

Topology and the Analysis of High-Dimensional Data

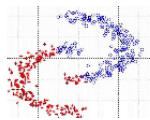
Workshop on Algorithms for Modern
Massive Data Sets

June 23, 2006
Stanford

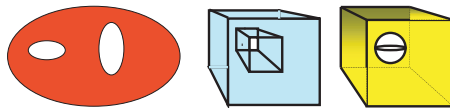
Gunnar Carlsson
Department of Mathematics
Stanford University
Stanford, California 94305
gunnar@math.stanford.edu

Qualitative Properties of Data Sets

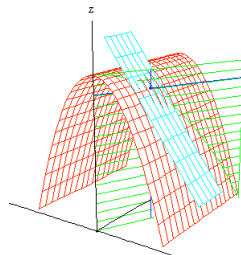
- Clustering



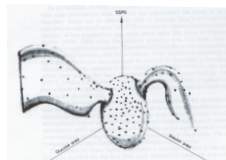
- “Loopiness” and “Holes” in the data



- Dimension



- Presence of Flares



Idea: Study Probability Density Functions Qualitatively

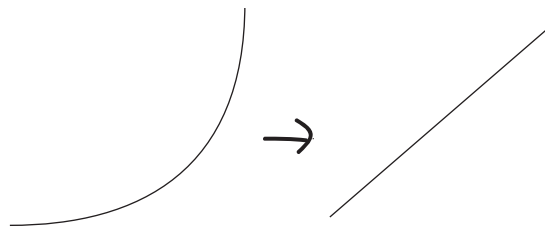
Think of data set as being sampled from
a probability distribution

Can be obtained from the data set using
a density estimator

How does one study density functions qual-
itatively?

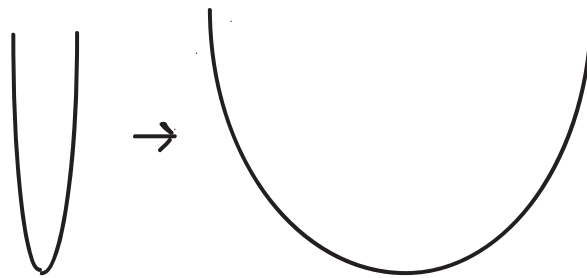
Qualitative Analysis of Functions in 1D

- In trying to find functional dependencies of one variable with another, often useful to change coördinates via transformations
- When a function grows rapidly, useful to consider its logarithm



- Amounts to applying coördinate change $y \rightarrow \log(y)$ to the y -coördinate

- Also useful to perform coordinate change in x -coördinate



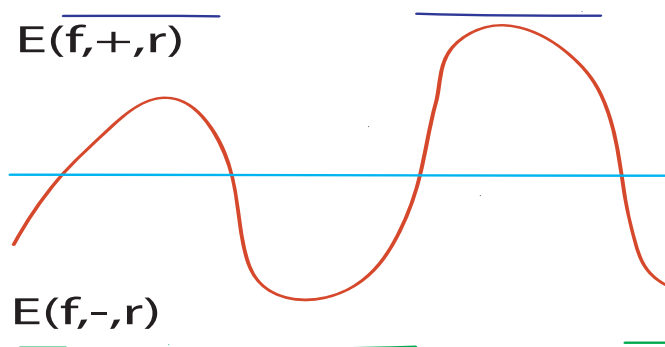
- Scales so that features and properties are more clearly visible
- Simultaneous coördinate changes also useful, e.g. log-log plots
- Need for coördinate changes suggests we don't have the "right" coördinates
- Right coördinates often clear in physics, perhaps less so in biology

