Topology and data
(Gunnar Carlsson, Bulletin of the AMS, 2009)

Dino Sejdinovic

Gatsby Unit MLJC

October 24, 2012

Introduction

- too many dimensions, too little time

Introduction

- too many dimensions, too little time
- dimensionality reduction / visualization

Introduction

- too many dimensions, too little time
- dimensionality reduction / visualization
- the meaning of coordinates / reparametrization? (is the intrinsic meaning of coordinates justified in your data?)

Introduction

- too many dimensions, too little time
- dimensionality reduction / visualization
- the meaning of coordinates / reparametrization? (is the intrinsic meaning of coordinates justified in your data?)
- what metric to use? e.g., distance between two DNA sequences? what is the significance of a measured distance?

Introduction

- too many dimensions, too little time
- dimensionality reduction / visualization
- the meaning of coordinates / reparametrization? (is the intrinsic meaning of coordinates justified in your data?)
- what metric to use? e.g., distance between two DNA sequences? what is the significance of a measured distance?
  - sometimes, only to reflect the intuitive notion of similarity: nearby data points are similar, far apart data points are different

Introduction

- too many dimensions, too little time
- dimensionality reduction / visualization
- the meaning of coordinates / reparametrization? (is the intrinsic meaning of coordinates justified in your data?)
- what metric to use? e.g., distance between two DNA sequences? what is the significance of a measured distance?
    - sometimes, only to reflect the intuitive notion of similarity: nearby data points are similar, far apart data points are different
    - we do not trust large distances (genomic sequences differing by 100/150 entries?)

Introduction

- too many dimensions, too little time
- dimensionality reduction / visualization
- the meaning of coordinates / reparametrization? (is the intrinsic meaning of coordinates justified in your data?)
- what metric to use? e.g., distance between two DNA sequences? what is the significance of a measured distance?
  - sometimes, only to reflect the intuitive notion of similarity: nearby data points are similar, far apart data points are different
  - we do not trust large distances (genomic sequences differing by 100/150 entries?)
  - we trust small distances only a little bit (strength of similarity as encoded by the distance may not be significant)

Introduction

- what is then the meaning of very refined notions we obtain from such "rough" distance notion, say curvature?

Introduction

- what is then the meaning of very refined notions we obtain from such "rough" distance notion, say curvature?
- asking qualitative (unsupervised) questions about data?

Introduction

- what is then the meaning of very refined notions we obtain from such "rough" distance notion, say curvature?
- asking qualitative (unsupervised) questions about data?
  - properties robust to changes in metrics?

Introduction

- what is then the meaning of very refined notions we obtain from such "rough" distance notion, say curvature?
- asking qualitative (unsupervised) questions about data?
  - properties robust to changes in metrics?
  - the study of idealized versions of such properties: topology

Introduction

- geometry studies metrics. topology studies what remains after one stretches and deforms without tearing it

Introduction

- geometry studies metrics. topology studies what remains after one stretches and deforms without tearing it
- replace the quantitative values fo distance functions with the notion of infinite nearness (i.e., metric understood in a coarse way only): what remains is “connectivity information” of your data

Introduction

- geometry studies metrics. topology studies what remains after one stretches and deforms without tearing it
- replace the quantitative values fo distance functions with the notion of infinite nearness (i.e., metric understood in a coarse way only): what remains is "connectivity information" of your data
- connected components / clusters: zeroth order topological information

Homotopy

- connectivity: $x, y \in \mathcal{X}$, say $x \sim y$ iff $\exists$ continuous map $f : [0,1] \to \mathcal{X}$, such that $f(0) = x$, $f(1) = y$

Homotopy

- connectivity: $x, y \in \mathcal{X}$, say $x \sim y$ iff $\exists$ continuous map $f : [0,1] \to \mathcal{X}$, such that $f(0) = x$, $f(1) = y$
- also, equivalence classes of maps: $f \sim g$ iff $\exists$ continuous map $F : [0,1]^2 \to \mathcal{X}$, such that $F(t,0) = f(t)$, $F(t,1) = g(t)$

Homotopy

- connectivity: $x, y \in \mathcal{X}$, say $x \sim y$ iff $\exists$ continuous map $f : [0,1] \to \mathcal{X}$, such that $f(0) = x$, $f(1) = y$
- also, equivalence classes of maps: $f \sim g$ iff $\exists$ continuous map $F : [0,1]^2 \to \mathcal{X}$, such that $F(t,0) = f(t)$, $F(t,1) = g(t)$
- In general, $f, g : \mathcal{Y} \to \mathcal{X}$, and $F : \mathcal{Y} \times [0,1] \to \mathcal{X}$. f and g are said to be *homotopic*

Homotopy

- connectivity: $x, y \in \mathcal{X}$, say $x \sim y$ iff $\exists$ continuous map $f : [0,1] \to \mathcal{X}$, such that $f(0) = x$, $f(1) = y$
- also, equivalence classes of maps: $f \sim g$ iff $\exists$ continuous map $F : [0,1]^2 \to \mathcal{X}$, such that $F(t,0) = f(t)$, $F(t,1) = g(t)$
- In general, $f, g : \mathcal{Y} \to \mathcal{X}$, and $F : \mathcal{Y} \times [0,1] \to \mathcal{X}$. f and g are said to be *homotopic*
- $\mathcal{X}$ and $\mathcal{Y}$ are homotopy equivalent if there are $f : \mathcal{X} \to \mathcal{Y}$ and $g : \mathcal{Y} \to \mathcal{X}$, s.t. $f \circ g$ is homotopic to $id_{\mathcal{Y}}$ and $g \circ f$ is homotopic to $id_{\mathcal{X}}$

