Topological data analysis and the construction of intrinsic circle-valued coordinates.

Mikael Vejdemo-Johansson

11-1-11

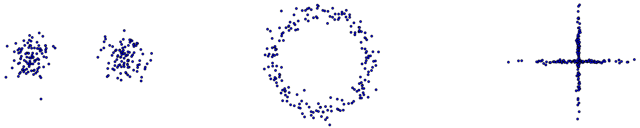
Data has shape

What is data?

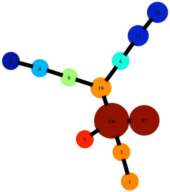
Data comes as numerical values: for instance physiological measurements from patients in a study.

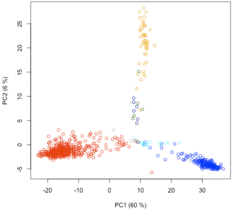
Captured as point clouds in $\mathbb{R}^d$.

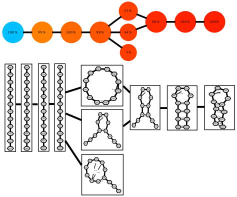
What is shape?



Shape matters







Homology

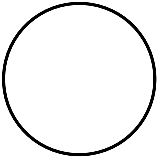
One major tool for describing these shapes comes from topology:

The ith homology with coefficients in a field k assigns to a topological space X a vector space $H_i(X; k)$.

Easiest description is through Betti numbers $\beta_i = \dim_k H_i(X; k)$.
Counts the number of i-dimensional voids. (almost)

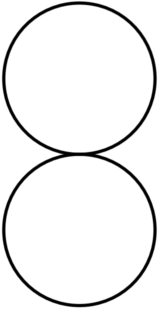
Pleasant to use because computable with matrix arithmetic.

Homology – intuitively



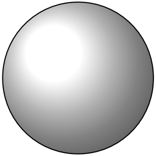
$\beta_0 = 1$

$\beta_1 = 1$



$\beta_0 = 1$

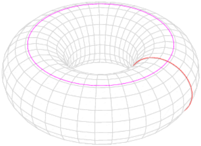
$\beta_1 = 2$



$\beta_0 = 1$

$\beta_1 = 0$

$\beta_2 = 1$



$\beta_0 = 1$

$\beta_1 = 2$

$\beta_2 = 1$

Homology — why algebra?

Even if we only work with β_i, the algebra provided by using vector spaces remains important.

At the core: Noether's principle. Along topological maps, the homology groups change with linear maps.

$$X \xrightarrow{f} Y$$

$$H_i(X; k) \xrightarrow{H_i(f;k)} H_i(Y; k)$$

Vector space structures carry additional information that can be leveraged for computation or analysis. This functoriality property will reappear later.

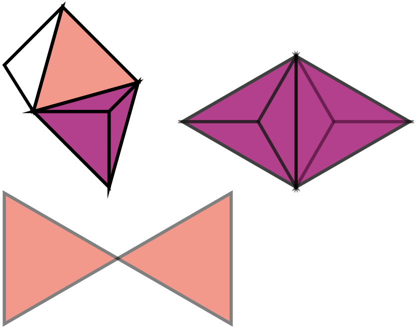
Simplicial topology: continuous made discrete

Definition

A simplicial complex is a family of simplices: vertices, edges, triangles, tetrahedra, . . . – such that any two simplices intersect in a subsimplex.

Definition

An abstract simplicial complex is a family of subsets of a given set V, such that all subsets of a member are members.



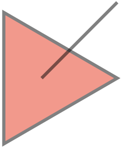
Simplicial topology: continuous made discrete

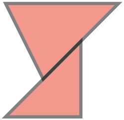
Definition

A simplicial complex is a family of simplices: vertices, edges, triangles, tetrahedra, . . . – such that any two simplices intersect in a subsimplex.

Definition

An abstract simplicial complex is a family of subsets of a given set V, such that all subsets of a member are members.



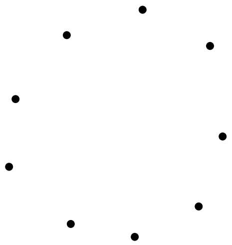


Simplicial complexes: discrete made continuous

Definition

The Vietoris-Rips complex is an abstract simplicial complex $VR_\epsilon(X)$ for $\epsilon \in \mathbb{R}_+$ and X a finite metric space:

- Contains one vertex for each element in X.
- Contains a simplex $(x_0, \ldots, x_k)$ exactly when $d(x_i, d_j) < \epsilon$ for all $i, j \in [k]$.

