Stabilizing the unstable output of persistent homology computations

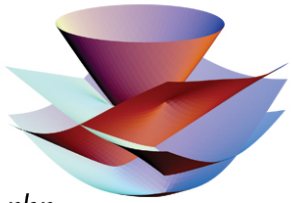
Peter Bubenik
with Paul Bendich (Duke) and Alex Wagner (Florida)

May 5, 2017

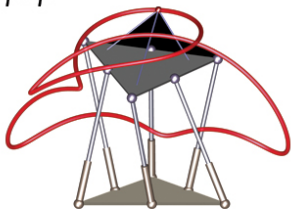
Conference on Applied and Computational Algebraic Topology
Hausdorff Research Institute for Mathematics
Bonn, Germany

University of Florida
http://people.clas.ufl.edu/peterbubenik/

Advertising

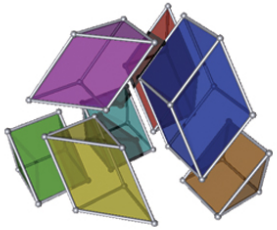


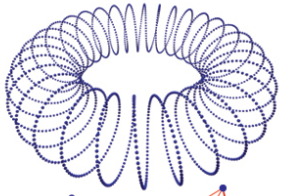
http://www.siam.org/journals/siaga.php

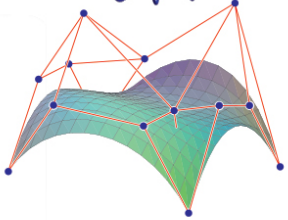


SIAM Journal on

Applied Algebra and Geometry







Now accepting research articles "on the development of algebraic, geometric, and topological methods with strong connection to applications."

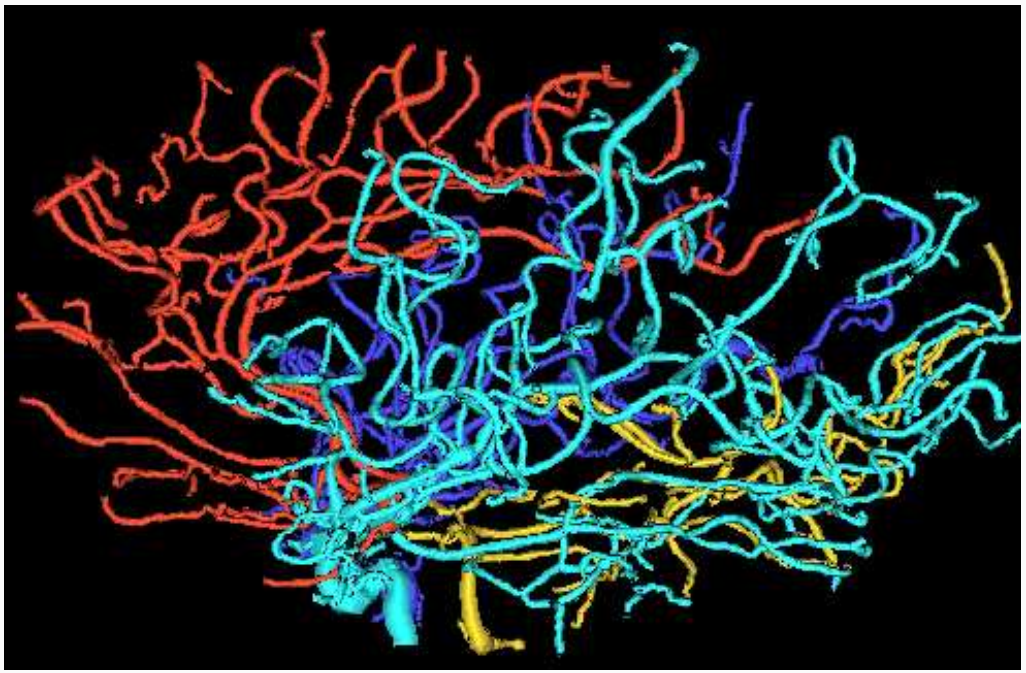
Outline

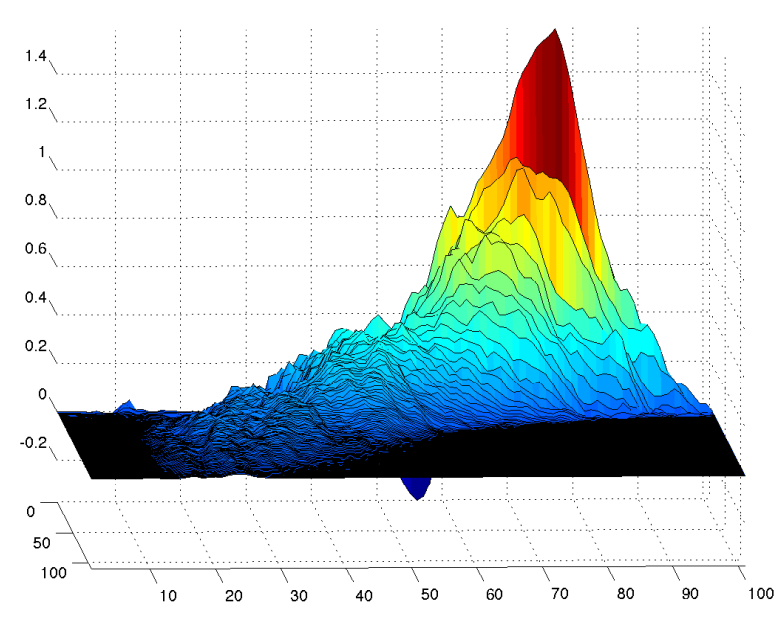
1. Motivation
2. Main Example
3. Theory
4. Other Examples

Motivation

Applied setting: Topological Data Analysis

Differences between two clinical groups

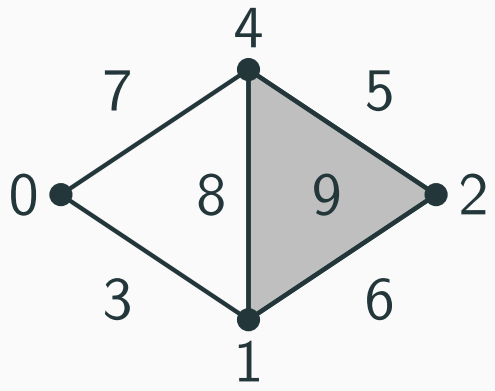




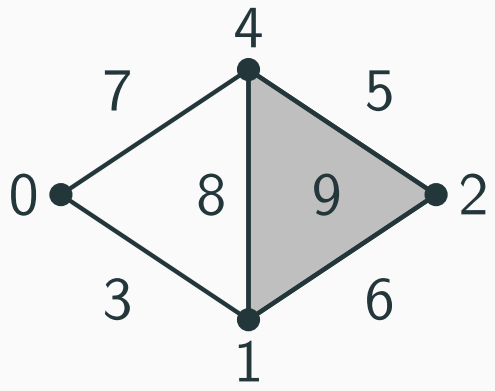
Where is the difference?

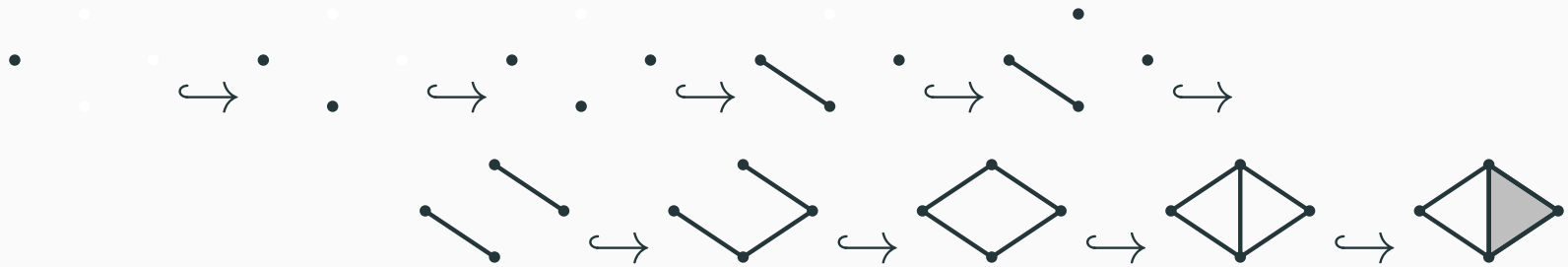
Discrete setting: Filtered simplicial complexes

