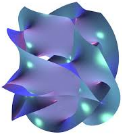
Mathematics of Data: Algebraic and Topological Methods

Federica Galluzzi

April 28, 2014

Browsing through Mathematics



Types of Data

- images,videos,speech waves, gene expression,financial data
- internet, biological/social networks
- documents and information flows

Problems

- How to capture variations of data distribution?
- How to distinguish significant features from noise?

Algebra and Topology may play a role

- Convert the data set into global topological objects
- Infer high dimensional structure from low dimensional representations

Networks or Point Cloud as undirected graphs

- Point cloud as vertices of a graph
- Connectivity data as edges

The graph ignores higher order features beyond clustering. Think of the graph as a scaffold: complete it to a *simplicial complex*

Simplicial Complexes

- K, a set
- $\mathcal{S}$, a collection of subsets (*simplices) in K*

such that

- for all $v \in K$, $\{v\} \in \mathcal{S}$
- for all $\sigma \in \mathcal{S}$ and $\tau \subset \sigma$, then $\tau \in \mathcal{S}$

- the sets $\{v\}$ are the *vertices of K.*
- $\sigma \in \mathcal{S}$ *is a* $k -$ *simplex* if $|\sigma| = k + 1$.
- a subset $\tau \subset \sigma$ is a *face* of σ

A simplicial complex is called *oriented* if it comes with a total order on its vertices. We denote the simplices $\sigma = [v_0, ..., v_n]$.

Standard simplices in $\mathbb{R}^3$

A simplex may be realized geometrically as the convex hull of $k+1$ affinely independent points in $\mathbb{R}^d$ with $d \geq k$.

Example

If K is a tethraedron, triangle faces are the $2-$simplices, edges are the $1-$simplices, vertices are the $0-$simplices.

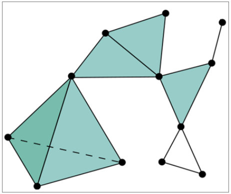


Figure: Simplicial complex

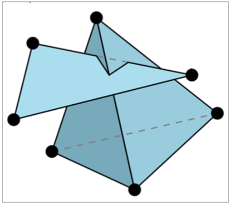


Figure: Invalid simplicial complex

From clouds to complexes

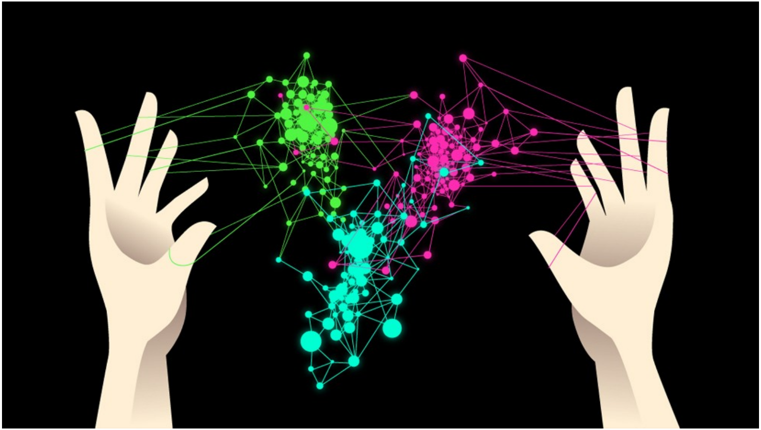


Figure: Tang Yau Hoon

Clique Complexes

A clique is a subset of vertices such that every two vertices are connected by an edge. The clique complex associated to a graph G has the vertices of G and the faces are the cliques of G.

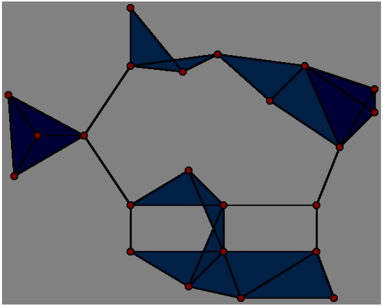


Figure: Wikipedia

Some Algebraic Topology

The $k-$th chain Group $C_k(K)$

A k-chain is a linear combination of $k-$simplices in K with integer coefficients. The $k-$th chain group is the set of all linear combinations