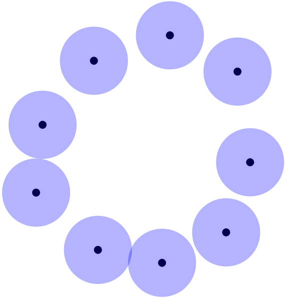


Simplicial complexes: discrete made continuous

Definition

The Vietoris-Rips complex is an abstract simplicial complex $VR_\epsilon(X)$ for $\epsilon \in \mathbb{R}_+$ and X a finite metric space:

- Contains one vertex for each element in X.
- Contains a simplex $(x_0, \ldots, x_k)$ exactly when $d(x_i, d_j) < \epsilon$ for all $i, j \in [k]$.

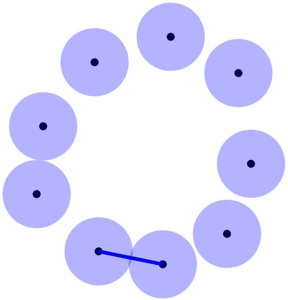


Simplicial complexes: discrete made continuous

Definition

The Vietoris-Rips complex is an abstract simplicial complex $VR_\epsilon(X)$ for $\epsilon \in \mathbb{R}_+$ and X a finite metric space:

- Contains one vertex for each element in X.
- Contains a simplex $(x_0, \ldots, x_k)$ exactly when $d(x_i, d_j) < \epsilon$ for all $i, j \in [k]$.

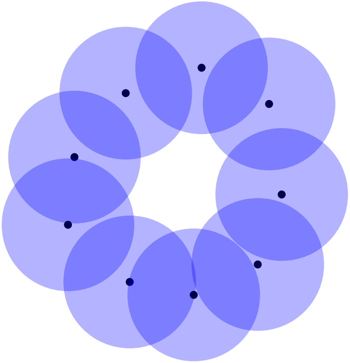


Simplicial complexes: discrete made continuous

Definition

The Vietoris-Rips complex is an abstract simplicial complex $VR_\epsilon(X)$ for $\epsilon \in \mathbb{R}_+$ and X a finite metric space:

- Contains one vertex for each element in X.
- Contains a simplex $(x_0, \ldots, x_k)$ exactly when $d(x_i, d_j) < \epsilon$ for all $i, j \in [k]$.

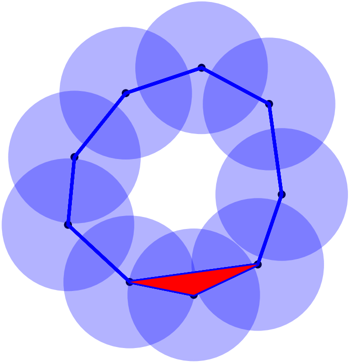


Simplicial complexes: discrete made continuous

Definition

The Vietoris-Rips complex is an abstract simplicial complex $VR_\epsilon(X)$ for $\epsilon \in \mathbb{R}_+$ and X a finite metric space:

- Contains one vertex for each element in X.
- Contains a simplex $(x_0, \ldots, x_k)$ exactly when $d(x_i, d_j) < \epsilon$ for all $i, j \in [k]$.



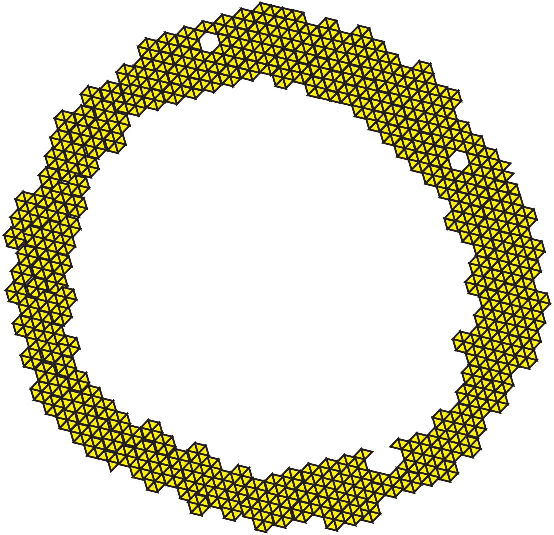
Computing homology

Given a simplicial complex S we can compute its homology using matrix operations.

