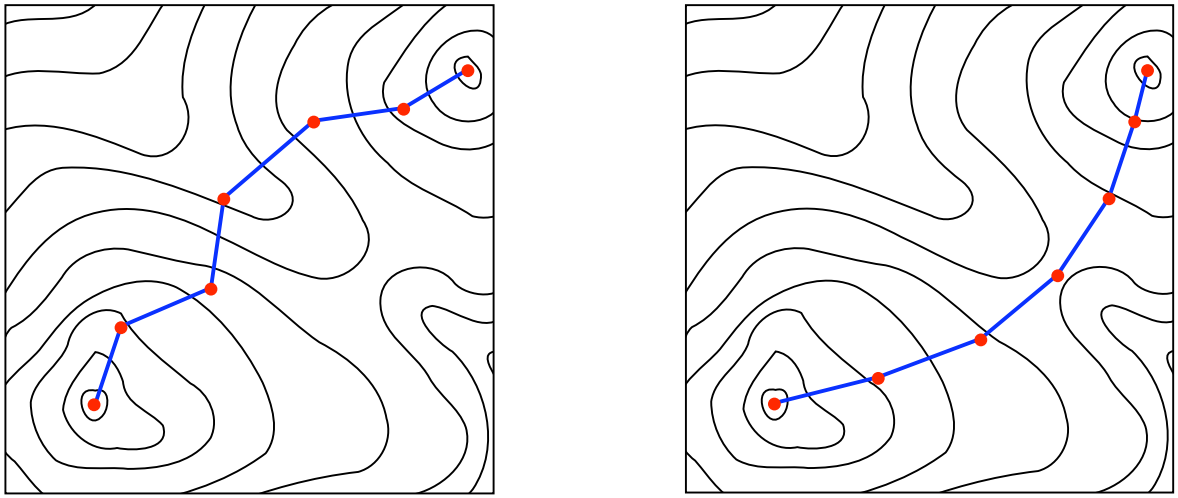
Nudged Elastic Band in Topological Data Analysis



Henry Adams, Atanas Atanasov, and Gunnar Carlsson

Comptop Seminar, Stanford University

January 29, 2010

Abstract

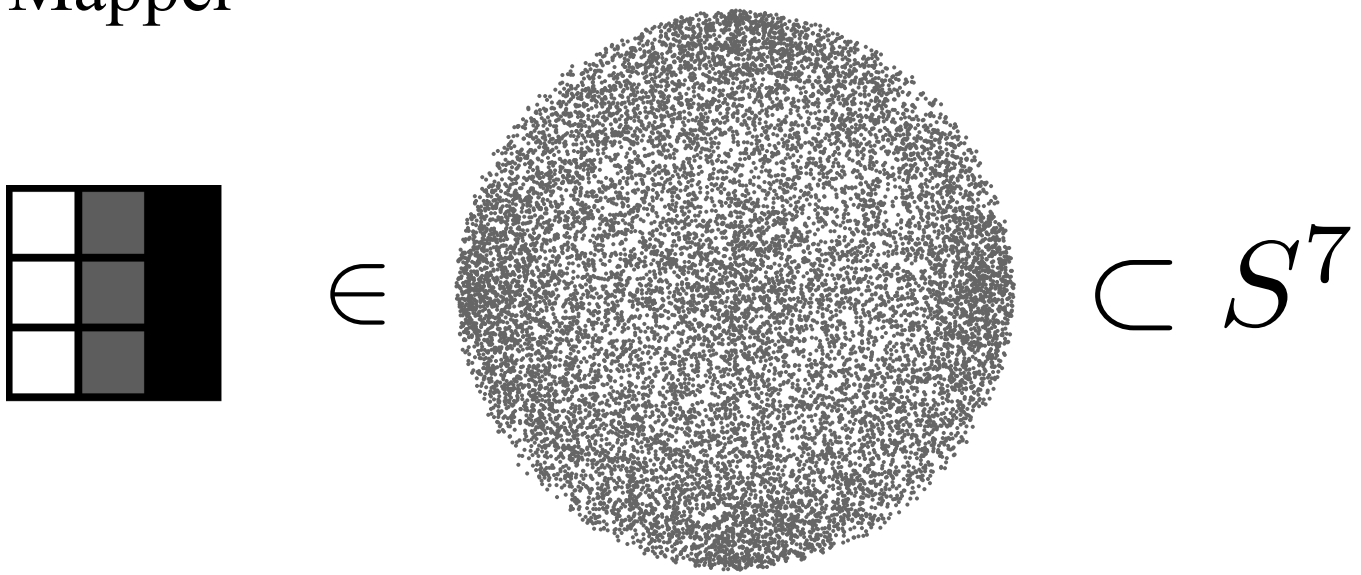
We introduce a method for analyzing high-dimensional data. Our approach is inspired by Morse theory and uses the nudged elastic band method from computational chemistry. As output, we produce an increasing sequence of cell complexes modeling the dense regions of the data. We test the method on data sets arising in social networks, in image processing, and in microarray analysis, and we obtain small cell complexes revealing informative topological structure.

Outline

1. Persistent homology $\Rightarrow$ motivating questions
2. Nudged elastic band (chemistry)
3. Nudged elastic band (data analysis)
4. Testing on datasets
   - 3x3 optical image patches
   - 5x5 range image patches
   - 3x3 optical flow patches
5. Kinks, higher dimensions, conclusions

Problem

- How do you find structure hidden in a high-dimensional point cloud dataset?
  - Persistent homology
  - Mapper

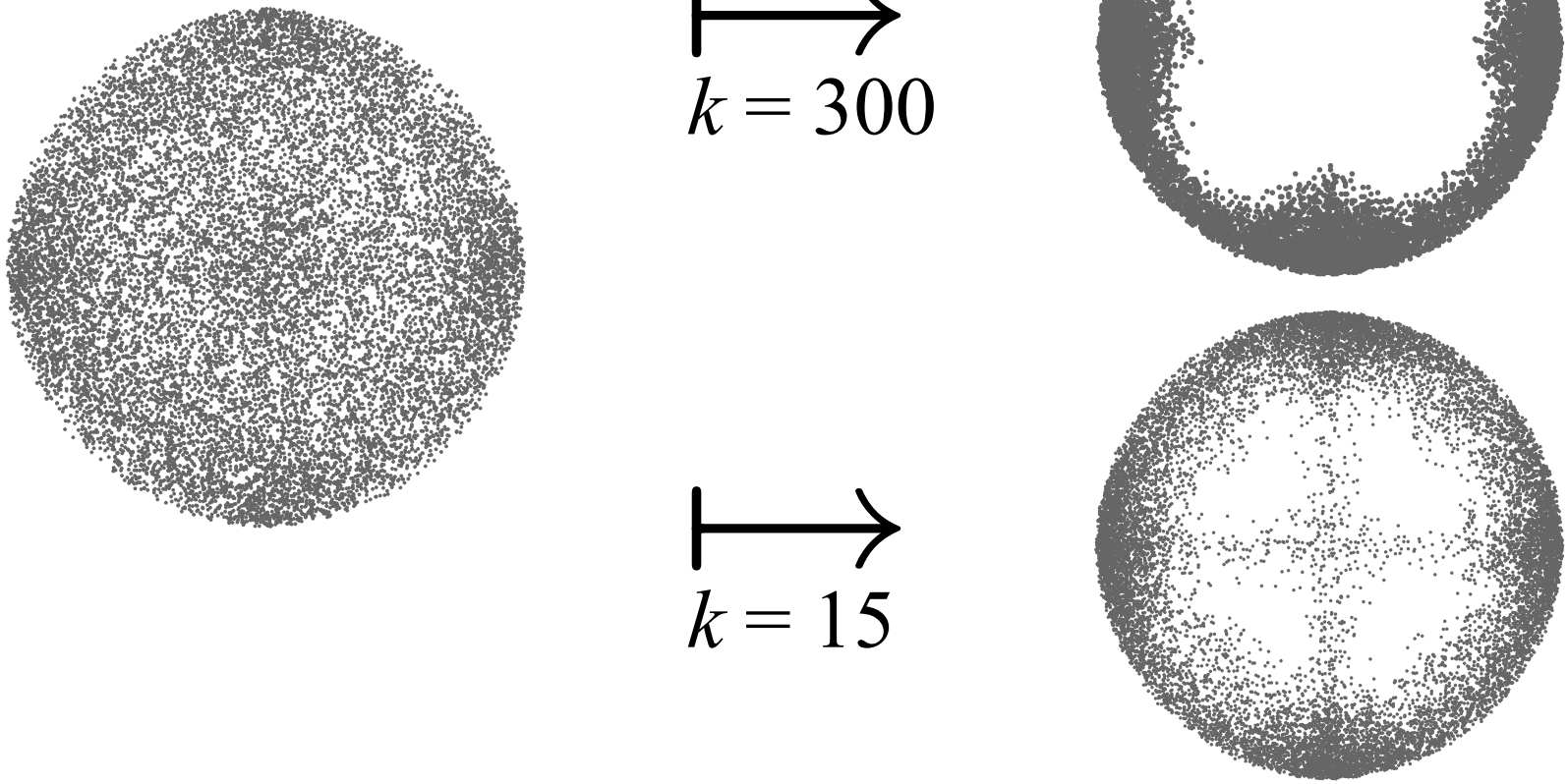


Example: 3x3 optical image patches, from *On the local behavior of spaces of natural images* by G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian

Persistent homology

1. Take dense core subset.

$$\rho_k(x) = \frac{1}{d(x,\ k\text{-th nearest neighbor})}$$

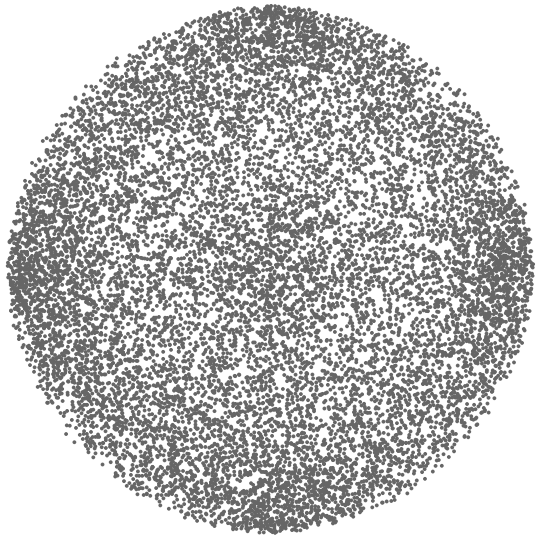


Persistent homology

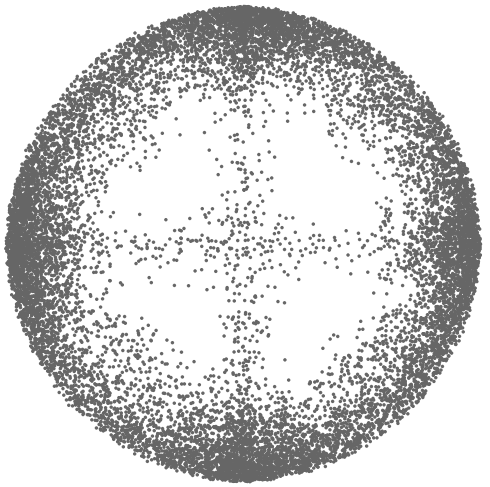
1. Take dense core subset.

$$\rho_k(x) = \frac{1}{d(x,\ k\text{-th nearest neighbor})}$$

Remark: you may not have these projections onto nice basis elements.