Homotopy

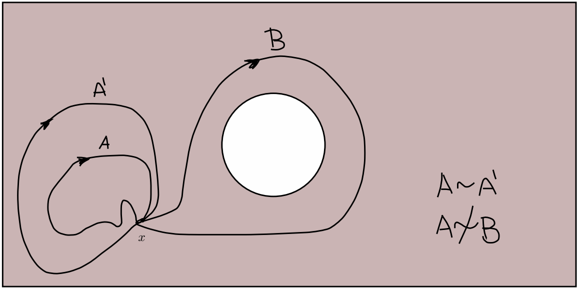
- connectivity: $x, y \in \mathcal{X}$, say $x \sim y$ iff $\exists$ continuous map $f : [0,1] \to \mathcal{X}$, such that $f(0) = x$, $f(1) = y$
- also, equivalence classes of maps: $f \sim g$ iff $\exists$ continuous map $F : [0,1]^2 \to \mathcal{X}$, such that $F(t,0) = f(t)$, $F(t,1) = g(t)$
- In general, $f, g : \mathcal{Y} \to \mathcal{X}$, and $F : \mathcal{Y} \times [0,1] \to \mathcal{X}$. f and g are said to be *homotopic*
- $\mathcal{X}$ and $\mathcal{Y}$ are homotopy equivalent if there are $f : \mathcal{X} \to \mathcal{Y}$ and $g : \mathcal{Y} \to \mathcal{X}$, s.t. $f \circ g$ is homotopic to $id_{\mathcal{Y}}$ and $g \circ f$ is homotopic to $id_{\mathcal{X}}$
- Every homeomorphism is a homotopy equivalence, but the converse is not true

Homotopy

- connectivity: $x, y \in \mathcal{X}$, say $x \sim y$ iff $\exists$ continuous map $f : [0,1] \to \mathcal{X}$, such that $f(0) = x$, $f(1) = y$
- also, equivalence classes of maps: $f \sim g$ iff $\exists$ continuous map $F : [0,1]^2 \to \mathcal{X}$, such that $F(t,0) = f(t)$, $F(t,1) = g(t)$
- In general, $f, g : \mathcal{Y} \to \mathcal{X}$, and $F : \mathcal{Y} \times [0,1] \to \mathcal{X}$. f and g are said to be *homotopic*
- $\mathcal{X}$ and $\mathcal{Y}$ are homotopy equivalent if there are $f : \mathcal{X} \to \mathcal{Y}$ and $g : \mathcal{Y} \to \mathcal{X}$, s.t. $f \circ g$ is homotopic to $id_{\mathcal{Y}}$ and $g \circ f$ is homotopic to $id_{\mathcal{X}}$
- Every homeomorphism is a homotopy equivalence, but the converse is not true
- click me

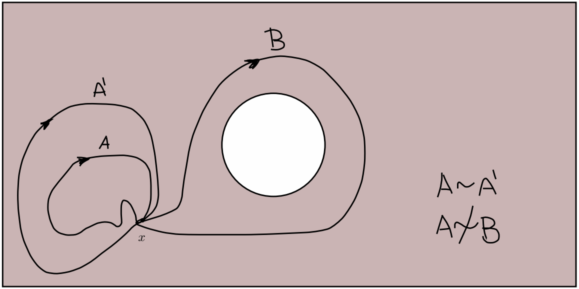
Homotopy groups

- n-th order topological information: homotopy classes of equivalence of continuous maps f from the n-dimensional sphere S^n to $\mathcal{X}$ s.t. $f(s) = x$



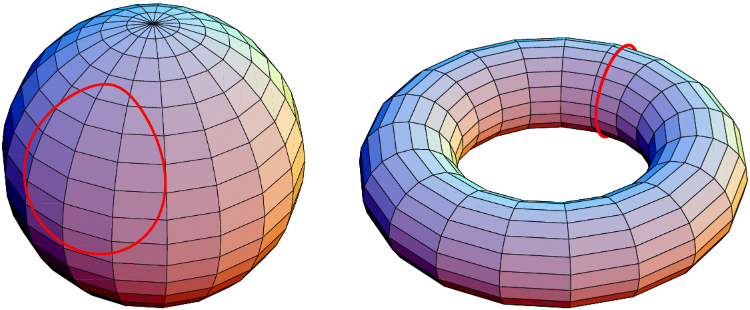
Homotopy groups

- n-th order topological information: homotopy classes of equivalence of continuous maps f from the n-dimensional sphere S^n to $\mathcal{X}$ s.t. $f(s) = x$



- Classes of equivalence form a group structure $\pi_n(\mathcal{X})$; for $n = 1$, **fundamental group**, e.g., $\pi_1(\mathbb{R}^n) = \{0\}$, $\pi_1(\mathbb{R}^n \backslash \{0\}) = \pi_1(S^1) = \mathbb{Z}$.

Homotopy groups



Homology

- higher-dimensional homotopy groups extremely difficult to compute, even $\pi_n(S^i)$ for $n > i$ is a difficult problem / *Hopf fibration*

Homology

- higher-dimensional homotopy groups extremely difficult to compute, even $\pi_n(S^i)$ for $n > i$ is a difficult problem / *Hopf fibration*
- a friendlier alternative: **homology groups**, with an extended equivalence relation

Homology

- higher-dimensional homotopy groups extremely difficult to compute, even $\pi_n(S^i)$ for $n > i$ is a difficult problem / *Hopf fibration*
- a friendlier alternative: **homology groups**, with an extended equivalence relation
- e.g., two loops are equivalent if there is a surface with boundary equal to the difference of two loops

Simplices and chains

- A 0-simplex is a point $[i]$

Simplices and chains

- A 0-simplex is a point $[i]$
- A 1-simplex is an edge $[ij]$

Simplices and chains

- A 0-simplex is a point $[i]$
- A 1-simplex is an edge $[ij]$
- An n-**simplex** is a hyperedge $\sigma = [i_0 \dots i_n]$. A face of σ is an $(n-1)$-dimensional simplex $[i_0 \dots i_{j-1} i_{j+1} \dots i_n] = \left[i_{\setminus j}\right]$

Simplices and chains

- A 0-simplex is a point $[i]$
- A 1-simplex is an edge $[ij]$
- An n-**simplex** is a hyperedge $\sigma = [i_0 \ldots i_n]$. A face of σ is an $(n-1)$-dimensional simplex $[i_0 \ldots i_{j-1} i_{j+1} \ldots i_n] = \left[i_{\setminus j} \right]$
- An n-**chain** c is a formal sum of n-simplices, e.g., $[12] + [23] + [34] \in C_1$ (may occur with a multiplicity or with an opposite orientation - winding numbers):

$$c = \sum_k \alpha_k \sigma_k, \qquad \alpha_k \in A, \sigma_k \in S_n$$

Simplices and chains

- A 0-simplex is a point $[i]$
- A 1-simplex is an edge $[ij]$
- An n-**simplex** is a hyperedge $\sigma = [i_0 \dots i_n]$. A face of σ is an $(n-1)$-dimensional simplex $[i_0 \dots i_{j-1} i_{j+1} \dots i_n] = [i_{\setminus j}]$
- An n-**chain** c is a formal sum of n-simplices, e.g., $[12]+[23]+[34] \in C_1$ (may occur with a multiplicity or with an opposite orientation - winding numbers):

$$c = \sum_k \alpha_k \sigma_k, \qquad \alpha_k \in A, \sigma_k \in S_n$$

- A is an abelian group (such as $\mathbb{Z}$,$\mathbb{F}_p$); S_n is a finite set of n-simplices