- To S we assign a vector space $CS = \bigoplus_{\sigma \in S} \sigma \cdot k$.
- On CS we define a linear boundary map $\partial : CS \to CS$. Each simplex is mapped to a (signed) sum of the maximal simplices on its boundary.
- From the algebra (and geometry) at hand follows $\partial(\partial\sigma) = 0$ for all simplices σ. So $\partial(S) \subseteq \ker \partial$.
- We define the homology $H(S; k) = \ker \partial / \partial(S)$.
- Restricting to i-dimensional simplices yields $H_i(S; k)$; the i-dimensional homology group.

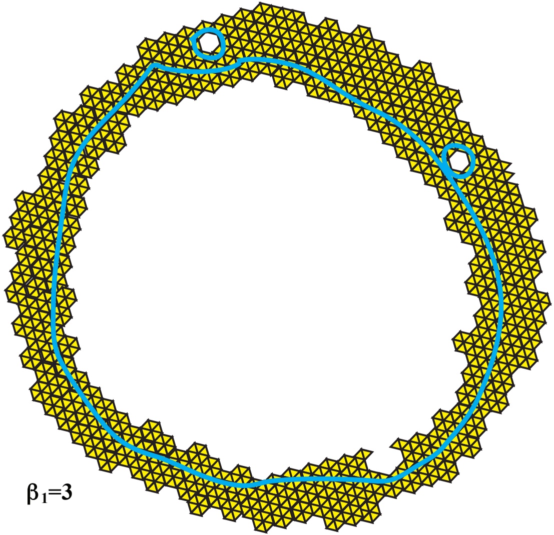
Example: pointcloud of a circle

First idea: pick some nice ϵ to work with $VR_\epsilon(X)$.



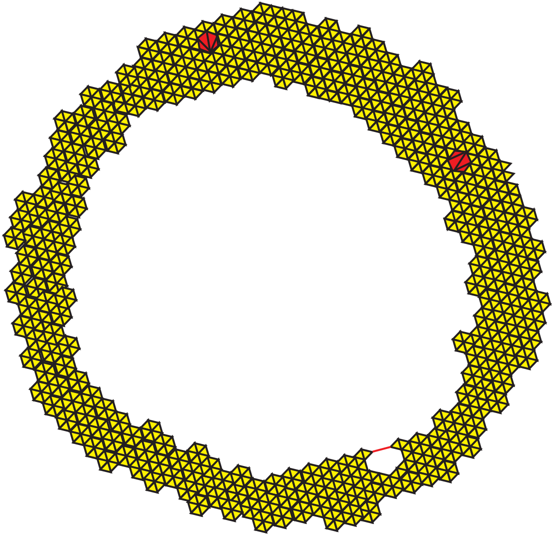
Example: pointcloud of a circle

First idea: pick some nice ϵ to work with $VR_\epsilon(X)$.



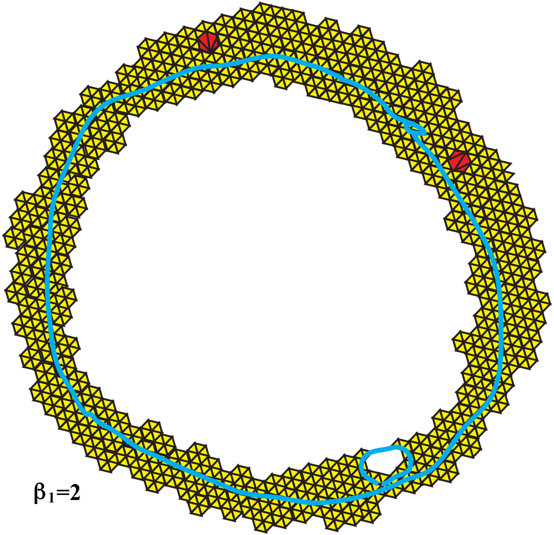
Example: pointcloud of a circle

First idea: pick some nice ϵ to work with $VR_\epsilon(X)$.



Example: pointcloud of a circle

First idea: pick some nice ϵ to work with $VR_\epsilon(X)$.



Example: pointcloud of a circle

Better approach: study changes in $H_i(VR_\epsilon(X); k)$ for different values of ϵ.

If $\epsilon < \epsilon'$, then $VR_\epsilon(X) \subset VR_{\epsilon'}(X)$. By functoriality of H_i, the inclusion map of simplicial complexes induces a map $H_i(VR_\epsilon(X); k) \to H_i(VR_{\epsilon'}(X); k)$.

