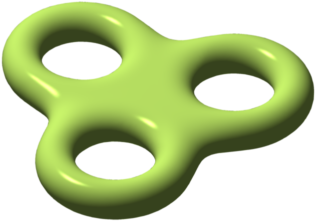
Learning Algebraic Varieties from Samples

Sara Kališnik

November 23, 2018



With Paul Breiding, Bernd Sturmfels and Madeleine Weinstein

Short overview:

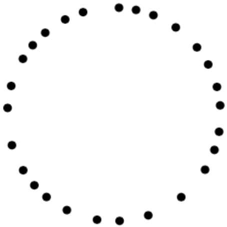
1. Introduction to Varieties (Basic Definitions, Examples, Applications)
2. Extracting Information from Samples: Dimension Estimates
3. Extracting Information from Samples: Persistent Homology
4. Extracting Information from Samples: Computing Polynomials, using Algebraic Geometry Software

Extracting Information from Samples

The Data

We are given a finite sample of points $\Omega = \{u^{(1)}, u^{(2)}, \ldots, u^{(m)}\}$ in $\mathbb{R}^n$ or $\mathbb{RP}^{n-1}$. These are sampled from an unknown variety.

Goal: Learn as much information about V as possible.

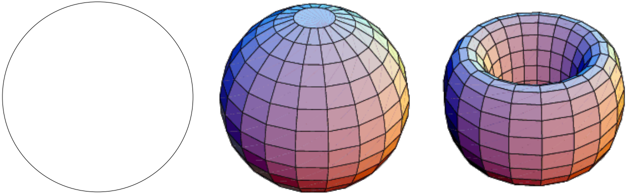


$$\longrightarrow \quad (x-3)^2 + (x-5)^2 - 100$$

Dimension, equations, degree, homology.

Measuring Shape

Homology is a formalism for measuring shape...



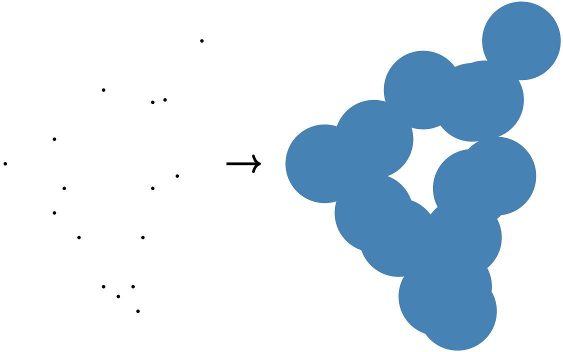
$b_1 = 1$ $b_2 = 0$ $\qquad$ $b_1 = 0$ $b_2 = 1$ $\qquad$ $b_1 = 2$ $b_2 = 1$

The extension of homology to more general setting including point clouds is called persistent homology.

The concept emerged independently in the work of Frosini, Ferri, and collaborators in Bologna, Italy, of Robins at Boulder, Colorado, and of Edelsbrunner, Letscher and Zomorodian at Duke, North Carolina.

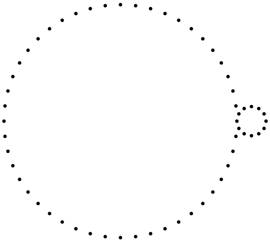
Persistent Homology

A finite metric space $\mathbb{X}$ has no interesting topology.

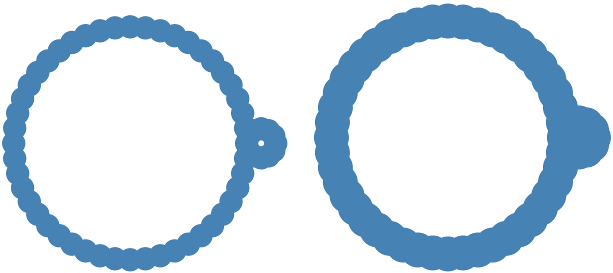


Let $U(\mathbb{X}, R)$ be the union of balls of radius R centered at the points of $\mathbb{X}$. For any $R > 0$ and $i \geq 0$, i-th Betti number of $U(\mathbb{X}, R)$ gives us a qualitative descriptor of $\mathbb{X}$.

Persistent Homology



Persistent Homology



$b_0 = 1$
$b_1 = 2$

$b_0 = 1$
$b_1 = 1$

Persistent Homology

Problems with this descriptor

- No canonical choice of R.
- Invariant is unstable with respect to perturbation of data or small changes in R.
- Does not distinguish ‘small’ holes from ‘big’ ones.

Persistent Homology

Persistent Homology

- Consider not only single reconstruction $U(\mathbb{X}, R)$ of $\mathbb{X}$, but a 1-parameter family of reconstructions

$$F(\mathbb{X}) = \{U(\mathbb{X}, r)\}_{r \in [0,\infty)}$$

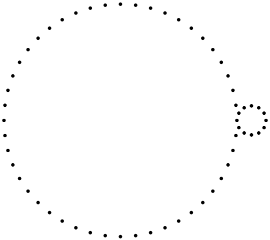
  and inclusion maps $U(\mathbb{X}, r) \hookrightarrow U(\mathbb{X}, r')$ whenever $r \leq r'$.
- Apply i-dimensional homology functor H_i with field coefficients
- Obtain a family of vector spaces $\{V_r\}_r$ and linear maps between them. Call such algebraic structures persistence vector spaces.