No Preferred Coördinates

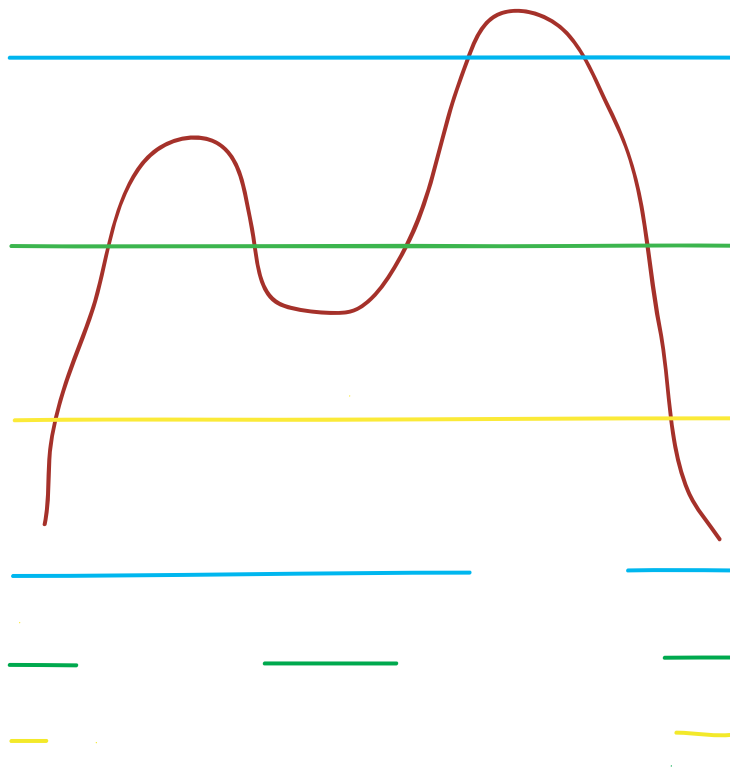
- Study properties or quantities which don't change under *arbitrary* changes of coördinates
- Such quantities have a discrete or combinatorial character
- Examples:
 1. Number of local maxima
 2. Number of local minima
 3. Total ordering (in the x -direction) on the collection of local maxima or minima
 4. Partial ordering in the y -direction of maxima or minima

Alternate Viewpoint

- Study *Excursion sets* for f
- $E(f, -, r) = \{x | f(x) \leq r\}$ and $E(f, +, r) = \{x | f(x) \geq r\}$
- *Connected components* of excursion sets reflect the presence of local maxima and minima, and therefore qualitative properties of f



Components of $E(f, -, r)$ vary with r



Components can be “born” or can merge
with increasing r

Combinatorial diagrams track components

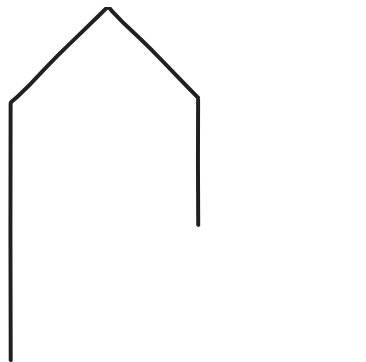
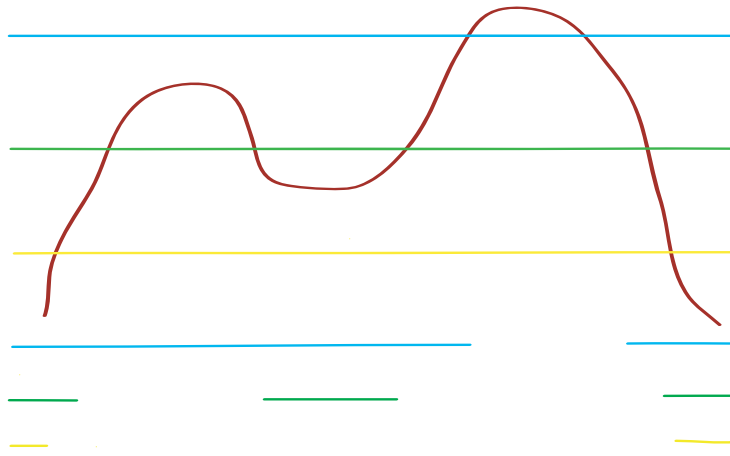
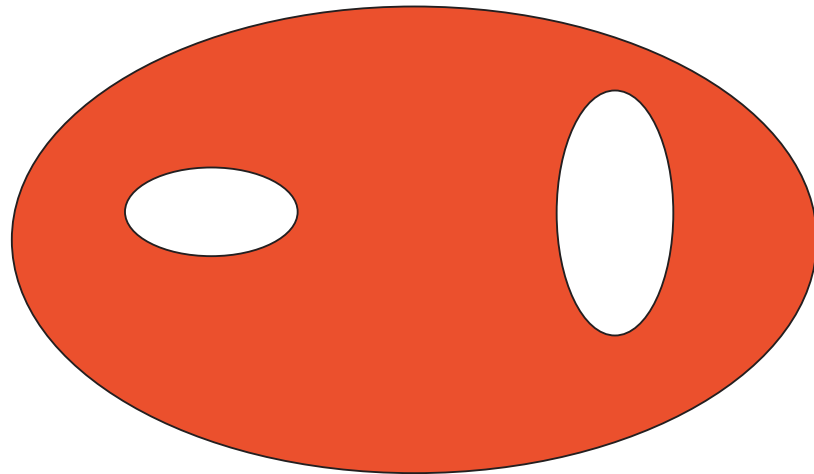