We can summarize with a diagram of vector spaces and linear maps

$$H_i(VR_{\epsilon_0}(X)) \to H_i(VR_{\epsilon_1}(X)) \to \cdots \to H_i(VR_{\epsilon_k}(X))$$

A diagram like this we'll call a persistent vector space.

Some algebra

There is an equivalence between persistent vector spaces and graded $k[t]$-modules.

$$V_0 \xrightarrow{\iota} V_1 \xrightarrow{\iota} \dots \xrightarrow{\iota} V_k \qquad \Rightarrow \qquad \bigoplus_i V_i \quad =: V_*$$

The module structure is given by determining the action of t.

$$t \cdot (v_0, v_1, \dots, v_k) = (0, \iota v_0, \iota v_1, \dots, \iota v_{k-1})$$

Some algebra

The ring $k[t]$ is a graded PID, and thus graded modules over $k[t]$ have unique decompositions:

$$V_* = \bigoplus_i t^{a_i} k[t] \oplus \bigoplus_j t^{b_j} k[t]/t^{c_j}$$

Some algebra

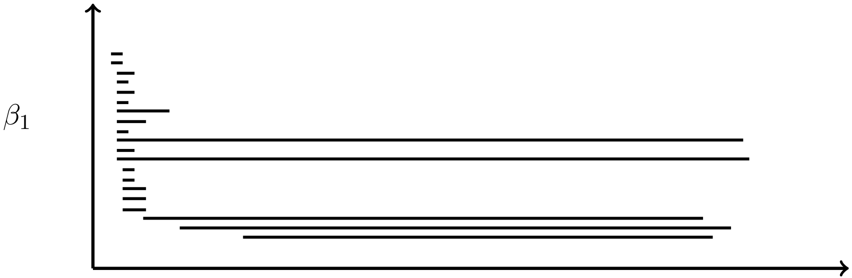
The ring $k[t]$ is a graded PID, and thus graded modules over $k[t]$ have unique decompositions:

$$V_* = \bigoplus_i \underbrace{t^{a_i} k[t]}_{[a_i,\infty)} \oplus \bigoplus_j \underbrace{t^{b_j} k[t]/t^{c_j}}_{[b_j,b_j+c_j)}$$

Some algebra

The ring $k[t]$ is a graded PID, and thus graded modules over $k[t]$ have unique decompositions:

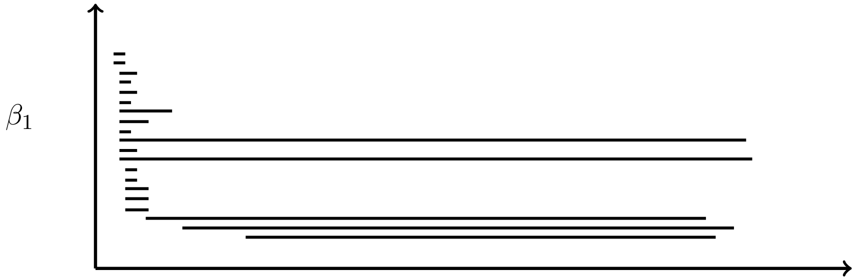
$$V_* = \underbrace{\bigoplus_i t^{a_i} k[t]}_{[a_i,\infty)} \oplus \underbrace{\bigoplus_j t^{b_j} k[t]/t^{c_j}}_{[b_j, b_j+c_j)}$$

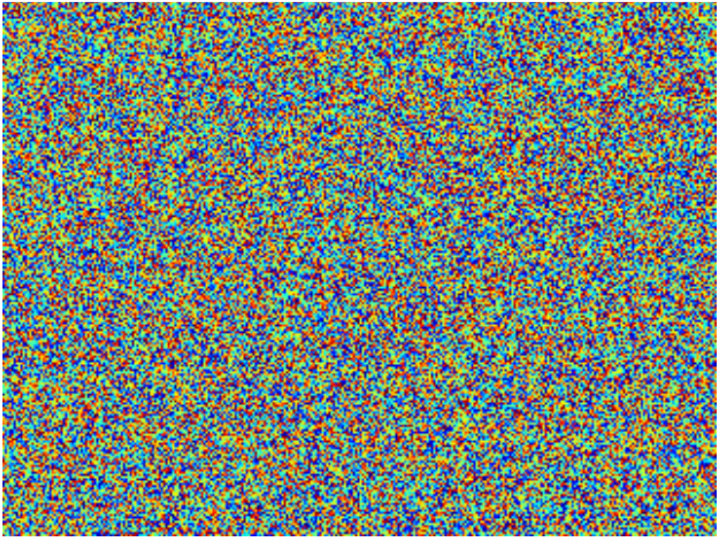


Interpreting the barcode

Barcodes of betti numbers of Vietoris-Rips complexes of point clouds tell us which homological properties are significant, and which result from noise.