Can we classify persistence vector spaces that arise from filtrations up to isomorphism?

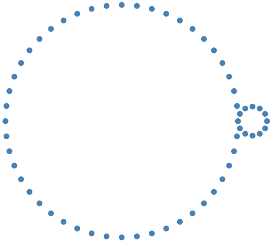
Yes, by barcodes.

Computing Persistent Homology by G. Carlsson and A. J. Zomorodian

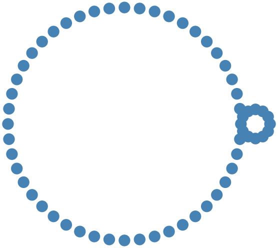
Persistent Homology



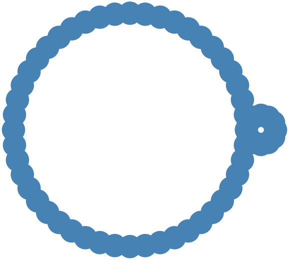
Persistent Homology



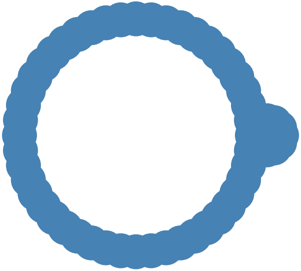
Persistent Homology



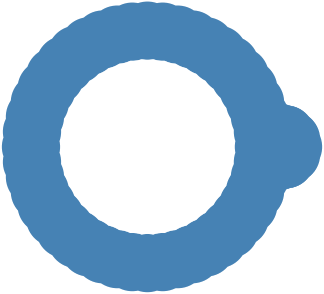
Persistent Homology



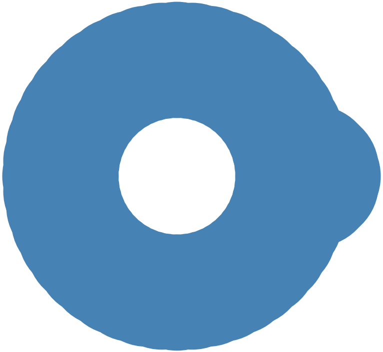
Persistent Homology



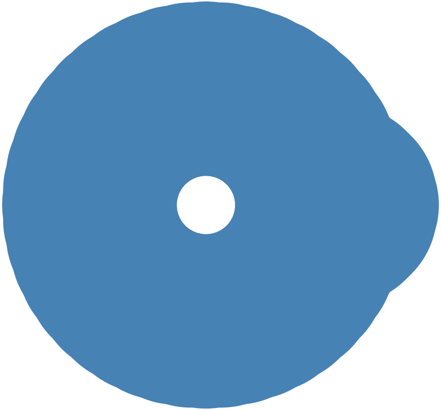
Persistent Homology



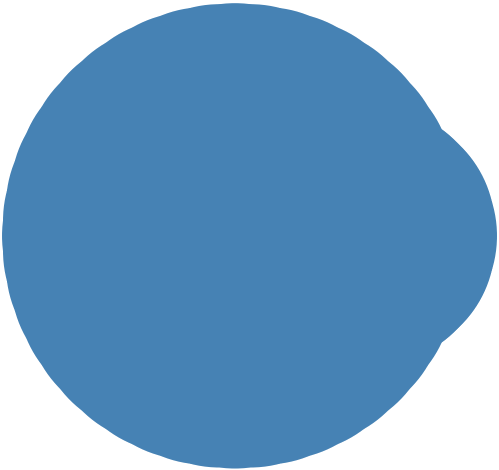
Persistent Homology



Persistent Homology

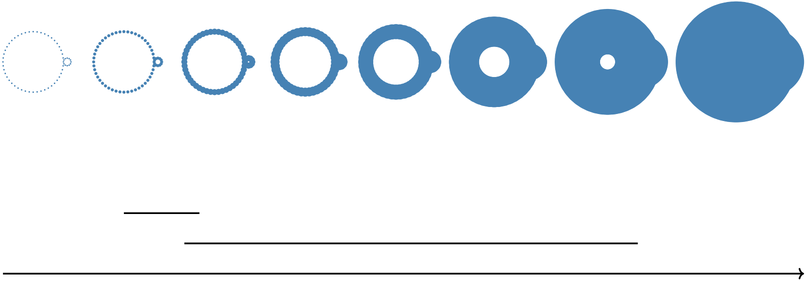


Persistent Homology



Persistent Homology

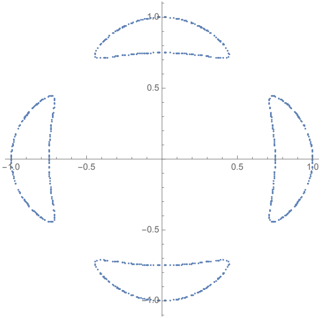
Barcode for H_1:

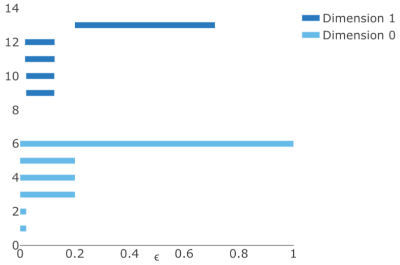


For each interval:

- Left endpoint is the index at which the hole is born
- Right endpoint is index at which hole dies
- Length of interval is the lifetime of a hole in filtration

Persistent Homology





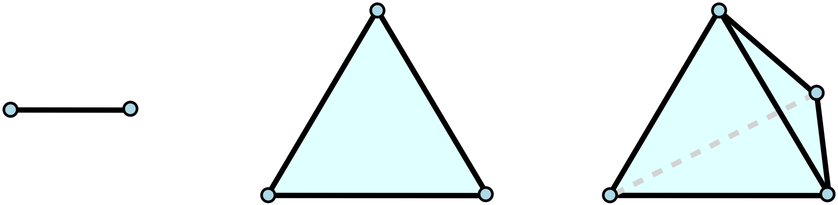
Persistent homology barcodes for the Trott curve.