Simplices and chains

- A 0-simplex is a point $[i]$
- A 1-simplex is an edge $[ij]$
- An n-**simplex** is a hyperedge $\sigma = [i_0 \dots i_n]$. A face of σ is an $(n-1)$-dimensional simplex $[i_0 \dots i_{j-1} i_{j+1} \dots i_n] = \left[i_{\setminus j}\right]$
- An n-**chain** c is a formal sum of n-simplices, e.g., $[12] + [23] + [34] \in C_1$ (may occur with a multiplicity or with an opposite orientation - winding numbers):

$$c = \sum_k \alpha_k \sigma_k, \qquad \alpha_k \in A, \sigma_k \in S_n$$

- A is an abelian group (such as $\mathbb{Z}$,$\mathbb{F}_p$); S_n is a finite set of n-simplices
- The set of all n-chains is denoted C_n; $(C_n, +)$ forms a free abelian group: $c + c' = \sum \left(\alpha_k + \alpha'_k\right) \sigma_k$ (abelian group with a "basis")

Boundary map

- **Simplicial complex** is a collection $\mathfrak{C}$ of simplices with a special structure:

$$\sigma \in \mathfrak{C} \quad \Rightarrow \quad \text{any face of } \sigma \quad \in \mathfrak{C}$$

Boundary map

- **Simplicial complex** is a collection $\mathfrak{C}$ of simplices with a special structure:

$$\sigma \in \mathfrak{C} \quad \Rightarrow \quad \text{any face of } \sigma \quad \in \mathfrak{C}$$

- Boundary of an n-simplex is defined to be the sum of its faces:

$$\partial_n \left[i_0 \ldots i_n \right] \quad = \quad \sum_{j=0}^{n-1} (-1_A)^j \left[i_{\setminus j} \right]$$

Boundary map

- **Simplicial complex** is a collection $\mathfrak{C}$ of simplices with a special structure:

$$\sigma \in \mathfrak{C} \quad \Rightarrow \quad \text{any face of } \sigma \quad \in \mathfrak{C}$$

- Boundary of an n-simplex is defined to be the sum of its faces:

$$\partial_n [i_0 \dots i_n] \quad = \quad \sum_{j=0}^{n-1} (-1_A)^j \left[i_{\setminus j} \right]$$

- Boundary of a general chain:

$$\partial_n \sum_k \alpha_k \sigma_k \quad \doteq \quad \sum_k \alpha_k \partial_n \sigma_k$$

Boundary map

- **Simplicial complex** is a collection $\mathfrak{C}$ of simplices with a special structure:

$$\sigma \in \mathfrak{C} \quad \Rightarrow \quad \text{any face of } \sigma \quad \in \mathfrak{C}$$

- Boundary of an n-simplex is defined to be the sum of its faces:

$$\partial_n \left[i_0 \ldots i_n\right] \;=\; \sum_{j=0}^{n-1} (-1_A)^j \left[i_{\setminus j}\right]$$

- Boundary of a general chain:

$$\partial_n \sum_k \alpha_k \sigma_k \;\doteq\; \sum_k \alpha_k \partial_n \sigma_k$$

- Boundary map $\partial_n : C_n \to C_{n-1}$ is a group homomorphism

Boundary map

- Example:

$$\partial_1 ([12] + [23] + [34]) = [2] - [1] + [3] - [2] + [4] - [3] = [4] - [1].$$

Fundamental Lemma of Homology

- The boundary of the boundary of a simplex is empty:

$$
\begin{aligned}
\partial_n \partial_{n+1} [i_0 \dots i_{n+1}] &= \partial_n \left(\sum_{j=0}^{n+1} (-1)^j \left[i_{\setminus j} \right] \right) \\
&= \sum_{j<l} \left[(-1)^{j+l-1} - (-1)^{l+j} \right] \left[i_{\setminus l,j} \right] \\
&= 0
\end{aligned}
$$

Fundamental Lemma of Homology

- The boundary of the boundary of a simplex is empty:

$$\begin{aligned}\partial_n \partial_{n+1} [i_0 \ldots i_{n+1}] &= \partial_n \left(\sum_{j=0}^{n+1} (-1)^j \left[i_{\setminus j} \right] \right) \\ &= \sum_{j<l} \left[(-1)^{j+l-1} - (-1)^{l+j} \right] \left[i_{\setminus l,j} \right] \\ &= 0\end{aligned}$$

- Therefore, the boundary of the boundary of a chain is also empty, i.e., $\partial_n \partial_{n+1} C_{n+1} \equiv 0 \Rightarrow \mathrm{im} \partial_{n+1} \subset \ker \partial_n$

Cycles and boundaries

- An **n-cycle** is a chain with no boundary, e.g., $[12] + [23] + [34] + [41]$. The set of cycles: $Z_n = \ker \partial_n$ is a subgroup of C_n

Cycles and boundaries

- An **n-cycle** is a chain with no boundary, e.g., $[12] + [23] + [34] + [41]$. The set of cycles: $Z_n = \ker \partial_n$ is a subgroup of C_n
- $\mathrm{im}\partial_{n+1} \subset \ker \partial_n$ means that all boundaries of higher order chains are cycles

Cycles and boundaries

- An **n-cycle** is a chain with no boundary, e.g., $[12] + [23] + [34] + [41]$. The set of cycles: $Z_n = \ker \partial_n$ is a subgroup of C_n
- $\mathrm{im}\partial_{n+1} \subset \ker \partial_n$ means that all boundaries of higher order chains are cycles
- Some cycles (not all) are **boundaries** of higher order chains, e.g., $[23] + [31] + [12] = \partial_2 [123]$. The set of **n-boundaries**: $B_n = \mathrm{im}\partial_{n+1}$ is a subgroup of Z_n