The length of an interval corresponds to the size of the corresponding feature.





Example: natural images

Lee-Mumford-Pedersen investigated whether a statistically significant difference exists between natural and random images.

Natural images form a “subspace” of all images. Dimension of ambient space e.g. $640 \times 480 = 307\,200$.

This space of natural images should have:

- high dimension: there are many different images.
- high codimension: random images look nothing like natural ones.

Natural 3x3 patches

Instead of studying entire images, we consider the distribution of 3×3 pixel patches.

Most of these will be approximately constant in natural images. Allowing these drowns out structure.

Lee-Mumford-Pedersen chose $8\,500\,000$ patches with high contrast from a collection of black-and-white images used in cognition research. Each 3×3-patch is considered a vector in $\mathbb{R}^9$.

Normalised brightness: $\mathbb{R}^9 \to \mathbb{R}^8$. Normalised contrast: $\mathbb{R}^8 \to S^7$.

Subsequent topological analysis by Carlsson–de Silva–Ishkanov–Zomorodian.

Pixel patches in S^7

The resulting patches are dense in S^7 – so we consider high-density regions.

Pick out 25% densest points. We can pick a parametrised method to measure density:

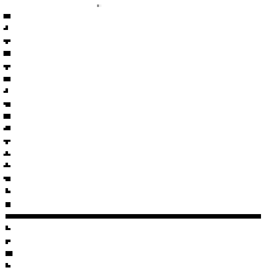
Definition

k-codensity $\delta_k(x)$ of a point x is the distance to its kth nearest neighbour.

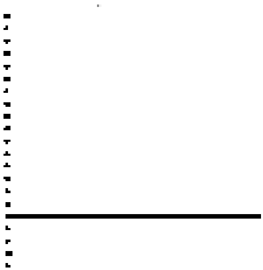
k-density $d_k(x)$ is $1/\delta_k(x)$.

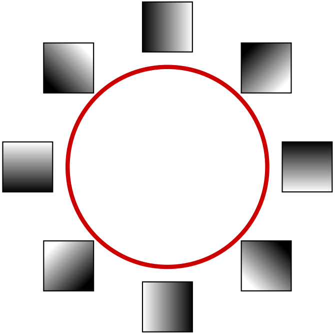
High k yields a smoothly changing density measure capturing global properties. Low k yields a wilder density measure capturing local properties. k acts as a kind of focus control.

