Persistent Homology and the Analysis of High Dimensional Data

Symposium on the Geometry of Very Large Data Sets

University of Ottawa

February 24, 2005

Gunnar Carlsson

Department of Mathematics

Stanford University

(gunnar@math.stanford.edu)

Webpage: http://math.stanford.edu/comptop/

Collaborators: P. Diaconis, L. Guibas, V. de Silva, A. Collins, A. Zomorodian, T. Ishkanov, S. Holmes, D. Ringach

Lecture I: Connectivity

Information for Point Cloud Data

Qualitative Properties of Data

- Regression : very important method for the analysis of various kinds of data
- Typically uses a theoretical model with various parameters, and finds optimal values of the parameters
- Other methods provide quantitative information about specific aspects of the data, such as measures of spread, average, significance, clustering, etc.

Sometimes certain kinds of **qualitative properties** are important in obtaining an overall understanding of the nature of the data because

- Can suggest the form of a theoretical model
- Precise quantitive models are too complicated to obtain
- Qualitative information is sometimes actually more important than more precise quantitative information

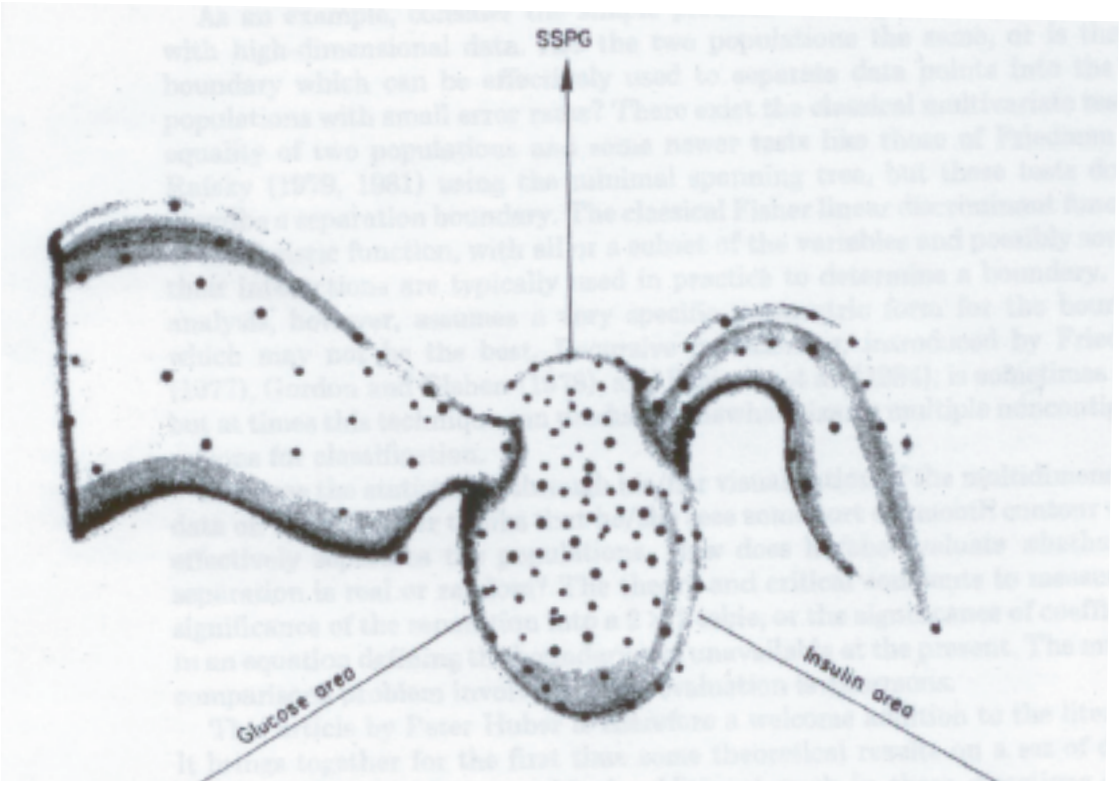
Example: Miller-Reaven Diabetes Study

- Study carried out in 1976, on 145 patients at Stanford Hospital
- Most patients had a form of diabetes, although some were normal
- For each patient, four metabolic variables (involving insulin response and glucose tolerance) along with relative weight were measured, giving a set of 145 points in 5D space

How to analyze the data to get information about the nature of diabetes? Two problems are:

- Data is too high-dimensional to visualize
- No accepted theoretical model to describe the data

Miller and Reaven used the **Projection Pursuit** method to obtain a useful projection of into 3D space



From: *Annals of Statistics*, Vol. 13, No. 2 June, 1985

Central core: Normal patients

Lobes: Type I and Type II diabetes, respectively

Conclusion: There are two essentially distinct forms of the disease, one early onset and the other adult onset

Families of Images as Data Sets

- An image taken by a black and white digital camera can be viewed as a point in **pixel space**, with a gray scale coördinate for each pixel

- A family of pictures taken of a dynamic scene gives a data set in pixel space. Its geometry should be related to the geometry of the "phase space" for the problem

- Images can be viewed as an *exotic coördinate system* of the phase space

- These coördinatizations are

  – Very high dimensional

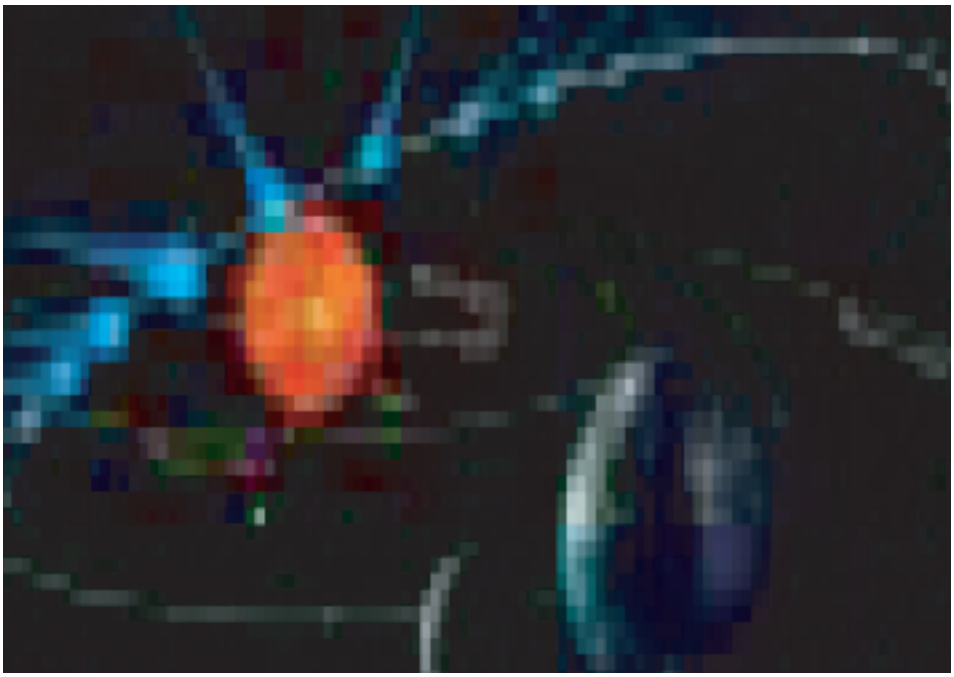
  – Highly non-linear

  – Not smooth

Question: Is it possible to obtain qualitative information from the family of images directly without reconstructing the geometry of the phase space?

Example: Periodic Motion in Space

- Data consists of pictures taken of a region in space with digital camera
- Don't have times at which the pictures are taken
- **Question:** Is there periodic motion going on in the region?
- Are **not** asking for the form of the orbit (circular, elliptical, etc.)



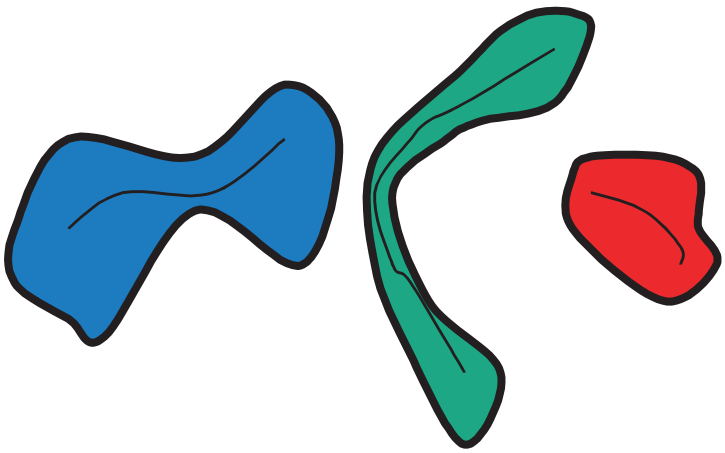
Example: Lung Cancer Imaging

- 3D radiological images of cancerous lungs show both tumors and blood vessels as areas of increased density
- Blood vessels show up as long **tunnels** in the image
- Tumors show up as balls
- **Question:** How to distinguish automatically between tumors and blood vessels?
- It is **not important** to know exact shape of tumor or blood vessel, at least initially
- The **qualitative** nature of the objects is crucial, not quantitative

What Kind of Qualitative Information Do We Want?