How do we implement this?

The problem is that we cannot feed unions of balls to a computer. We must find combinatorial objects whose shape is the same as that of this union ('homotopy equivalence').

Simplicial Complexes

The building blocks are simplices, for example, *vertices* (0-simplices), *edges* (1-simplices), *triangles* (2-simplices), *tetrahedra* (3-simplices) and higher dimensional equivalents glued together along a common faces.

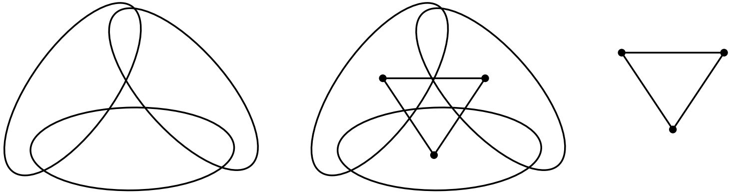


From Topological Spaces to Simplicial Complexes

To get from the continuous world of open and closed sets set to simplicial complexes, we use a construction called the nerve of a covering.

Nerve of a Covering

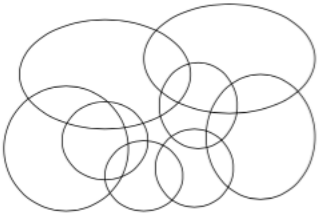
Let X be a topological space, and let $\mathcal{U} = \{U_i\}_{i \in I}$ be any covering of X. The *nerve* of $\mathcal{U}$, denoted by $\mathcal{N}\mathcal{U}$, is the abstract simplicial complex with vertex set I, where a family $\{i_0, \ldots, i_k\}$ spans a k-simplex if and only if $U_{i_0} \cap \ldots \cap U_{i_k} \neq \emptyset$.



Triangulating Point Clouds

Nerve of a Covering

Let X be a topological space, and let $\mathcal{U} = \{U_i\}_{i \in I}$ be any covering of X. The *nerve* of $\mathcal{U}$, denoted by $\mathcal{N}\mathcal{U}$, is the abstract simplicial complex with vertex set I, where a family $\{i_0, \ldots, i_k\}$ spans a k-simplex if and only if $U_{i_0} \cap \ldots \cap U_{i_k} \neq \emptyset$.



Triangulating Point Clouds

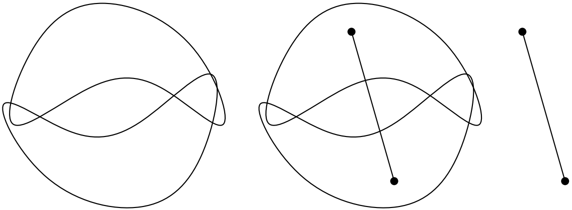
NERVE THEOREM

Suppose that X and $\mathcal{U}$ are as above, and suppose that the covering consists of finitely many open sets. Suppose further every nonempty intersection of sets in $\mathcal{U}$ is contractible. Then $\mathcal{N}\mathcal{U}$ is homotopy equivalent to X.

Triangulating Point Clouds

NERVE THEOREM

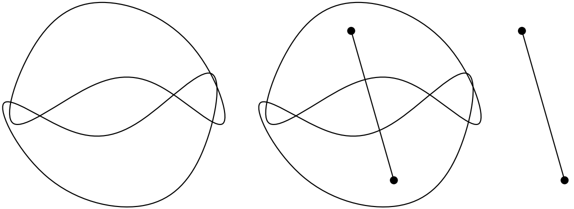
Suppose that X and $\mathcal{U}$ are as above, and suppose that the covering consists of finitely many open sets. Suppose further every nonempty intersection of sets in $\mathcal{U}$ is contractible. Then $\mathcal{N}\mathcal{U}$ is homotopy equivalent to X.



Triangulating Point Clouds

NERVE THEOREM

Suppose that X and $\mathcal{U}$ are as above, and suppose that the covering consists of finitely many open sets. Suppose further every nonempty intersection of sets in $\mathcal{U}$ is contractible. Then $\mathcal{N}\mathcal{U}$ is homotopy equivalent to X.



This implies that under certain conditions the nerve has homotopy groups isomorphic to the underlying space. One now needs methods for generating coverings.

Generating Coverings

There are a variety of ways to triangulate a collection of points.

Čech Complex

When the space in question is a finite metric space $\mathcal{C}$, one covering is given by the family $B_r(X) = \{B_r(x)\}_{x \in X}$, for some $r > 0$. We will denote this construction by $\check{C}(\mathcal{C}, r)$, and refer to it as the Čech complex attached to $\mathcal{C}$ and r.

- Since the sets $B_r(x)$ are all convex, the Nerve Theorem applies,.
- Given two radii $r < r'$, we have the inclusion $\check{C}(\mathcal{C}, r) \subseteq \check{C}(\mathcal{C}, r')$.