Cycles and boundaries

- n-th Homology group: $H_n = Z_n/B_n = \ker \partial_n / \mathrm{im}\partial_{n+1}$, i.e., it is a factor group of equivalence classes, given by:

$$z \cong z' \quad \textit{iff} \quad z' - z \in B_n$$

Cycles and boundaries

- n-th Homology group: $H_n = Z_n/B_n = \ker \partial_n / \mathrm{im}\partial_{n+1}$, i.e., it is a factor group of equivalence classes, given by:

$$z \cong z' \quad iff \quad z' - z \in B_n$$

- two cycles are equivalent (**homologous**) if they differ by a boundary, say $z_1 = [12] + [23] + [34] + [41]$, $z_2 = [12] + [23] + [34] + [45] + [51]$, then:

$$\begin{aligned} z_2 - z_1 &= [45] + [51] + [14] \\ &= \partial_2 [145] \end{aligned}$$

Cycles and boundaries

- n-th Homology group: $H_n = Z_n/B_n = \ker \partial_n / \mathrm{im} \partial_{n+1}$, i.e., it is a factor group of equivalence classes, given by:

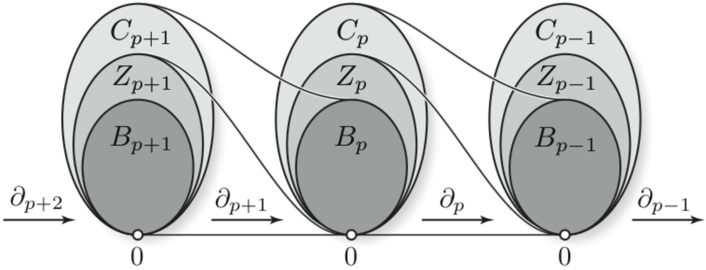
$$z \cong z' \quad iff \quad z' - z \in B_n$$

- two cycles are equivalent (**homologous**) if they differ by a boundary, say $z_1 = [12] + [23] + [34] + [41]$, $z_2 = [12] + [23] + [34] + [45] + [51]$, then:

$$\begin{aligned} z_2 - z_1 &= [45] + [51] + [14] \\ &= \partial_2 [145] \end{aligned}$$

- rank of H_n (roughly) counts the number of n-dimensional holes in the space

Chains, cycles and boundaries



Homology groups

- In general $H_n = H_n(\mathcal{X}, A)$ is a **free abelian group** which depends on the underlying topological space **and** the choice of the underlying abelian group A of winding numbers in the definition of the chain

Homology groups

- In general $H_n = H_n(\mathcal{X}, A)$ is a **free abelian group** which depends on the underlying topological space **and** the choice of the underlying abelian group A of winding numbers in the definition of the chain
- **Functoriality**: tranforming topological problems into algebraic problems. If $f : \mathcal{X} \to \mathcal{Y}$ is continuous then there is an induced homomorphism $H_n(f, A) : H_n(\mathcal{X}, A) \to H_n(\mathcal{Y}, A)$, with
  - $H_n(id_{\mathcal{X}}, A) = id_{H_n(\mathcal{X}, A)}$
  - $H_n(f, A) \circ H_n(g, A) = H_n(f \circ g, A)$

Homology groups

- In general $H_n = H_n(\mathcal{X}, A)$ is a **free abelian group** which depends on the underlying topological space **and** the choice of the underlying abelian group A of winding numbers in the definition of the chain
- **Functoriality**: tranforming topological problems into algebraic problems. If $f : \mathcal{X} \to \mathcal{Y}$ is continuous then there is an induced homomorphism $H_n(f, A) : H_n(\mathcal{X}, A) \to H_n(\mathcal{Y}, A)$, with
  - $H_n(id_{\mathcal{X}}, A) = id_{H_n(\mathcal{X},A)}$
  - $H_n(f, A) \circ H_n(g, A) = H_n(f \circ g, A)$
- If f and g are homotopic then $H_n(f, A) = H_n(g, A)$, i.e., **if topological spaces** $\mathcal{X}$ and $\mathcal{Y}$ **are homotopy equivalent then** $H_n(g, A) \circ H_n(f, A) = H_n(g \circ f, A) = H_n(id_{\mathcal{X}}, A) = id_{H_n(\mathcal{X},A)}$, i.e., **their homology groups** $H_n(\mathcal{X}, A)$ and $H_n(\mathcal{Y}, A)$ **are isomorphic**

Homology vector spaces

- If underlying group of winding numbers is a field $A = F$, then $H_n(\mathcal{X}, F)$ is a **vector space** over F

Homology vector spaces

- If underlying group of winding numbers is a field $A = F$, then $H_n(\mathcal{X}, F)$ is a **vector space** over F
- $\beta_n(\mathcal{X}, F) = \dim H_n(\mathcal{X}, F)$ is called the n-th **Betti number** of $\mathcal{X}$ w.r.t. F

Homology vector spaces

- If underlying group of winding numbers is a field $A = F$, then $H_n(\mathcal{X}, F)$ is a **vector space** over F
- $\beta_n(\mathcal{X}, F) = \dim H_n(\mathcal{X}, F)$ is called the n-th **Betti number** of $\mathcal{X}$ w.r.t. F
- If two spaces are homotopy equivalent, then all their Betti numbers are equal

Homology vector spaces

- Given the sets of n-simplices S_n, we form

Homology vector spaces

- Given the sets of n-simplices S_n, we form
  - the chain finite-dimensional vector spaces C_n

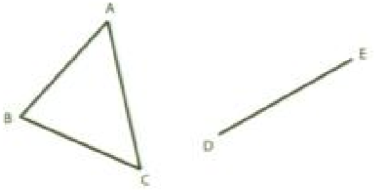
Homology vector spaces

- Given the sets of n-simplices S_n, we form
  - the chain finite-dimensional vector spaces C_n
  - boundary homomorphisms (linear maps) $\partial_n : C_n \to C_{n-1}$, which can be expressed as a sequence of matrices D_n, with

$$(D_n)_{\tau\sigma} = \begin{cases} (-1)^j & \tau \text{ is a face of } \sigma \\ 0 & \text{otherwise} \end{cases}$$