Connected components

- Two points x, y in a space X are connected by a path if there is a continuous map $\varphi : [0, 1] \to X$ so that $\varphi(0) = x$ and $\varphi(1) = y$.
- The **connected component** of a point $x \in X$ is the collection of all points connected to x by a path
- The collection of connected components of points in X forms a **partition** of X



- The number of connected components of X is a **discrete invariant** of X

- Call it **zeroth order connectivity information**

- **Connection to the diabetes problem:**

  If we remove the normal patients, the remaining patients decompose into two connected components

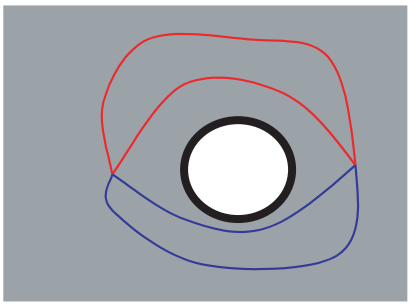
- Forming the collection of connected components is idealized form of **clustering**

- A space is (path-)connected if it consists of one component

First Order Connectivity Information

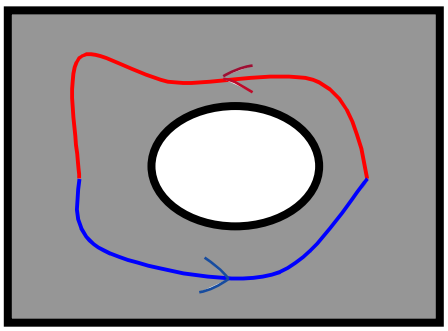
- When a space is connected, pairs of points may be connected in different ways

- We say one path is *essentially the same* as another if it can be deformed to it



The two red paths are essentially the same, as are the two blue paths, but the red paths are essentially distinct from the blue paths

Useful Reformulation: The existence of essentially distinct paths is reflected in the presence of *essentially distinct loops*



- The two distinct paths between points are "glued together" at their endpoints to form a loop in the space

- In this case, the essentially distinct loops are parametrized by the integers

- Parametrization is via *winding number*

- First order connectivity information can be applied to the periodic motion problem

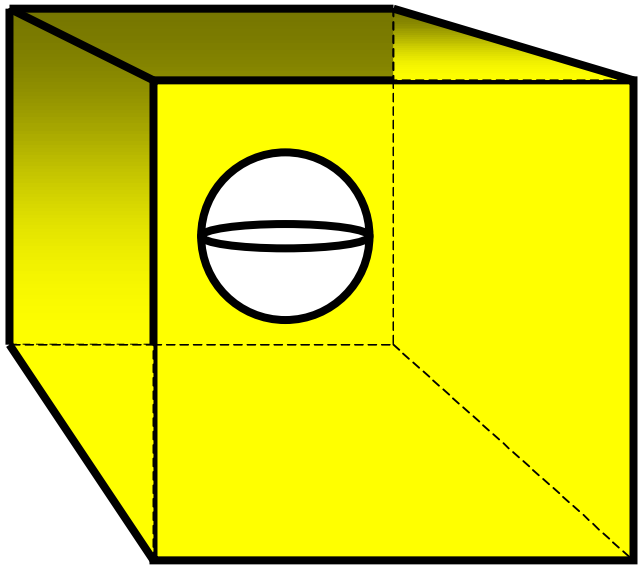
- If there is an object undergoing periodic motion in the family of images, there will be an essential closed loop in the phase space

- If the phase space is captured by the family of images, we find that the essential closed loops **in the space of images** are also parametrized by the integers

- If there are several objects undergoing periodic motion, the closed loops will be parametrized by vectors of integers

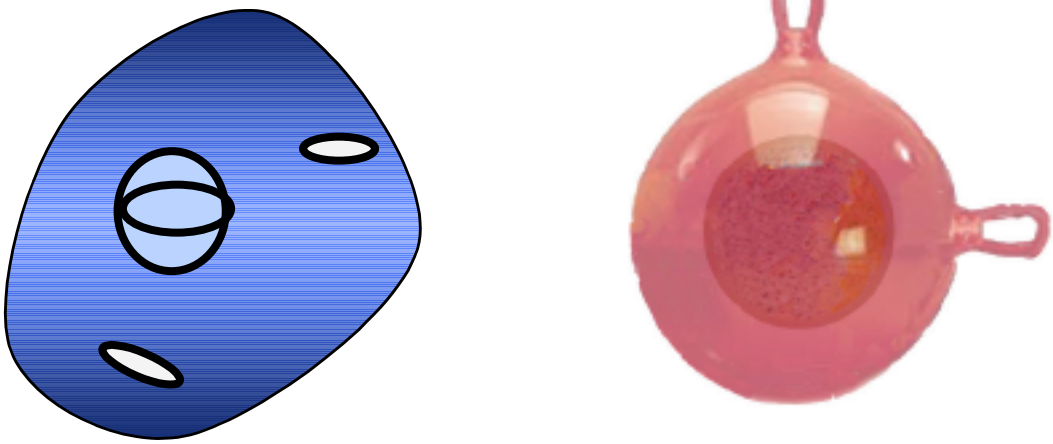
Second and Higher Order Connectivity Information

Suppose we have a void drilled out of a solid rectangular region



- The remaining space has only one component
- Any path is essentially equivalent to any other

In this case, a there is a 2D surface or *cycle* surrounding the void



The cycle is essential in that it cannot be dragged around and off the obstacle, just like an essential loop could not be dragged around the obstacle

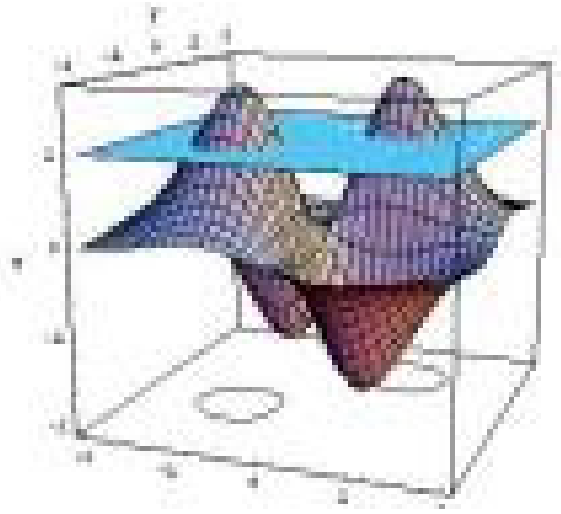
Connection with the lung cancer problem:

- If we remove the regions in the lung with higher density, the tumors are voids and the blood vessels are tunnels
- Tunnels are detected by presence of essential loops
- Voids are detected the presence of essential cycles

So, if no essential cycles, then no tumors

Connectivity Information and Optimization

Let X be a space, and let $f : X \to \mathbb{R}$ be a continuous function. For $r \in \mathbb{R}$, the *excursion set* for f at r, $E_r(f)$, is the set $\{x \in X | f(x) \leq r\}$.



Connectivity information in $E_r(f)$ reflects the presence of local maxima for f at which the value of f is $> r$

This idea is called **Morse theory**, f is called a *Morse function* if it satisfies certain non-degeneracy conditions

How to Make Connectivity Information into Precise Mathematics?