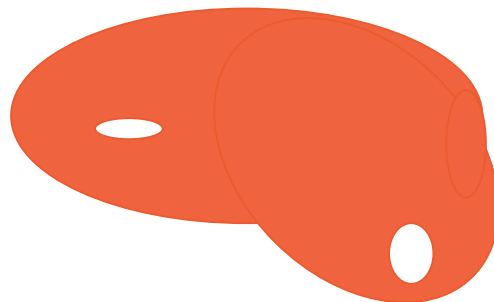
Diagram encodes all information required
to reconstruct the function up to contin-
uous coördinate change

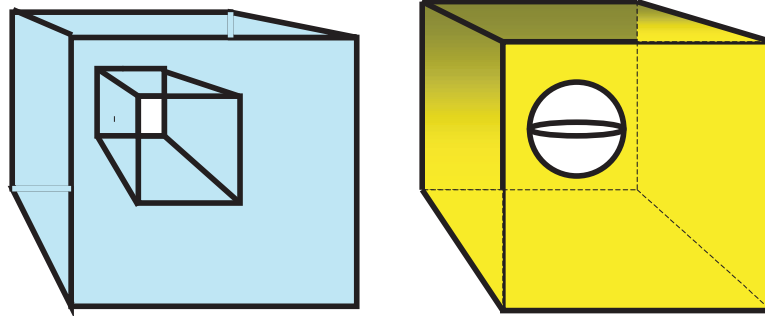
Topological Properties

- The number of connected components of a space is a *topological property*
- Don't change if we re-coordinate the space using any continuous coordinate change
- Don't change if I stretch and deform the space without tearing it
- Insensitive to size but encodes an abstract closeness relationship
- In 1D, components are main topological property
- In higher dimensions, wide variety of behaviors possible



- Region is connected, but has “holes”
- Don't vanish if I stretch or deform without tearing