Homology vector spaces

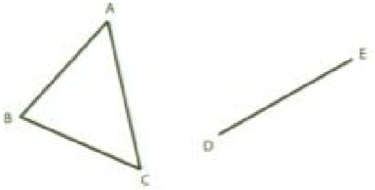
$$\begin{aligned}
\beta_n(\mathcal{X}, F) &= \dim H_n(\mathcal{X}, F) \\
&= \dim \ker \partial_n - \dim \mathrm{im}\partial_{n+1} \\
&= \dim C_n(\mathcal{X}, F) - \dim \mathrm{im}\partial_n - \dim \mathrm{im}\partial_{n+1} \\
&= \dim C_n(\mathcal{X}, F) - \mathrm{rank} D_n - \mathrm{rank} D_{n+1}
\end{aligned}$$



	AB	AC	BC	DE
A	1	1	0	0
B	1	0	1	0
C	0	1	1	0
D	0	0	0	1
E	0	0	0	1

- D_1 = incidence matrix, S_0-vertices, S_1-edges

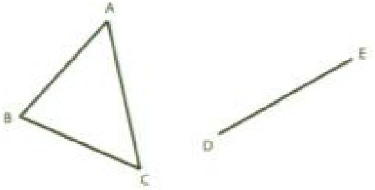
rank-nullity in graph theory



	AB	AC	BC	DE
A	1	1	0	0
B	1	0	1	0
C	0	1	1	0
D	0	0	0	1
E	0	0	0	1

- D_1 = incidence matrix, S_0-vertices, S_1-edges
- #connected components=#nodes-rank(D_1)

rank-nullity in graph theory



	AB	AC	BC	DE
A	1	1	0	0
B	1	0	1	0
C	0	1	1	0
D	0	0	0	1
E	0	0	0	1

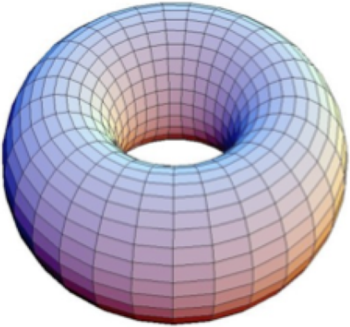
- D_1 = incidence matrix, S_0-vertices, S_1-edges
- #connected components=#nodes-rank(D_1)
- #loops=#edges-rank(D_1)

Betti numbers

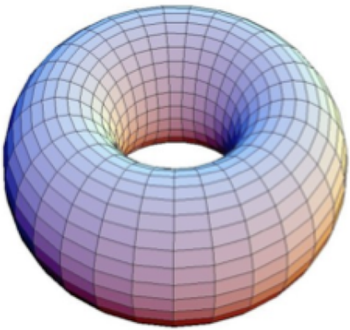
Betti numbers

$$\beta_0 = 1,\ \beta_1 = 1,\ \beta_k = 0, \text{ for } k \geq 2$$

Betti numbers



Betti numbers



$$\beta_0 = 1,\ \beta_1 = 2,\ \beta_2 = 1,\ \beta_k = 0 \text{ for } k \geq 3$$

The story so far

- If someone gave us the topological space $\mathcal{X}$ which consists of sets of points, edges, triangles, ..., n-simplices, we can compute its Betti numbers over, say, $\mathbb{F}_2$ using linear algebra (simplicial homology)

The story so far

- If someone gave us the topological space $\mathcal{X}$ which consists of sets of points, edges, triangles, ..., n-simplices, we can compute its Betti numbers over, say, $\mathbb{F}_2$ using linear algebra (simplicial homology)
- Only got 0-simplices - we have to build the higher order structure into data, i.e. form the **simplicial complex**

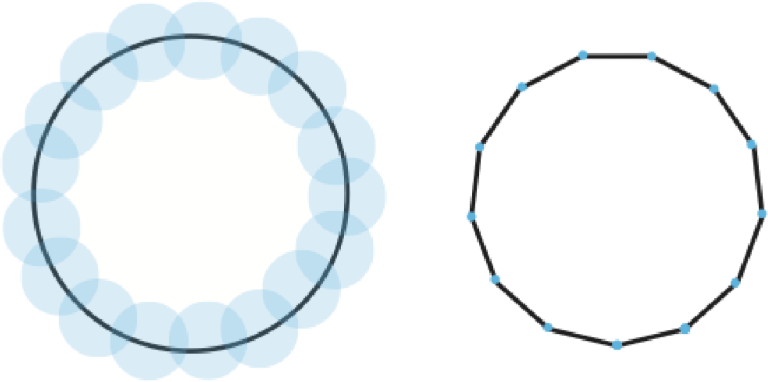
Čech complex

- Čech complex (**nerve**) $\check{\mathfrak{C}}(\epsilon)$ of data $\{Y_i\}_{i=1}^N$ contains:
  - 0-simplices $[i]$
  - 1-simplices $[ij]$ whenever $\|Y_i - Y_j\| \leq \epsilon$
  - n-simplices $[i_0 \ldots i_n]$ whenever $\bigcap_{j=0}^n U_{i_j} \neq \emptyset$,
    $U_{i_j} = \{y \in \mathcal{Y} \ : \ \|y - Y_{i_j}\| \leq \epsilon\}$

Čech complex

- Čech complex (**nerve**) $\check{\mathfrak{C}}(\epsilon)$ of data $\{Y_i\}_{i=1}^N$ contains:
  - 0-simplices $[i]$
  - 1-simplices $[ij]$ whenever $\|Y_i - Y_j\| \leq \epsilon$
  - n-simplices $[i_0 \ldots i_n]$ whenever $\bigcap_{j=0}^n U_{i_j} \neq \emptyset$,
    $U_{i_j} = \{y \in \mathcal{Y} \,:\, \|y - Y_{i_j}\| \leq \epsilon\}$
- **The nerve theorem**: In a general topological space $\mathcal{X}$, the nerve $N(\mathcal{U})$ is associated to an open covering $\mathcal{U} = \{U_i\}_{i \in I}$. $N(\mathcal{U})$ is homotopy equivalent to $\mathcal{X}$ whenever every U_i is contractible (homotopy equivalent to a point).