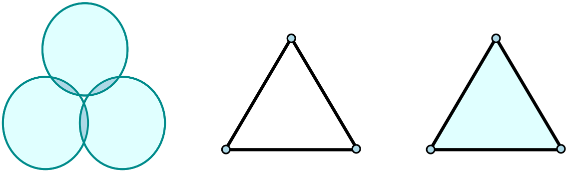
Efficiency versus Precision

Vietoris Rips Complex

Given a point cloud $\mathcal{C}$ and a fixed number $r \geq 0$, we define the Vietoris-Rips complex of $\mathcal{C}$ and r to be:

$$\mathrm{VR}(\mathcal{C}, r) = \{\sigma \subseteq X \mid B_r(x) \cap B_r(x) \neq \emptyset, \forall x, y \in \sigma\}.$$

For $r < r'$, we again have the inclusion $\mathrm{VR}(\mathcal{C}, r) \subseteq \mathrm{VR}(\mathcal{C}, r')$.



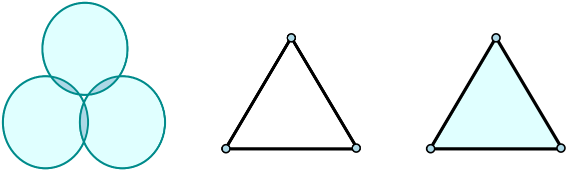
Efficiency versus Precision

Vietoris Rips Complex

Given a point cloud $\mathcal{C}$ and a fixed number $r \geq 0$, we define the Vietoris-Rips complex of $\mathcal{C}$ and r to be:

$$\mathrm{VR}(\mathcal{C}, r) = \{\sigma \subseteq X \mid B_r(x) \cap B_r(x) \neq \emptyset, \forall x, y \in \sigma\}.$$

For $r < r'$, we again have the inclusion $\mathrm{VR}(\mathcal{C}, r) \subseteq \mathrm{VR}(\mathcal{C}, r')$.



Saving grace: $\check{C}(\mathcal{C}, r) \subseteq \mathrm{VR}(\mathcal{C}, r) \subseteq \check{C}(\mathcal{C}, 2r)$.

Delaunay Triangulation

Even the Vietoris-Rips complex is computationally expensive, though, due to the fact that its vertex set consists of the entire metric space in question.

Voronoi Cell

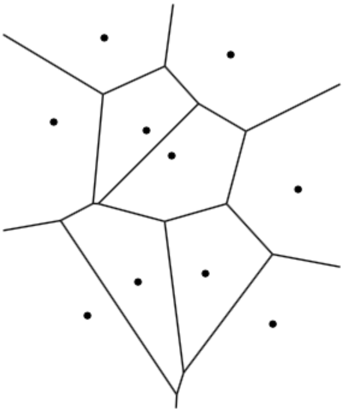
Given a finite point set $X \subseteq \mathbb{R}^d$, we define the Voronoi cell of a point $p \in X$ to be:

$$V_p = \{x \in \mathbb{R}^d \,|\, \mathrm{d}(x, p) \leq d(x, q), \forall q \in X\}.$$

Delaunay Triangulation

Voronoi Diagram

The collection of all Voronoi cells is called the Voronoi diagram of X; we note that it covers the entire ambient space $\mathbb{R}^d$.



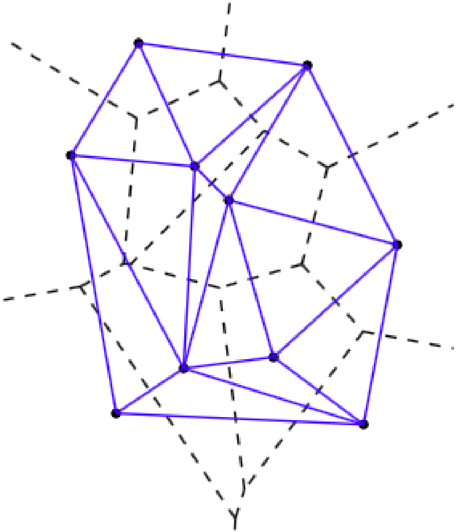
Delaunay Triangulation

The Delaunay triangulation of S to be (isomorphic to) the nerve of the collection of Voronoi cells; more precisely,

$$\mathrm{Del}(X) = \{\sigma \subseteq X \mid \bigcap_{p \in \sigma} V_p \neq \emptyset\}.$$

A set of vertices $\sigma \subseteq X$ forms a simplex in $\mathrm{Del}(S)$ iff these vertices all lie on a common $(d-1)$-sphere in $\mathbb{R}^d$. Assuming general position, we do in fact get a simplicial complex.