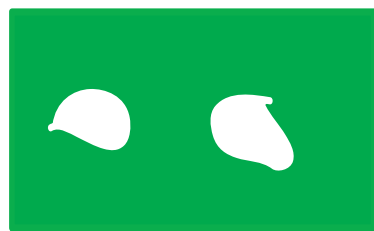
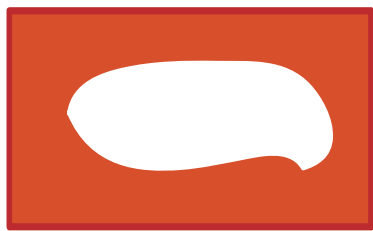
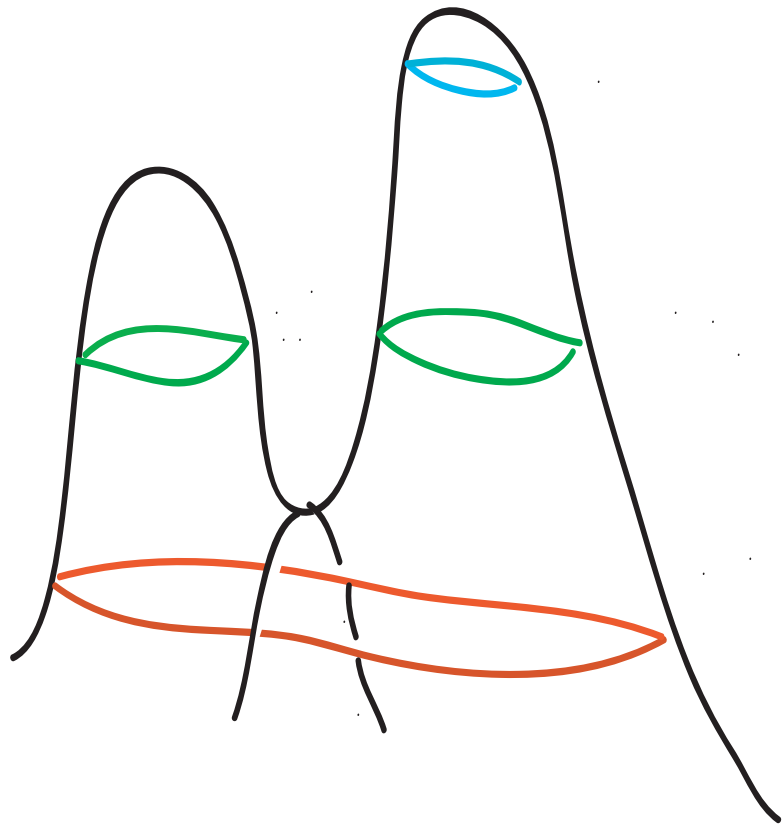




- In 3D, more variety is possible
- Tunnels and voids are distinct phenomena
- Spaces with a knotted circle removed are different from those with a standard circle removed

Qualitative Analysis in Higher Dimensions

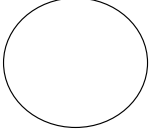
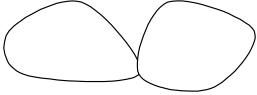
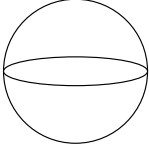
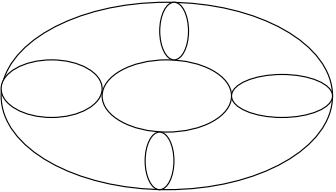
- Functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ studied qualitatively using topological properties of excursion sets
- In 2D, holes in excursion set reflect minima
- Saddles reflected in dividing of a hole in excursion set into two smaller holes
- Maxima occur when holes are filled in



How to make precise sense of presence of holes?

- Formalism called *algebraic topology* makes rigorous counts and classification of holes
- Uses linear algebra to obtain hole counts in various dimensions from ranks of certain matrices
- For each space X , creates *Betti numbers* $\beta_k(X)$, one for each $k \geq 0$
- $\beta_k(X)$ counts the number of k -dimensional holes
- When X given in closed form, homology calculations can be carried out effectively by hand

Examples

	$\begin{aligned} \beta_0 = \beta_1 &= 1 \\ \beta_i &= 0 \text{ for } i > 1 \end{aligned}$
	$\begin{aligned} \beta_0 &= 1 \quad \beta_1 = 2 \\ \beta_i &= 0 \text{ for } i > 1 \end{aligned}$
	$\begin{aligned} \beta_0 = \beta_2 &= 1 \\ \beta_i &= 0 \text{ otherwise} \end{aligned}$
S^n	$\begin{aligned} \beta_0 = \beta_n &= 1 \\ \beta_i &= 0 \text{ otherwise} \end{aligned}$
	$\begin{aligned} \beta_0 = \beta_2 &= 1, \quad \beta_1 = 2 \\ \beta_i &= 0 \text{ for } i > 2 \end{aligned}$

Remarks

- $\beta_0(X)$ counts the number of connected components in X
- Betti numbers of excursion sets can be used to predict presence of critical points of various types, i.e saddles, maxima, etc.
- Unlike the 1D case, homology does not tell the whole story. Many different spaces have same Betti numbers
- The rigorous definition of homology requires highly infinite methods. Finite calculation methods are possible when the space is presented in a combinatorial way, but they do not apply in general
- Adaptation of the ideas to situation where we only have points sampled from an object is necessary if the methods are to be applied to real world situations

Homology of Point Clouds

Point cloud data: finite but large set of points X sampled from Euclidean space, or even a more general metric space.