A filtered simplicial complex

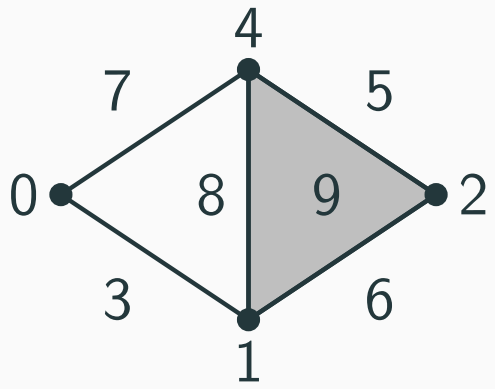


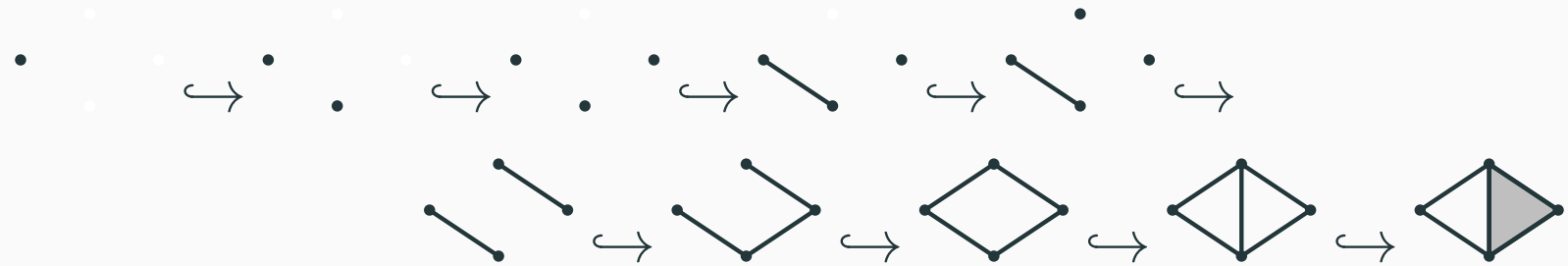
A filtered simplicial complex

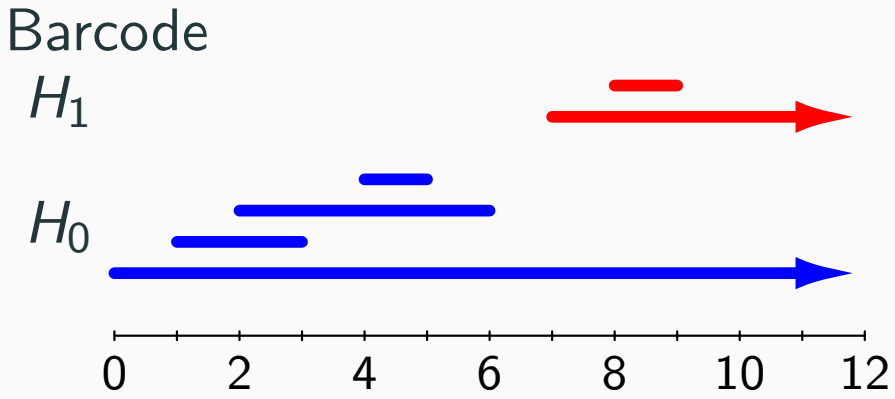


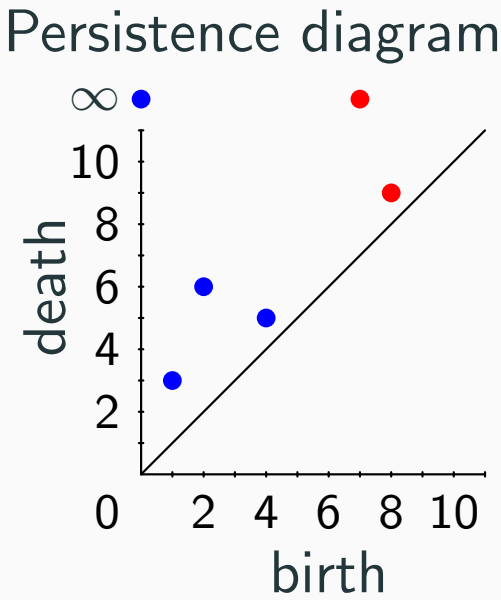


A filtered simplicial complex

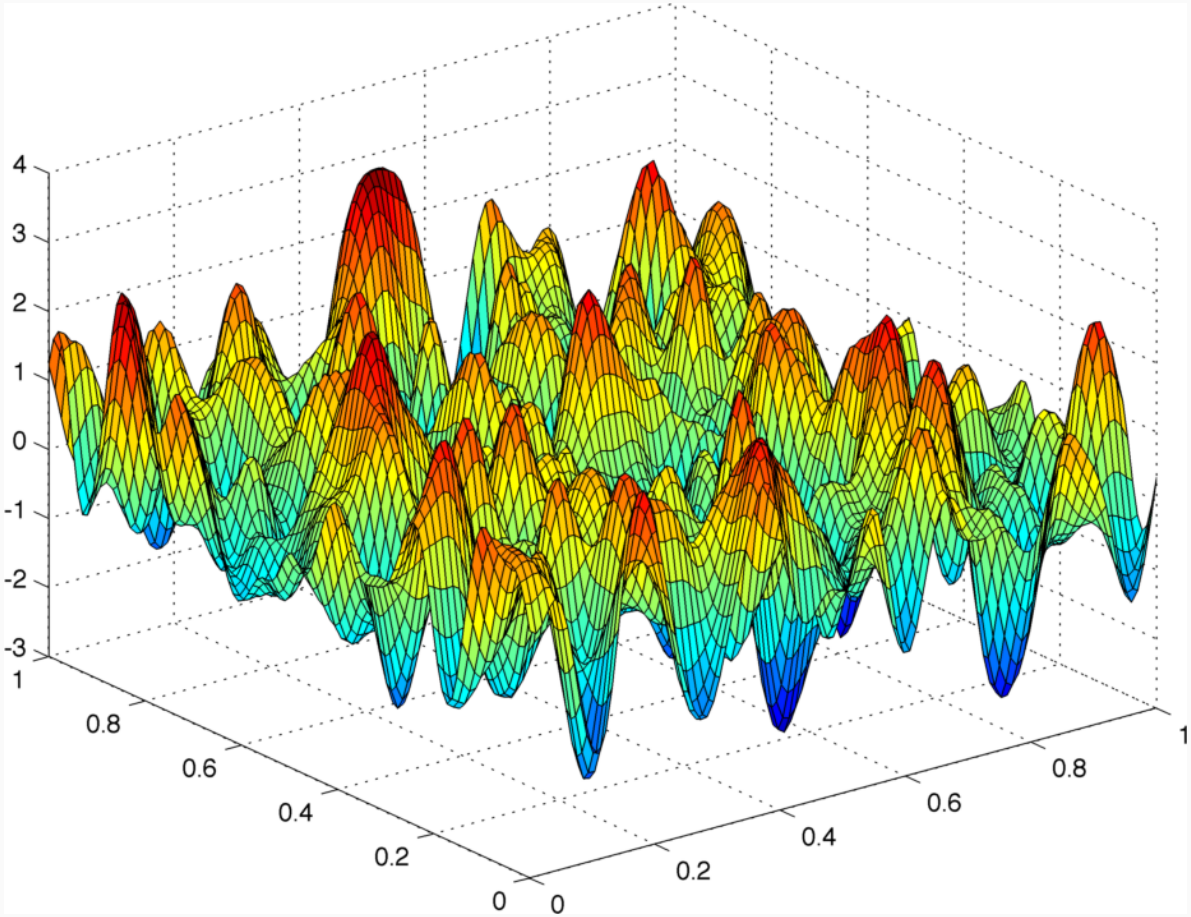