Čech complex



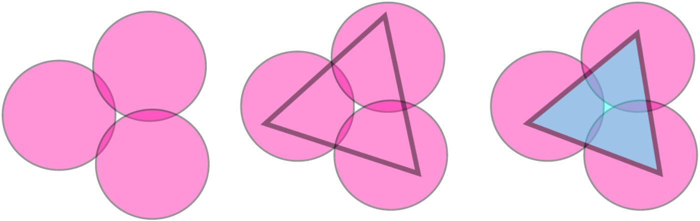
Vietoris-Rips complex

- VR complex $VR(\epsilon)$ contains:
  - 0-simplices $[i]$
  - 1-simplices $[ij]$ whenever $\|Y_i - Y_j\| \leq \epsilon$
  - n-simplices $[i_0 \ldots i_n]$ whenever all its faces are in $VR(\epsilon)$.

Vietoris-Rips complex

- VR complex $VR(\epsilon)$ contains:
  - 0-simplices $[i]$
  - 1-simplices $[ij]$ whenever $\|Y_i - Y_j\| \leq \epsilon$
  - n-simplices $[i_0 \dots i_n]$ whenever all its faces are in $VR(\epsilon)$.
- $\check{\mathfrak{C}}(\epsilon) \subseteq VR(2\epsilon) \subseteq \check{\mathfrak{C}}(2\epsilon)$

$\check{\mathfrak{C}}(\epsilon)$ vs $VR(\epsilon)$



- Vietoris-Rips is the maximal simplicial complex that can be built on top of the 1-simplicial skeleton (*flag complex*)

Witness complexes

- Choose a set of **landmark points** $\mathcal{L} \subset \{Y_i\}_{i=1}^N$ - this is the set of 0-simplices
- Strong witness complex:
  - $[l_0 \dots l_n] \in W^s(\epsilon)$ iff $\exists Y$ (a strong witness): $d(Y, l_j) \leq d(Y, \mathcal{L}) + \epsilon$, $\forall j = 0, \dots, n$
- Weak witness complex:
  - $[l_0 \dots l_n] \in W^w(\epsilon)$ iff $\exists Y$ (a weak witness): $d(Y, l_j) \leq d(Y, \mathcal{L} \backslash \{l_0 \dots l_n\}) + \epsilon$, $\forall j = 0, \dots, n$

How to choose ϵ?

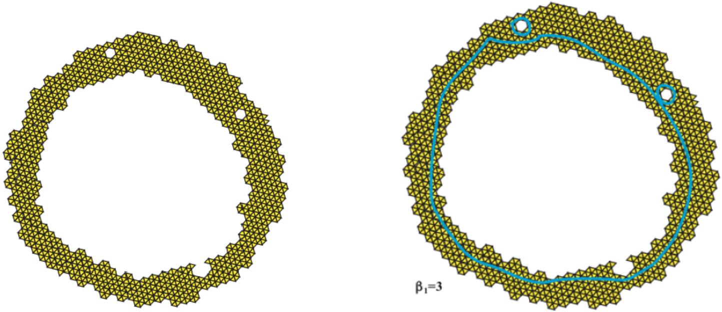


Figure: Scale ϵ_1: $\beta_0 = 1$, $\beta_1 = 3$

How to choose ϵ?

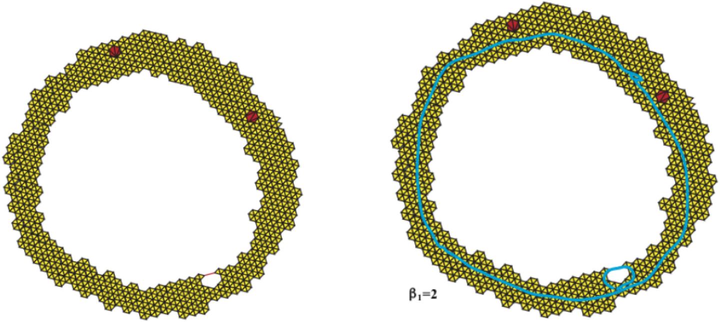


Figure: Scale ϵ_1: $\beta_0 = 1$, $\beta_1 = 2$

Persistence

- $\mathfrak{C}(\epsilon) \subset \mathfrak{C}(\epsilon')$ whenever $\epsilon \leq \epsilon'$

Persistence

- $\mathfrak{C}(\epsilon) \subset \mathfrak{C}(\epsilon')$ whenever $\epsilon \leq \epsilon'$
- Using inclusion $\iota : \mathfrak{C}(\epsilon) \to \mathfrak{C}(\epsilon')$, we get a homomorphism $H_n(\iota, F) : H_n(\mathfrak{C}(\epsilon), F) \to H_n(\mathfrak{C}(\epsilon'), F)$ (and can study the image of the homology of a smaller complex in the homology of a larger complex)

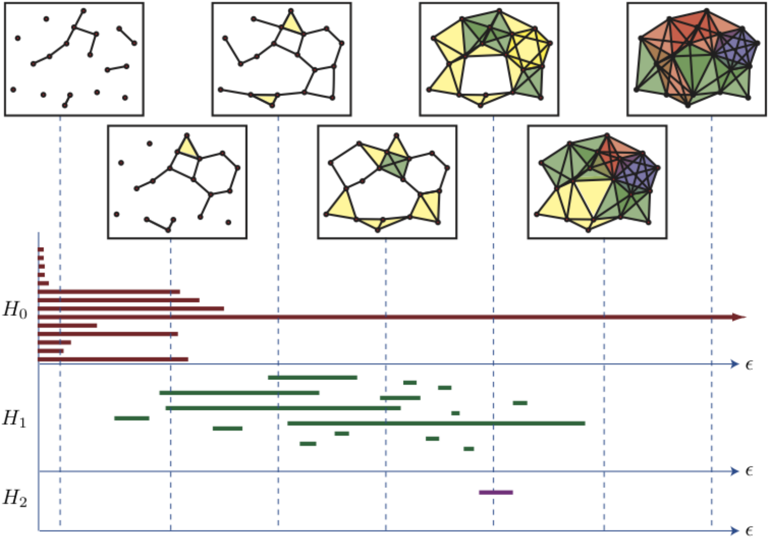
Persistence

- $\mathfrak{C}(\epsilon) \subset \mathfrak{C}(\epsilon')$ whenever $\epsilon \leq \epsilon'$
- Using inclusion $\iota : \mathfrak{C}(\epsilon) \to \mathfrak{C}(\epsilon')$, we get a homomorphism $H_n(\iota, F) : H_n(\mathfrak{C}(\epsilon), F) \to H_n(\mathfrak{C}(\epsilon'), F)$ (and can study the image of the homology of a smaller complex in the homology of a larger complex)
- two small cycles in the smaller complex vanish in the larger complex, the small cycle in the larger complex is not in the image of $H_n(\iota, F)$, only the largest cycle persists