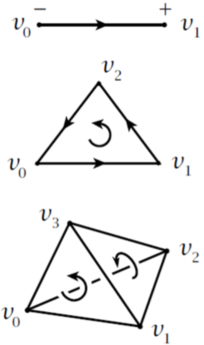
$$C_k(K) := \sum_i n_i \sigma_i, \quad n_i \in \mathbb{Z}, \ \sigma_i \, k - \textit{simplex in } K$$

The boundary operator $\partial_k : C_k(K) \to C_{k-1}(K)$

The boundary operator is a homomorphism defined on a $k-$simplex by:

$$\partial_k([v_0, ..., v_{k+1}]) = \sum_i (-1)^i [v_0, ..., \widehat{v}_i, ..., v_{k+1}]$$

and on a $k-$chain by linearity.



$$\partial[v_0, v_1] = [v_1] - [v_0]$$

$$\partial[v_0, v_1, v_2] = [v_1, v_2] - [v_0, v_2] + [v_0, v_1]$$

$$\begin{aligned}\partial[v_0, v_1, v_2, v_3] = {} & [v_1, v_2, v_3] - [v_0, v_2, v_3] \\ & + [v_0, v_1, v_3] - [v_0, v_1, v_2]\end{aligned}$$

Figure: Hatcher's book

The boundary of a boundary is zero

The operator ∂ connects chain groups

$$\dots \longrightarrow C_{k+1}(K) \xrightarrow{\partial_{k+1}} C_k(K) \xrightarrow{\partial_k} C_{k-1}(K) \to \dots$$

It has the important property that

$$\partial_k \circ \partial_{k+1} = 0$$

Cycles and Boundaries in $C_k(K)$

A *cycle* is a chain with zero boundary.

- $Z_k(K) := \ker \partial_k$ the $k-$th cycle group
- $B_k(K) := im\ \partial_{k+1}$ the $k-$th boundary group
- $\partial \circ \partial = 0 \implies B_k \subseteq Z_k$

These groups are nested

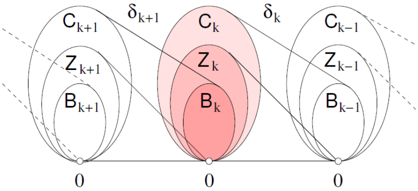
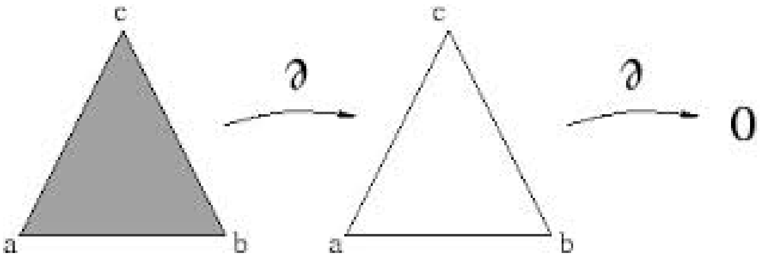


Figure 4. A chain complex with its internals: chain, cycle, and boundary groups, and their images under the boundary operators.

Boundaries of higher order chains are uninteresting



$$\partial(\partial[a, b, c]) = \partial([b, c] - [a, c] + [a, b]) = c - b - (c - a) + b - a = 0$$

Use Homology to identify interesting cycles

The $k-$th homology group is the quotient group of cycles over boundaries

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

A element $\alpha \in H_k(K)$ is a *homology class*.

Betti numbers

- β_k the $k-$th Betti number : rank of $H_k(K)$

Holes = Interesting cycles

Homology can identify

- clusters (β_0 is the number of connected components)
- holes (1st order holes),
- voids or cavities (2nd order holes, the inside of a balloon)

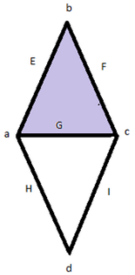


Figure: Wikipedia

a, b, c, d : 0−simplices; E, F, G, H, I : 1−simplices; shaded region: 2−simplex. $\beta_0 = 1$. One hole: $\beta_1 = 1$. No voids: $\beta_2 = 0$.

How can homology track the evolution of a data set?