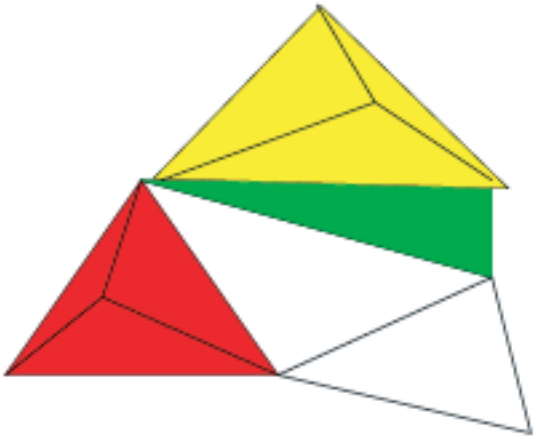
Simplicial Complexes

- **n-Simplex:** Convex hull of $(n+1)$ points in general position

- **Simplicial Complex:** A space written as a union of simplices which intersect each other in faces (subsimplices)

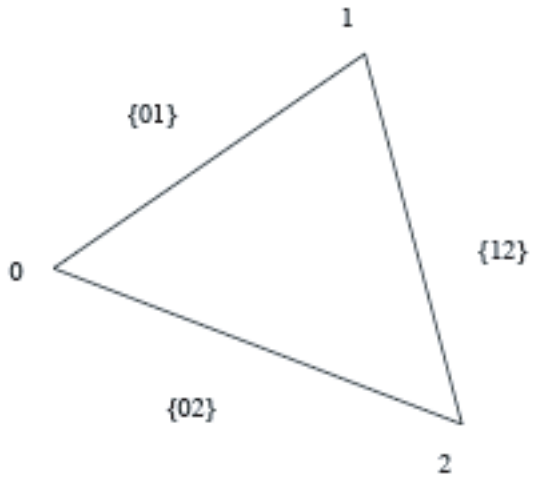
- **Abstract Simplicial Complex:** A pair (V, Σ), where V is a finite set, and Σ is a family of non-empty subsets of V, so that $\sigma \subseteq \tau \in \Sigma$ implies $\sigma \in \Sigma$.

- Abstract simplicial complexes determine actual simplicial complexes



Connected Components and Essential Loops for Graphs

- **Graph:** One-dimensional simplicial complex
- Set up a matrix ("boundary matrix") corresponding to graph
- Rows correspond to vertices
- Columns correspond to edges



Entries in the Boundary Matrix

- Entry in intersection of a row and a column depends on relationship of corresponding v and edge e
- Entry is 0 if $v \notin e$
- Entry is -1 if v is initial vertex in e
- Entry is +1 if v is terminal vertex in e

$$\begin{array}{c|cccc} & (01) & (02) & (12) \\ 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & -1 \\ 2 & 0 & 1 & 1 \end{array}$$

Connectivity Information and Boundary Matrix

- Perform Gaussian elimination on boundary matrix

- Number of essentially different loops

  = Dim(null space of boundary matrix)

  = # (free variables)

- Number of distinct components

  = Dim(complement to column space of boundary matrix)

  = # (rows) - # (pivot rows)

Higher Dimensional Complexes

- Construct boundary matrices ∂_k for each $k \geq 0$, with $\partial_k \partial_{k+1} = 0$
- Rows of ∂_k correspond to k-simplices of simplicial complex X
- Columns of ∂_k correspond to $(k+1)$-simplices of X
- Essential k- dimensional cycles are now counted as

Dim(Null space of ∂_{k-1}) -

Dim(Complement of column space of ∂_k)

- Precise mathematical formulation of "higher dimensional cycles" for spaces given as simplicial complex

Homology

- **Observation of E. Noether:** #(Essential cycles) can be viewed as the dimension of a vector space

- Proposed that there is value in studying the vector spaces rather than just their dimensions

- For every $k \geq 0$, obtain a vector space $H_k(X)$ for any simplicial complex

- Example of "categorification"

- Dimension of $H_0(X)$ is the number of connected components of X

- Dimension of $H_1(X)$ is the number of essentially different loops

- H_i defined for any topological space X, using the *singular complex*, roughly a simplicial complex with a k-simplex for every continuous map $\Delta^k \to X$

Properties of Homology

- **Functoriality:** For a map of simplicial complexes $f : X \to Y$, there is an induced linear transformation (matrix)

$$H_k(f) : H_k(X) \to H_k(Y)$$

for each $k \geq 0$

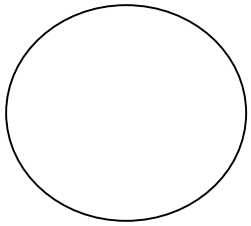
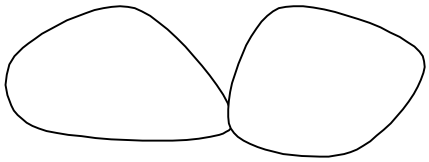
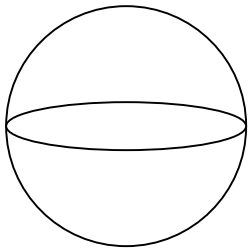
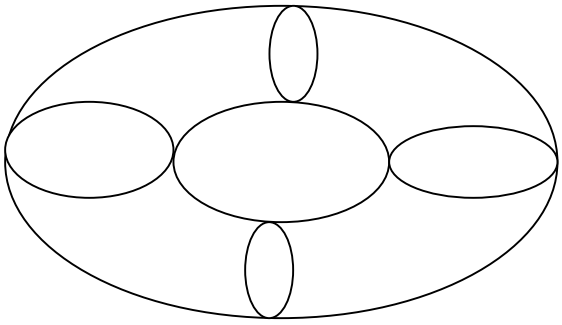
- **Homotopy invariance:** For every pair of *homotopic maps* $f, g : X \to Y$ (i.e. so that there is a map

$$\mathcal{H} : X \times [0, 1] \to Y$$

such that $\mathcal{H}(x, 0) = f(x)$ and $\mathcal{H}(x, 1) = g(x)$), $H_k(f) = H_k(g)$ for all $k \geq 0$

- **Importance of properties:**
  - Key tools for computing homology
  - Critical for making computations when spaces are not given in closed form

Examples

	$\beta_0 = \beta_1 = 1$ $\beta_i = 0 \quad for \quad i > 1$
	$\beta_0 = 1 \quad \beta_1 = 2$ $\beta_i = 0 \quad for \quad i > 1$
	$\beta_0 = \beta_2 = 1$ $\beta_i = 0 \quad otherwise$
S^n	$\beta_0 = \beta_n = 1$ $\beta_i = 0 \quad otherwise$
	$\beta_0 = \beta_2 = 1, \ \beta_1 = 2$ $\beta_i = 0 \quad for \quad i > 2$

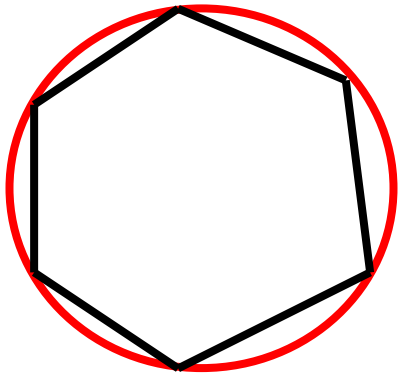
Homology and Data

- **Question:** How to obtain qualitative information about spaces underlying sets of data?

- Problems:
  - Finite sets of points are always *discrete*, carry no higher dimensional homology
  - Real world data is generally noisy, so don't have points exactly on the underlying space
  - Homology is an integer valued invariant, can't recognize error vs. real phenomena

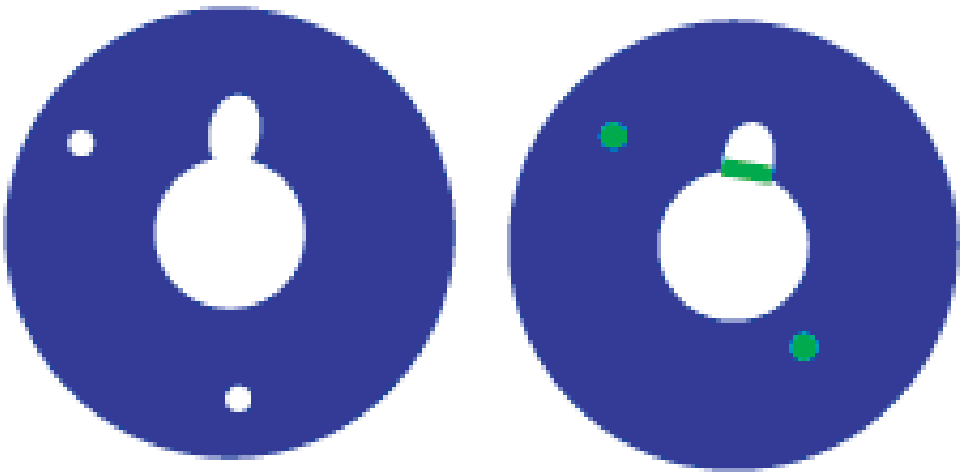
Rips Complex

- $\mathcal{X}$ a set of data in $\mathbb{R}^n$
- $\varepsilon \geq 0$ a parameter
- $R(\mathcal{X}, \varepsilon)$ has vertex set $\mathcal{X}$, and $\{x_0, x_1, \ldots, x_k\}$ span a k-simplex iff $d(x_i, x_j) \leq \varepsilon$ for all i, j
- For suitable values of ε, $R(\mathcal{X}, \varepsilon)$ is often homotopy equivalent to the underlying space, when $\mathcal{X}$ is obtained by sampling from a space in $\mathbb{R}^n$