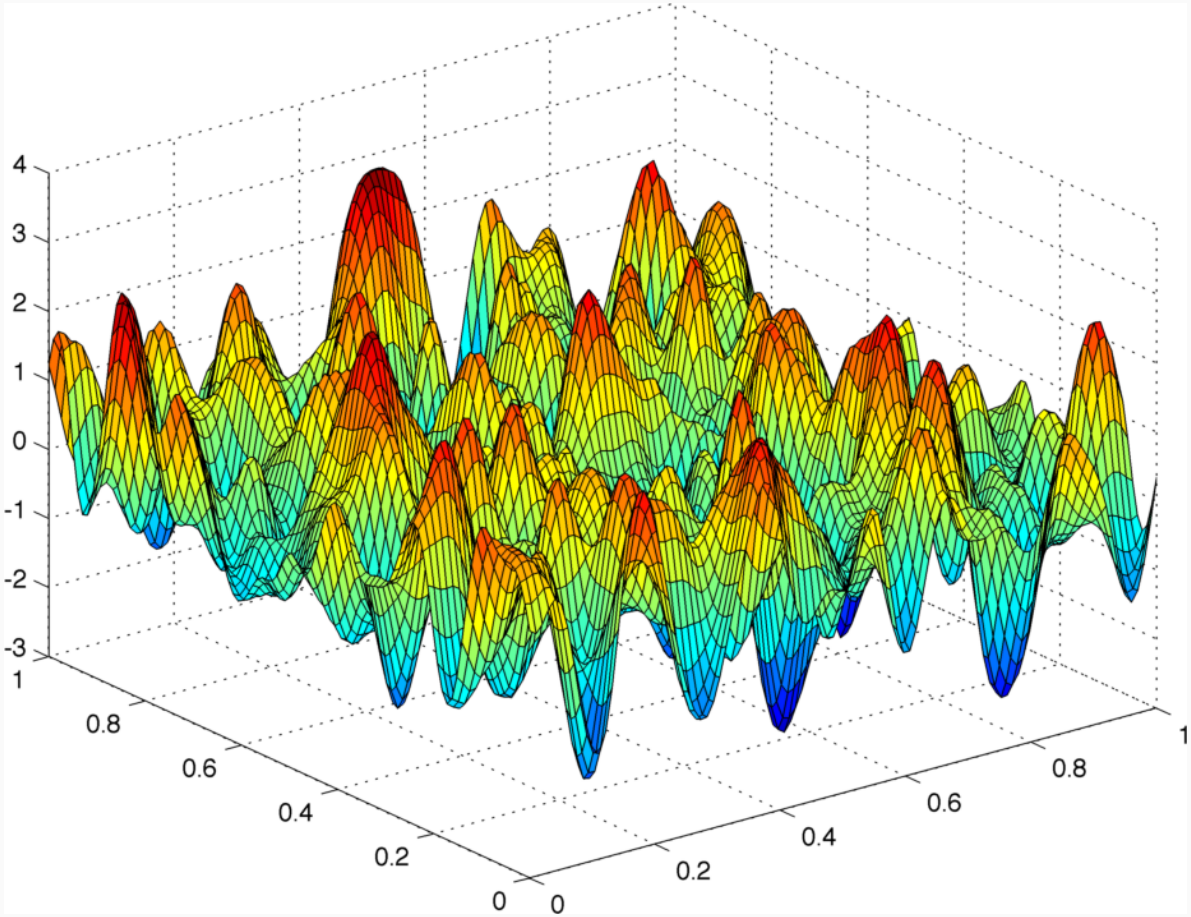




Continuous setting: Morse functions



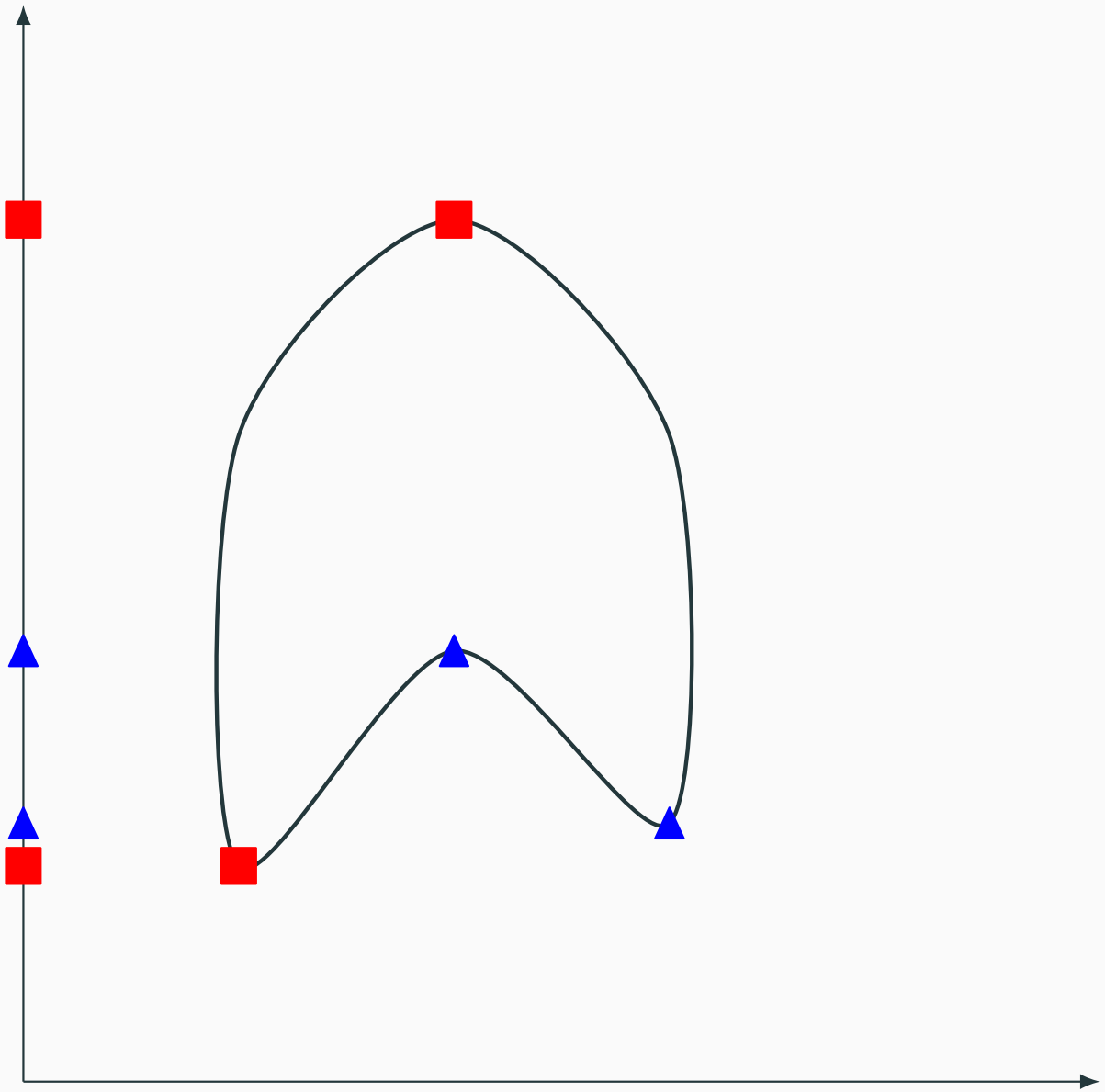
Continuous setting: Morse functions



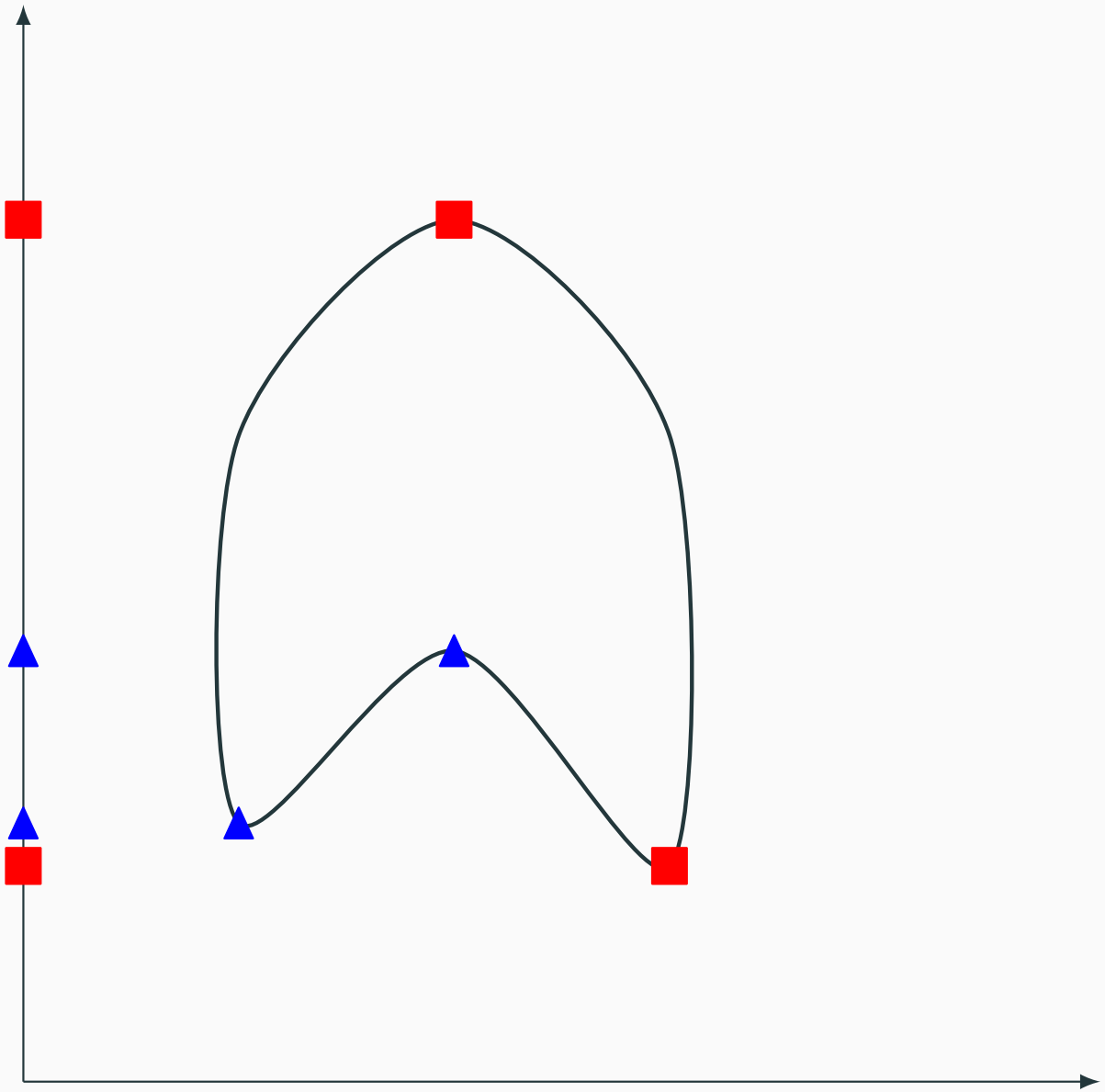
Apply persistent homology:

- The pairing of critical values is stable.
- The pairing of critical points is unstable.