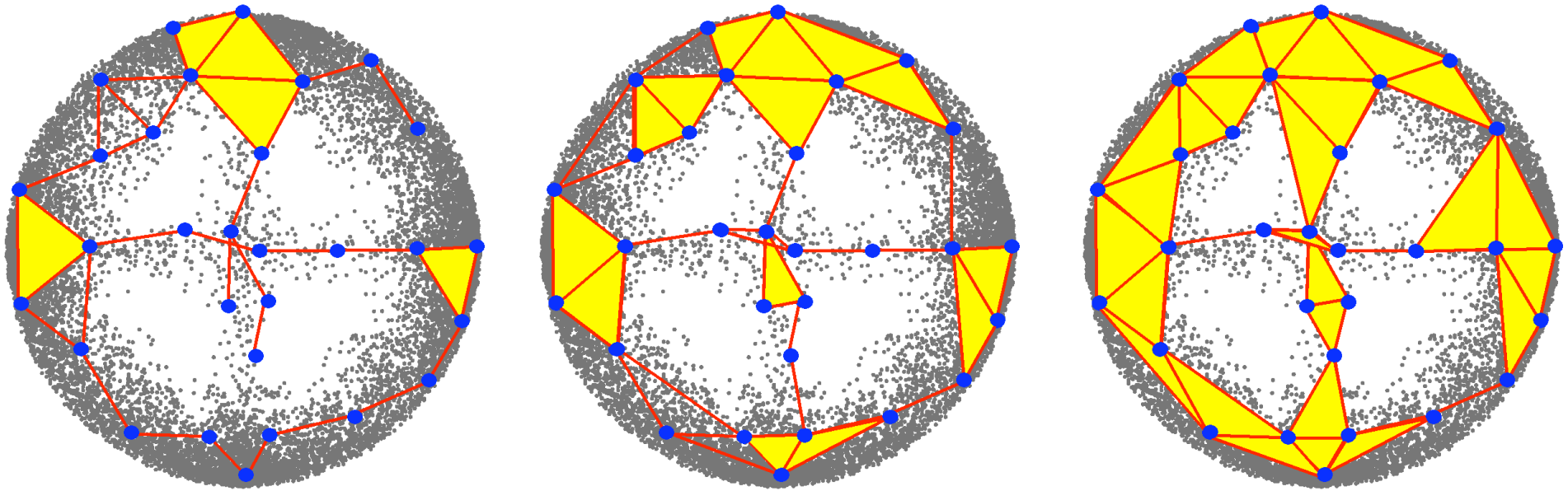


$k = 15$



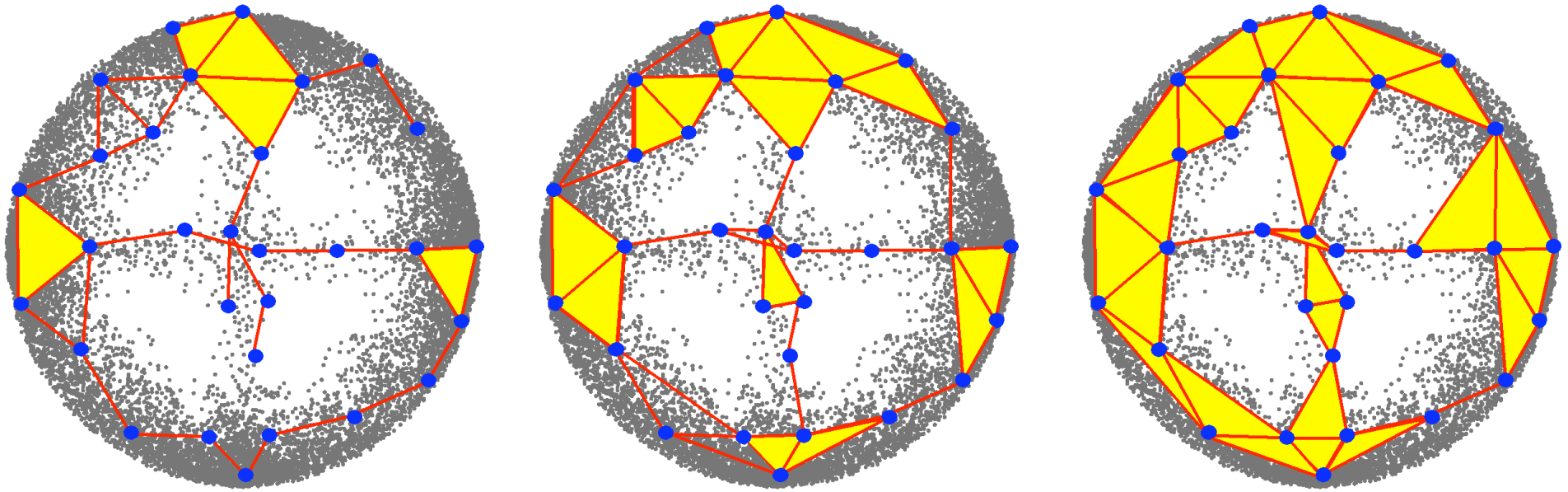
Persistent homology

1. Take dense core subset.
2. Build increasing sequence of simplicial complexes.



Persistent homology

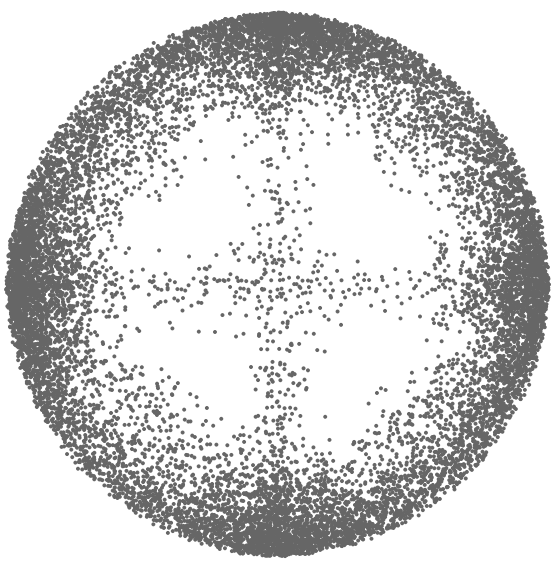
1. Take dense core subset.
2. Build increasing sequence of simplicial complexes.
3. Compute Betti barcodes.

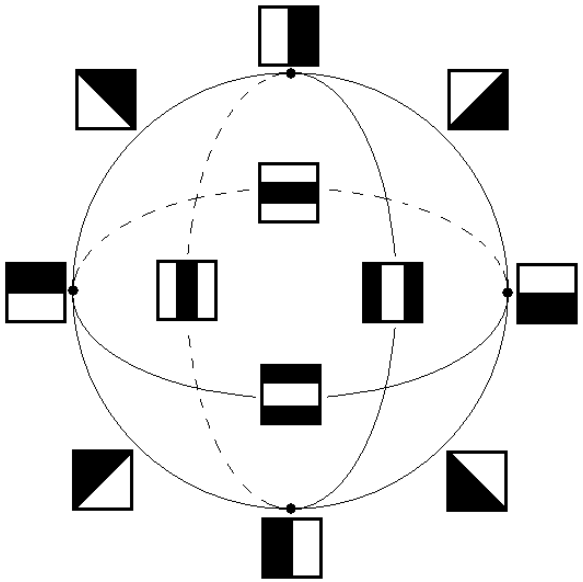


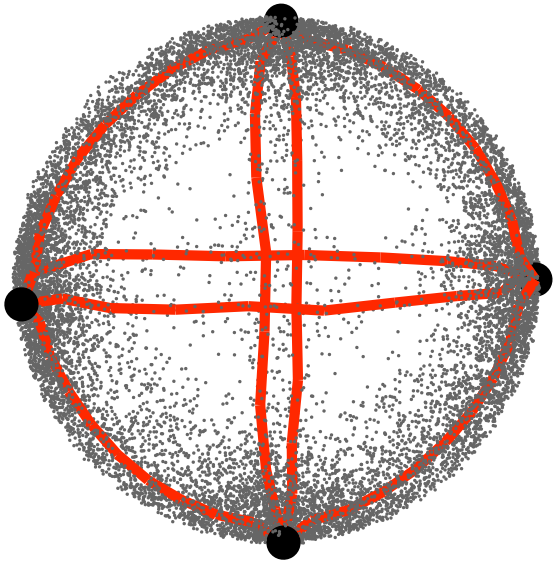
Persistent homology

4. Identify model.

   - Usually with your bare hands, not automated

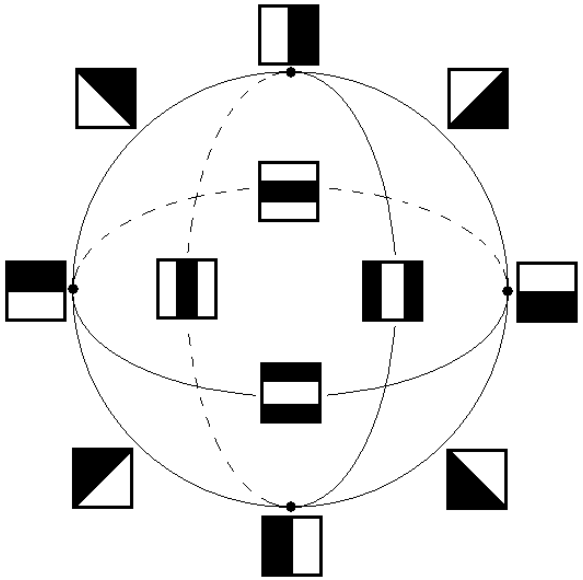






Motivating questions

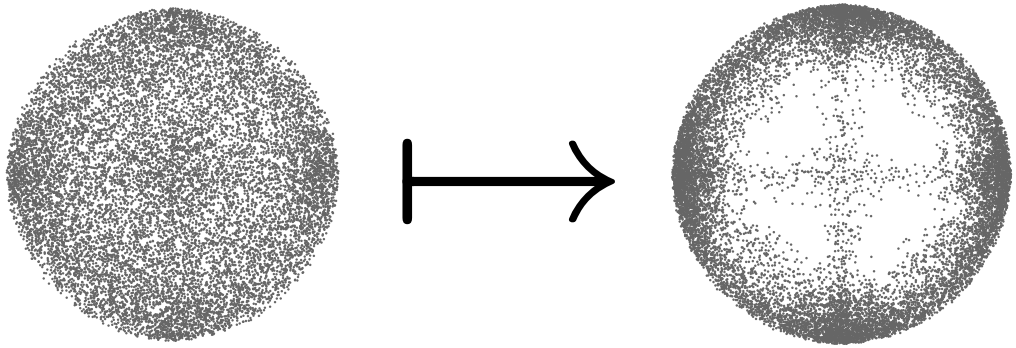
(A) How do you find a simple model matching the Betti barcodes?



- Your simplicial complexes are not simple.
- Nice homology generators?
- Localized homology generators?
- Cell complexes are often nice models.

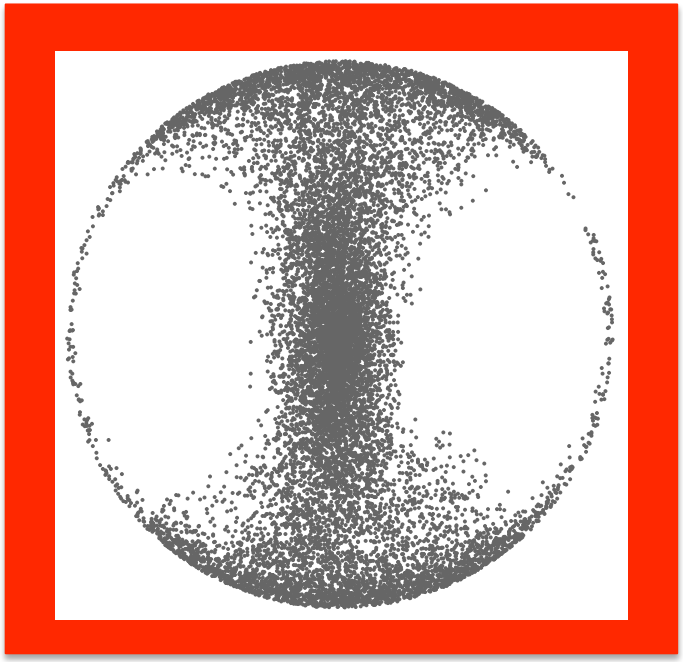
Motivating questions

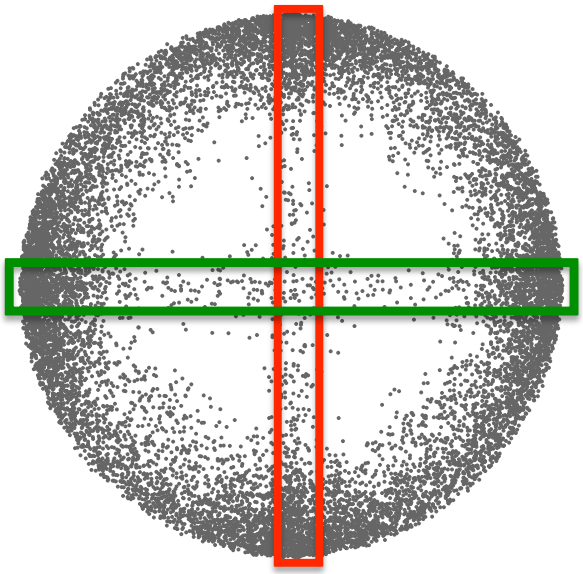
(B) Can we be more robust to noise?

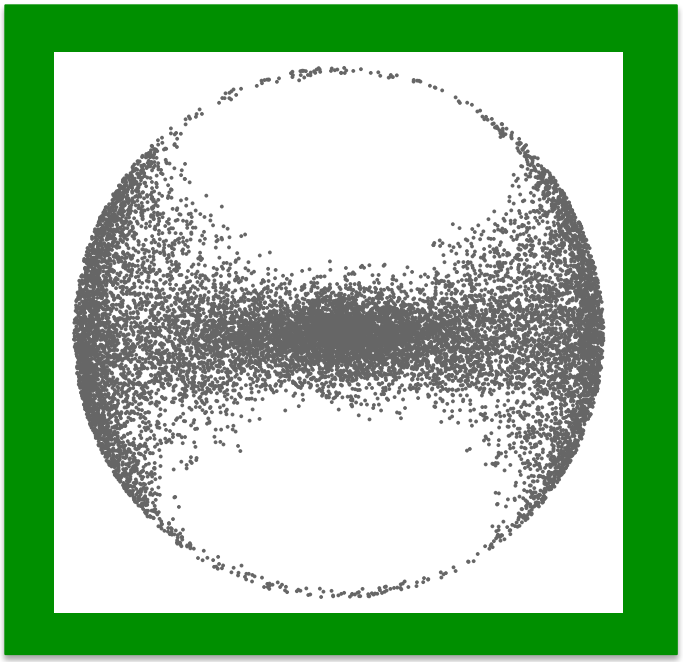


- For persistence pipeline, removing noise, say by taking dense core subsets, is usually necessary.
- Rips & Witness simplicial complexes greatly affected by non-Hausdorff noise.
- However, removing noise but not features is hard.
- “Noise” is a misnomer: all data points contain information.
- Instead of cutting out the noise, can we see through the noise? (Analogy from our reading group)

Aside: Removing noise but not features is hard.