However -

- Difficult to know how to choose ε
- In many cases, there is **no** ε which works



Solution: Persistence

(Edelsbrunner, Letscher, Zomorodian)

- Study all $R(\mathcal{X}, \varepsilon)$ at once

- Whenever $\varepsilon < \varepsilon'$, have a simplicial inclusion

$$R(\mathcal{X}, \varepsilon) \hookrightarrow R(\mathcal{X}, \varepsilon')$$

- System of complexes and inclusions carries more information than the individual complexes

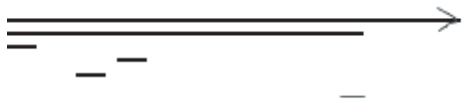
- The maps induced by the inclusions using *functoriality* are the key additional piece of information

- Obtain a *persistence vector space*, i.e. a family of vector spaces V_ε together with linear transformations $V_\varepsilon \to V_{\varepsilon'}$ whenever $\varepsilon \leq \varepsilon'$

Classification

- Ordinary vector spaces are classified by an integer, the dimension

- Persistence vector spaces are classified by a **barcode**, i.e. a finite family of intervals on the non-negative real line

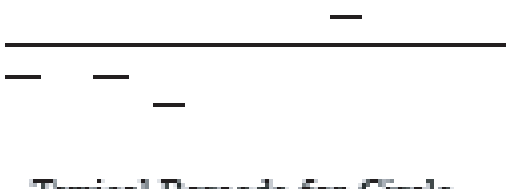


- Barcodes are computationally tractable - roughly as computable as a single complex

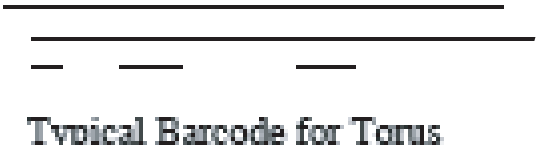
- By taking Rips complexes, any set of point cloud data gives a barcode in each dimension k

Measuring Homology

- Homology is difficult to measure, since it is integer valued, so very sensitive to noise
- Persistence solves this problem
- Long intervals correspond to "geometric" classes, from the underlying space from which we sample
- Short intervals correspond to noise



Typical Barcode for Circle



Typical Barcode for Torus

We are now in a position to *measure* connectivity information, i.e to evaluate it in situations with noise and with incomplete information

Goal: Utilize these techniques to obtain qualitative information about real world data

Lecture II: Applications

The Mumford-Lee-Pedersen Set

- Pictures with digital camera can be viewed as vectors in high-dimensional vector space (pixel space)

- One coördinate for each pixel, the coördinate is the corresponding gray scale value for that pixel

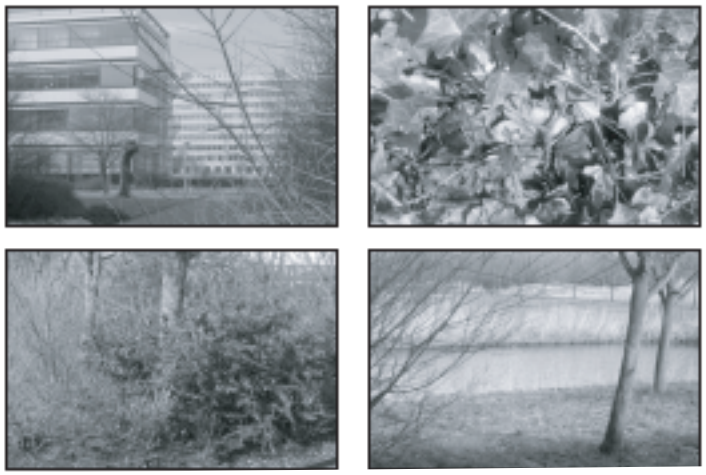
- **Mumford's first question:** If we take many pictures (with no particular subject in mind), what can be said statistically about the set of vectors in pixel space obtained from photos?

- **Mumford's second question:** Can anything be said about the projections of the vectors on coördinates corresponding to pixels in 3×3 square pixel arrays?

- One is studying a set of vectors in 9-dimensional space

- In *The nonlinear statistics of high-contrast patches in natural images*, International Journal of Computer Vision, Vol. 54, 83-103, 2003, D. Mumford, A. Lee, and K. Pedersen construct and study a data set of such pixel patches

- They work from a database of natural images compiled by J. van Hateren and A. van der Schaaf



- **First observation:** Most patches will be essentially constant, since there are many solid regions in most images

- Mumford et al remove **low contrast patches**, i.e. patches in which all coördinates are within a threshhold of their mean

- They also normalize the patches so that their mean value is zero, by subtracting the mean from all coördinates in the patch, obtaining a vector in 8-dimensional space

- They normalize so that the length of the vectors is 1. Possible since low-contrast patches have been removed, so the points lie away from the origin in $\mathbb{R}^8$

- Result is a data set $\mathcal{M}$ of over 8×10^6 points in $S^7 \subseteq \mathbb{R}^8$

Question: What can be said about the set $\mathcal{M}$? Does it fill out the 7-sphere?

Initial answer: Yes, in the sense that points of $\mathcal{M}$ occur throughout S^7

- However, not all regions in S^7 are equally densely populated by points in $\mathcal{M}$
- Suggests that one should study qualitatively the points of *highest density*, suitably defined
- Ultimately, information of this kind should be useful for thinking about compression of images

Density Estimation

- Highly developed area in statistics, with many interesting methods

- Our choice is *nearest neighbor estimation*, but would be interesting to consider other methods

Definition: For any subset $X \subseteq \mathbb{R}^n$, and any $k \geq 0$, define a function $\delta_k : X \to [0, +\infty)$ by

$\delta_k(x) =$ distance from x to its k-th nearest neighbor in X

- Large values of δ_k indicate *sparse* points, small values *dense* points

- Large values of k mean we are measuring density by considering density of large neighborhoods of x, smaller values mean we are using smaller neighborhoods

- Density function corresponding to large k should be viewed as a "smoothed out version" of density function corresponding to smaller values of k

- We have a parametrized family of density measures on our set X, each one giving potentially different answers

- Statistics has criteria to suggest which values of k are "best"; we want to use all the measures to get information about the data set

What we do

1. Select a value of k

2. Select a percentage threshhold T, and construct the subset $\mathcal{M}_T^k$ of all points which are among the $T\%$ densest points as measured by δ_k

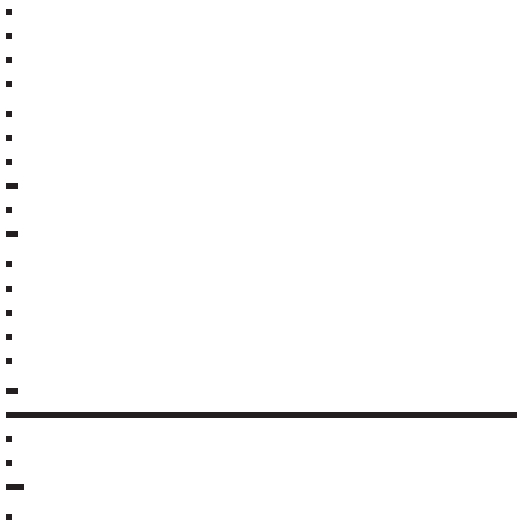
3. Build the family of Rips complexes on $\mathcal{M}_T^k$

4. Apply certain techniques we have developed (landmarking, witness complexes) to shrink the size of the complexes and minimize the number of small intervals in the barcode output

We have carried this out for various values of k and T, and obtained interesting barcodes and interpretations of these barcodes

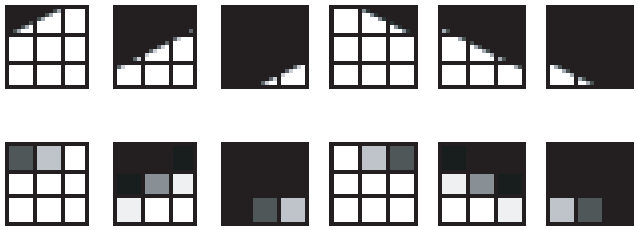
The case of large k

Applying this methodology for the case of $k = 300$, with 5×10^4 points sampled at random from $\mathcal{M}$, and selecting $T = 25$, we obtain the following barcode for one-dimensional homology