Persistence

- $\mathfrak{C}(\epsilon) \subset \mathfrak{C}(\epsilon')$ whenever $\epsilon \leq \epsilon'$
- Using inclusion $\iota : \mathfrak{C}(\epsilon) \to \mathfrak{C}(\epsilon')$, we get a homomorphism $H_n(\iota, F) : H_n(\mathfrak{C}(\epsilon), F) \to H_n(\mathfrak{C}(\epsilon'), F)$ (and can study the image of the homology of a smaller complex in the homology of a larger complex)
- two small cycles in the smaller complex vanish in the larger complex, the small cycle in the larger complex is not in the image of $H_n(\iota, F)$, only the largest cycle persists
- incremental computation of Betti numbers

Persistent homology barcode



Natural image statistics

- 3x3 patches from a database of black and white images - each datapoint is a vector in $\mathbb{R}^9$

Carlsson et al, *On the local behaviour of spaces of natural images*, International Journal of Computer Vision 2008

Natural image statistics

- 3x3 patches from a database of black and white images - each datapoint is a vector in $\mathbb{R}^9$
- remove the low contrast (nearly constant) patches

Carlsson et al, *On the local behaviour of spaces of natural images*, International Journal of Computer Vision 2008

Natural image statistics

- 3x3 patches from a database of black and white images - each datapoint is a vector in $\mathbb{R}^9$
- remove the low contrast (nearly constant) patches
- mean-center - "turning the brightness knob"

Carlsson et al, *On the local behaviour of spaces of natural images*, International Journal of Computer Vision 2008

Natural image statistics

- 3x3 patches from a database of black and white images - each datapoint is a vector in $\mathbb{R}^9$
- remove the low contrast (nearly constant) patches
- mean-center - "turning the brightness knob"
- normalize the contrast - "turning the contrast knob"

Carlsson et al, *On the local behaviour of spaces of natural images*, International Journal of Computer Vision 2008

Natural image statistics

- 3x3 patches from a database of black and white images - each datapoint is a vector in $\mathbb{R}^9$
- remove the low contrast (nearly constant) patches
- mean-center - “turning the brightness knob”
- normalize the contrast - “turning the contrast knob”
- the points sit on a 7D ellipsoid in $\mathbb{R}^8$, but not uniformly

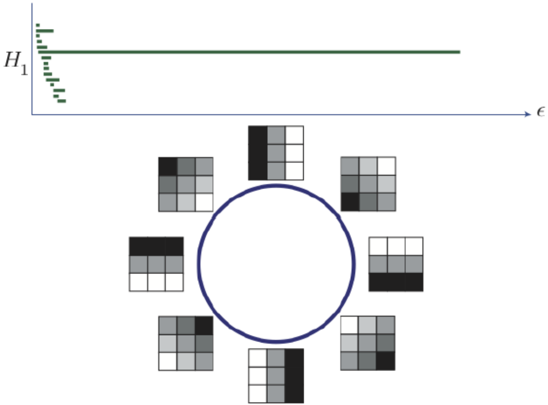
Carlsson et al, *On the local behaviour of spaces of natural images*, International Journal of Computer Vision 2008

Natural image statistics

- 3x3 patches from a database of black and white images - each datapoint is a vector in $\mathbb{R}^9$
- remove the low contrast (nearly constant) patches
- mean-center - "turning the brightness knob"
- normalize the contrast - "turning the contrast knob"
- the points sit on a 7D ellipsoid in $\mathbb{R}^8$, but not uniformly
- exploring the high-density regions, using the k-codensity proxy $\delta_k(x) = \|x - \nu_k(x)\|$

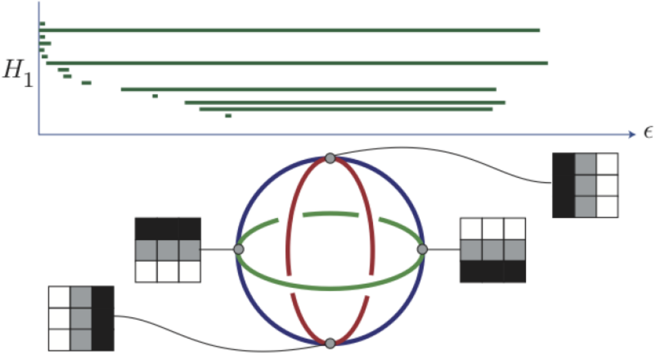
Carlsson et al, *On the local behaviour of spaces of natural images*, International Journal of Computer Vision 2008

Natural image statistics



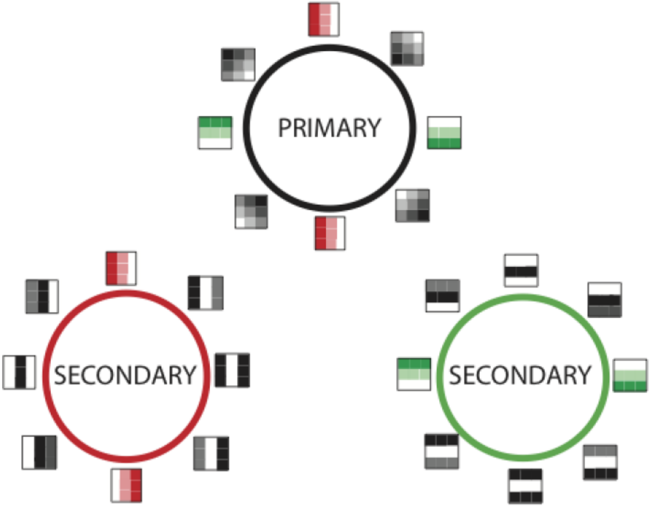
- $k = 300$, top 25% “densest points” - the underlying space appears to form a circle