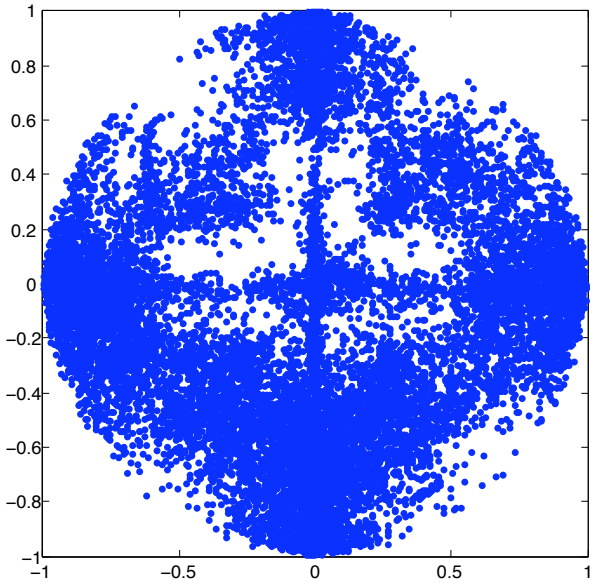




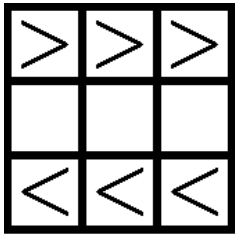


- For optical image patches, dense core subsets separate primary and secondary features from noise…

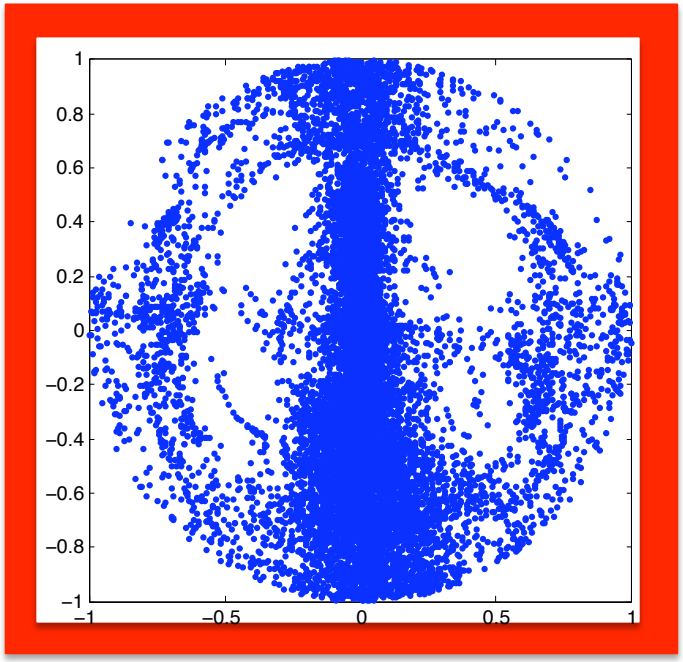
Aside: Removing noise but not features is hard.

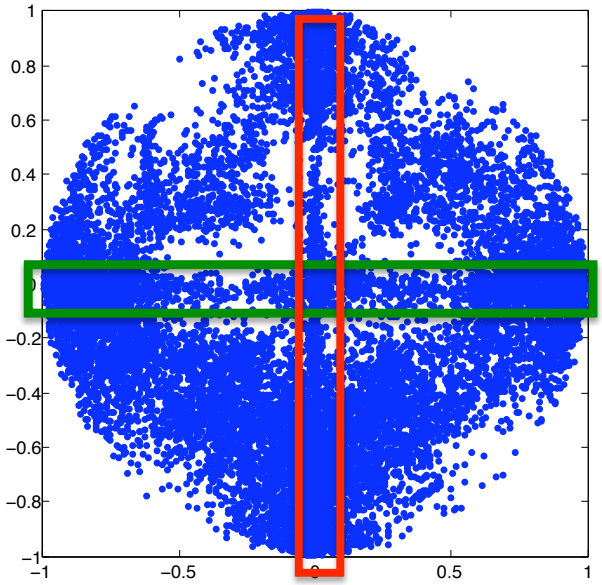


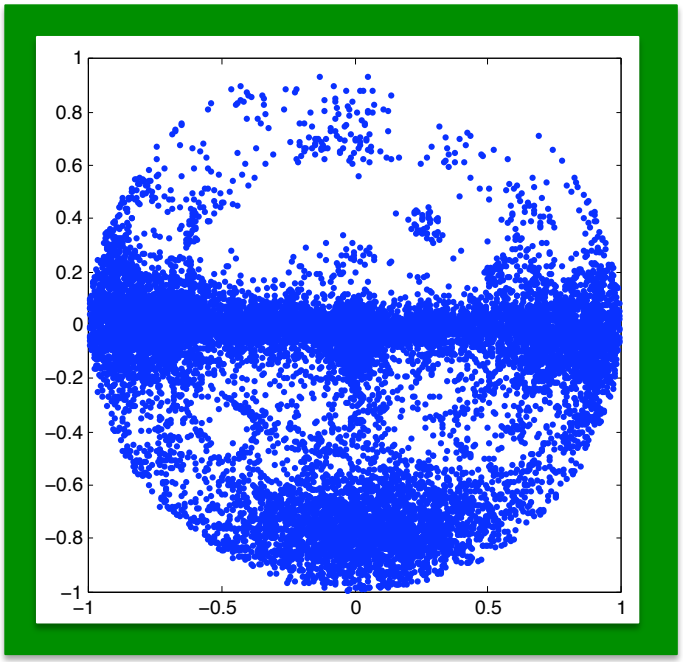
- For optical image patches, dense core subsets separate primary and secondary features from noise…
- Not so lucky with optical flow patches.



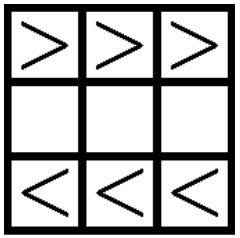
Aside: Removing noise but not features is hard.



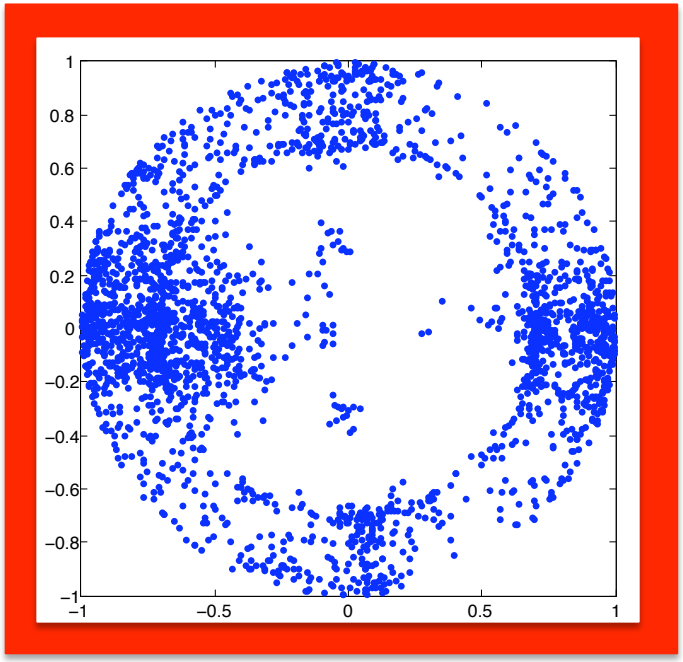


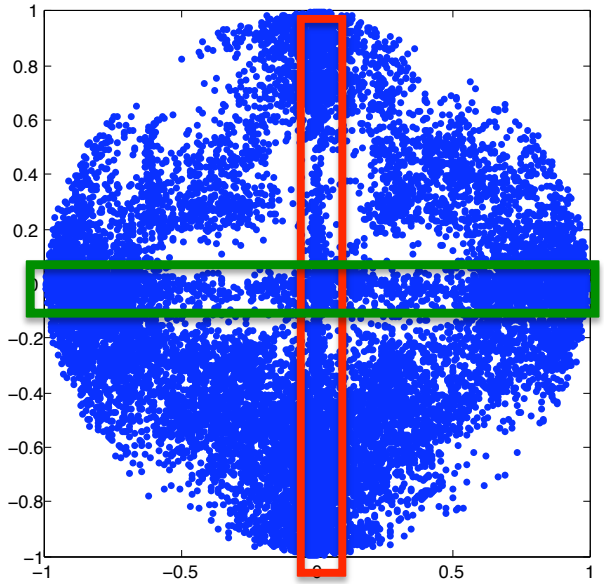


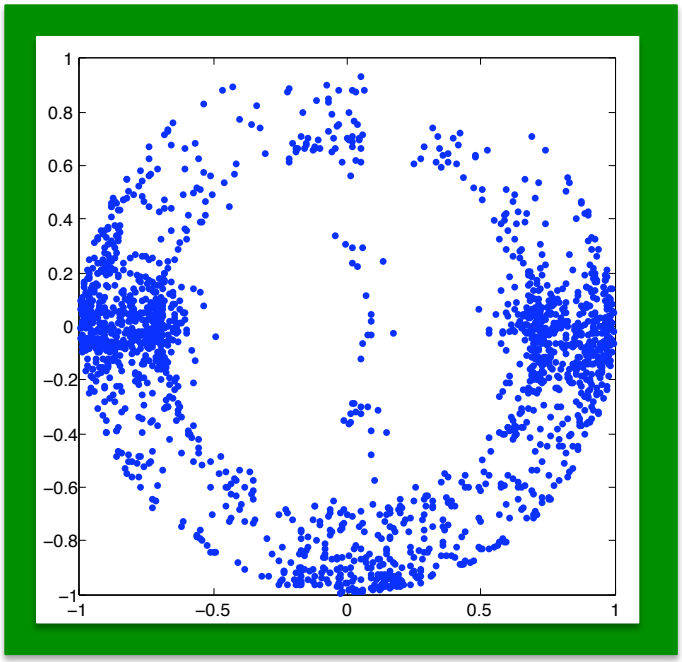
- For optical image patches, dense core subsets separate primary and secondary features from noise…
- Not so lucky with optical flow patches.



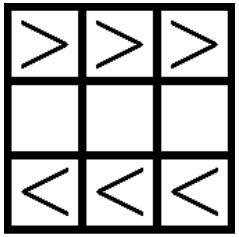
Aside: Removing noise but not features is hard.





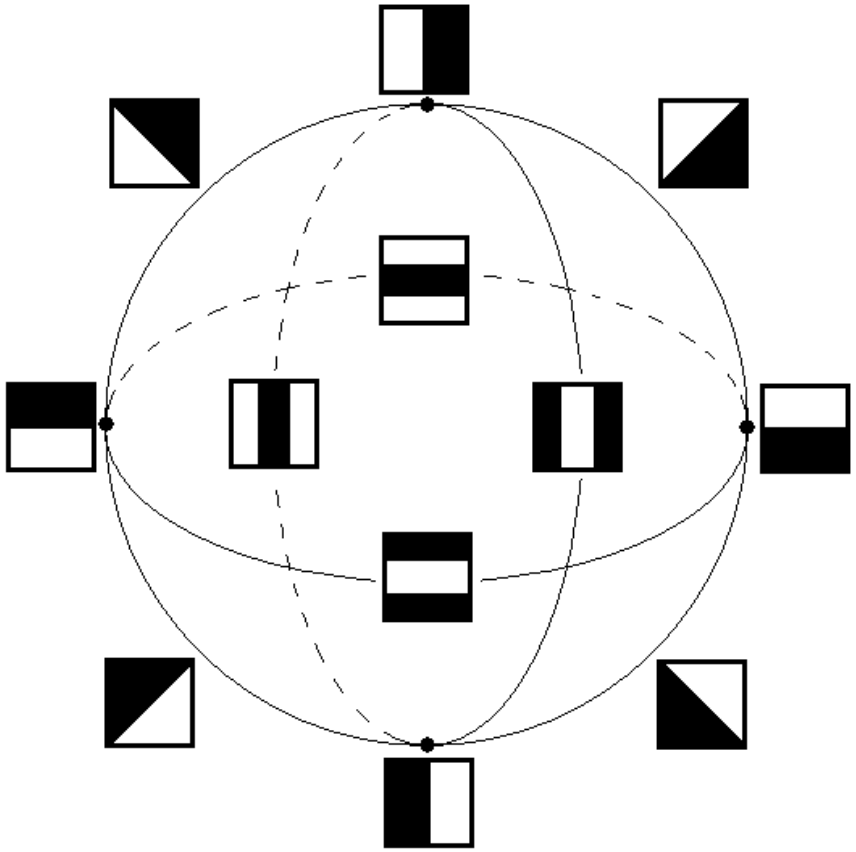


- For optical image patches, dense core subsets separate primary and secondary features from noise…
- Not so lucky with optical flow patches.



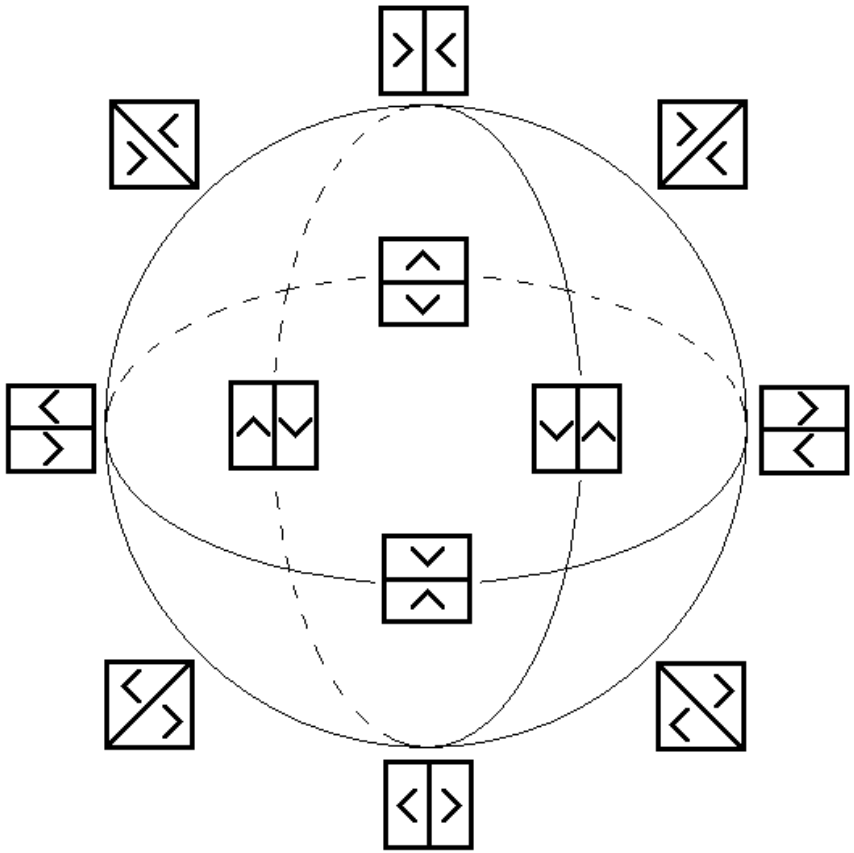
Aside: Removing noise but not features is hard.