Interpretation

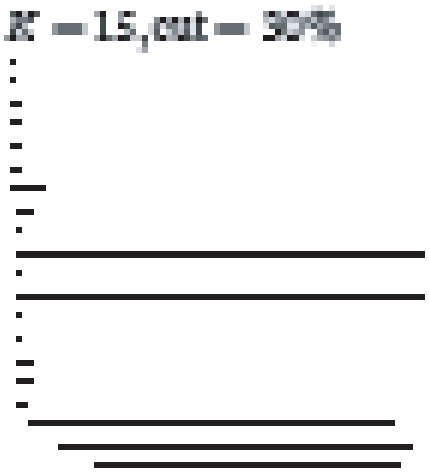
- Suggests the possibility that the space is a circle
- **Plausible explanation:** The space of densest points are obtained by evaluating a non-trivial function in the two space variables at the nine pixels in the patch



The space is in fact more like and annulus, with the angular part being the angle of the line between the light and dark regions, and the radial variable being the distance of that line from the origin

The case of small k

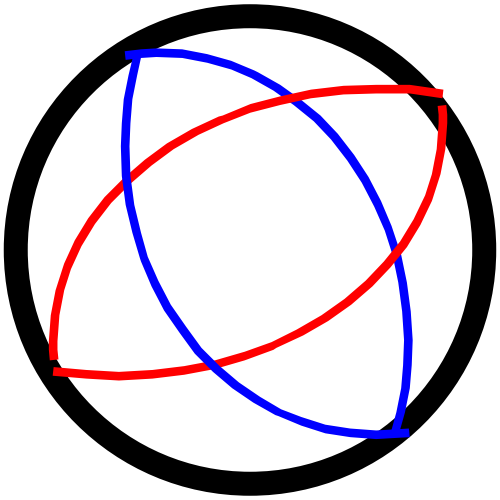
When we apply the same method to the case $k = 15$, with $T = 30$, and sampling 5×10^4 points from $\mathcal{M}_T^k$, we obtain the following barcode



Note that there are actually five long lines in this picture, suggesting presence of 5 loops. Result is robust, in that it recurs after resampling many times

Geometric Explanation

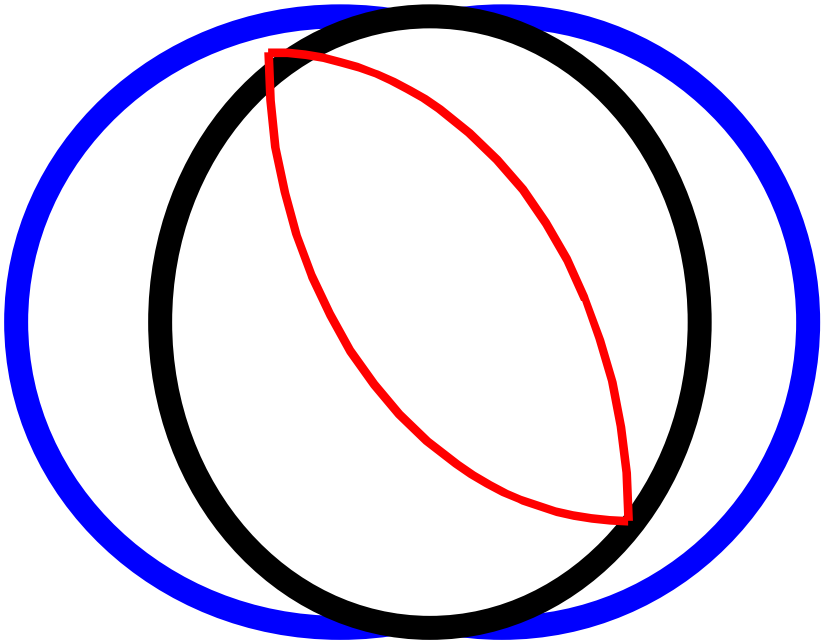
Picture below provides explanation for the observedbarcode



Note that the red and blue circles do not intersect each other

Why does this figure give a barcode with 5 intervals?

Fatten out the picture to put it in the plane

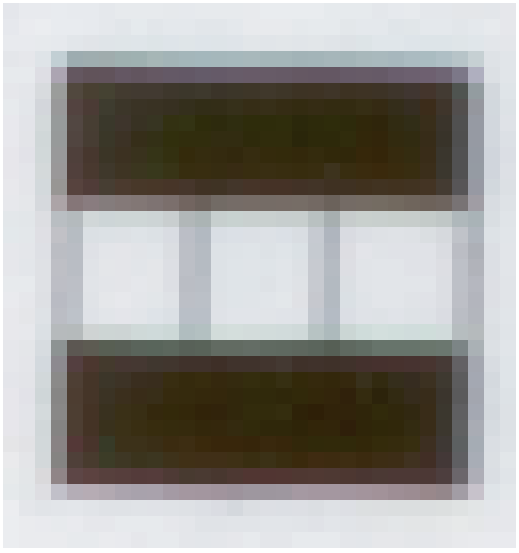


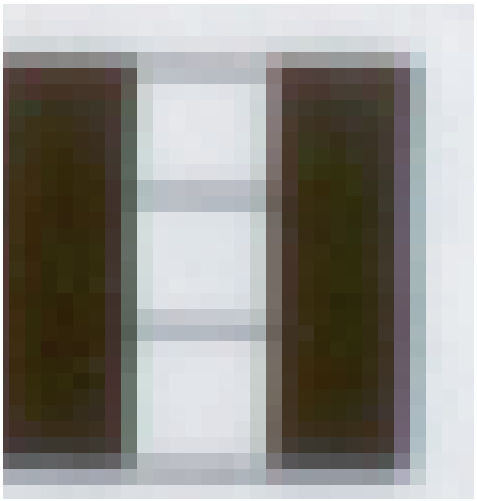
Note the five essential loops, or the five connected regions of the plane interior to the picture

Interpretation

- Main circle (black one) corresponds to the linear intensity functions described in the case $k = 300$

- Red and blue circles correspond to quadratic functions in one of the coördinate variables





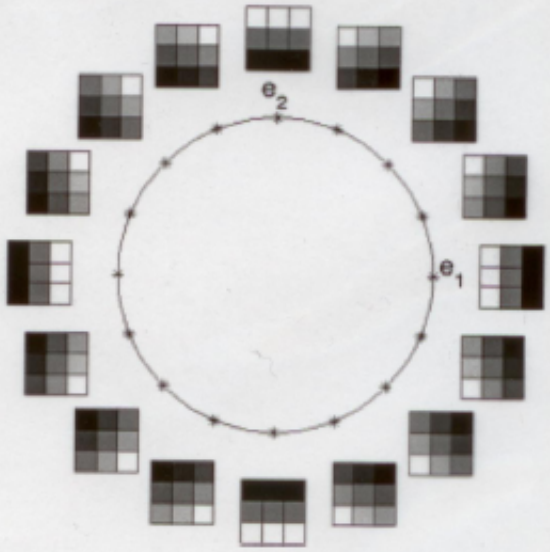


Figure 14: The e_1-e_2 circle in image-patch space.

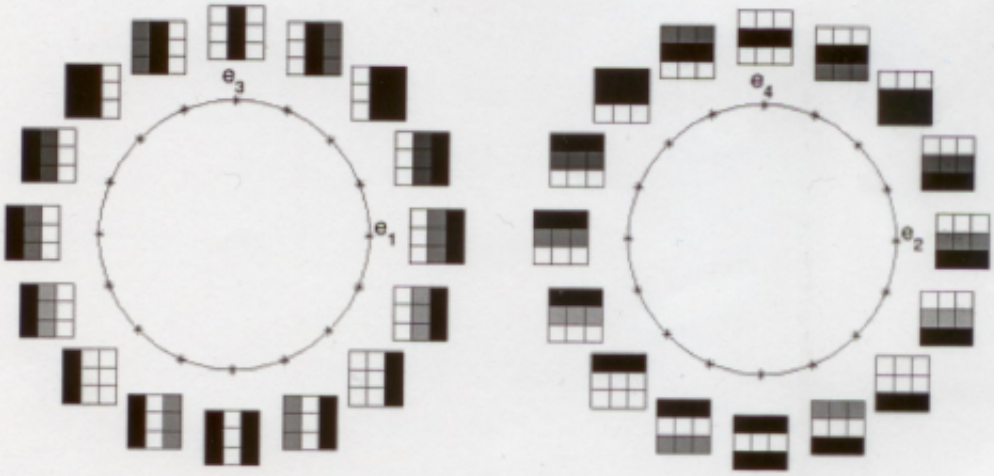


Figure 15: The e_1-e_3 and e_2-e_4 circles in image-patch space.

Summary

- Homology detects the preference toward linear intensity functions using the smoother density estimation (k= 300)

- Using more local density estimation ($k = 15$), we see clearly the preference for intensity functions depending on a single coördinate (vertical or horizontal lines separating dark from light)

- These are competing preferences

- Interesting to contemplate what happens as T grows. Set should grow into a 2D object - conjecturally a Klein bottle.

