Robust Topological Inference

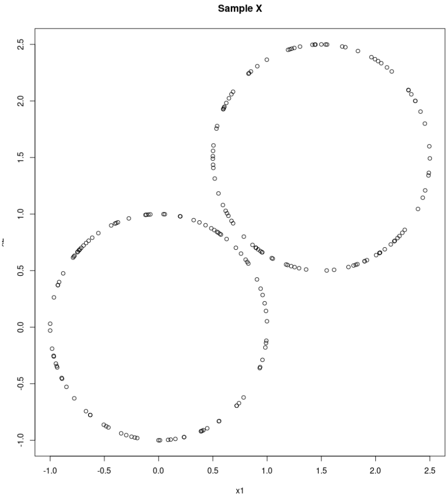
Tom Norman

July 1, 2020

Overview

Distance Function Offset

Using a distance function to discover the persistent homology of the data



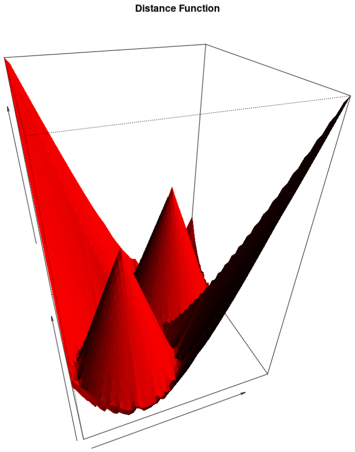


Figure 1: 2 circles and L2 distance of every point in $\mathbb{R}^2$ to the nearest sample

- Given data $X = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^d$
- $\forall x \in \mathbb{R}^d$ compute $d_X(x) = \min_{x_i \in X} \|x_i - x\|_2$
- Run persistent homology for fucntions to get filtarations of sublevel sets $L_t = \{x : d_X(x) \leq t\}$

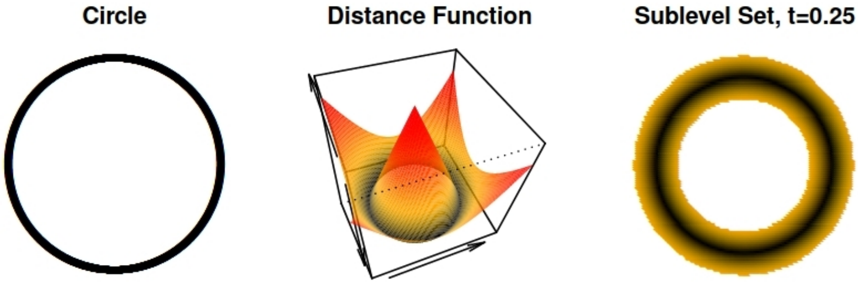


Figure 2: A circle, its L2 distance function and the sublevel set for t=0.25

Persistence Diagram

A way to visualize persistent homology

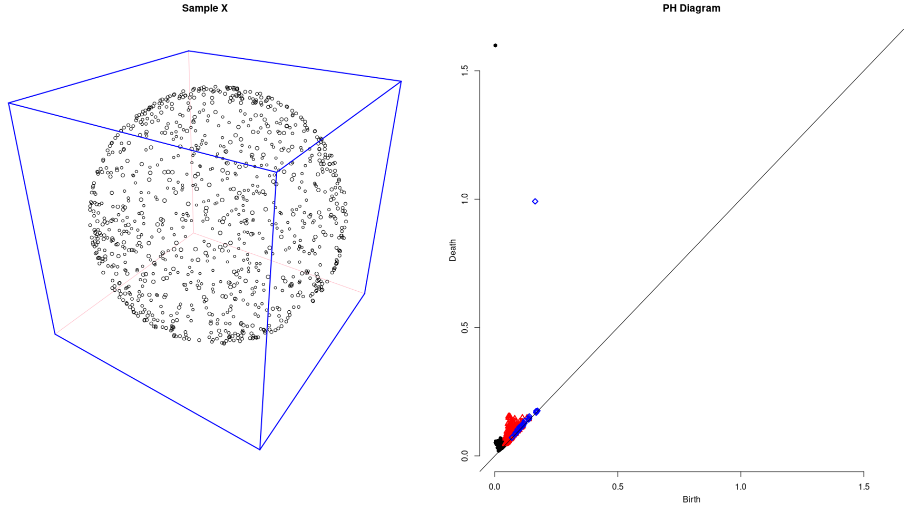


Figure 3: Persistence diagram of the unit sphere

- For $t : 0 \to \infty$ find birth and death (denoted b_i and d_i) of each homology feature (connected component, hole, void, etc.)
- Create 2D graph from those b_i, d_i
- Note that $b_i \leq d_i$

