 1), p_3 = (0, 1)$$

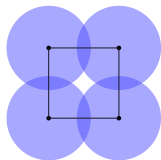


Figure: The Rips complex for the square for filtration level 1.2

## Definition

A **persistence vector space** consists of

- 1 a family of real vector spaces  $\{V^r\}_{r \in \mathbb{R}_{\geq 0}}$
- 2 for  $r' \geq r$ , linear maps  $f^{r,r'} : V^r \rightarrow V^{r'}$  satisfying

$$f^{r_2,r_3} \circ f^{r_1,r_2} = f^{r_1,r_3}$$

Clearly for each  $k$ , the vector space from the Rips complex,  $\{C_k^r\}_{r \in \mathbb{R}_{\geq 0}}$  and inclusion maps  $f^{r,r'}$ , satisfy the required properties.

## Theorem

*The Rips-complex forms a persistence vector space.*

## More on persistence vector spaces

A basic example is the following

$$\{\mathcal{P}(a, b)\}^r = \begin{cases} F & r \in [a, b) \\ 0 & \text{otherwise.} \end{cases}$$

with maps  $f^{r_1, r_2}$  satisfying

$$f^{r_1, r_2} = \begin{cases} id & r_1, r_2 \in [a, b) \\ 0 & \text{otherwise.} \end{cases}$$

We can efficiently encode by drawing the interval  $[a, b)$ . We can obviously take direct sums of these spaces, and these can be encoded by *barcodes*.

The Rips complex for the “square” can now be written as

$$C_0 = \mathcal{P}(0, \infty)\{0\} \oplus \mathcal{P}(0, \infty)\{1\} \oplus \mathcal{P}(0, \infty)\{2\} \oplus \mathcal{P}(0, \infty)\{3\}.$$

The vector space of 1-chains is given by

$$C_1 = \mathcal{P}(1, \infty)\{0, 1\} \oplus \mathcal{P}(1, \infty)\{1, 2\} \oplus \mathcal{P}(1, \infty)\{2, 3\} \\ \oplus \mathcal{P}(1, \infty)\{0, 3\} \oplus \mathcal{P}(\sqrt{2}, \infty)\{0, 2\} \oplus \mathcal{P}(\sqrt{2}, \infty)\{1, 3\}.$$

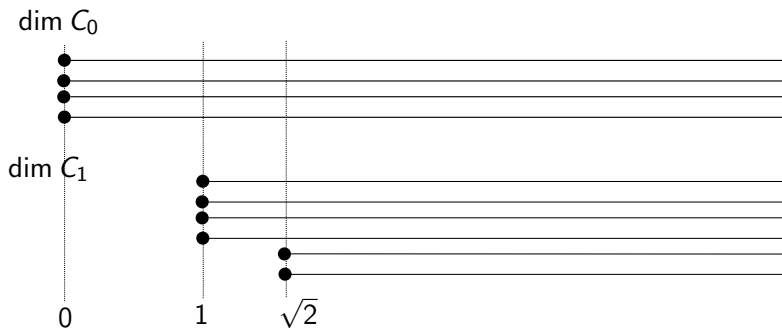


Figure: The barcodes for the previous example

## Example

We consider again the data points

$p_0 = (0, 0)$ ,  $p_1 = (1, 0)$ ,  $p_2 = (1, 1)$ ,  $p_3 = (0, 1)$ , and consider the Vietoris-Rips complex. The persistence homology vector spaces are given by

$$H_0^r \cong (\mathcal{P}(0, 1) \oplus \mathcal{P}(0, 1) \oplus \mathcal{P}(0, 1) \oplus \mathcal{P}(0, \infty))^r$$

$$H_1^r \cong \left( \mathcal{P}(1, \sqrt{2}) \right)^r$$

We find the following barcode for the “square”



Figure: Barcode for the square

## Persistent homology of noisy data

What shape do the following noisy looking points have?



Figure: Random points?

We compute the persistent homology and obtain the following barcodes. The long bars are similar to those of the square.

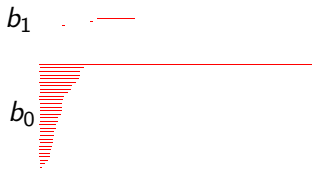


Figure: Barcode for the noisy circle

## Computations

At first sight computations seem daunting.

- computing homology at one threshold value involves computing ranks of fairly big matrices (depending on the data).
- many threshold parameters seem even harder
- we also need to compute the effect of the inclusion maps

There is both good and bad news. First the good news:

- Zomorodian and Carlsson have developed an algorithm to compute persistent homology for all threshold parameters at the same time. The (simplified) idea is to do reduction (as in Gauss elimination) in a more general setting, namely not over  $F$ , but over the polynomial ring  $F[t]$ , where the power  $a$  in  $t^a$  is the filtration degree.
- complexity of Gauss elimination is cubic number of field operations depending on the number of simplices

## Drawbacks

- 1 the main weakness is that the Vietoris-Rips complex is very large.
- 2 if there are  $n$  data points, then the full complex has  $2^n - 1$  simplices
- 3 even if we truncate the complex (after all, who cares about  $H_{100}(X)$ ?), the complex is still too large

There are ways to deal with this issue. One idea is to use the so-called witness complex. Select a few data points that are representative of the data, and work with those rather than with the full set.

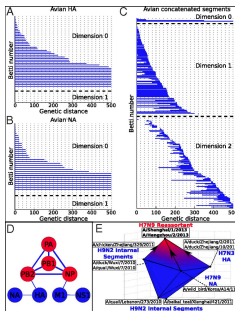
### Remark

*Despite these difficulties, there are remarkable applications.*

# Applications

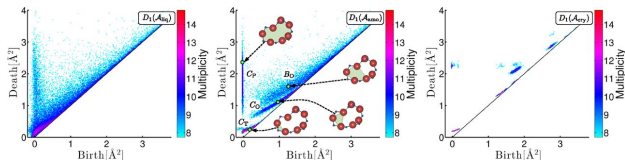
In *Topology of viral evolution* by Joseph Minhow Chan, Gunnar Carlsson, and Raul Rabadan, PNAS November 12, 2013. 110 (46) 18566-18571

- the “tree of life” turns out to be the “graph of life”. In addition to vertical evolution, there is also horizontal evolution in viruses leading to loops in persistent homology.



# Applications

*Hierarchical structures of amorphous solids characterized by persistent homology* by Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G. Escobar, Kaname Matsue, and Yasumasa Nishiura, PNAS June 28, 2016. 113 (26) 7035-7040;



From left to right: liquids, amorphous solids (glasses) and crystals

Thank you  
감사 합니다.