Optical image 3-circle model



Fills to Klein bottle.

Optical flow 3-circle model



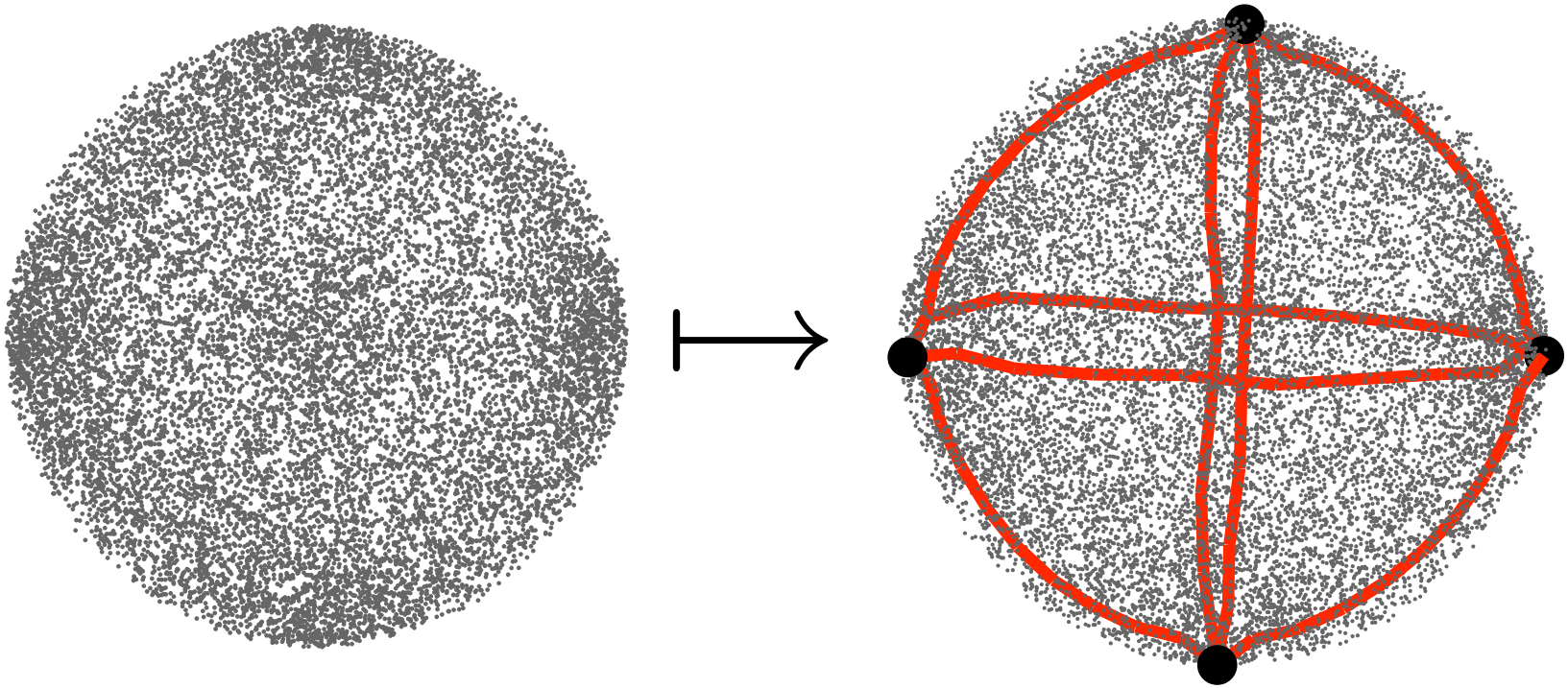
Fills to torus.

Motivating questions

(A) How do you find a simple model matching the Betti barcodes?

(B) Can we be more robust to noise?

We'd like:

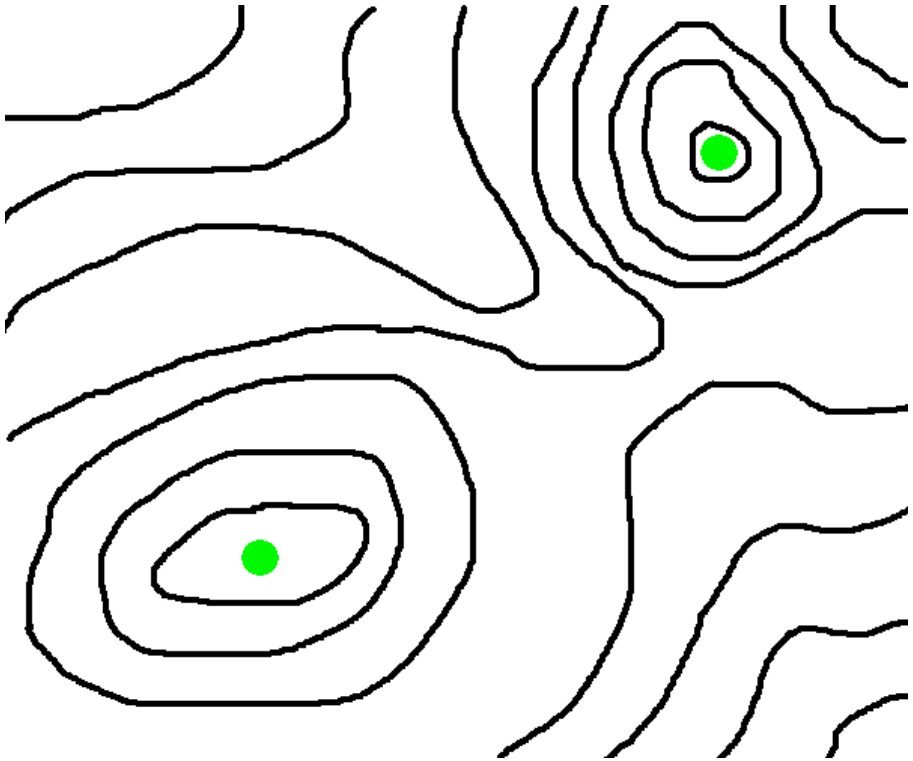


Input

Output

Nudged elastic band (chemistry)

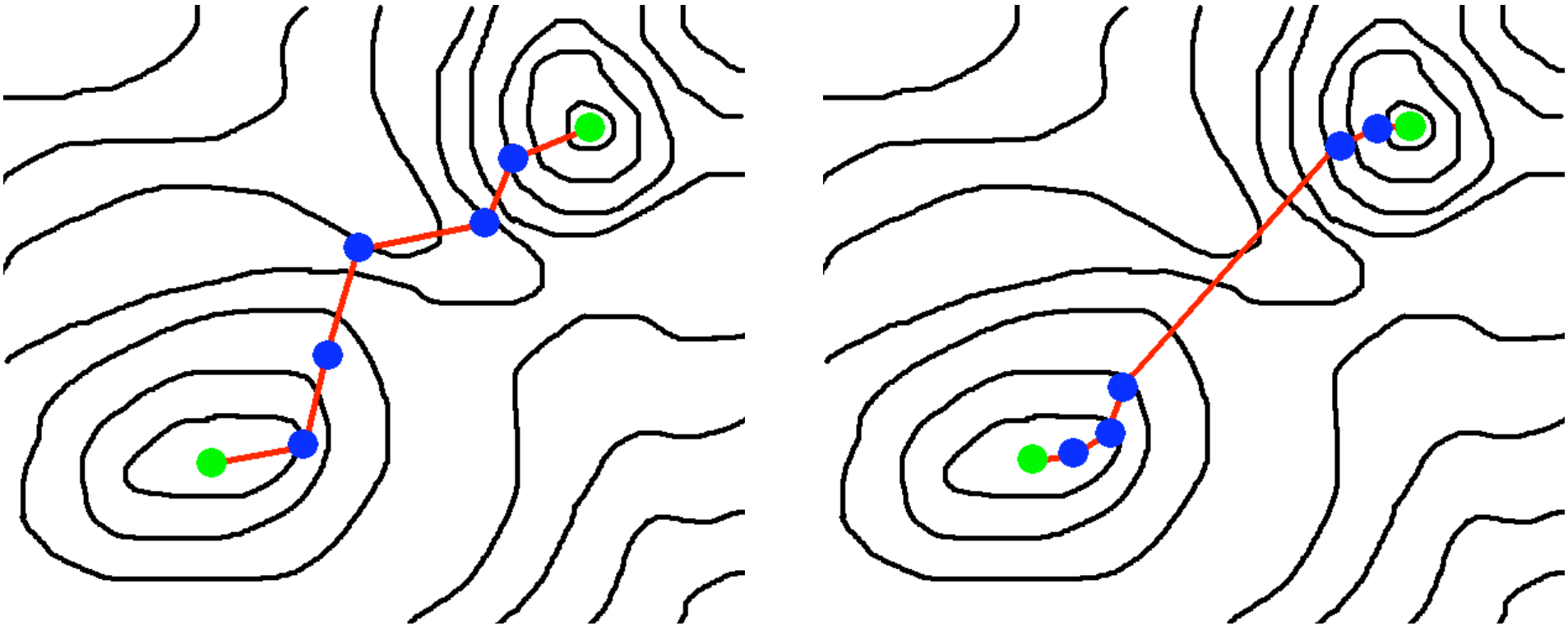
- Energy landscape for molecule configurations
- Local minima are stable configurations
- Minimum energy paths are transitions



Nudged elastic band method for finding minimum energy paths of transitions by H. Jónsson, G. Mills, and K. W. Jacobsen (1998)

Nudged elastic band (chemistry)

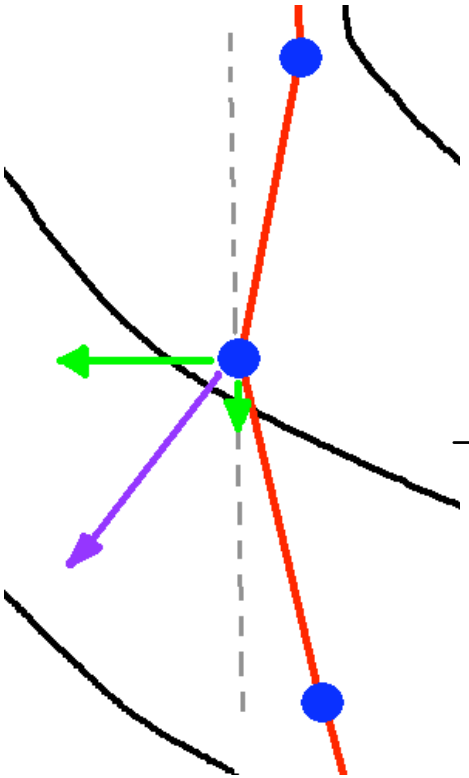
- Start with piecewise-linear band
- Evolve it towards the minimum energy path
- First guess: move each node according to $-\nabla$Energy. This fails.



Nudged elastic band (chemistry)

- Instead, use an energy and a "spring" force.
- The force acting on a node is

$$\text{Force} = -c\nabla\text{Energy}|_{\perp} + (||u^{+}|| - ||u^{-}||)\tau$$



$u^{+},\ u^{-}$ are the adjacent vectors in the band.

τ is the tangent approximation. Naïve choice is $\tau = \frac{u^{+}+u^{-}}{2}$.

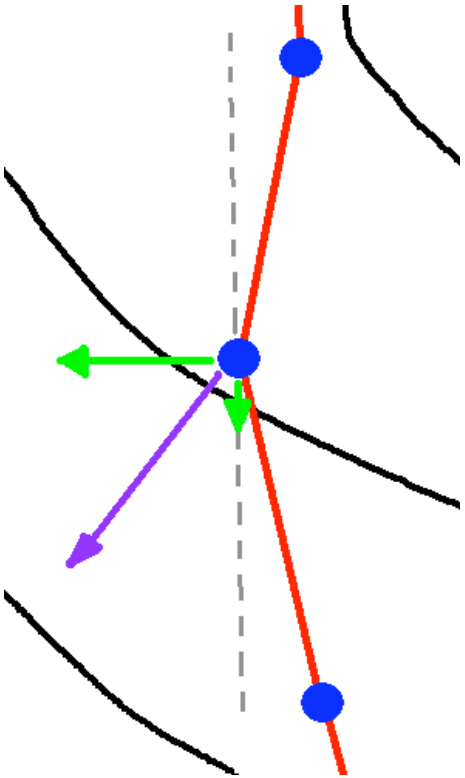
$-\nabla\text{Energy}|_{\perp}$ is the component of the negative energy gradient perpendicular to τ.

c is the constant of proportionality between the forces.