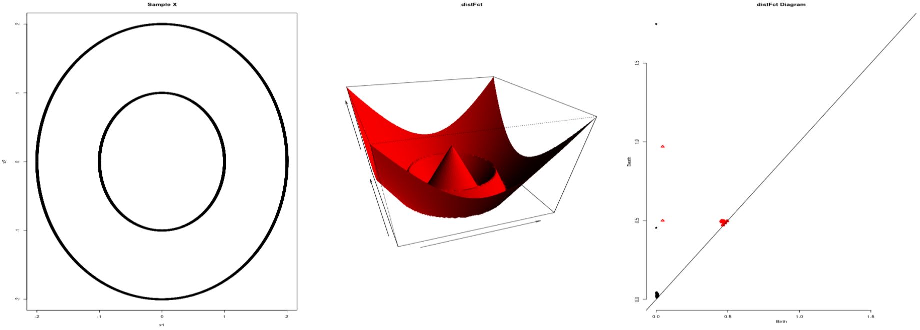


Figure 4: Persistence diagram of 2 circles

Outliers (Motivation)

L2 distance is very sensitive to outliers

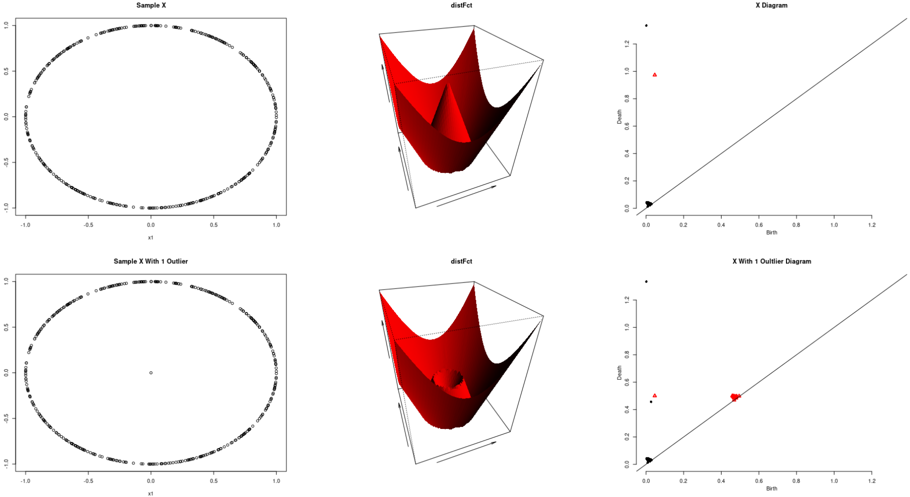


Figure 5: 400 samples of the unit circle, the lower plot with 1 extra outlier

Distance to (Probability) Measure

Smooth function which is robust to outliers and noise

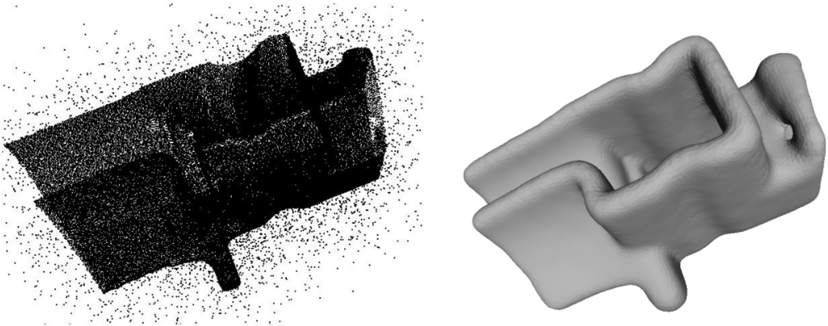


Figure 6: Left: Point cloud of a mechanical part with noise, Right: Reconstruction using DTM

- Define $G_x(t) = \mathbb{P}(\|X - x\| \leq t)$- probability of a ball with radius t around point x
- Given $m_0 \in (0,1)$ (smoothing parameter)

$$DTM_X^{m_0}(x) = \sqrt{\frac{1}{m_0}\int_0^{m_0}\left(G_x^{-1}(u)\right)^2 du}$$