A Morse function, $f : S^1 \to \mathbb{R}$

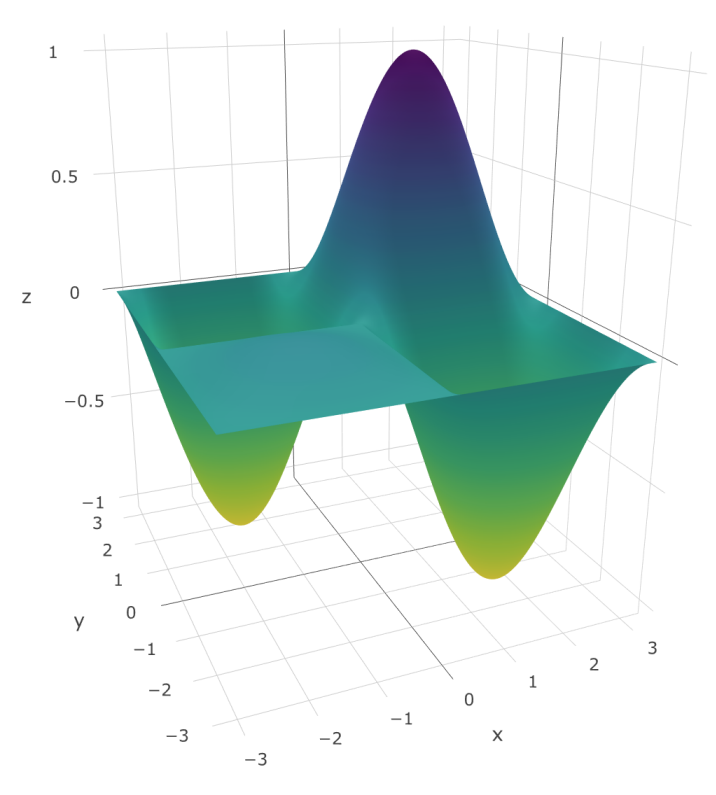


A Morse function, $f : S^1 \to \mathbb{R}$

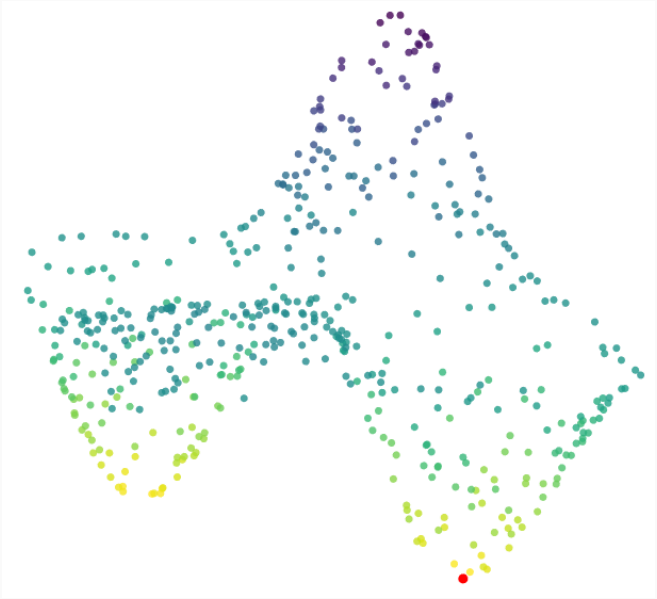


Statistical setting: discrete sample from continuous function

Want: properties of some unknown function



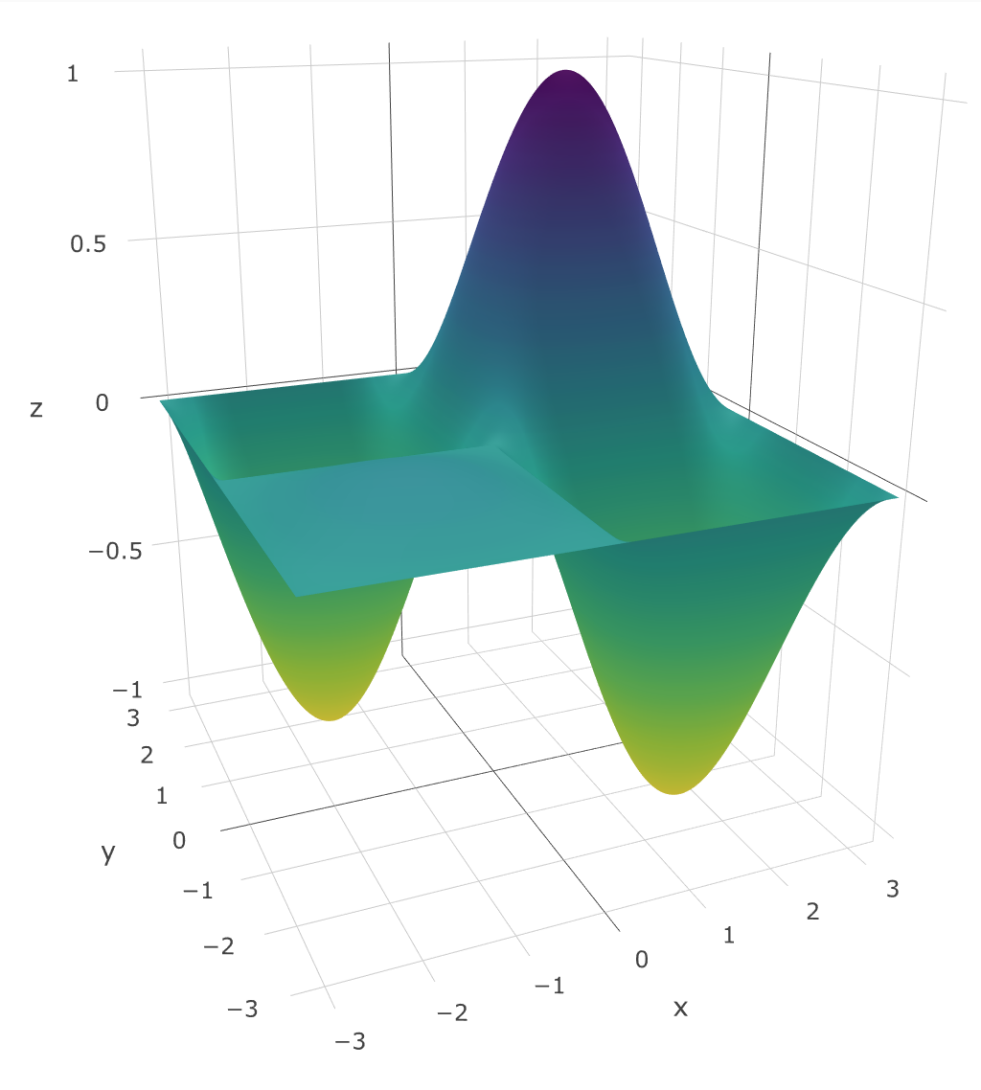
Have: a finite sample



Main Example

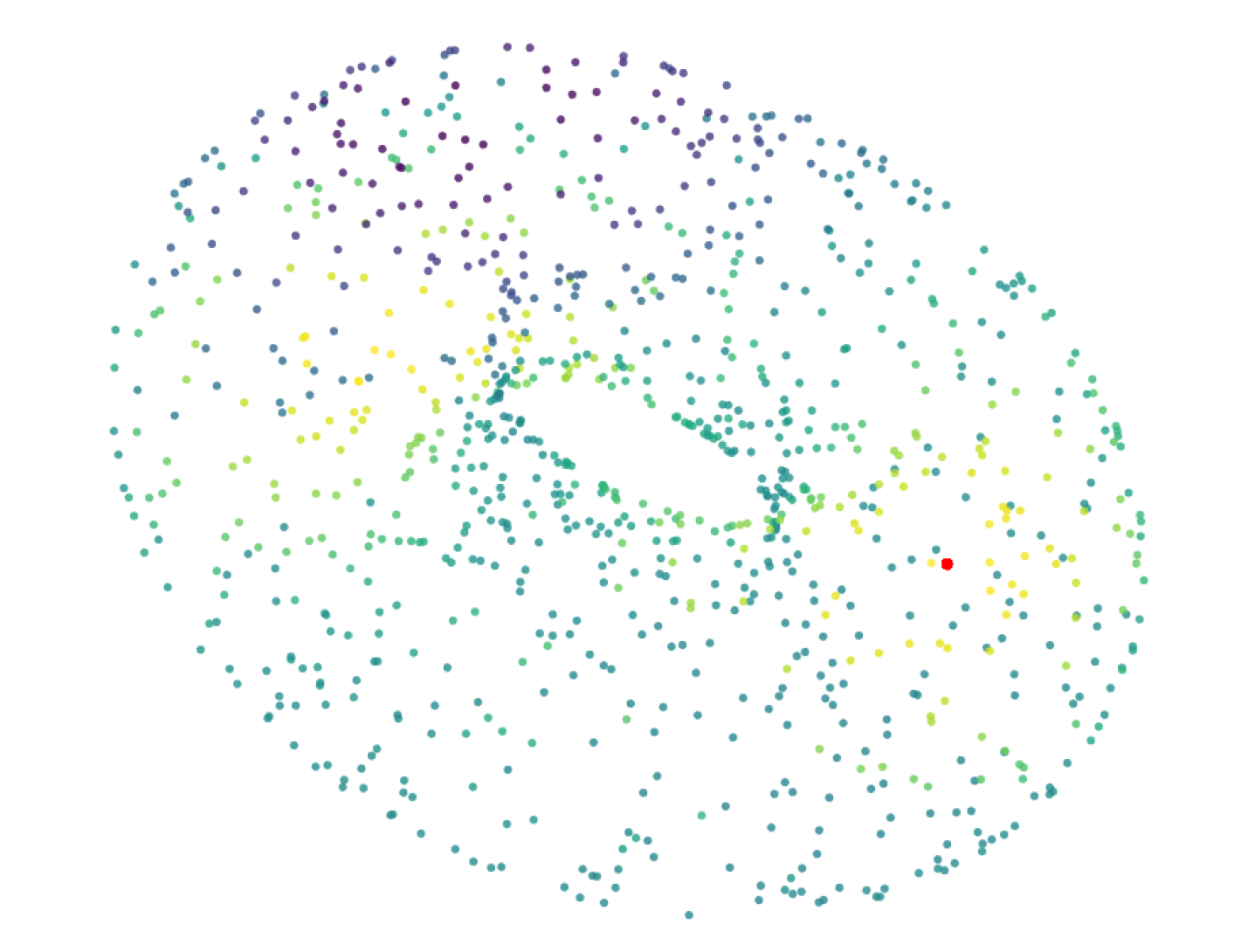
Main example

This function on the square induces a function f on the torus.