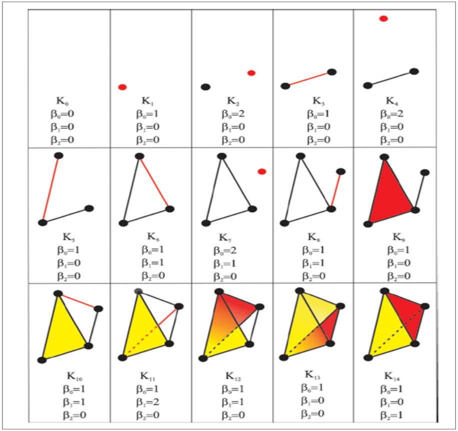


Figure: D.Horak "Persistence Homology of Complex Networks"

Adding or removing simplices

Filtrations

A *filtration* of a complex K is a nested sequence of subcomplexes

$$\emptyset = K^0 \subseteq K^1 \subseteq K^2 \subseteq K^3 \subseteq \ldots \subseteq K^m = K$$

Birth and death of a homology class

The filtration induces maps on the homology groups

$$\ldots \to H_k(K^{i-1}) \to H_k(K^i) \to H_k(K^{i+1}) \to \ldots$$

If a class α is born in $H_k(K^i)$ and dies in $H_k(K^j)$, the *persistence* (lifetime) of α is $l = j - i - 1$

Persistent homology

The $p-$persistent $k-$th homology group of K^i is

$$H_k^{i,p} := Z_k^i/(B_k^{i+p} \cap Z_k^i)$$

Homology classes of K^i that are still alive in K^{i+p}

Persistent Betti numbers

- $\beta_k^{i,p}$ the $p-$persistent $k-$th Betti number : rank of $H_k^{i,p}$

Independent homology classes in K^i that are still alive and independent in K^{i+p}

Persistent homology tracks homology classes along the filtration: for which value of p a hole appears, and how long it persists till it is filled in.

Visualize persistent homology: barcodes

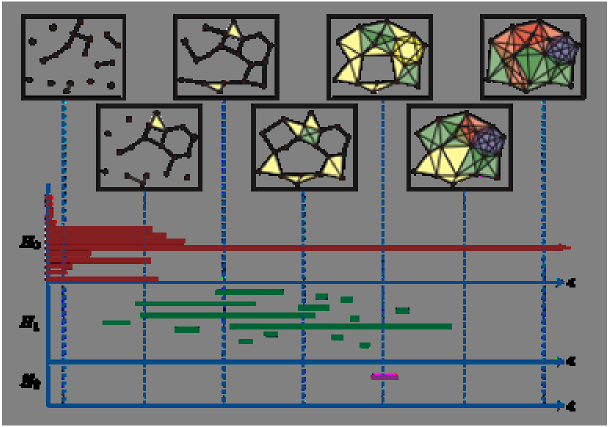


Figure: R.Ghrist "The Persistent Topology of Data"

- The horizontal axis is p
- The vertical axis represents ordered homology generators for the H_k
- Each horizontal bar represents the birth death of a separate homology class

- Longer bars correspond to more robust topological structure in the data.
- Shorter bars have short lifetimes and may be considered as topological noise.

Applications

- Separate topological signal from topological noise
- Give important information about robustness of networks against addition or removal of nodes
- Exhibit the highest topological resilience to change in the addition or removal of nodes
- Try to detect hierarchies in a (social, infrastructural, biological) network
- Process motion capture data to distinguish significant features
- ....

Other Approaches

- Complexes associate to graphs : Cech complex, Rips complex,
- Persistence Complexes : maps $f^i : K^i \to K^{i+i}$ instead of inclusions $K^i \subset K^{i+1}$
- Random networks

Computational aspects

JavaPlex, Java library for persistent homology (CompTop, Stanford) http://code.google.com/p/javaplex

Short Bibliography

- G.Carlsson, A.Zomorodian "Computing Persistence Homology" , Discrete Comput. Geom. (2005)
- H. Edelsbrunner, J. Harer "Persistent Homology. A Survey", Contemporary Mathematics (2008)
- F. Cagliari, M. Ferri, P.Pozzi "Size functions from the categorical viewpoint", Acta Appl. Math. (2001).
- P. Frosini, C. Landi "Size theory as a topological tool for computer vision", Pattern Recognition and Image Analysis (1999)
- R. Ghrist "Barcodes: The persistent topology of data". B. AM. Math. Soc. (2008)

THANK YOU!