Nudged elastic band (chemistry)

- Instead, use an energy and a "spring" force.
- The force acting on a node is

$$\text{Force} = -c\nabla\text{Energy}|_{\perp} + (||u^{+}|| - ||u^{-}||)\tau$$

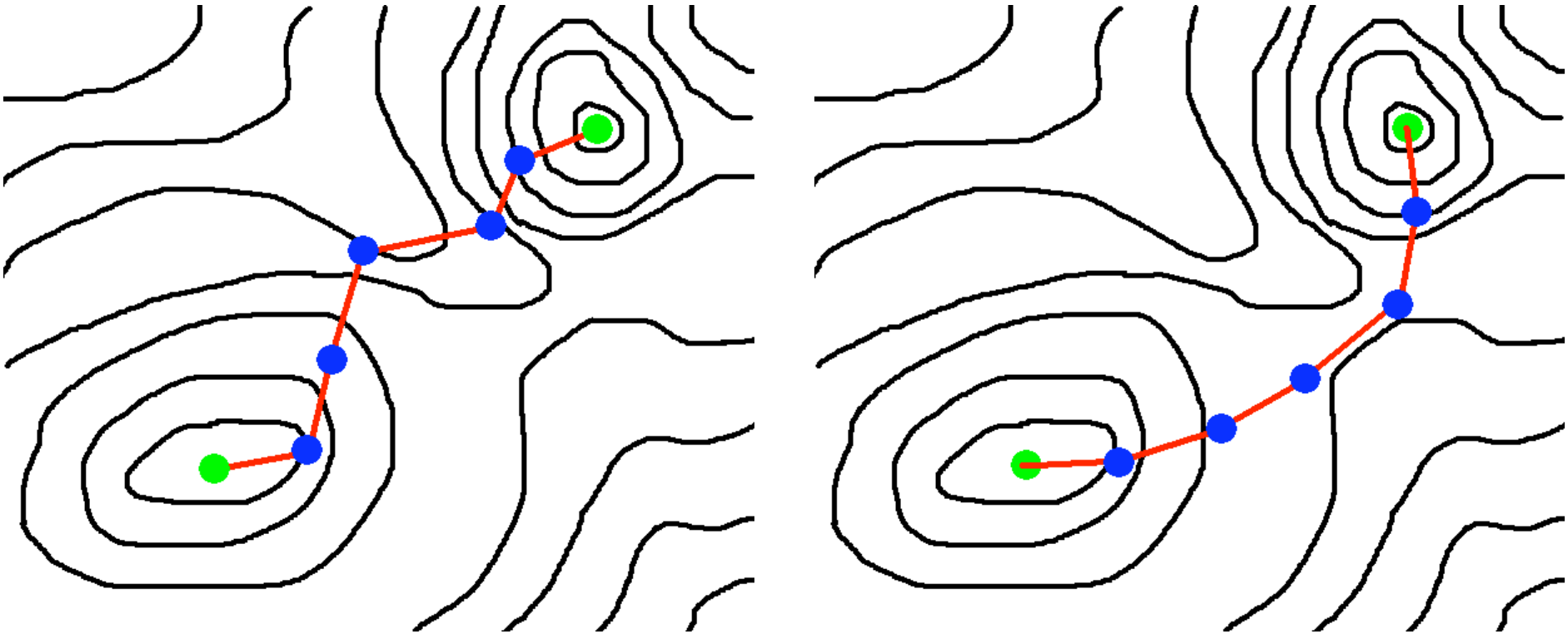


- Simulate the differential equation: take small step, recalculate forces.
- Stop when forces become very small.
- May be multiple transition paths.
- More sophisticated options.

Nudged elastic band (chemistry)

- Instead, use an energy and a "spring" force.
- The force acting on a node is

$$\text{Force} = -c\nabla\text{Energy}|_{\perp} + (||u^{+}|| - ||u^{-}||)\tau$$



Nudged elastic band (data analysis)

- Maximize density instead of minimizing energy.
- Local maxima are vertices.
- Maximum density paths are edges.

Nudged elastic band (data analysis)

- Maximize density instead of minimizing energy.
- Local maxima are vertices.
- Maximum density paths are edges.
- Higher dimensional cells?
- Inductively, we build a cell complex model.

Density function

- We want a differentiable density function built from our point cloud $X \subset \mathbb{R}^n$.
- A natural choice is

$$\text{Density}(x) = \frac{1}{|X|} \sum_{y \in X} \phi_{y,\sigma}(x)$$

where $\phi_{y,\sigma}$ is the probability density function for a normal distribution, centered at y, with standard deviation σ.

Density function

- We want a differentiable density function built from our point cloud $X \subset \mathbb{R}^n$.
- A natural choice is

$$\text{Density}(x) = \frac{1}{|X|} \sum_{y \in X} \phi_{y,\sigma}(x)$$

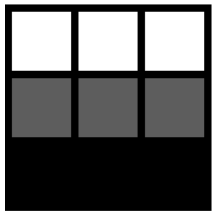
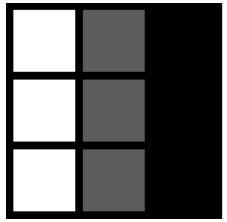
where $\phi_{y,\sigma}$ is the probability density function for a normal distribution, centered at y, with standard deviation σ.

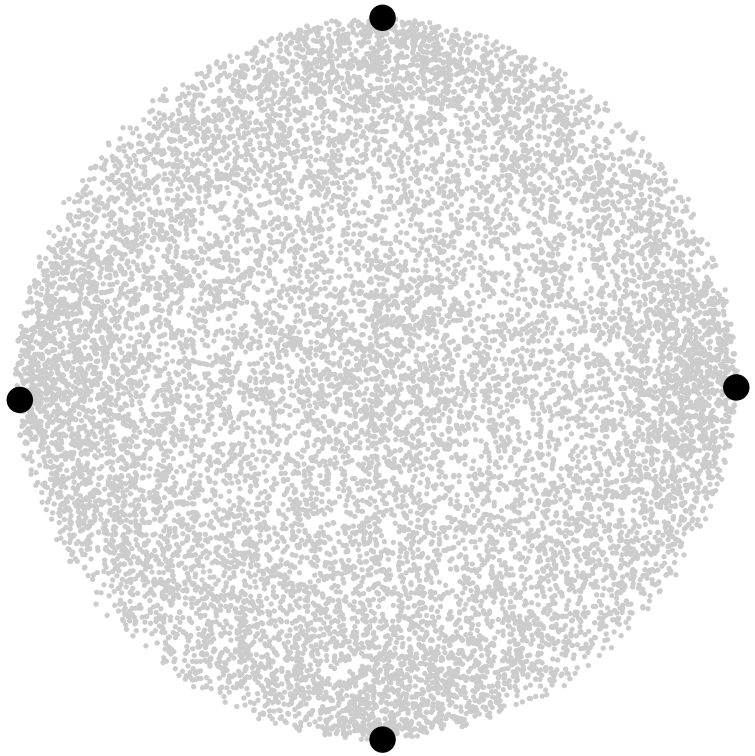
- The choice of σ is analogous to the choice of k in ρ_k: dictates scale of recovered features. In my opinion, this is the essential parameter. Others are easier to choose.

Finding vertices

- Many possibilities.
- Approach analogous to bands would flow vertices in $\mathbb{R}^n$ according to $\nabla \mathrm{Density}$.
- We have better luck as follows:
  - pick an initial seed in the point cloud, step to the point in its neighborhood (100 closest) with highest density.
  - Keep stepping until stabilize at terminal point.
  - Do this for many random seeds.
  - Cluster terminal points, pick one vertex from each large cluster.

Testing on optical data

- 15,000 random points; no dense core subset.
- Four vertices: $\pm$  and $\pm$ 



Finding edges

$$\text{Force} = -c\nabla\text{Energy}|_{\perp} + (||u^{+}|| - ||u^{-}||)\tau$$

becomes

$$\text{Velocity} = c\nabla\text{Density}|_{\perp} + (||u^{+}|| - ||u^{-}||)\tau$$

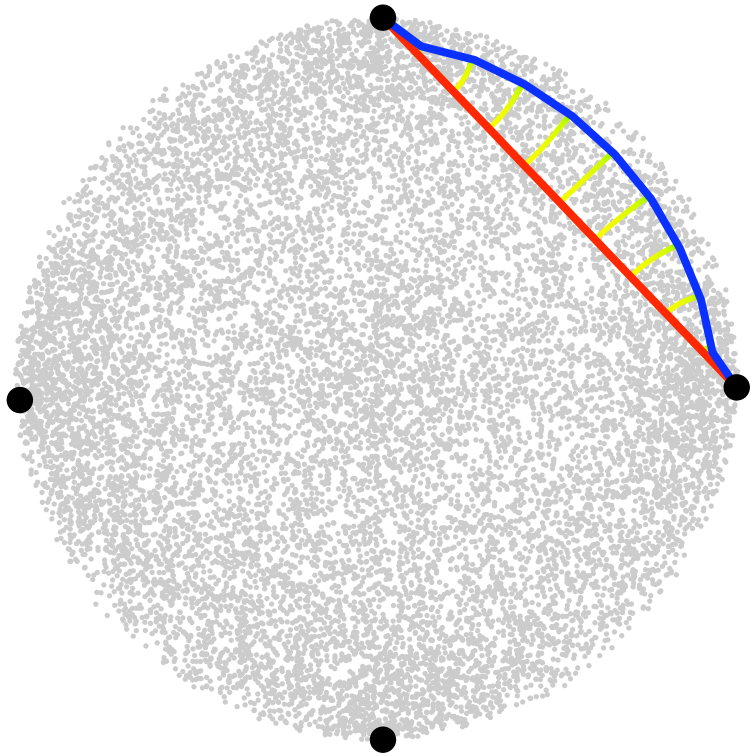
- For us, swapping force/acceleration with velocity is a matter of preference.
- Naïve tangent, naïve simulator (parameters)
- Choose c so that $||\nabla\phi_{\sigma}||_{\infty} = 1$. Depends on dimension.
- Angle force.

Finding edges

- Between every pair of vertices we throw a collection of random initial edges.
- Discard bands that don’t converge or that lie near a non-endpoint vertex.
- Cluster remaining bands (simple metric on bands sharing endpoints), pick one edge from each large cluster.

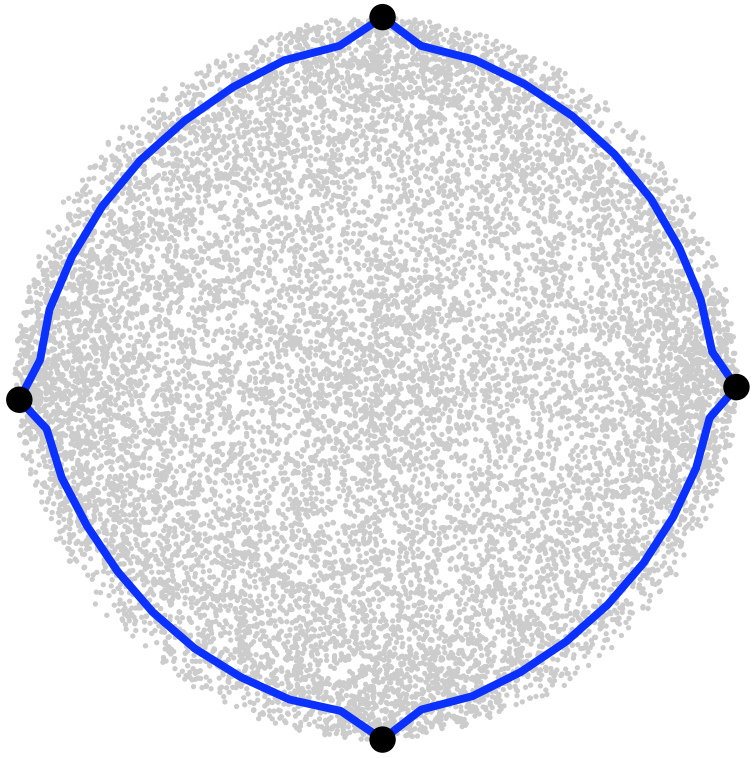
Testing on optical data

- Four vertices
- Adjacent vertices: four edges on primary circle



Testing on optical data

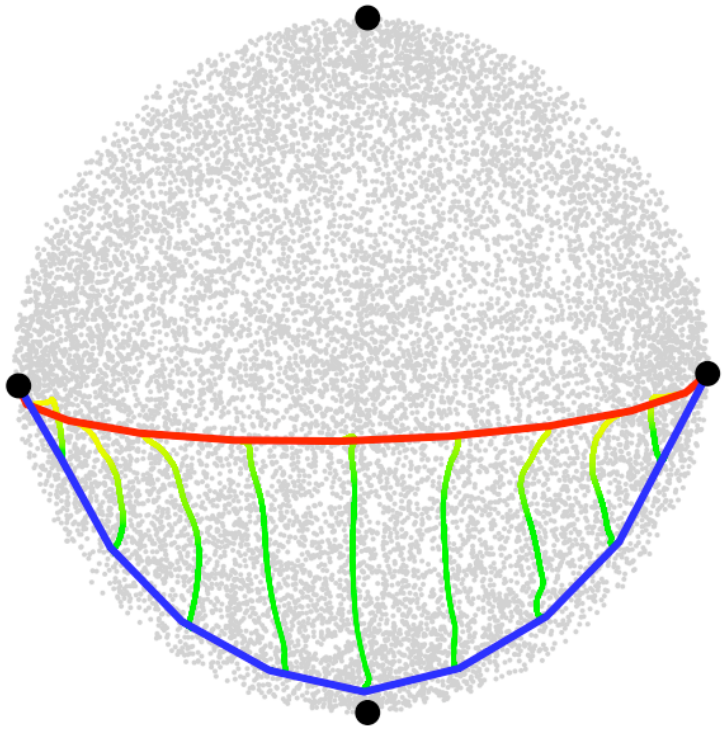
- Four vertices
- Adjacent vertices: four edges on primary circle