Three-circle model

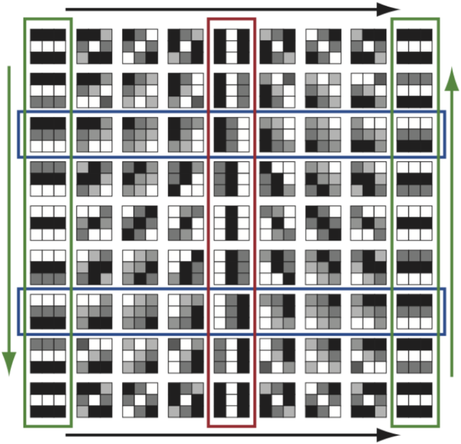


- $k = 15$, top 25% "densest points" leads to $\beta_1 = 5$
- green and red circles do not touch, each touches the blue circle

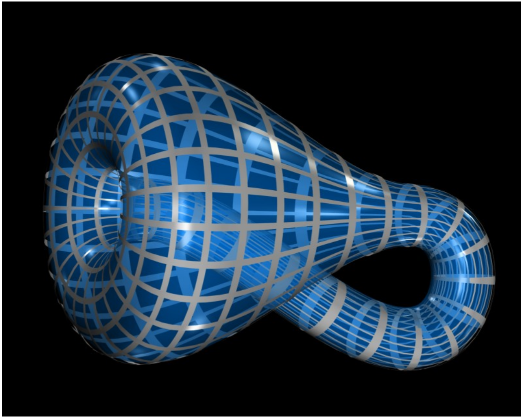
Three-circle model



Three-circle model



Klein bottle!



A mathematician named Klein

Thought the Möbius band was divine.

Said he: "If you glue

The edges of two,

You'll get a weird bottle like mine."

V1 data

- recordings from 10x10 electrode arrays from the V1 in Macaque monkeys (20-30 minutes):
  - **spontaneuous** / no stimulus presented
  - **evoked** / video sequences presented

Singh et al, *Topological Structure of Population Activity in Primary Visual Cortex*, Journal of Vision 2008

V1 data

- recordings from 10x10 electrode arrays from the V1 in Macaque monkeys (20-30 minutes):
  - **spontaneuous** / no stimulus presented
  - **evoked** / video sequences presented
- each data segment consists of 200 50ms bins - for each neuron a firing count within such bin is recorded

Singh et al, *Topological Structure of Population Activity in Primary Visual Cortex*, Journal of Vision 2008

V1 data

- recordings from 10x10 electrode arrays from the V1 in Macaque monkeys (20-30 minutes):
    - **spontaneuous** / no stimulus presented
    - **evoked** / video sequences presented
- each data segment consists of 200 50ms bins - for each neuron a firing count within such bin is recorded
- five neurons with highest firing rate: data point cloud $\mathcal{X}$ is 200 points in $\mathbb{R}^5$

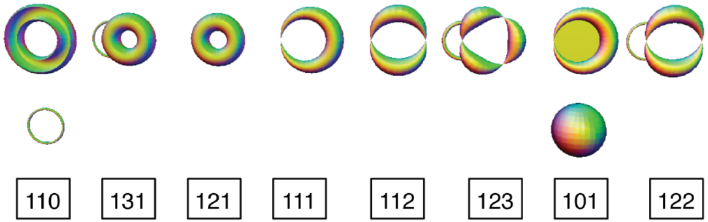
Singh et al, *Topological Structure of Population Activity in Primary Visual Cortex*, Journal of Vision 2008

V1 data

- recordings from 10x10 electrode arrays from the V1 in Macaque monkeys (20-30 minutes):
  - **spontaneuous** / no stimulus presented
  - **evoked** / video sequences presented
- each data segment consists of 200 50ms bins - for each neuron a firing count within such bin is recorded
- five neurons with highest firing rate: data point cloud $\mathcal{X}$ is 200 points in $\mathbb{R}^5$
- For each data segment, construct a *witness complex*, and obtain its Betti signature $(\beta_0, \beta_1, \beta_2)$

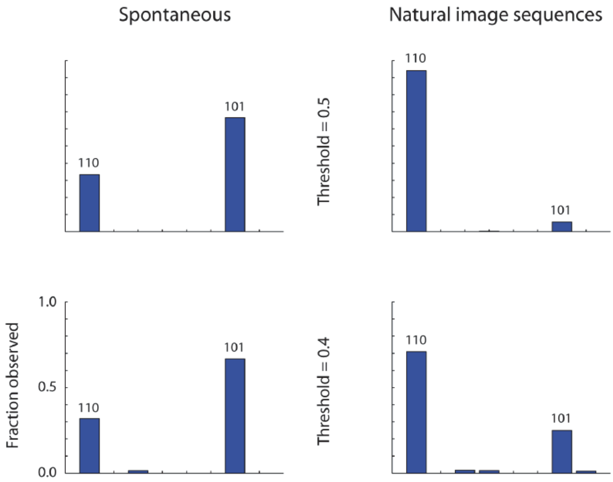
Singh et al, *Topological Structure of Population Activity in Primary Visual Cortex*, Journal of Vision 2008

V1 data - the observed signatures



- the most frequently occurring signatures are 110 (circle) and 101 (sphere)

V1 data - the observed signatures



V1 data - the observed signatures

- the data sets under two regimes are topologically different

V1 data - the observed signatures

- the data sets under two regimes are topologically different
- significance validation of observed Betti numbers:
  - simulate firings from a Poisson model
  - frequency of obtaining persistent segments of β_1, β_2 is $< .005$

V1 data - the observed signatures

- the data sets under two regimes are topologically different
- significance validation of observed Betti numbers:
  - simulate firings from a Poisson model
  - frequency of obtaining persistent segments of β_1, β_2 is $< .005$
- topology distinguishes both data sets from the Poisson model and from each other; the nature of the "circular" topological phenomenon?

V1 data - the observed signatures

- the data sets under two regimes are topologically different
- significance validation of observed Betti numbers:
  - simulate firings from a Poisson model
  - frequency of obtaining persistent segments of β_1, β_2 is $< .005$
- topology distinguishes both data sets from the Poisson model and from each other; the nature of the "circular" topological phenomenon?
- not likely due to periodicity of body's natural rhytms - no peaks in the amplitude spectrum observed

Code

- Toolbox: JPlex (http://comptop.stanford.edu/)
    - Java version of Plex, work with Matlab
    - Rips, Witness complex, Persistence Homology, barcodes
- Other Choices: Plex 2.5/Matlab (not maintained any more), Dionysus (Dimitry Morozov)