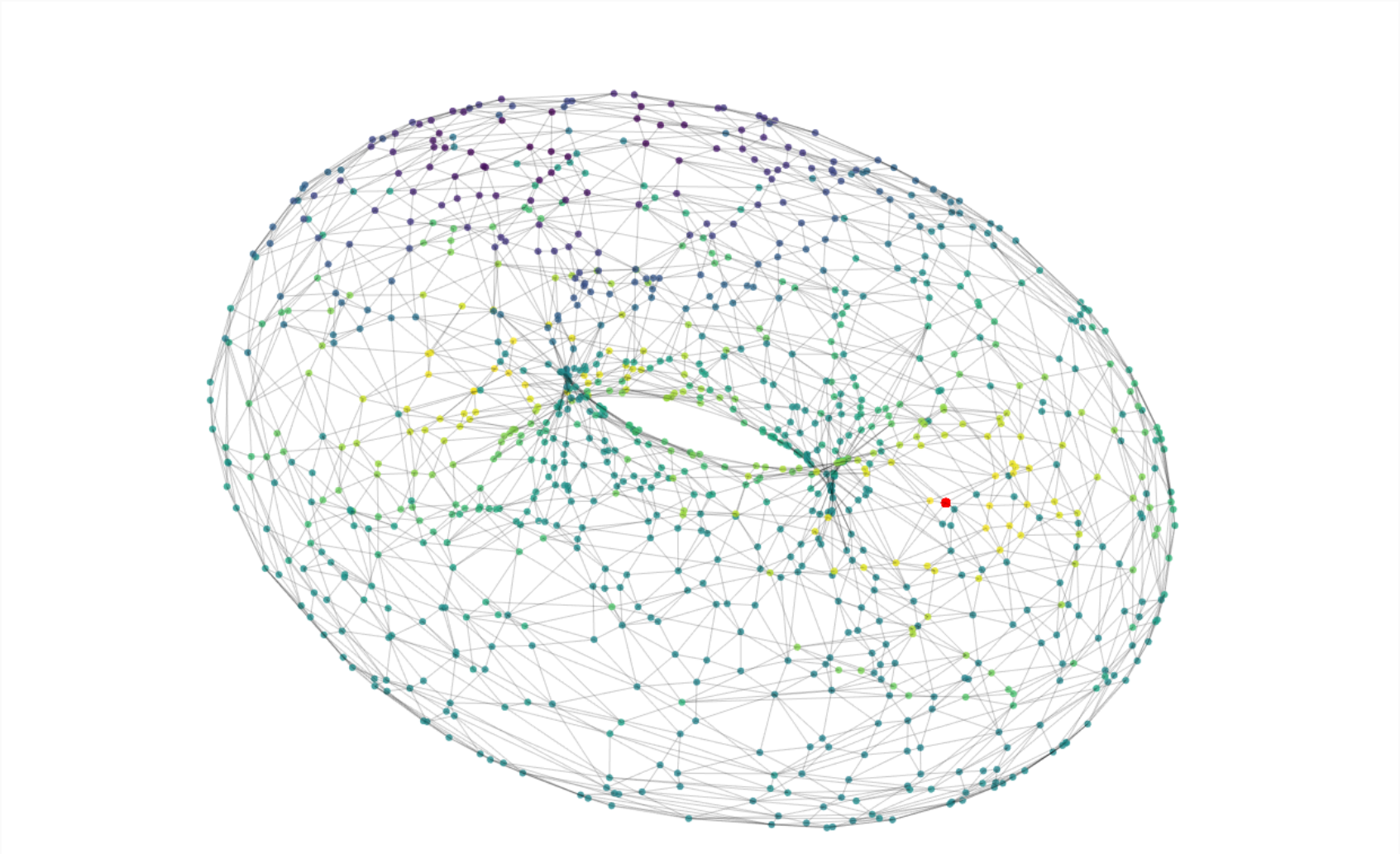
Main example

Suppose we are given a sample $X \subseteq \{(x, f(x)) : x \in T^2\}$ of size N.

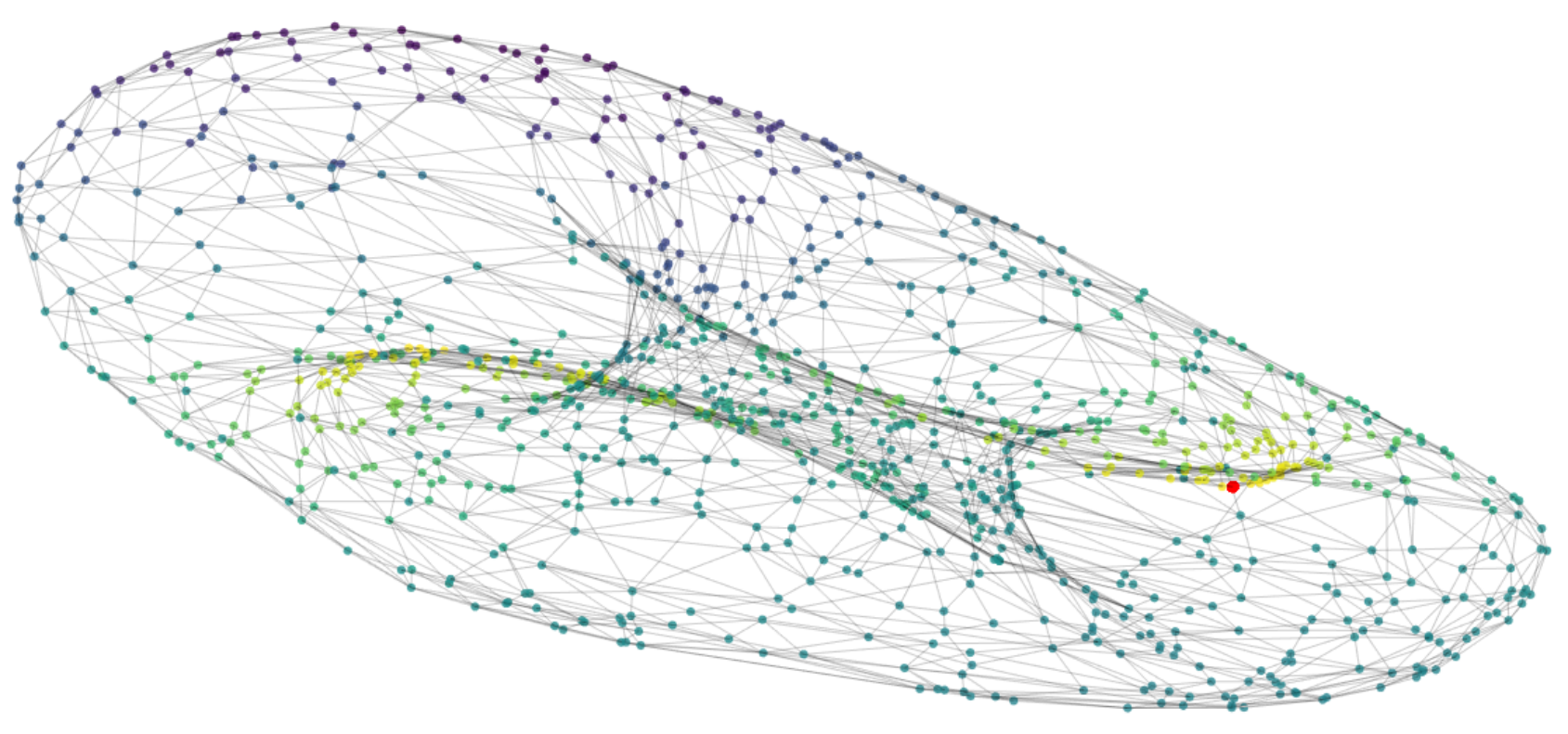


Main example

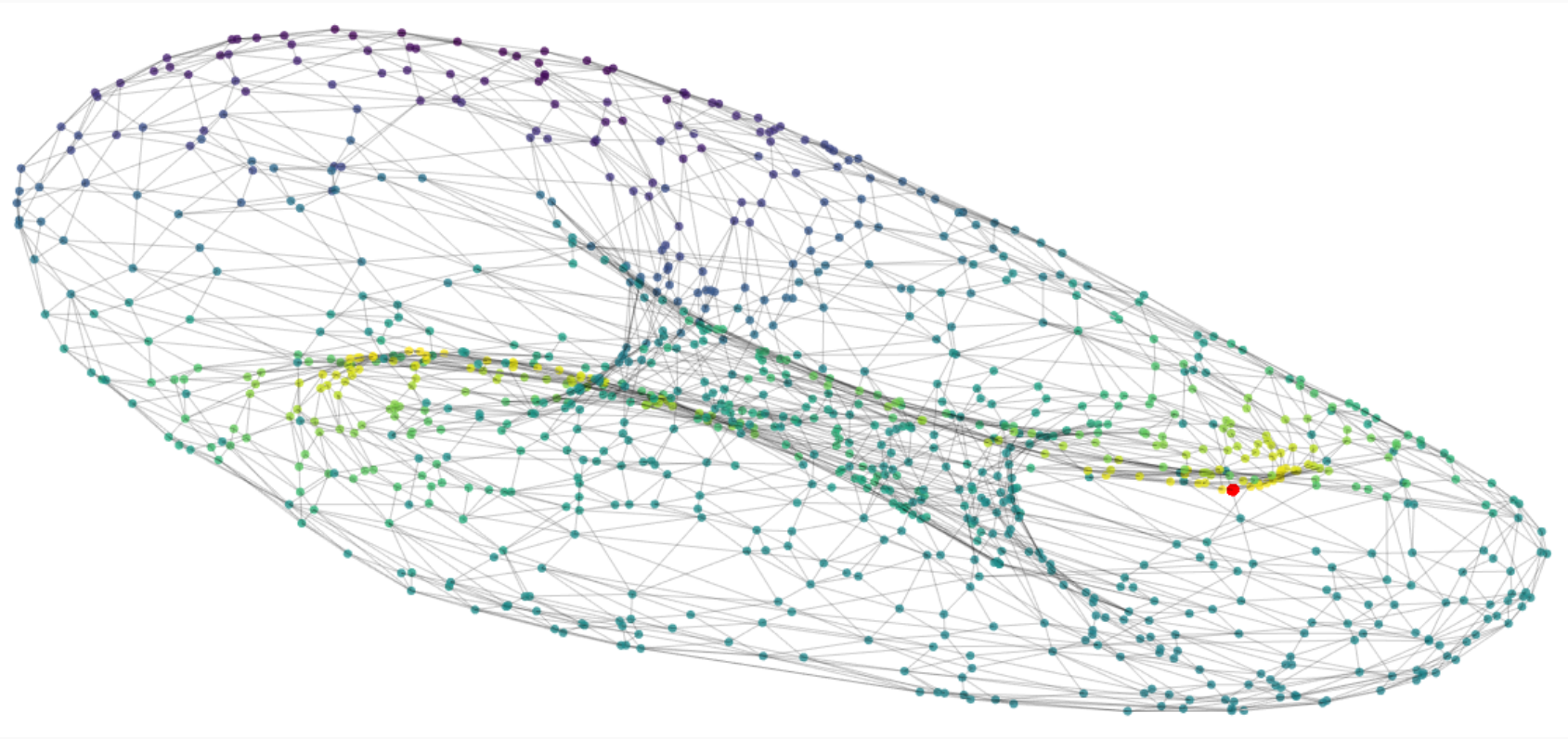
Use the sample to get a Delaunay triangulation of the torus.