Testing on optical data

- Four vertices
- Adjacent vertices: four edges on primary circle

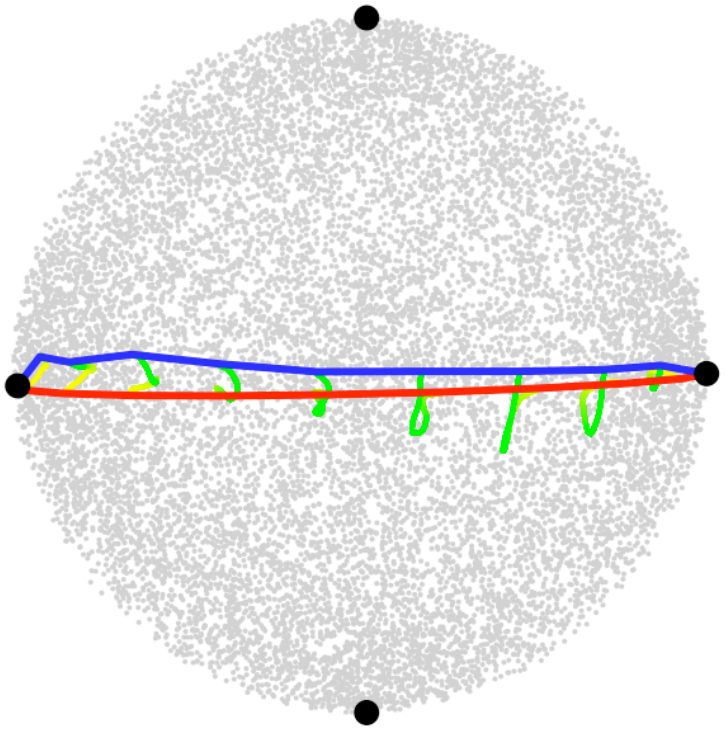
90% on primary circle



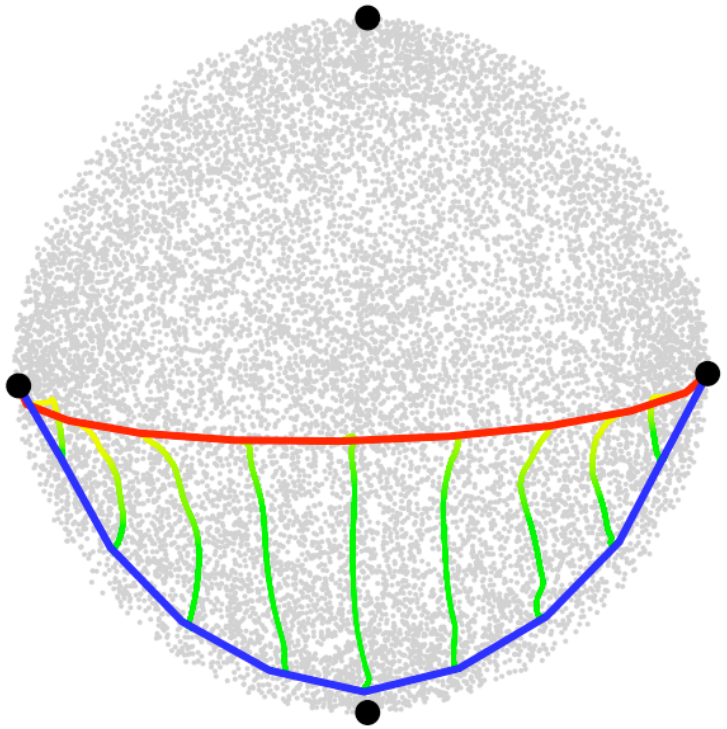
Testing on optical data

- Four vertices
- Adjacent vertices: four edges on primary circle

10% on secondary circles



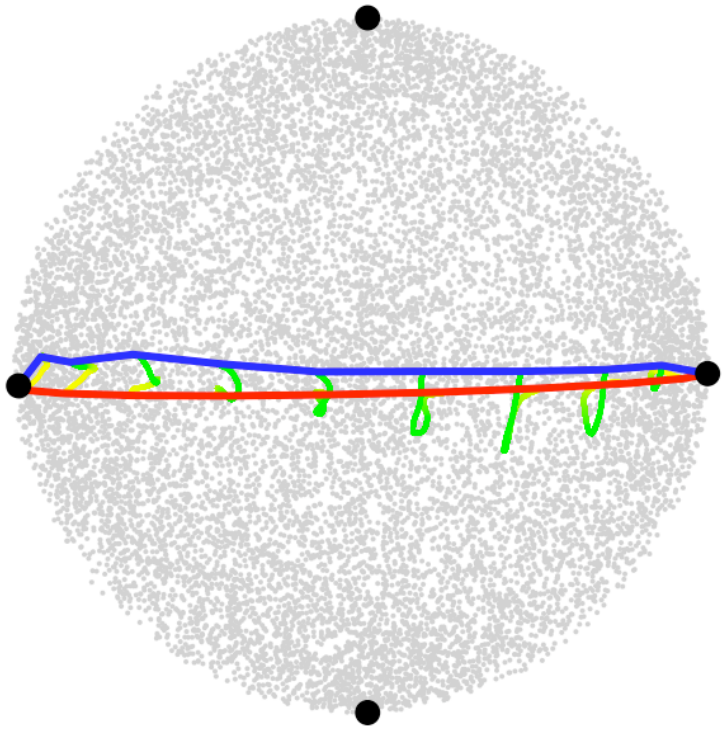
90% on primary circle



Testing on optical data

- Four vertices
- Adjacent vertices: four edges on primary circle

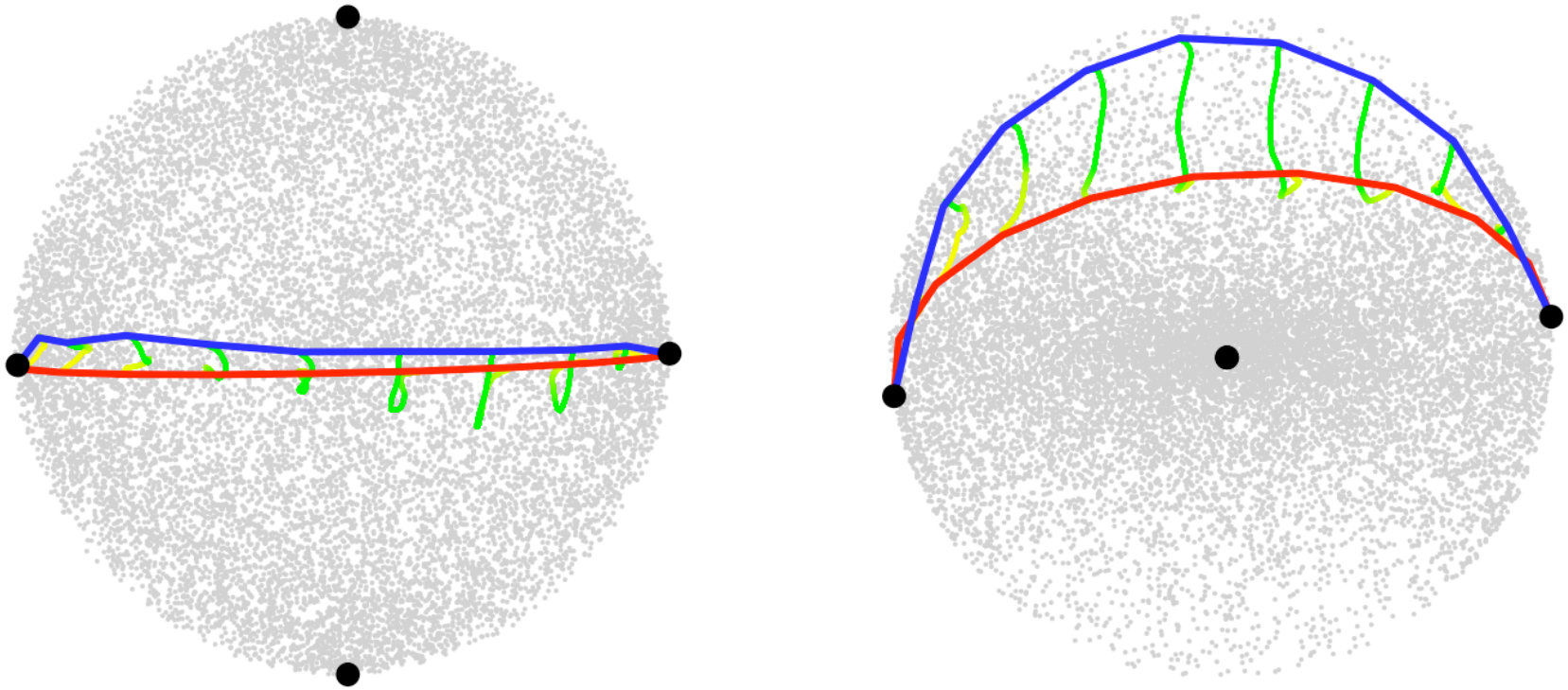
10% on secondary circles



Testing on optical data

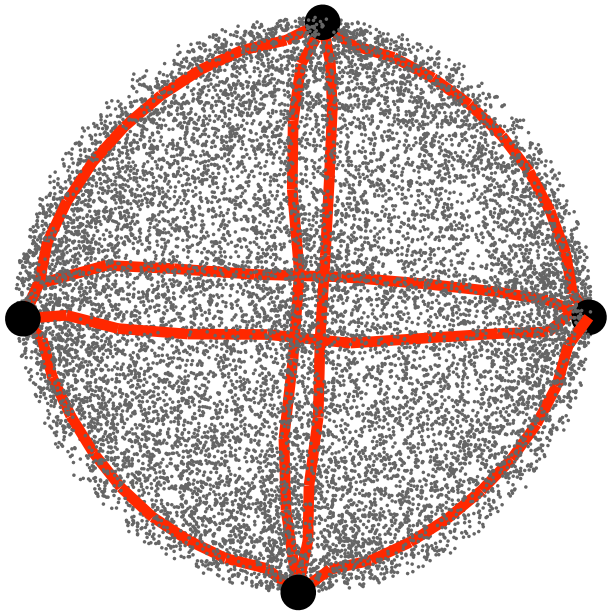
- Four vertices
- Adjacent vertices: four edges on primary circle
- Antipodal vertices: Two edges on each secondary circle

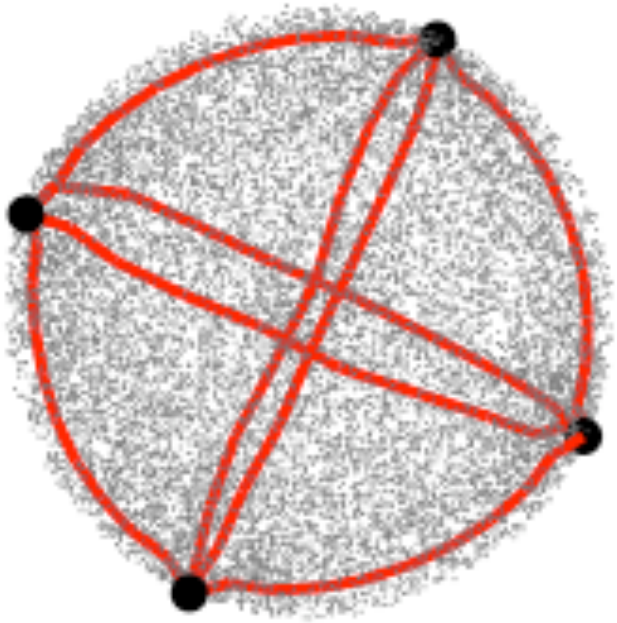
10% on secondary circles



Testing on optical data

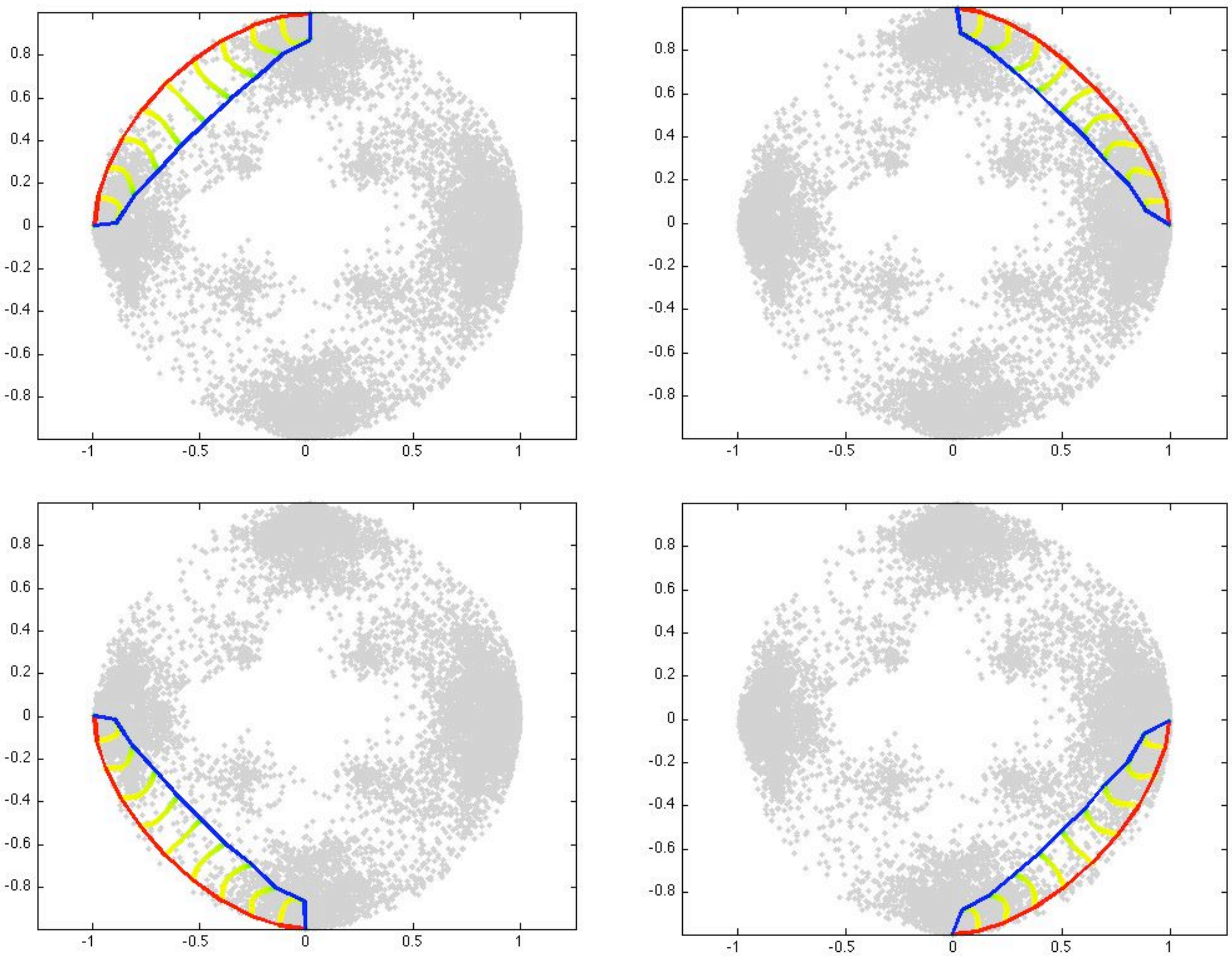
- Four vertices
- Adjacent vertices: four edges on primary circle
- Antipodal vertices: Two edges on each secondary circle