If we believe points of X are obtained by sampling (perhaps with noise) from a geometric object \mathbb{X} , how to build a space using only X which is believably a good representation of \mathbb{X} ?

Solution: Persistent homology (Edelsbrunner, Letscher, and Zomorodian). Builds an increasing family of simplicial complexes attached to X .

Uses all complexes in the family, together with the inclusions of one in the next

Vietoris-Rips Complexes

$VR(X, \epsilon)$ is a simplicial complex (union of intervals, triangles, and higher dimensional analogues) with vertex set equal to X , and where $\{x_0, x_1, \dots, x_n\}$ spans a n -simplex if and only if

$$d(x_i, x_j) \leq \epsilon \text{ for all } 0 \leq i, j, \leq n$$

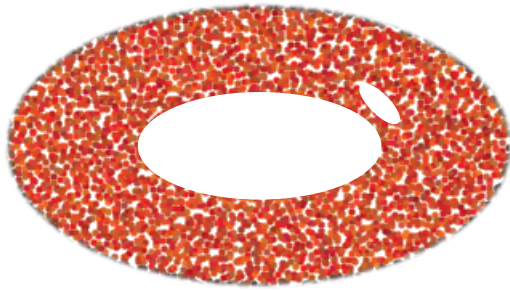
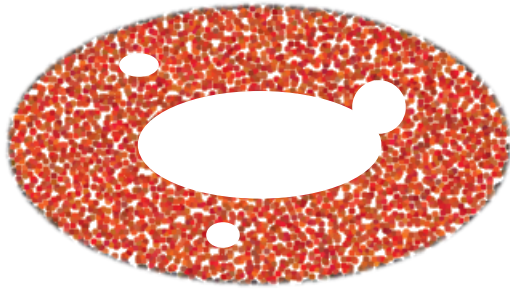


Note that the complexes grow as ϵ does

Note that the construction depends only on the choice of a metric. Could come from anywhere

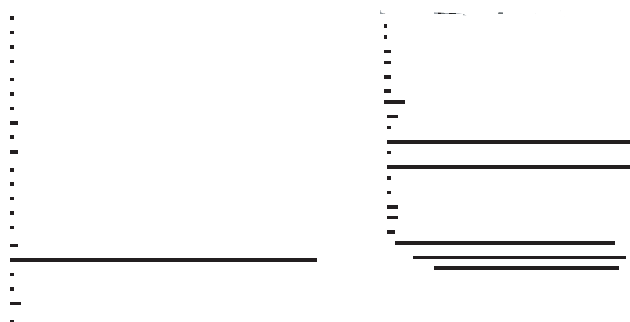
Persistence

- Just as in the analysis of birth times and merging times for components, it is possible to define analogues for higher dimensional homology
- One tracks the presence of holes in each complex $V(X, \epsilon)$ as well as the values of ϵ when they are “born” and when they “die”
- Sometimes, no single value of ϵ gives the correct answer
- Holes which have a very short lifetime are considered artifacts, coming from small irregularities of sampling or noise
- Holes with a longer lifetime represent actual geometric properties of the space X



Output

- These informal ideas have been formalized into software
- The analogues to Betti numbers in non-persistent homology are *bar codes*, i.e. finite unions of intervals



- 1D barcodes - left represents a circle, right represents an object with 5 holes. Note: vertical placement of intervals has no significance, for display only
- 0D barcodes would simply count number of components, i.e. would count clusters

- The only input required for the method is a set with a metric
- The metric can be arbitrary, not necessarily Euclidean. For example, “edit” style metrics which are used in the analysis of sequences in genetics would work, as would correlation metrics
- Even a metric isn’t necessary, any measure of dissimilarity could work