- Analysis should clarify the relationship between three competing preferences: linearity, vertical/horizontal, dependence on a single linear function

Patterns of Application of Topology

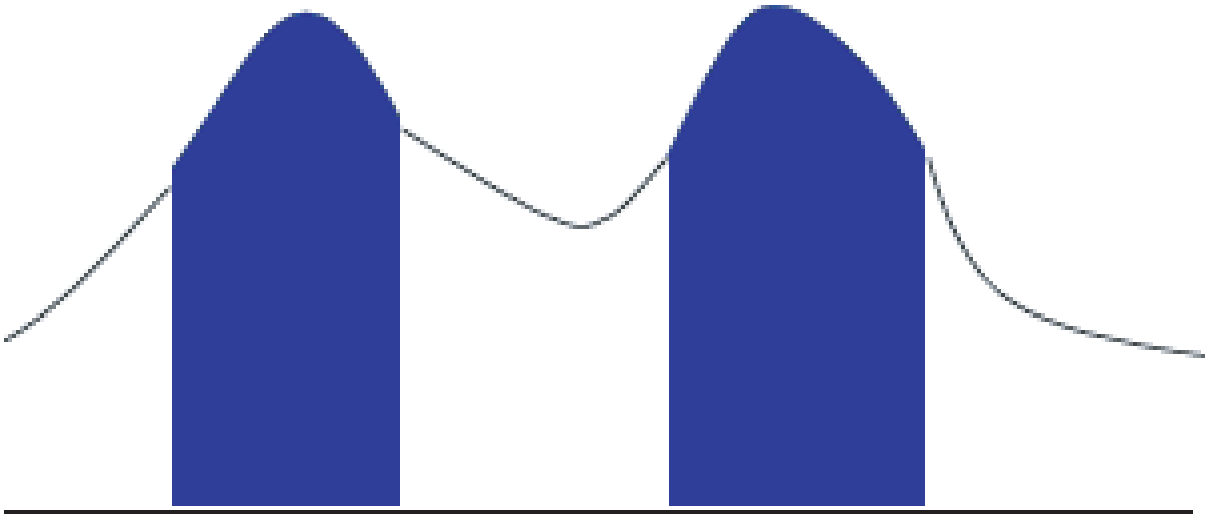
- Idealized mathematical model for point cloud data should be a space equipped with a function, to be used as a *Morse function*

- Each form of density estimation gives a discrete version of such a function

- Connectivity information about excursion sets of the density function helps describe the distribution of the data in a qualitative way. **One should do Morse theory on the density function**

- Elementary example of this is the location of modes in one-dimensional distributions. Connected components in the space of points of high density correspond to the number of modes

Bimodal Distribution



The regions on the line lying under the blue regions are the excursion set - it breaks up into two components

In higher dimensions, one needs more sophisticated ways to describe the distribution qualitatively

Data from Neuroscientists

- Neuroscientists are obtaining increasing amounts of very interesting data from various portions of the brain

- Two important mechanisms for obtaining data
  - Firing patterns of arrays of neurons. This data can now be obtained by implanting arrays of electrodes in the relevant portions of the brain
  - Optical imaging of various portions of the brain, which reflects the chemical activity in this region

- Both methods give rise to very high dimensional data, one because of the number of neurons used, the other since image data is intrinsically high dimensional

- Data is very noisy

Spaces from Firing Data

- Data can be viewed as an array of *point processes*, i.e. collections of firing times for each neuron

- Choose "bins" (i.e. short intervals) in the time direction, perhaps overlapping

- For each bin β and each neuron ν, let $\varphi(\beta, \nu)$ denote the number of firings of neuron ν during bin β

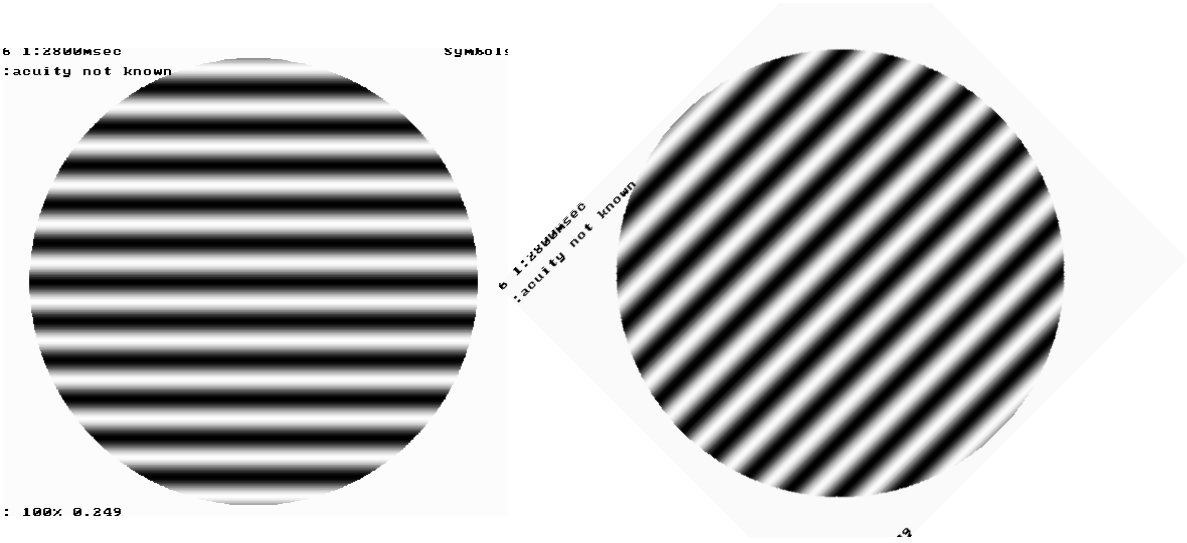
- Each bin β now gives rise to a vector $\vec{x}_\beta = \{\varphi(\beta, \nu)\}_\nu$

- Creates a data set in $\mathbb{R}^N$, where N is the number of neurons

- By using Rips complexes, can create simplicial complexes. Do they reflect what is happening in the brain in a qualitative way?

Ringach's Experiment

- Experiment performed by Dario Ringach's group at UCLA
- 20 electrodes implanted in the primary visual cortex of Macaque monkeys
- Monkeys are shown a family of images determined by two parameters, one a phase and one an angle



- "Phase space" of all such stimuli is two-dimensional, a torus

- Image are shown sequentially and very quickly, in a way which approaches every point in this torus

- **Question:** How can one read off the fact that it is a torus using only the firing pattern data?

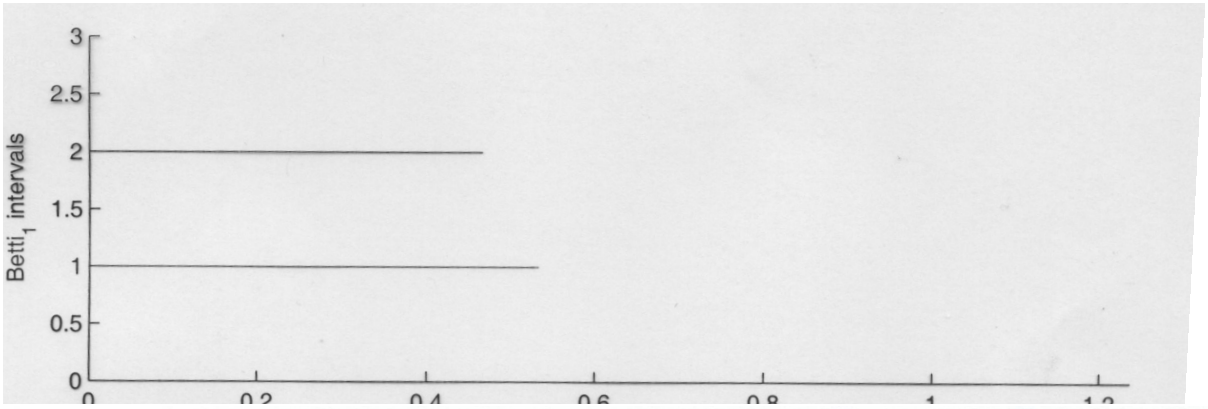
- We try by using the space we have constructed from the firing pattern data