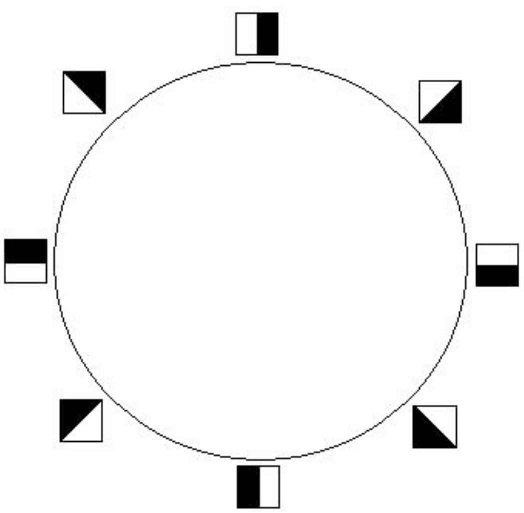




Testing on 5x5 range image patches

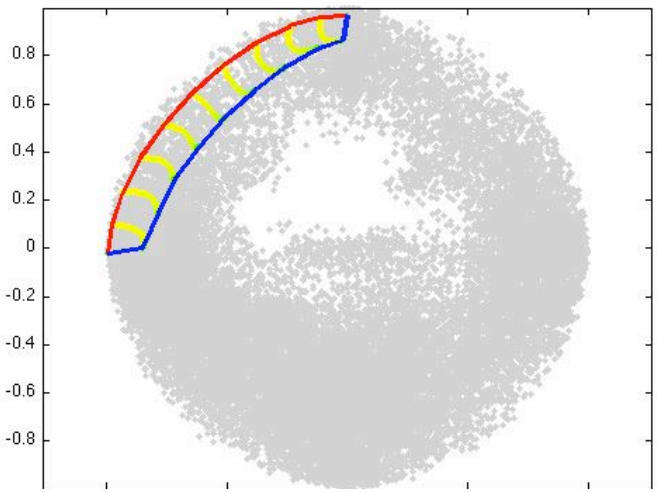
- Subset of 23-sphere
- Find range primary circle

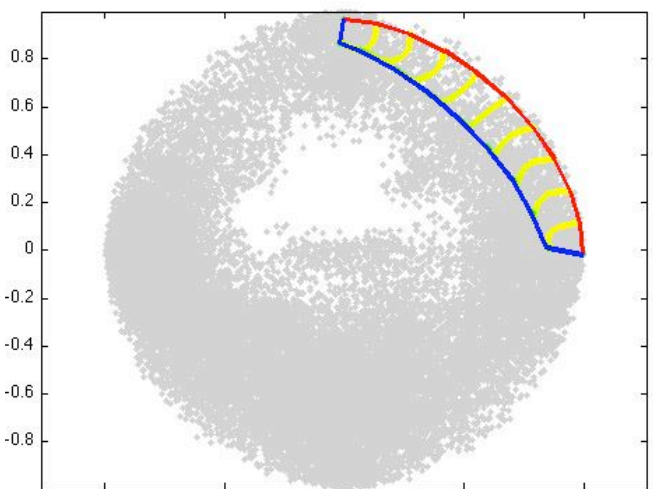


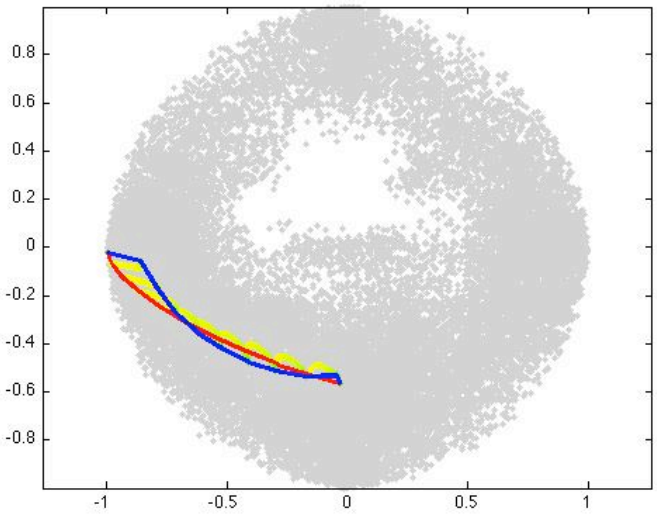


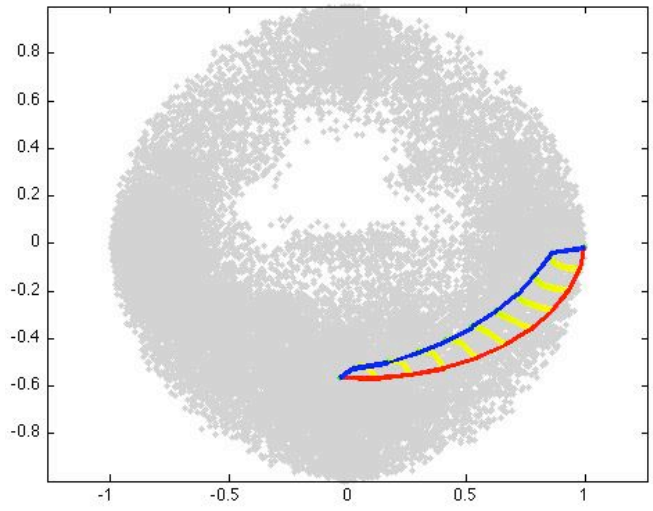
Testing on optical flow patches

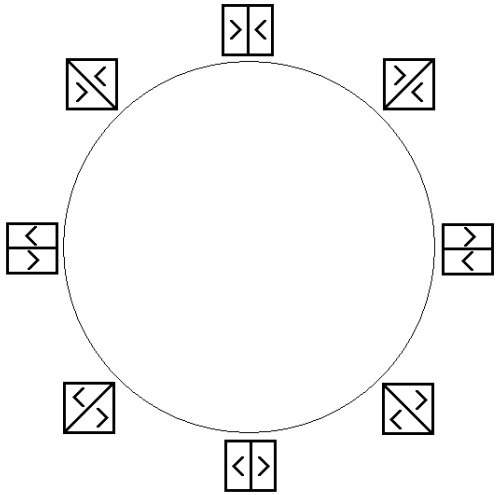
- Subset of 16-sphere
- Find horizontal flow circle





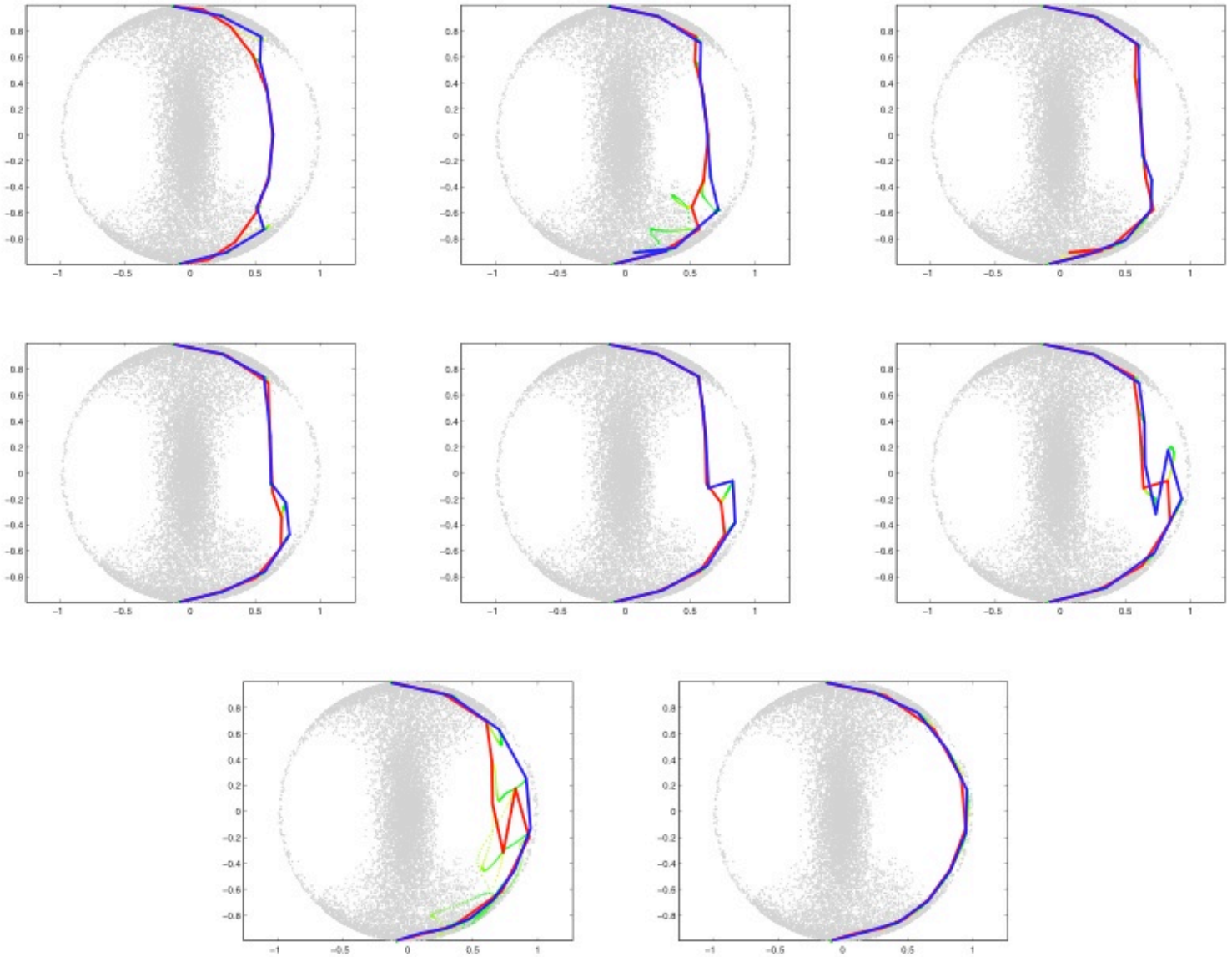






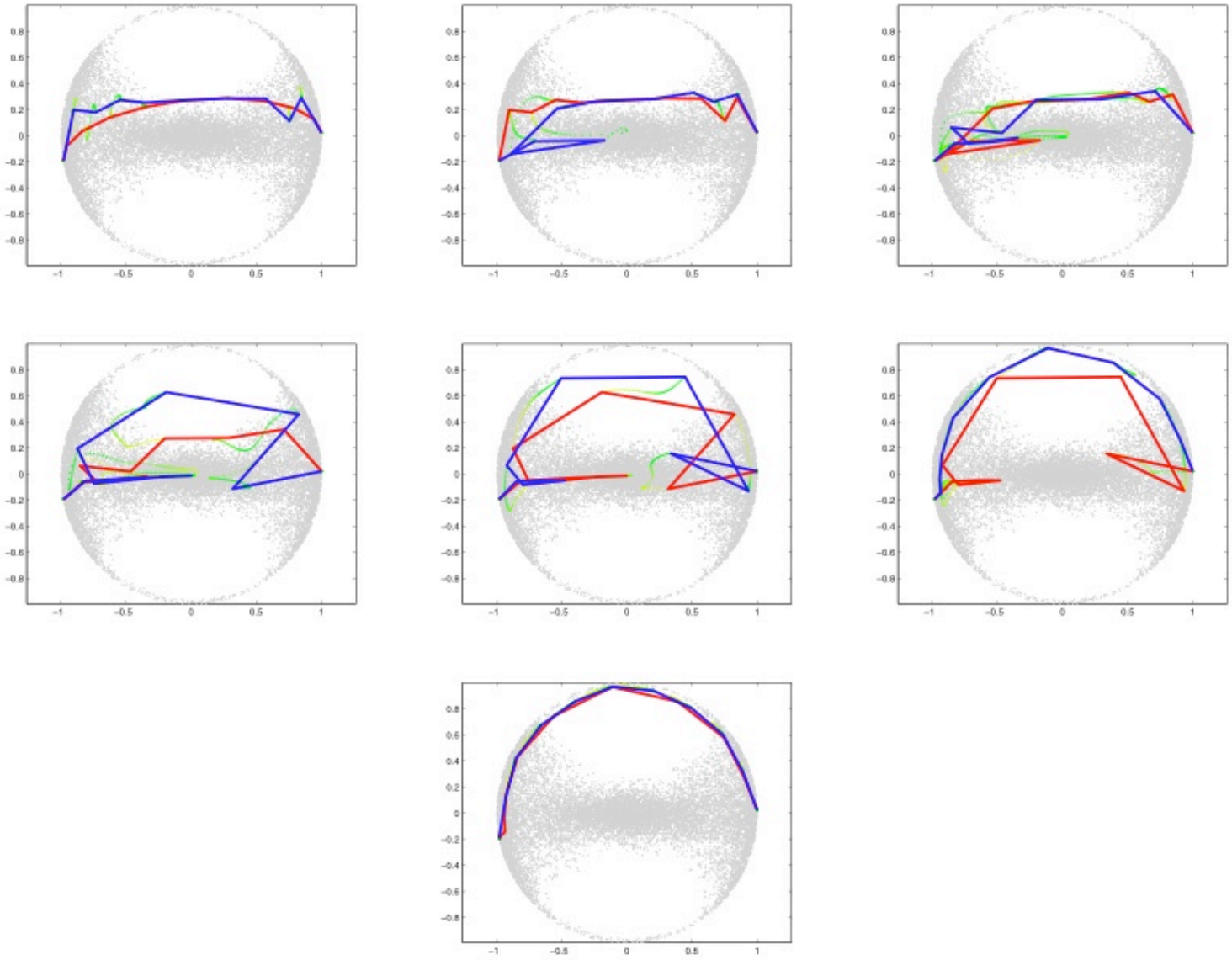
Before angle force, we had kinks!