Delaunay Triangulation



Alpha Complex

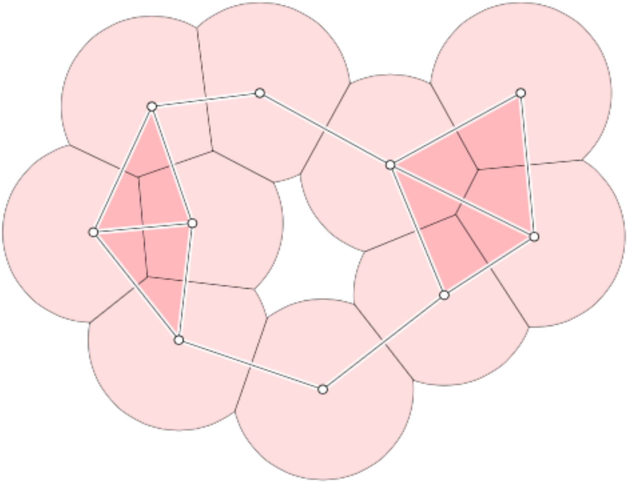
We again let X be a finite set of points in $\mathbb{R}^d$ and fix some radius r. As seen above, the complex $\check{\mathsf{C}}(X, r)$ has the same homotopy type as the union of r-balls X_r, but requires far too many simplices for large r. We now define a much smaller complex, $\mathrm{Alpha}(X, r)$, which is geometrically realizable in $\mathbb{R}^d$, and gives the correct homotopy type.

Alpha Complex

For each $p \in X$, we intersect the r-ball around p with its Voronoi region, to form $R_r(p) = B_r(p) \cap V_p$. These sets are convex and their union still equals X_r. We then define the Alpha complex of X and r to be the nerve of the collection of these sets:

$$\mathrm{Alpha}(X, r) = \{\sigma \subseteq X \mid \bigcap_{p \in \sigma} R_r(p) \neq \emptyset\}$$

Alpha Complex



Persistent Homology for Algebraic Varieties

Theoretical Guarantees

The Čech complex of a covering $U = \bigcup_{i=1}^{m} U_i$ has the homology of the union of balls U. But, can we give conditions on the sample $\Omega \subset M$ under which a covering reveals the true homology of M? A result due to Niyogi, Smale and Weinberger offers an answer in some circumstances. These involve the concept of the **reach**, which is an important metric invariant of a manifold M.

Persistent Homology for Algebraic Varieties

Theoretical Guarantees

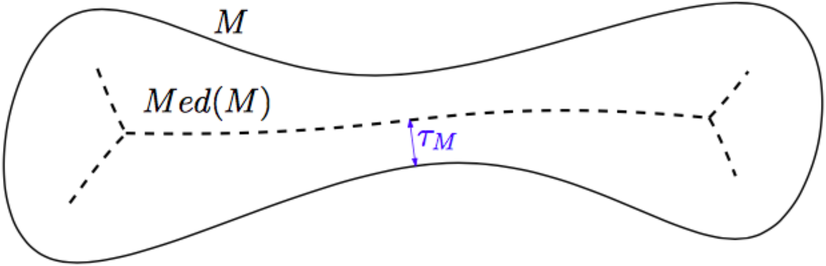
The **medial axis** of M is the set $Med(M)$ of all points $u \in \mathbb{R}^n$ such that the minimum distance from M to u is attained by two distinct points. The **reach** τ_M is the shortest distance from any point in the variety M to any point in its medial axis $Med(M)$.

Niyogi, Smale and Weinberger refer to $1/\tau_M$ as the **condition number of** M.

Finding the homology of submanifolds with high confidence from random samples by P. Niyogi, S. Smale and S. Weinberger

Persistent Homology for Algebraic Varieties

Theoretical Guarantees



Persistent Homology for Algebraic Varieties

Theoretical Guarantees

Let $M \subset \mathbb{R}^n$ be a compact manifold of dimension $d \leq 17$, with reach $\tau = \tau_M$ and d-dimensional Euclidean volume $\nu = \mathrm{vol}(M)$. Let $\Omega = \{u^{(1)}, \ldots, u^{(m)}\}$ be i.i.d. samples drawn from the uniform probability measure on M. Fix $\epsilon = \frac{\tau}{4}$ and $\beta = 16^d \tau^{-d} \nu$. For any desired $\delta > 0$, fix the sample size at

$$m \; > \; \beta \cdot \big(\log(\beta) + d + \log(\frac{1}{\delta})\big).$$

With probability $\geq 1 - \delta$, the homology groups of the following set coincide with those of M:

$$U(\epsilon) \; = \; \bigcup_{i=1}^{m} \big\{x \in \mathbb{R}^n : \|x - u^{(i)}\| < \epsilon\big\}.$$

Persistent Homology for Algebraic Varieties

Theoretical Guarantees

This theorem gives the asymptotics of a sample size m that suffices to reveal all topological features of M. For concrete parameter values it is less useful, though. For example, suppose that M has dimension 4, reach $\tau = 1$, and volume $V = 1000$. If we desire a 90% guarantee that $U(\epsilon)$ has the same homology as M, so $\delta = \frac{1}{10}$, then m must exceed 1592570365. In addition to that, the theorem assumes that the sample was drawn from the uniform distribution on M.

Persistent Homology for Algebraic Varieties

Tangent Spaces and Ellipsoids

Let us now return to our initial problem of a sample from a variety. Suppose that in addition to knowing Ω as a finite metric space, we also know some polynomials that vanish on Ω and the variety V that we are sampling from. Using their Jacobian, we can estimate the tangent space of V at each point $u^{(i)}$. We use ϵ-ellipsoids that are adjusted to these tangent spaces instead of ϵ-balls when computing the Vietoris-Rips complex for persistent homology in `Eirene`.