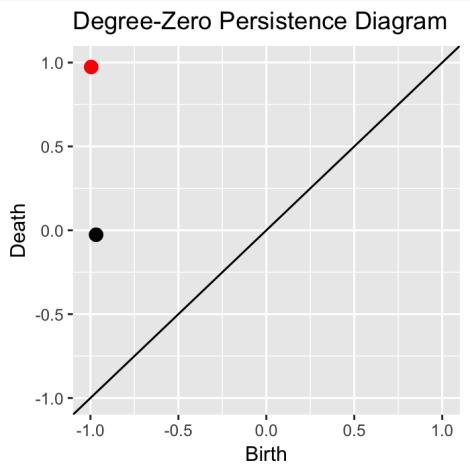


Main example



Main example





Return the length of the longest bar if it is born in the second quadrant, otherwise return 0.

The unstable function and how to stabilize it

Encoding the function

We encode this computation as a function, $h : \mathbb{R}^{3N} \to \mathbb{R}$.

Our sample corresponds to an input, $a \in \mathbb{R}^{3N}$.

Problem

$h(a) = 0$, but for nearby a', $h(a') \sim 2$.

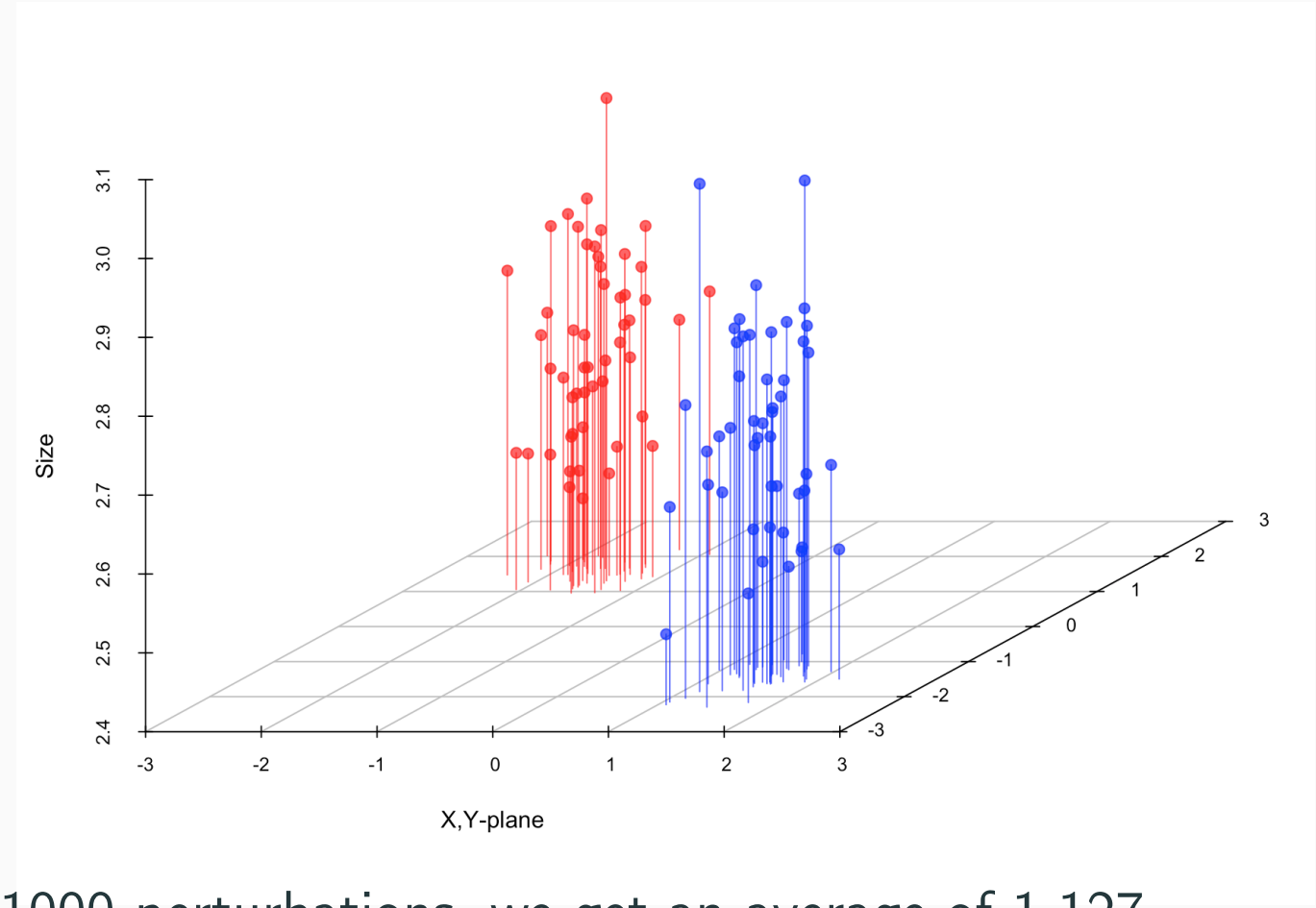
Solution

Randomly perturb the input $a \in \mathbb{R}^{3N}$, M times,

compute and average.

Main example

Here are the lengths of the longest bars and their birth locations from 100 perturbations of the input.



From 1000 perturbations, we get an average of 1.127.

Theory

Stability and Convolutions

Definition

We say that a function $g : \mathbb{R}^d \to \mathbb{R}$ is stable if it is Lipschitz.
That is, $\|g(x) - g(y)\| \leq C\|x - y\|$ for some constant C.

Definition

For $h, K : \mathbb{R}^d \to \mathbb{R}$, their convolution is

$$(h * K)(t) = \int_{\mathbb{R}^d} h(s)K(t-s)ds = \int_{\mathbb{R}^d} h(t-s)K(s)ds.$$

Kernel functions

We stabilize our unstable function h by convolving it with certain stable functions called kernels.

The triangular kernel

$$K(x) = \max(1 - \|x\|, 0)$$

The Epanechnikov kernel

$$K(x) = \max(1 - \|x\|^2, 0)$$

The Gaussian kernel

$$K(x) = e^{-\|x\|^2/2}$$

Stability Theorems

Theorem

If h is locally bounded, and K is the triangular kernel or the Epanechnikov kernel, then $h * K$ is locally Lipschitz.

Theorem

If h is bounded and K is the Gaussian kernel, then $h * K$ is Lipschitz.

Convolution and "perturb and average"

Procedure

- Start with an unstable $h : \mathbb{R}^d \to \mathbb{R}$ and observation $a \in \mathbb{R}^d$.
- Choose a kernel K. View K as a probability density.
- Sample $\varepsilon_1, \ldots, \varepsilon_M$ from K.
- Compute $\frac{1}{M} \sum_{i=1}^{M} h(a - \varepsilon_i)$.

Theorem

By the law of large numbers, this converges (almost surely) to $E[h(a - x)] = \int_{\mathbb{R}^d} h(a - x)K(x)dx = (h * K)(a)$.

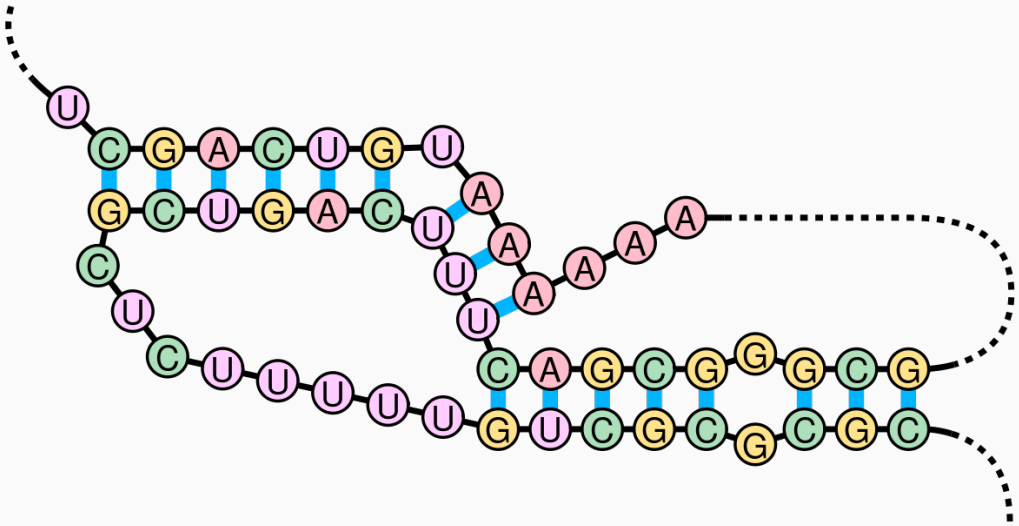
Stability of the convolution

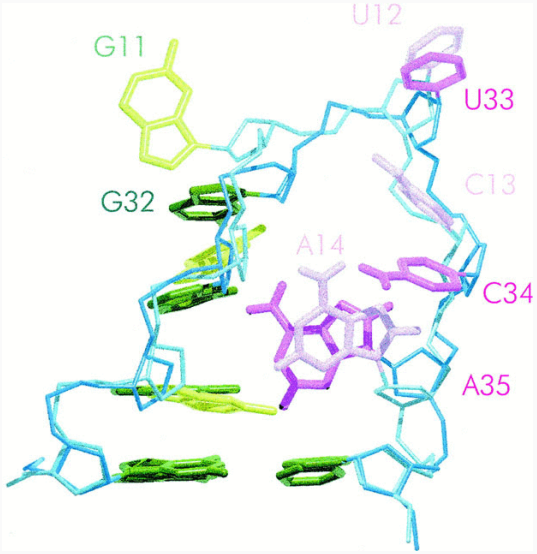
Theorem

The map $K \to h * K$ is stable
(as a map from $L^1(\mathbb{R}^d)$ to $L^\infty(\mathbb{R}^d)$).