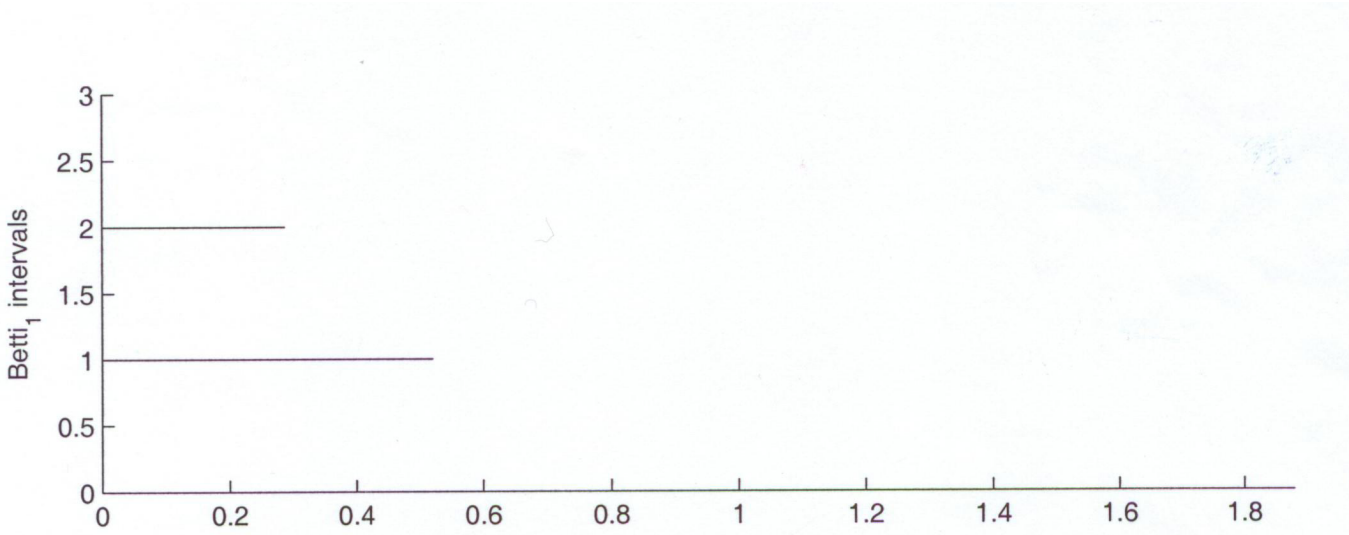
- First Betti number should be 2, i.e. the first Betti number of the torus

- Important first reduction: There are many vectors in this data set where no neuron has significant activity

- To remove these, only study vectors in which one neuron has activity greater than some threshhold. Analogous to removing low contrast patches from Mumford et al set

Results





The two intervals indicate the presence of two independent loops, as in the torus

Remarks

- We did many trials. Mostly we obtained the two intervals, but occasionally we obtained three. We suspect that it would be even more consistently two if we had data from more neurons.

- With less restrictive threshholding, we obtained consistently first Betti numbers of 5 or 6. This seems quite interesting, it may reflect that we will now be studying some states which occur in the transition from one image to another

- Neuroscience seems a very interesting area for potential application of topological technique. One often has phenomena or data which one expects are comparable qualitatively, but will not be so quantitatively

Other Kinds of Qualitative Information

We have shown how one can begin to make sense of the notion of connectivity information in the context of noise and incomplete information. Many kinds of qualitative information are not directly topological in nature.

Example: Letter Recognition

- The letters "A" and "B" can be distinguished by connectivity information - A has one loop, B has two

- "U" and "V" can be distinguished on the basis of a qualitative clue, the presence of a corner point. The cue is not topological, since U and V are identical topologically

- "C" and "I" are distinguished by the presence of a curved portion. They are topologically identical

Example: Geometric Primitives

- Geometric objects such as triangles, circles, tetrahedra, cubes, spheres, cones, etc. are all recognized by qualitative cues
- The cues include presence of corners, edges, cone points, etc., and the number of each
- Using qualitative cues means one can distinguish between two such objects even if given in "bad coördinates", or if they have been deformed or stretched

Example: Bottles, Glasses and Bowls

- Bottles, glasses, and bowls can be distinguished from each other by qualitative attributes: presence of a neck, presence of a crease at the bottom, etc.
- Better to distinguish using attributes than comparing with a large database of all objects of given kind

Making Non-Topological Attributes Topological

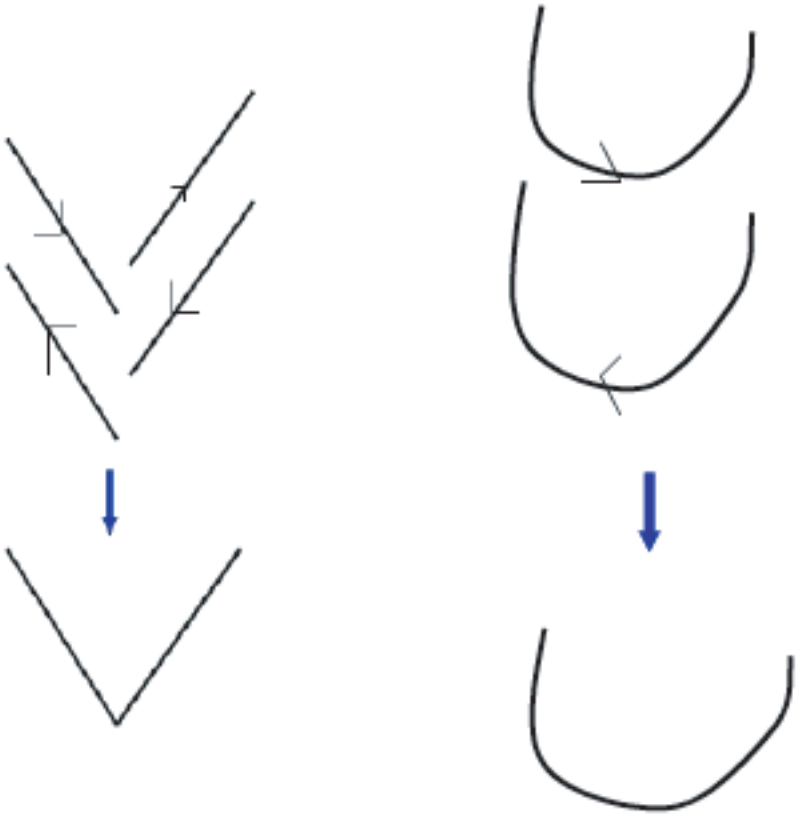
Definition: For any subset $X \subseteq \mathbb{R}^n$, we define the subset $T^0(X) \subseteq X \times S^{n-1}$ by

$$T^0(X) = \{(x, \zeta) | \lim_{t \to 0} \frac{d(x + t\zeta, X)}{t} = 0\}$$

T(X), the **tangent complex** of X, is the closure of $T^0(X)$ in $X \times S^{n-1}$

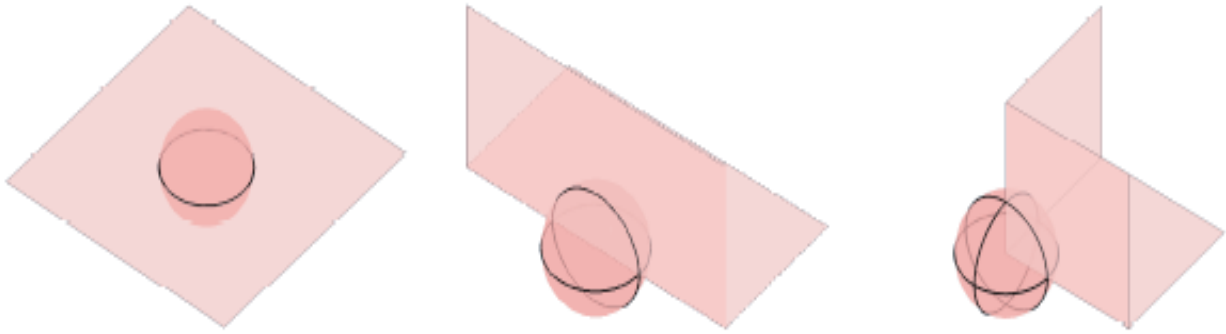
- There is a projection $\pi : T(X) \to X$, projection on the X-coördinate.
- $T(X)_x = \pi^{-1}(x)$ is called the *fiber* over x
- Topologicaly unchanged under *smooth* deformations of the ambient Euclidean space

Example



Connectivity information about TX distinguishes these two

2-Dimensional Fibers



(a) Smooth point: one circle

(b) Edge point: two circles

(c) Corner point: three circles

Here one dimensional homology is used to distinguish the tangent complexes

Homology of the tangent complex allows us to distinguish on the basis of presence of *hard features*, i.e edges, corners, etc., in other words, via the presence of singular points

Soft features

- Many qualitative features do not involve either connectivity information or the presence of singular points, but are "soft"

- The letter "C" and the letter "I" are distinguished by the soft feature that one is curved and the other is not

- Bottles and glasses are distinguished from each other by the presence of a neck in the bottle. This is a soft feature, they both have the same singular points

- **Strategy:** Impose filtration on the tangent complex based on curvature, and use this filtered space to create persistent homology. Barcode is now a signature for the qualitative aspects of the space

Filtered tangent complex

Let $X \subseteq \mathbb{R}^n$ be a *hypersurface*. For each point $(x, \zeta) \subseteq T(X)$, define $\rho(x, \zeta)$ to be the radius of the *osculating circle* to X within the plane spanned by ζ and the normal vector to X. ρ may be infinite.