Persistent Homology for Algebraic Varieties

Tangent Spaces and Ellipsoids

In practice, we perform the following procedure. Let $f = (f_1, \ldots, f_k)$ be a vector of polynomials that vanish on V, derived from the sample $\Omega \subset \mathbb{R}^n$. An estimator for the tangent space $\mathrm{T}_{u^{(i)}}V$ is the kernel of the Jacobian matrix of f at $u^{(i)}$. In symbols,

$$\widehat{\mathrm{T}}_{u^{(i)}}V \;:=\; \ker Jf(u^{(i)}).$$

Persistent Homology for Algebraic Varieties

Tangent Spaces and Ellipsoids

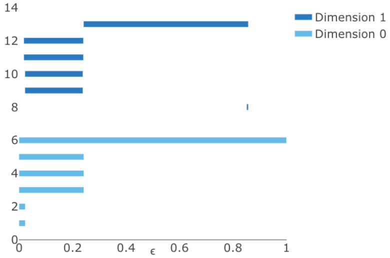
Let q_i denote the quadratic form on $\mathbb{R}^n$ that takes value 1 on $\widehat{\mathrm{T}}_{u^{(i)}} V \cap \mathbb{S}^{n-1}$ and value λ on the orthogonal complement of $\widehat{\mathrm{T}}_{u^{(i)}} V$ in the sphere $\mathbb{S}^{n-1}$. Then, the q_i specify the ellipsoids

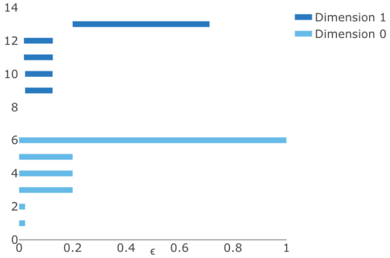
$$E_i \;:=\; \left\{ \sqrt{q_i(x)}\, x \in \mathbb{R}^n \,:\, \|x\| \leq 1 \right\}.$$

The role of the ϵ-ball enclosing the ith sample point is now played by $U_i(\epsilon) := u^{(i)} + \epsilon E_i$. These ellipsoids determine the covering $U(\epsilon) = \bigcup_{i=1}^m U_i(\epsilon)$ of the given point cloud Ω. From this covering we construct the associated Čech or Vietoris-Rips complexes.

Persistent Homology for Algebraic Varieties

Tangent Spaces and Ellipsoids





The left picture shows the barcode constructed from the ellipsoid-driven simplicial complex with $\lambda = 0.01$ and the right picture with standard Vietoris-Rips. All relevant topological features persist longer in the left plot.

Numerical Linear Algebra

Let $\mathcal{M}$ be a set of monomials in $S = \mathbb{R}[x_1, \ldots, x_n]$. Write $S_{\mathcal{M}}$ for the subspace with basis $\mathcal{M}$. Examples are all monomials of degree d resp. $\leq d$. The corresponding subspaces $S_{\mathcal{M}}$ satisfy

$$\dim(S_d) = \binom{n+d-1}{d} \quad \text{and} \quad \dim(S_{\leq d}) = \binom{n+d}{d}.$$

Write $U_{\mathcal{M}}(\Omega)$ for the multivariate Vandermonde matrix of format $m \times |\mathcal{M}|$: in the ith row are the values of the monomials in $\mathcal{M}$ at the point $u^{(i)}$. For example, if $n = 1$, $m = 3$, $\Omega = \{u, v, w\}$ then

$$U_{\leq d}(\Omega) = \begin{pmatrix} u^d & u^{d-1} & \cdots & u^2 & u & 1 \\ v^d & v^{d-1} & \cdots & v^2 & v & 1 \\ w^d & w^{d-1} & \cdots & w^2 & w & 1 \end{pmatrix}.$$

Remark: The kernel of $U_{\mathcal{M}}(\Omega)$ is the space $I_{\Omega} \cap S_{\mathcal{M}}$ of $\mathbb{R}$-linear combinations of $\mathcal{M}$ and that vanish on Ω.

Goal: Learn the ideal I_V of the unknown variety V.

Numerical Linear Algebra

Desirable properties in making an educated guess for $\mathcal{M}$:

a) *The ideal I_V is generated by its subspace $I_V \cap S_{\mathcal{M}}$.*

b) *Inclusion of $I_V \cap S_{\mathcal{M}}$ in $I_\Omega \cap S_{\mathcal{M}} = \ker(U_{\mathcal{M}}(\Omega))$ is an equality.*

Note: If $\mathcal{M}$ is too small then (a) fails. If $\mathcal{M}$ is too large then (b) fails.

Requirement (b) imposes a lower bound on the sample size:

$$m \geq |\mathcal{M}| - \dim(I_V \cap S_{\mathcal{M}}).$$

Example: It takes $m \geq \binom{n+2}{2}$ samples to learn quadrics in I_V.

We implemented

three methods for the kernel of the Vandermonde matrix $U_{\mathcal{M}}(\Omega)$

SVD	accurate, fast, but returns orthonormal and hence dense basis.
QR	slightly less accurate and fast than SVD, yields some sparsity.
RREF	no accuracy guarantees, not as fast as the others, gives sparse basis.

Computational Algebraic Geometry

We now have a set $\mathcal{P}$ of polynomials that vanish on Ω, and we hope that it defines the true variety V . What to do with $\mathcal{P}$?

Use symbolic or numerical methods to answer these questions:

1. What is the dimension of V ?
2. What is the degree of V ?
3. Find the irreducible components of V.
   Determine their dimensions and degrees.