- We'll assign each data point $\frac{1}{n}$ probability mass and take $m_0 = \frac{k}{n}$

$$DTM_X^{m_0}(x) = d_X^k(x) = \sqrt{\frac{1}{k}\sum_{x_i \in kNN_X(x)} \|x_i - x\|^2}$$

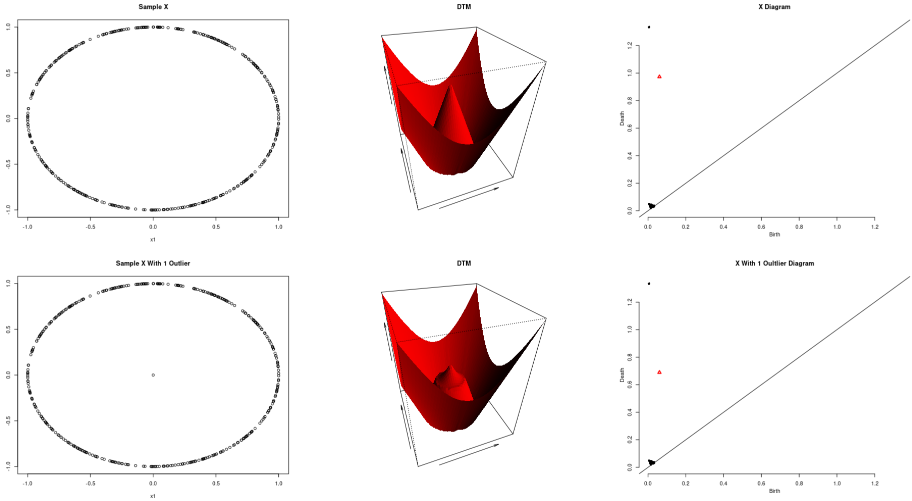


Figure 7: 400 samples, k=2

Bottleneck Distance

Distance between 2 persistence diagrams

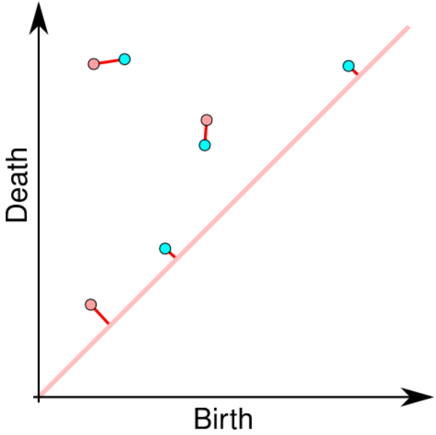


Figure 8: The bottleneck distance is the longest red edge

- Given 2 persistent diagrams D_1, D_2 (including the diagonal, where birth==death) denote bottleneck distance as
$$W_\infty(D_1, D_2) = \min_{g:D_1 \to D2} \sup_{z \in D_1} \|z - g(z)\|_\infty$$
(g is a bijection)
- In words: the maximum distance between the features of the 2 diagrams, after minimizing over all possible pairings of the features (including the diagonal)

Significance of Features Using Bootstrapping

Quantify the confidence that the homology feature came from the underlying shape

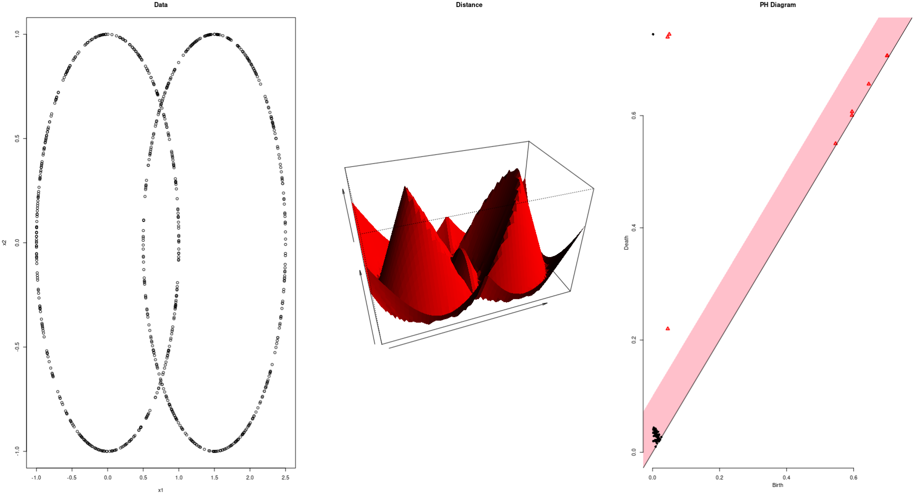


Figure 9: 2 overlapping unit circles. Only 3 holes are statistically significant.