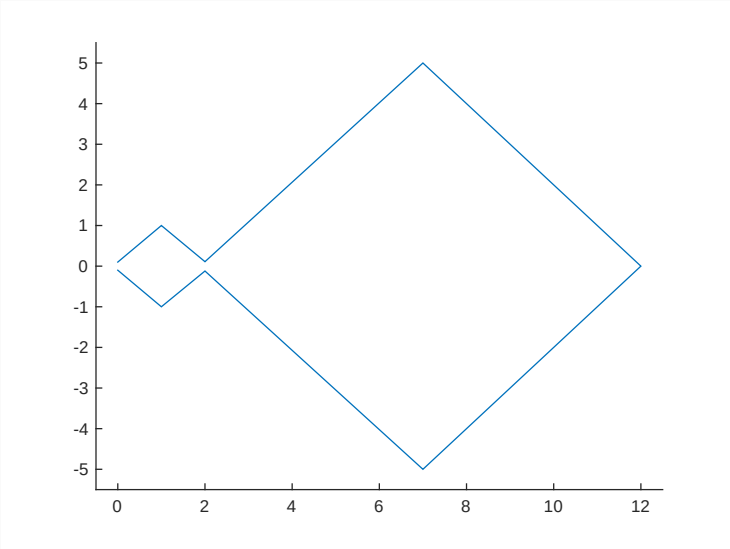
Other examples

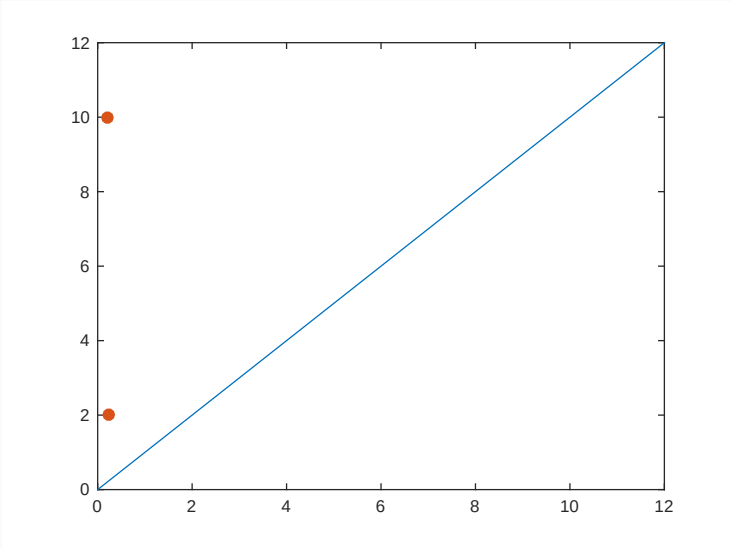
Detecting long-range pseudoknots in RNA



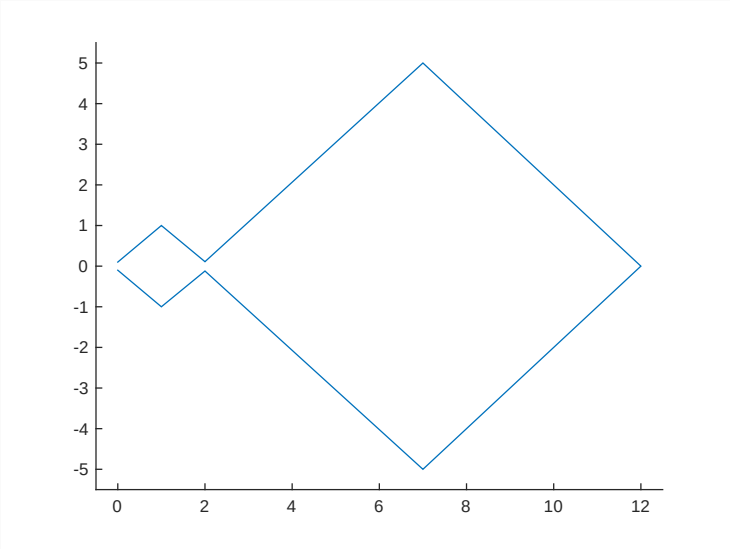


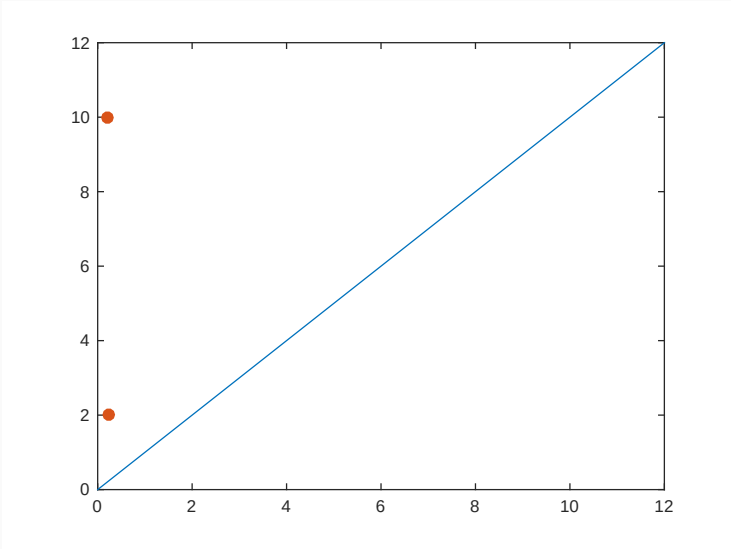
Edges responsible for degree-one persistent homology

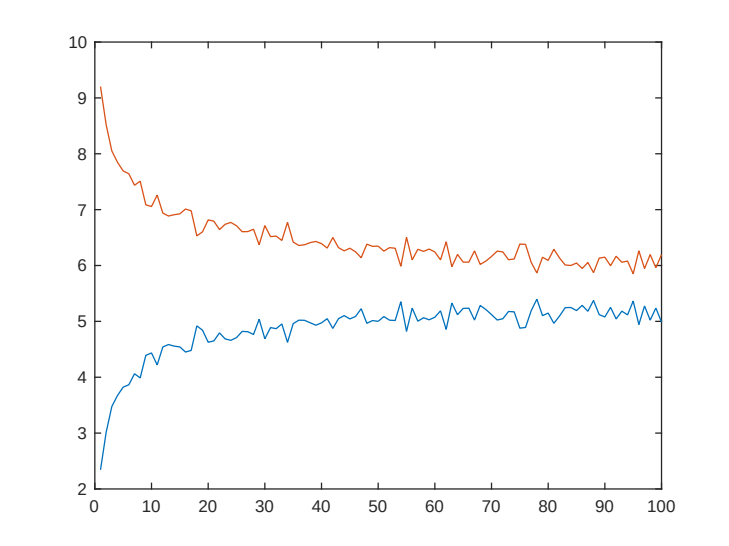




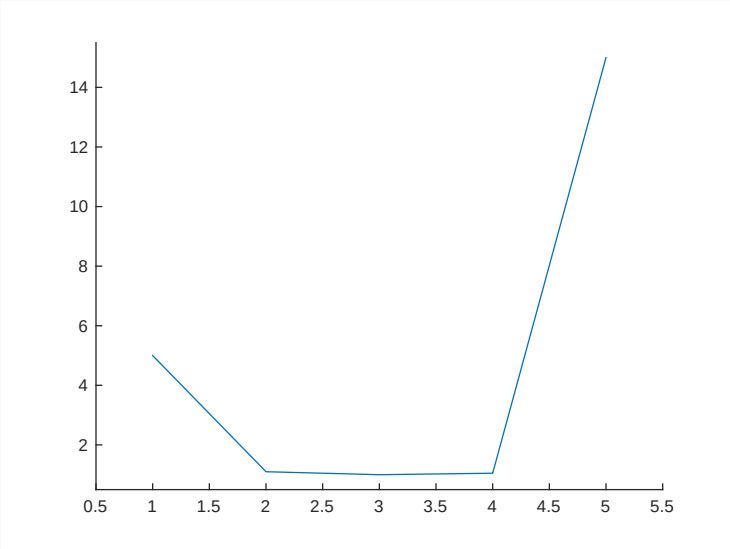
Edges responsible for degree-one persistent homology