Analyzing a Sample from SO(3)

Finding Equations

If we type

```
f = FindEquations(data, :with_qr, 2, false)
```

we get a list of 20 polynomials that vanish on the sample, the first two being

$$x_1x_4 + x_2x_5 + x_3x_6 = 0,$$
$$x_1x_7 + x_2x_8 + x_3x_9 = 0.$$

Analyzing a Sample from SO(3)

Finding Equations

d	method	number of linear independent equations
1	SVD	0
2	SVD	20
2	QR	20
2	RREF	20
3	SVD	136
4	SVD	550

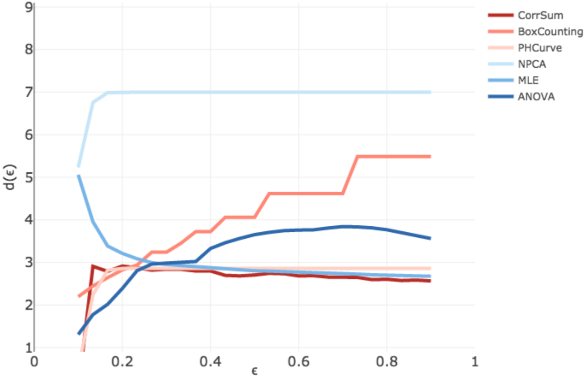
Analyzing a Sample from SO(3)

```
import Bertini: bertini
bertini(round.(f), TrackType = 1, bertini_path = p1)
```

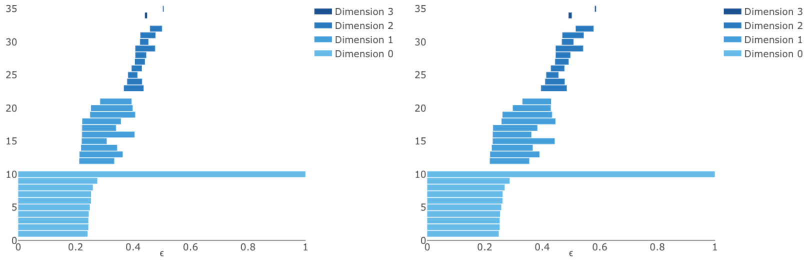
where p_1 is the path to the `Bertini` binary. `Bertini` tells us that the dimension is 3. We also learn that $\mathrm{SO}(3)$ is irreducible and that the degree is 8.

Analyzing a Sample from SO(3)

Recall also the diagram from the morning session, where the diagrams indicated that the dimension is 3.



Analyzing a Sample from SO(3)



Barcodes for a 250-point-subsample of $\mathrm{SO}(3)$ in dimensions 0, 1, 2 and 3. The left picture shows the standard Vietoris-Rips complex, while the barcode on the right is constructed from the ellipsoid-driven complex. Neither reveals any structures in dimension 3, though $\mathrm{SO}(3)$ is diffeomorphic to $\mathbb{RP}^3$ and has a non-vanishing 3-dimensional homology group.

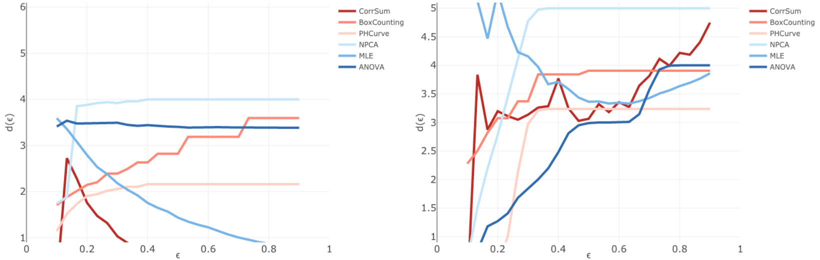
Analyzing a Sample from the Segre Variety

The next step is to find polynomials that vanish. We set `homogeneous_eqnarrays` to **`true`** and $d = 2$: `f = FindEquations(data, method, 2,true)`. All three methods, SVD, QR and RREF, correctly report the existence of three quadrics. The equations obtained with QR after rounding are as desired:

$$x_1x_4 - x_2x_3 = 0, \quad x_1x_6 - x_2x_5 = 0, \quad x_3x_6 - x_4x_5 = 0.$$

Running `Bertini` we verify that V is an irreducible variety of dimension 3 and degree 3.

Analyzing a Sample from the Segre Variety



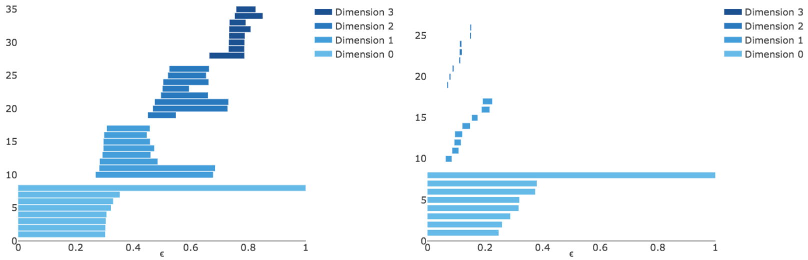
Dimension diagrams for 200 points on the variety of 2×3 matrices of rank 1. The left picture shows dimension diagrams for the estimates in $\mathbb{R}^6$. The right picture shows those for projective space $\mathbb{RP}^5$. The true dimension is 3.