(Core subset here, not a random subset)

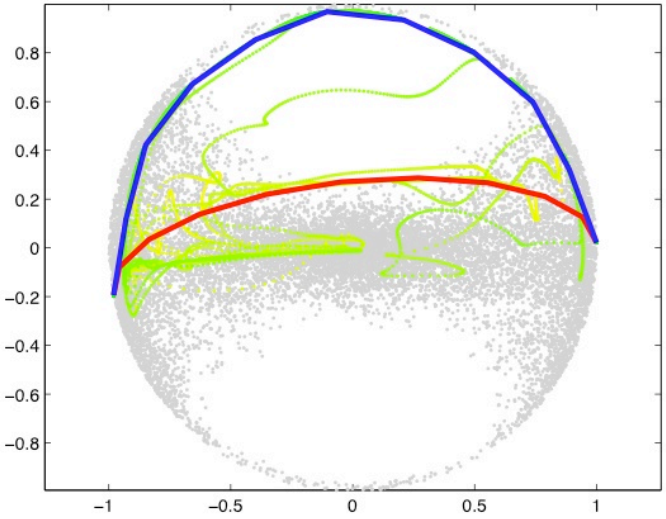


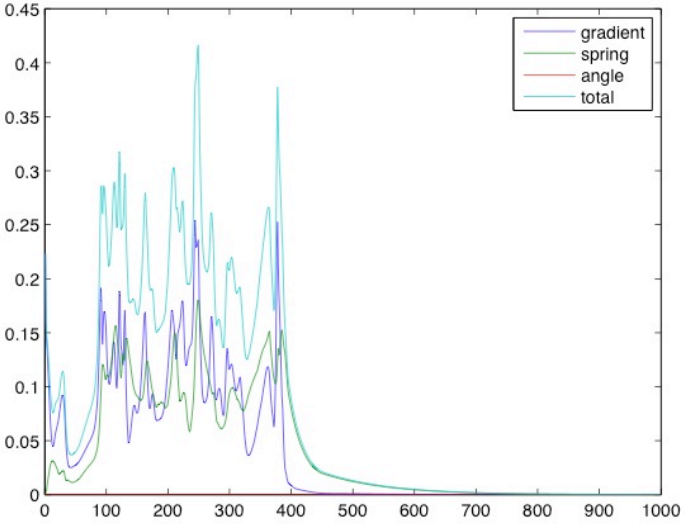
Before angle force, we had kinks!

(Core subset here, not a random subset)

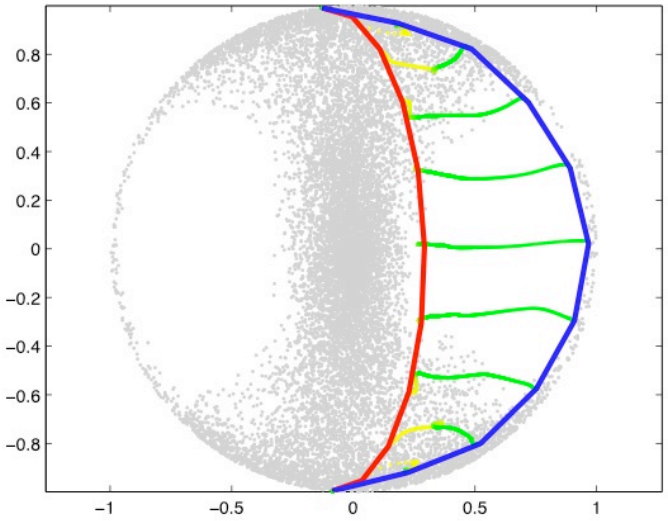


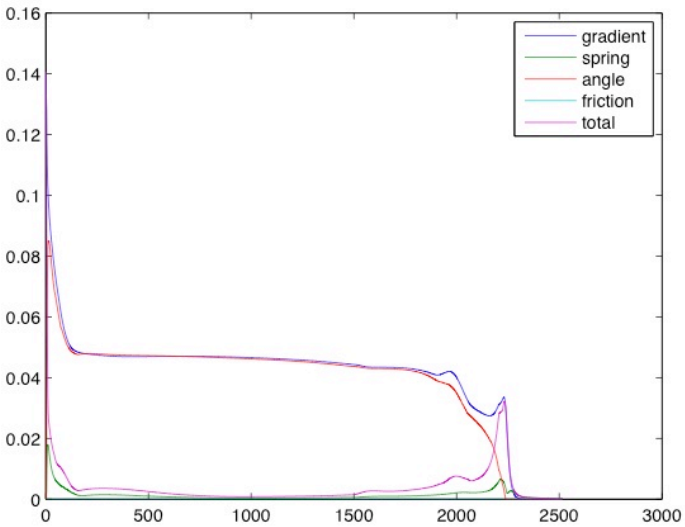
Without angle force



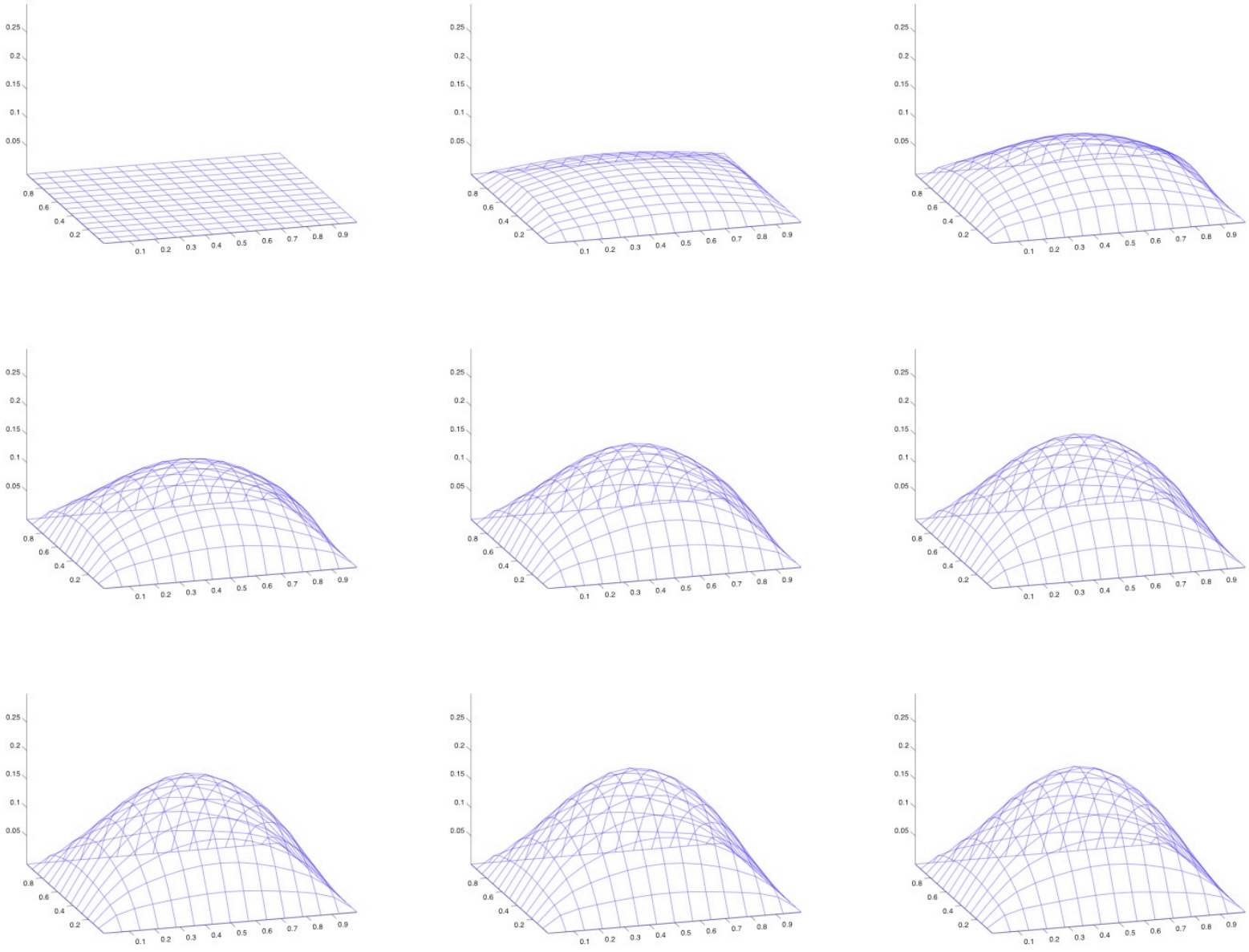


With angle force





Higher dimensional cells



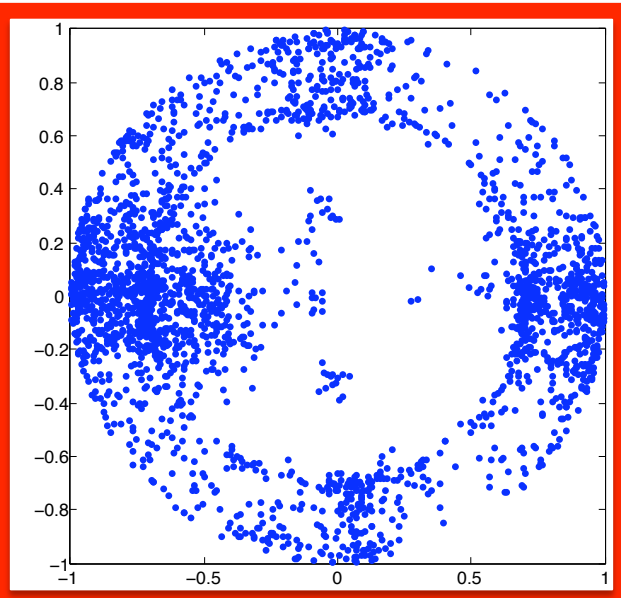
Higher dimensional cells

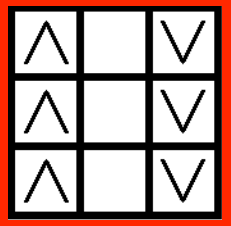
- Idea: find a short word of edges, throw a random 2-cell with that word as its boundary
- Challenges:
  - No canonical meshing of 2-cell
  - Estimating tangent plane
    - Principle component analysis
  - What should spring force be?
    - each edge pulls on adjacent vertices to try to achieve the current average length of all edges.
    - natural spring
  - In need of test datasets to motivate development

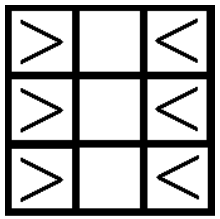
Conclusions

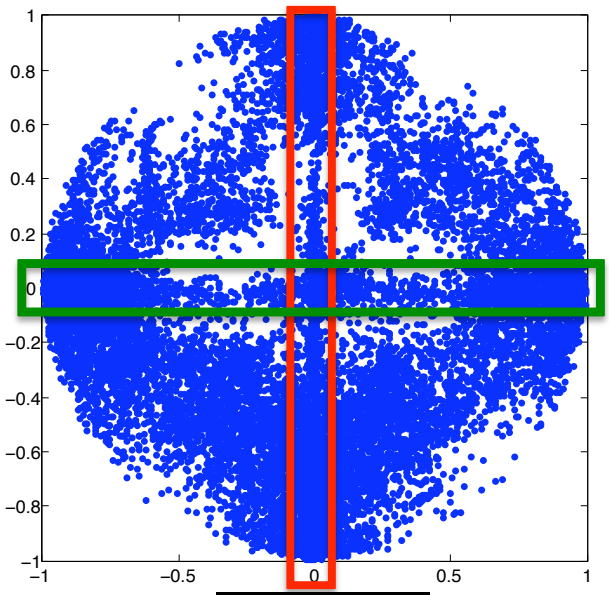
- Nudged elastic band has the potential to locate models for datasets.
- Can be used in concert with persistent homology.
- Shows some tolerance to noise.
- What's next?
  - New datasets

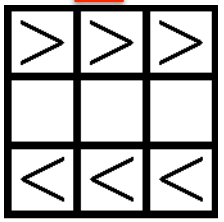
e2 circle











e1 circle