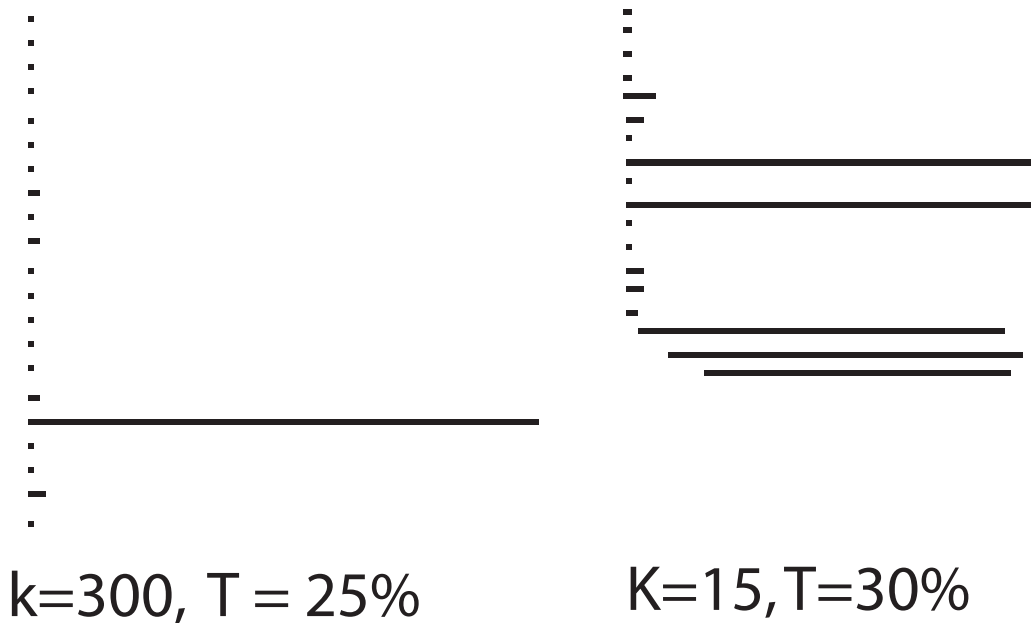
When are these techniques useful?

- Whenever one does not have an obvious choice of preferred coördinates
- When the data comes in terms of coördinates which have little relation to any underlying theory, and the transformation laws are non manageable
- When the data is coördinatized in a non-smooth way, so that standard linear approximation methods do not apply
- When the data is intrinsically non-Euclidean, such as in spaces of genetic sequences with an edit distance metric

Example 1: High Contrast Patches in Natural Images

- **Question:** What can be said statistically about 3×3 patches occurring in images taken with a digital camera?
- Only “high contrast” patches are interesting, rest are essentially constant
- Lee, Pedersen, and Mumford constructed a data set of 8.5×10^6 such high contrast patches, working from a database of c:a 4500 images taken by van Hateren and van der Schaaf
- Data set sits in \mathbb{R}^9 . After normalizations to set the mean intensity and total contrast to fixed values, on $S^7 \subseteq \mathbb{R}^8$
- What can be said about this set, or about the regions of highest density?

Results (de Silva, Ishkanov, Zomorodian, C.)



- k is a parameter determining the density estimator. Large k means a smooth, unlocalized estimator, small k means a more localized estimator
- T is a percentage threshold, i.e. we consider the $T\%$ densest points as measured by the estimator.