- Given X and k compute d_X^k
- Sample n points at random from X with replacement (denoted X_i^*), and compute $\theta_i^* = \sqrt{n}\|d_X^k(x) - d_{X_i^*}^k(x)\|_\infty \propto W_\infty\left(D_X, D_{X_i^*}\right)$
- Repeat last step B times
- Given α compute $t_\alpha = \min_t \left\{\sum_{i=1}^{B} 1\{\theta_i^* \geq t\} \leq \alpha B\right\}$:
  minimum t for which there are at most αB bigger θ_i^*'s.
- $\forall$ feature i:

  if $|b_i - d_i| > \frac{2t_\alpha}{\sqrt{n}}$, then the feature is α-significant

Choosing Smoothing Parameter

Choose m_0 that maximizes the total amount of significant information

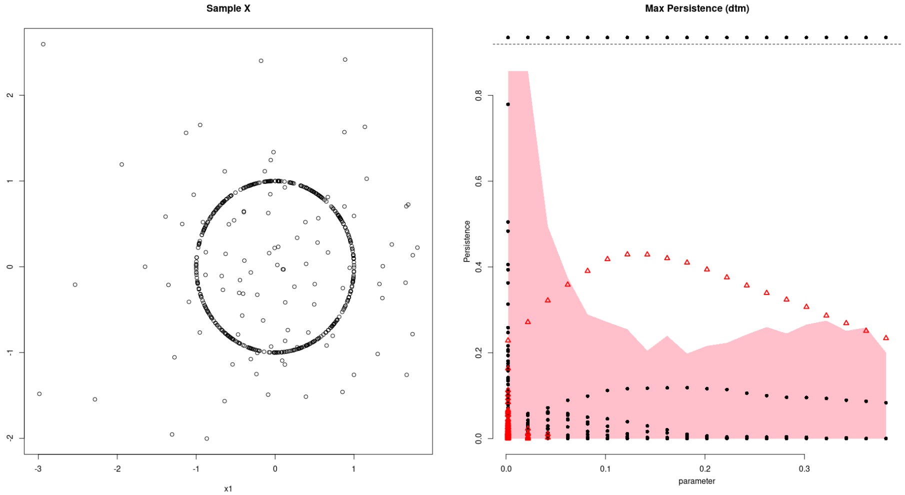


Figure 10: 400 samples from unit circle + 100 samples from normal distribution, k = [1:10:191]

- Let $\ell_i(m_0)$ be the lifetime of feature i with smoothing parameter m_0
- Given $\frac{t_\alpha}{\sqrt{n}}$, define:

$$N(m_0) = \#\left\{i : \ell_i(m_0) > \frac{2t_\alpha}{\sqrt{n}}\right\}$$

(Number of α-significant features for given m_0)

and:

$$S(m_0) = \sum_i \left[\ell_i(m_0) - \frac{2t_\alpha}{\sqrt{n}}\right]_+$$

(Sum of distances from α-significance for given m_0)

- $m_{opt} = \arg\max_{m_0} N(m_0)$ **or** $\arg\max_{m_0} S(m_0)$

Further Work

Would like to test DTM vs. L2 distance on real world data:

- Point cloud MNIST
- Find safe neighborhoods in Vancouver and Boston
- Movies - what is uncommon length, rating, etc for each genere
- Ideas?

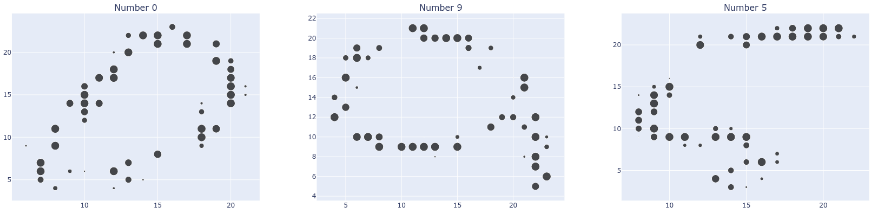


Figure 11: Point cloud MNIST