Definition: For a *hypersurface* $X \subseteq \mathbb{R}^n$ and $\delta \geq 0$, we define a subspace $T_\delta(X)$ by

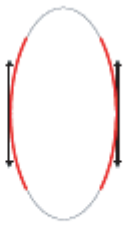
$$T_\delta(X) = \{(x, \zeta) | \frac{1}{\rho(x, \zeta)} \leq \delta\}$$

The barcodes for this filtered space often distinguish well between different kinds of shapes

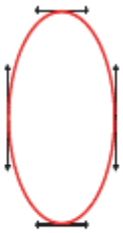
Examples

(a) $0 \leq \delta < \frac{a}{b^2}$

(b) $\delta = \frac{a}{b^2}$



(c) $\frac{a}{b^2} < \delta < \frac{b}{a^2}$

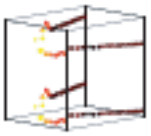


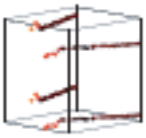
(d) $\delta \geq \frac{b}{a^2}$

(a) $\delta = 0$

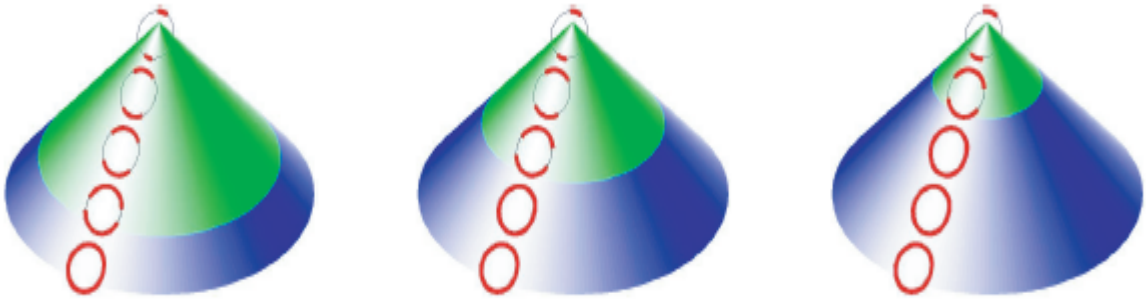
(b) $0 < \delta < \kappa_a$

(c) $\delta \geq \kappa_a$

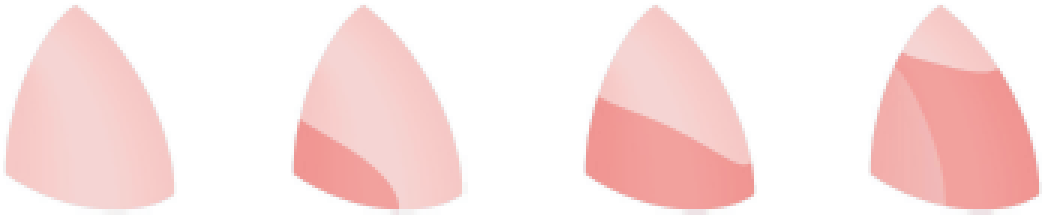




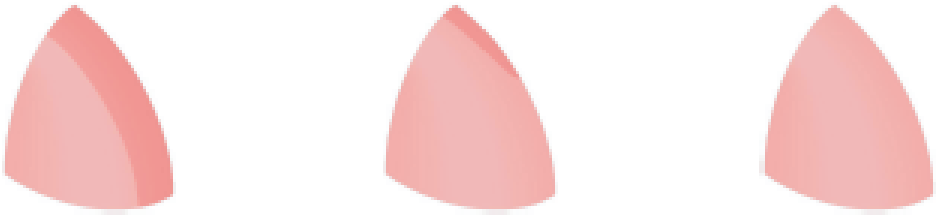
Further Examples



(a) $\delta = 0$ (b) $0 < \delta < \kappa_+$ (c) $\kappa_- \leq \delta < \kappa_0$ (d) $\kappa_0 < \delta < \kappa_H$ (e) $\delta \geq \kappa_H$



(a) $\delta < \lambda_1$ (b) $\lambda_1 < \delta < \mu_1$ (c) $\mu_1 < \delta < \lambda_2$ (d) $\lambda_2 < \delta < \nu_1$

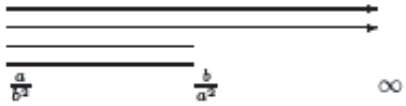


Barcodes for Filtered Tangent Complex

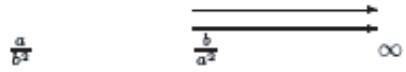
(a) β_0

(b) β_1

Figure 14: Barcodes for the circle.



(a) β_0



(b) β_1

gure 15: Barcodes for the ellipse.

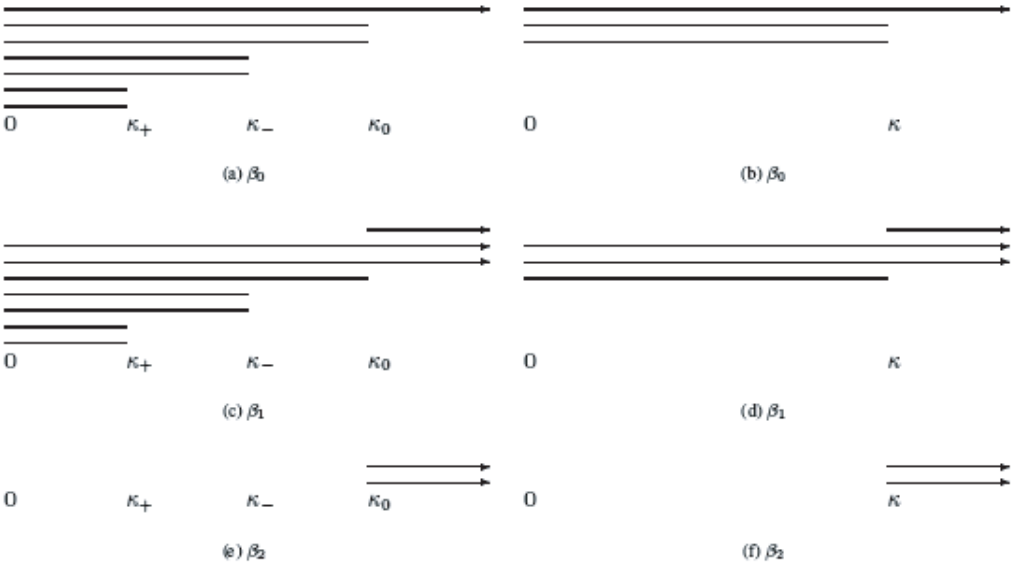


Figure 20: Barcodes for the bottle (left) and the glass (right.)

Metric on the Space of Barcodes

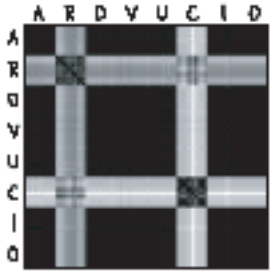
- There is a metric on the space of barcodes
- The distance between a pair of intervals is the measure of their "symmetric difference"
- A "partial matching" between a pair of barcodes is a one-one and onto correspondence between a collection of intervals in one code with a collection in the other
- For each partial matching, we compute a number D, which is the sum of the distances between each pair of matched intervals added to the sum of the lengths of all matched intervals
- Metric is now the minimum value of D over all possible partial matchings

-barcodes for the tangent com-
$'(P)$ filtered by curvature

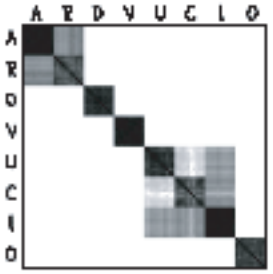
(b) A mask matrix, where distance is 0 if the letters have the same β_1, and ∞, otherwise

The combination of (a) and (b)

(d) β_0 of original space filtered top-down distinguishes 'U' from 'C'



of original space filtered right-
istinguishes pairs {'A','R'} and
'}



(f) The combination of all four distance functions distinguishes all letters