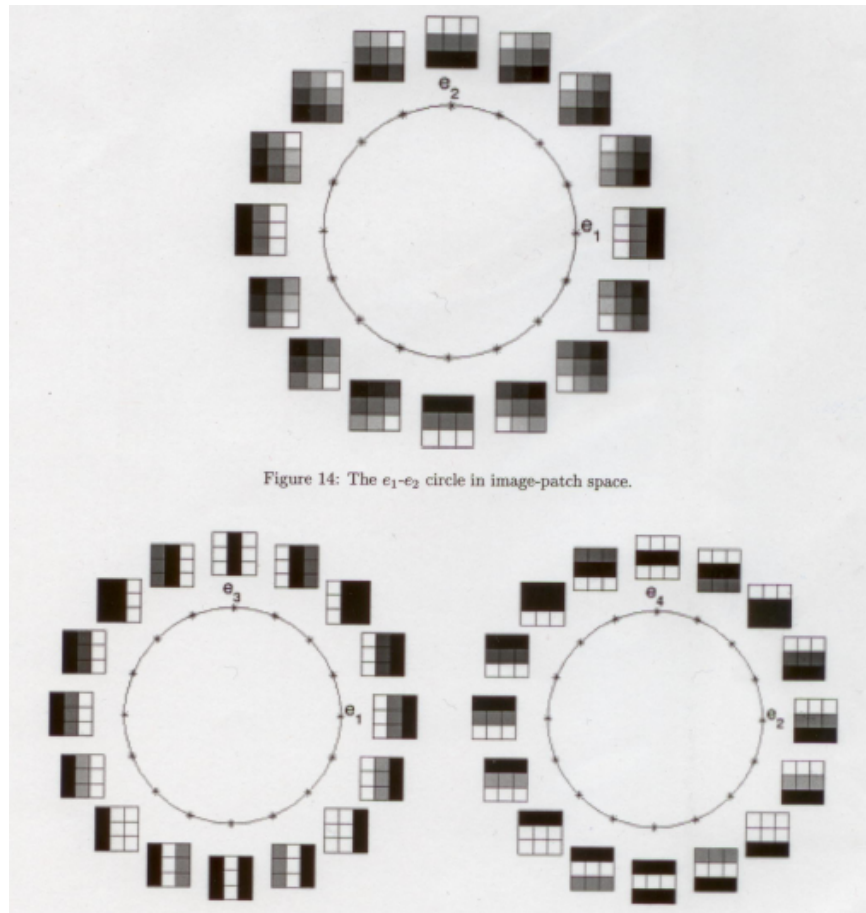
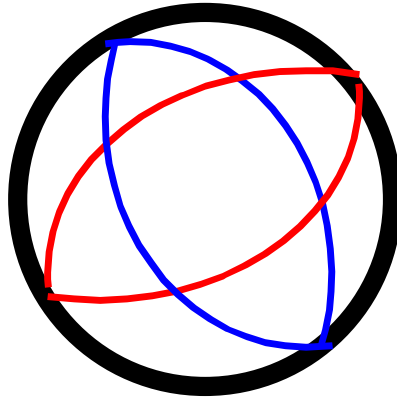
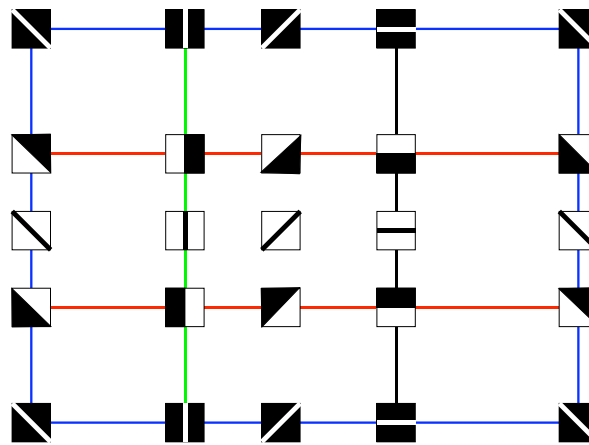


Figure 14: The e_1 - e_2 circle in image-patch space.