Analyzing a Sample from the Segre Variety

We type the following to produce a persistence diagram for homological dimensions 0, 1, 2, 3. We display 8 longest barcodes from each barcode.

```
# sample 250 random points
i = rand(1:887, 250)
# compute the scaled Euclidean distances
dists = ScaledEuclidean(data[:,i])
# pass the distance matrix to Eirene and plot the barcodes in dimensions up
   to 3
C = eirene(dists, maxdim = 3)
barcode_plot(C, [0,1,2,3], [8,8,8,8])
```

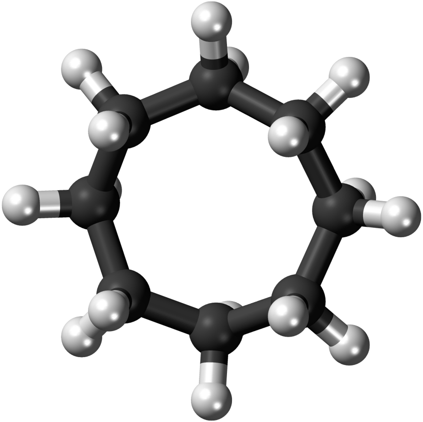
Analyzing a Sample from the Segre Variety



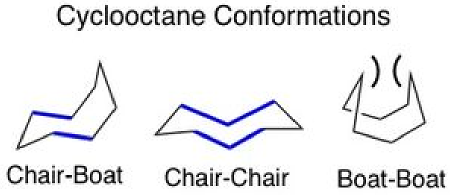
The left picture shows the barcodes for the usual Vietoris-Rips complex computed using the scaled Fubini-Study distance and the right picture shows the barcodes using the scaled Euclidean distance. Because the variety of 2×3 rank one matrices has the same topology as $\mathbb{RP}^1 \times \mathbb{RP}^2$ the Betti numbers are $1, 2, 2, 1$ in mod 2 coefficients for dimension $0, 1, 2, 3$, respectively.

Conformation Space of Cyclo-Octane

The third sample, which is 6040 points from the conformation space of cyclo-octane, is taken from the Javaplex Tutorial by Adams and Tausz.

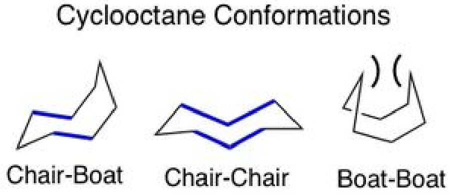


Conformation Space of Cyclo-Octane



Cyclo-octane consists of 8 carbon atoms arranged in a ring and each bonded to a pair of hydrogen atoms. The location of the hydrogen atoms is determined by that of the carbon atoms due to energy minimization. Hence, the conformation space of a cyclo-octane consists of all possible spatial arrangements, up to rotation and translation, of the ring of carbon atoms. Each datum is a point in $\mathbb{R}^{24} = \mathbb{R}^{8\cdot 3}$, which represents the coordinates for $\{z_0, \ldots, z_7\} \subset \mathbb{R}^3$, the locations of carbon atoms in cyclo-octane.

Conformation Space of Cyclo-Octane

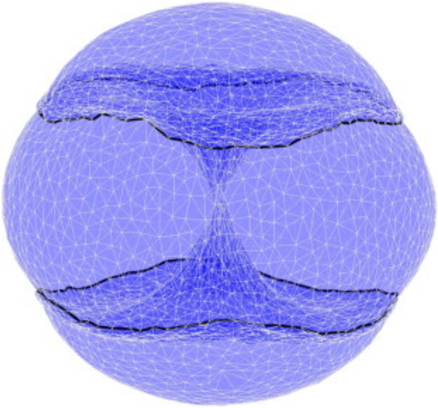


Each carbon atom z forms an isosceles triangle with its two neighbors in a way that that the angle at z is $\frac{2\pi}{3}$. Therefore, by the law of cosines there is a constant $c > 0$ such that the squared distances $d_{i,j} = \|z_i - z_j\|^2$ satisfy $d_{i,i+1} = c$ and $d_{i,i+2} = \frac{8}{3}c$ for all i (here i is modulo 8).

Conformation Space of Cyclo-Octane

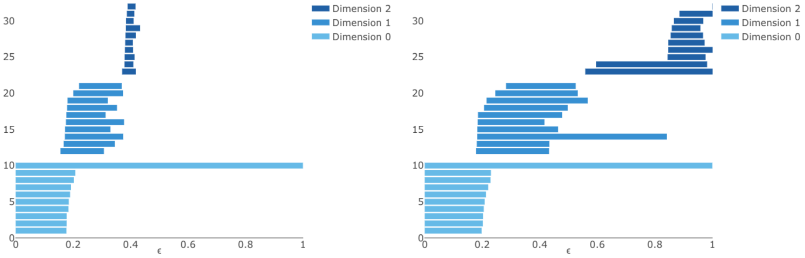
M.W. Brown, S. Martin et. al: *Algorithmic dimensionality reduction for molecular structure analysis*

S. Martin, A. Thompson, E. A. Coutsias, and J. P. Watson: *Topology of cyclo-octane energy landscape*.



The conformation space of cyclo-octane is the union of a sphere with a Klein bottle, glued together along two circles of singularities.

Conformation Space of Cyclo-Octane



Barcodes for a subsample of 500 points from the cyclo-octane dataset. The left plot shows the barcodes for the usual Vietoris-Rips complex. The right picture shows barcodes for the ellipsoid complex. The right barcode captures the correct homology.