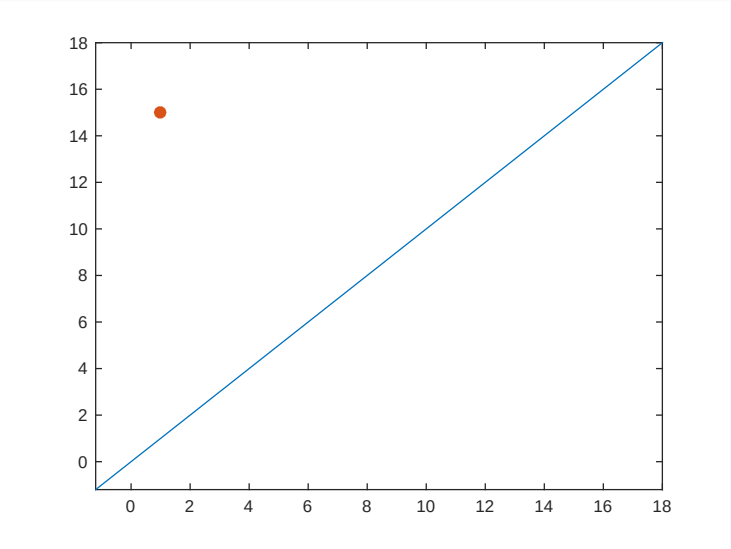




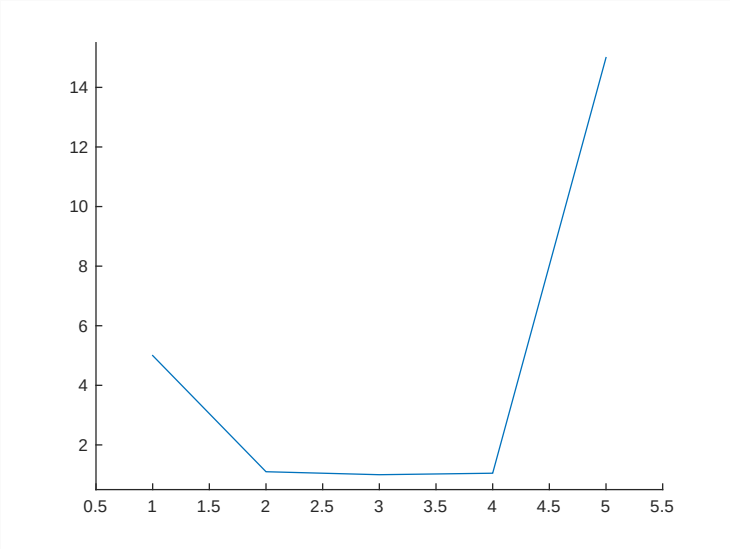


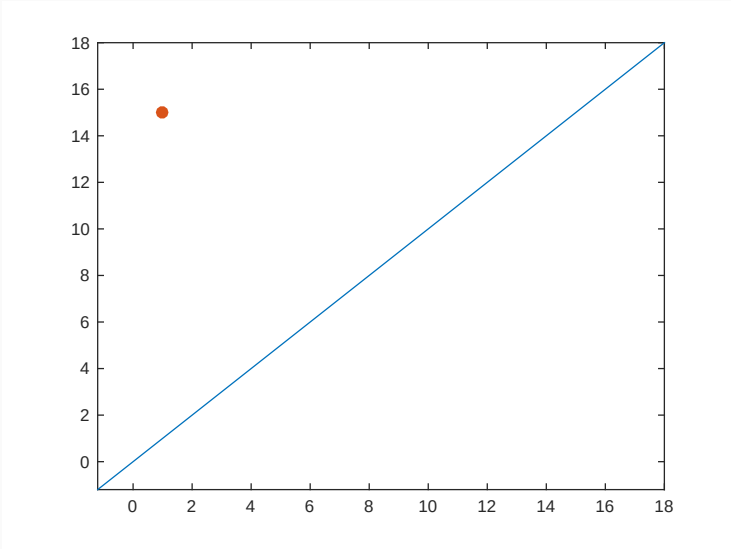
Spreading out a minimum

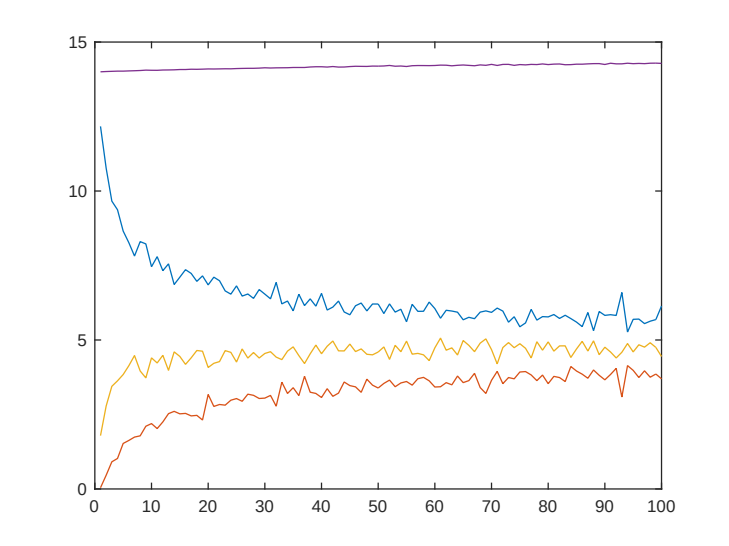




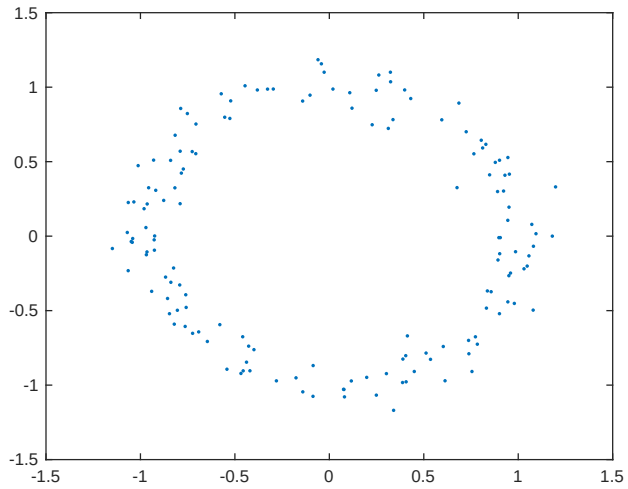
Spreading out a minimum

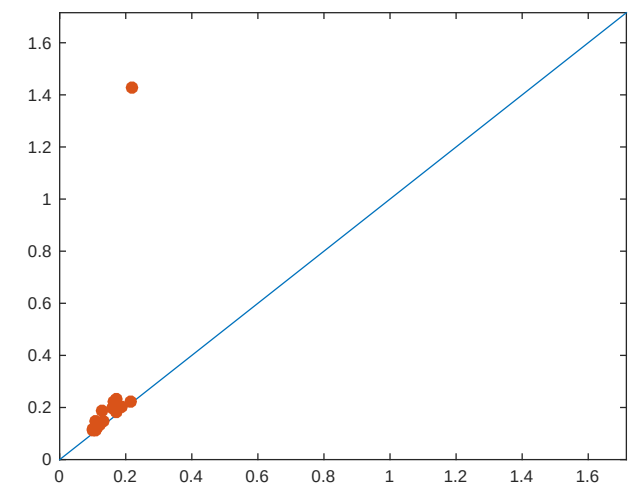


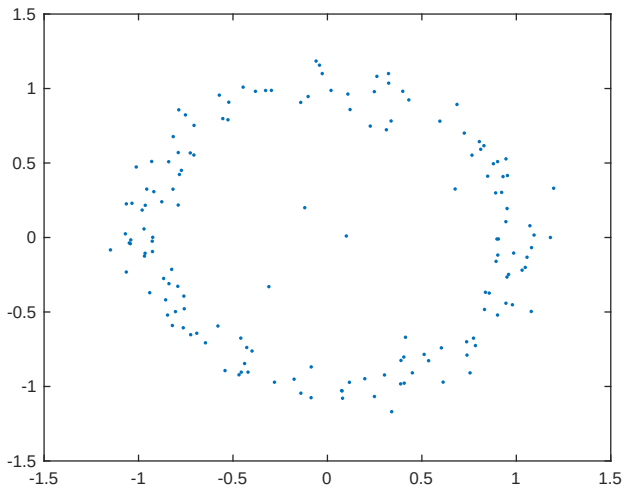


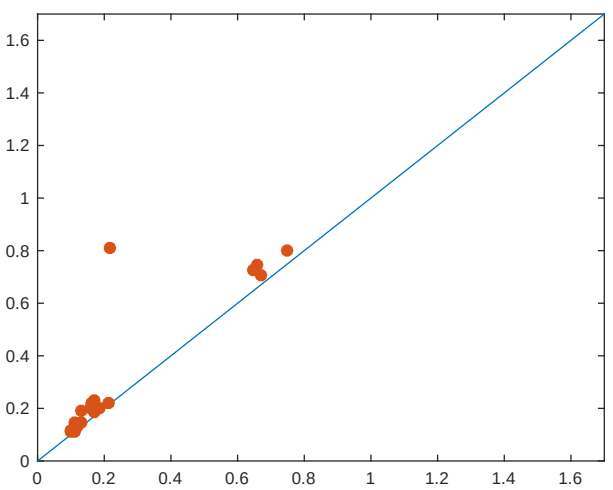


Density Threshold









Parameters used for pre-processing

Persistent homology computations may be unstable with respect to parameters used in pre-processing.

Suppose we have input data $a \in \mathbb{R}^d$ and parameters $b \in \mathbb{R}^e$.

We may consider our computation as $h : \mathbb{R}^d \to \mathbb{R}$, $h : \mathbb{R}^e \to \mathbb{R}$, or $h : \mathbb{R}^{d+e} \to \mathbb{R}$.

In each case, $h * K$ is stable.

Conclusion

Conclusion

- Practitioners would like to point to a spot in their data which is responsible for significant persistent homology features.
- The corresponding critical points, birth simplices or generating cycles are unstable.
- Parameters used in pre-processing are also unstable.
- A simple perturb and average procedure provides stability.
- This stability is obtained by convolving with a stable kernel.
- For more details see our preprint on the arXiv.

Thank you!