- Points are concentrated around a 2D object whose topology is that of a *Klein bottle*



- Identifications are made on the boundary
- Very highest density occurs on a circle, labelled in red in the picture

- Suggests a theoretical model in which patches have intensity functions given by quadratic polynomials in two variables of the form $q(\lambda(x, y))$, where q is quadratic in one variable and λ is linear in x and y
- The topology of this space of quadratics is correct
- Even if one guesses such a model, it doesn't fit the data well
- Due to the fact that we don't know what the right coördinatization of the the gray scale values should be
- Also possible that polynomial functions are too smooth, and that more piecewise linear functions should be used

Example 2: Images

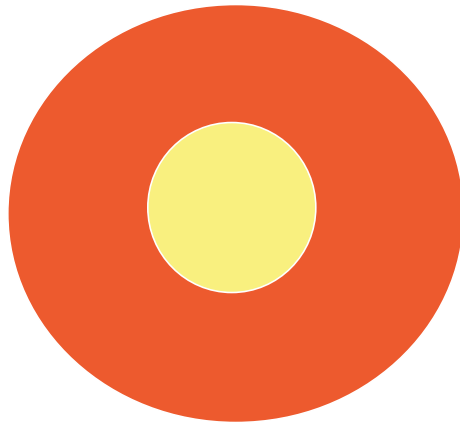
- Each of the faces of a regular tetrahedron is given a gray scale shading, different for each face
- Tetrahedron undergoes rotations through the entire family of 3D rotations, and images of it are taken by a black/white digital camera
- Each image taken can be viewed as a point in “pixel space”, i.e. as a vector with a gray scale value as a coordinate at each pixel
- **Question:** Can one recover the stimulus space (the space of 3D rotations) from this high dimensional data set?

Comments

- The images can be viewed as an exotic coördinatization of the space of rotations
- No individual coördinate has any particular significance for the structure of the space. If a single pixel is off, we can still detect the nature of the picture. Only the coördinates in the aggregate matter
- No reasonable theoretical model for the coördinate transformation relating more standard coördinates to the pixel coördinates
- The coördinatization is highly non-smooth. Standard linear methods for dimension estimation do not apply directly

Results (Rannaud, Shen, Wei, de Silva, C.)

- This experiment was carried out “synthetically”, creating a data set of 5K points
- The dimension was estimated topologically by computing the Betti numbers of a small neighborhood, with a smaller neighborhood removed



- Dimension of three was observed, agreeing with actual dimension of the space of 3D rotations

Example 3: Neuroscience

- Study of the responses to primary visual cortex to stimuli
- Each neuron creates a *spike train*, i.e. collection of firing times
- Observe spike trains from arrays of neurons
- Different ways to create metrics on spike train arrays
- Binning creates Euclidean vectors of spike counts - get Euclidean metric
- Distance between pairs of spike trains can also be computed by minimizing penalty associated to family of moves, involving moving a firing or deleting/adding one (essentially non-Euclidean). “Time code” metric

Responses to families of stimuli

- Given a family of stimuli parametrized by a geometric object, can one recover the geometry of the family using the resulting spike trains?
- In particular, neurons have both *phase* and *orientation* sensitivity
- If one presents stimuli with varying orientation or varying phase, or with both varying simultaneously, can one see that detected in the spike trains
- Orientation typically provides a “stronger signal”
- Our group has had substantial success in performing this when the orientation and phase are presented simultaneously. Geometry of stimulus space is a “Klein bottle”

Comparison with other methods

- Aronov, Reich, Mechler, and Victor confront the “weakness” of the phase signal directly
- Macaque monkey is presented with a moving grating with fixed orientation
- ARMV obtain a data set, which is given a time code metric
- Question: Is the resulting space a circle, as is the space of stimuli, due to the periodicity?

The ARMV procedure

- Find an optimal embedding of the data set in Euclidean space using MDS. Optimal means that the metric is distorted as little as possible relative to the Euclidean metric
- Find the best fitting ellipse to the embedded Euclidean data
- If the length of the minor axis is sufficiently large relative to the length of the major axis, they conclude that they are seeing a circle
- Their results confirm the hypothesis that the space is a circle

What we would do with ARMV data

- Construct our V-R complex on the data with the time code metric directly
- Compute the barcodes and confirm that the first Betti number is one
- Comments:
 1. Doesn't involve embedding in Euclidean space, which introduces distortion
 2. Doesn't involve fitting to a particular geometric model, which can be deceiving
 3. Gets directly at the core question, which is whether the space has a loop in it, given by periodicity