300-density

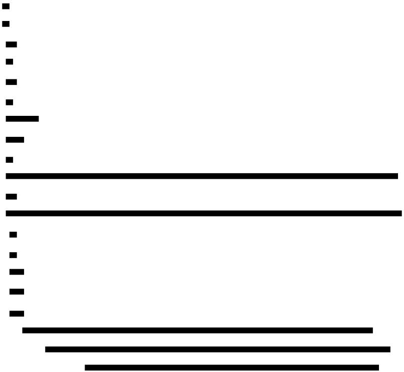


300-density

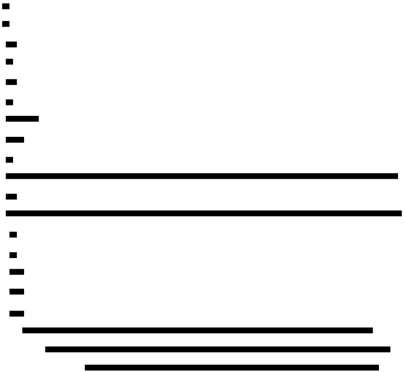


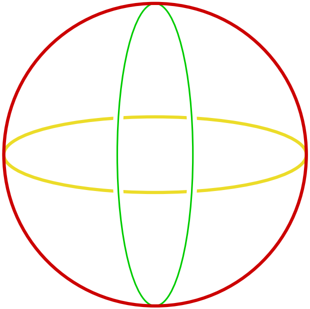


15-density

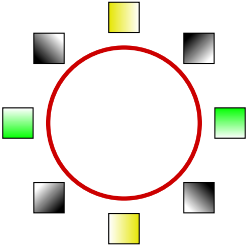


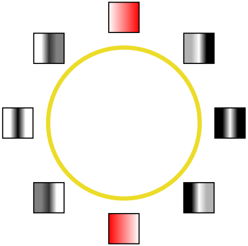
15-density

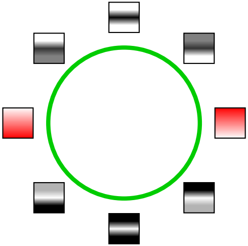




Three circles





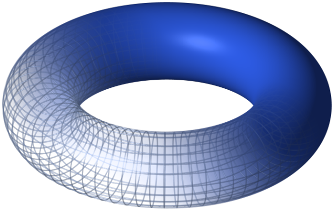


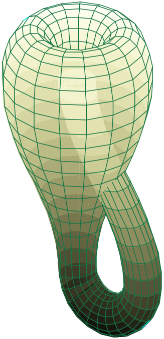
Identifying the subspace of natural pixel patches

Raising the cut-off bar yields, with coefficients in $\mathbb{F}_2$

$$\beta_0 = 1 \qquad \beta_1 = 2 \qquad \beta_2 = 1$$

Assuming the shape is a surface, this corresponds to one of



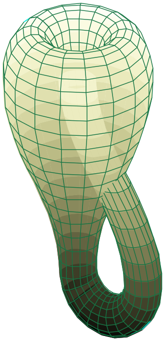


Identifying the subspace of natural pixel patches

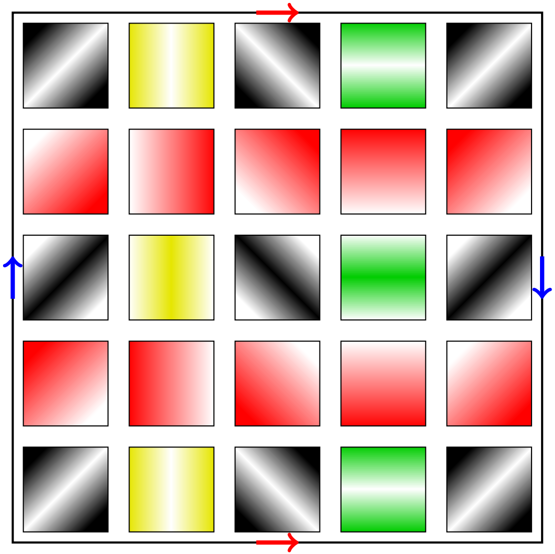
Raising the cut-off bar yields, with coefficients in $\mathbb{F}_3$

$$\beta_0 = 1 \qquad \beta_1 = 1$$

Thus, the relevant shape is:



Klein bottle of pixel patches



Applications of this analysis

Image compression

A 3×3-cluster may be described using 4 values:

- Position of its projection onto the Klein bottle
- Original brightness
- Original contrast

Texture analysis

Textures yield distributions of occuring patches on the Klein bottle. Rotating the texture corresponds to translating the distribution. [J Perea]

Coordinatization methods

My own work is on automating the above process by finding topological methods to recover intrinsic coordinate maps.

Idea

- Starting from a dataset X: compute its persistent homology $H(VR_*(X); k)$
- Guess a simplicial complex Y with corresponding homology
- Find maps $X \to Y$ or $VR_*(X) \to Y$ that lift to the expected correspondance.

First results

Joint with Vin de Silva and Dmitriy Morozov.

We can use that the circle is an Eilenberg-Mac Lane space, and thus

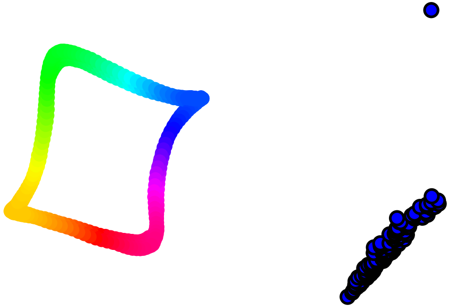
$$H^1(X;\mathbb{Z}) = [X, S^1]$$

We have established a definition of persistent cohomology, and produced techniques, algorithms and software for computing circle-valued coordinate functions using cohomology and a smoothing step.

Future directions

- Approach more generic coordinatizations by studying optimal chains in $H_0(\hom(CX, CY)) = \bigoplus_p \hom(H_pX, H_pY)$.
- Apply the circular coordinates work to periodic and recurrent systems and signals. Currently looking at data sets from: meteorology, climate research, gait research, music.
- Use circular coordinates for quality control on existing analysis methods for periodic signals.

Questions?



Delay embedding of a window from a clarinet tone, using circular coordinates and a persistence diagram to quality control the delay embedding.