SAMSI, February 5, 2014

Statistical Inference for Persistence Diagrams

Fabrizio Lecci

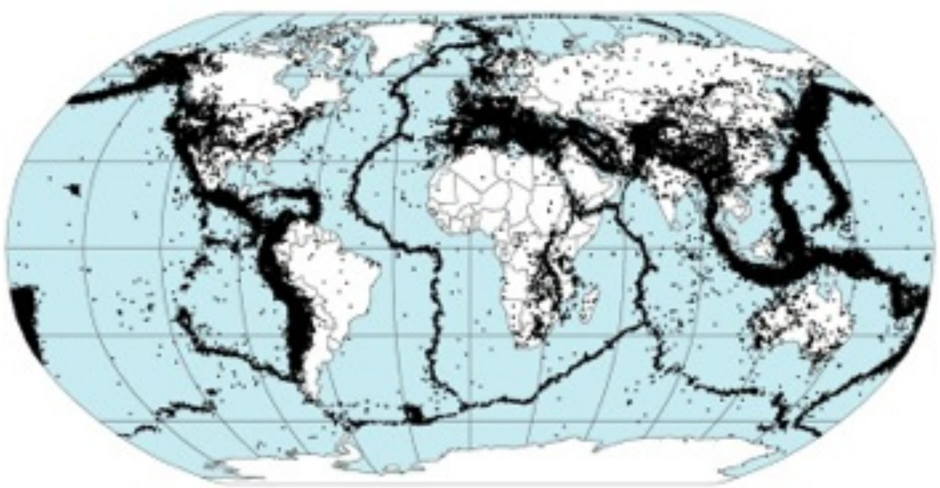
CMU Topstat Group

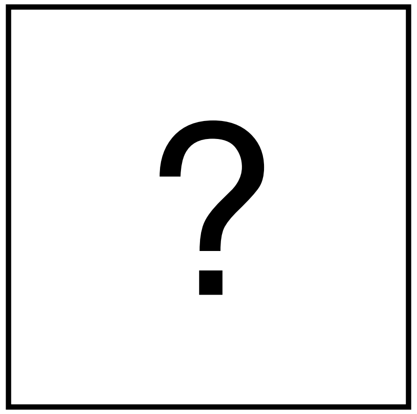
Carnegie Mellon University

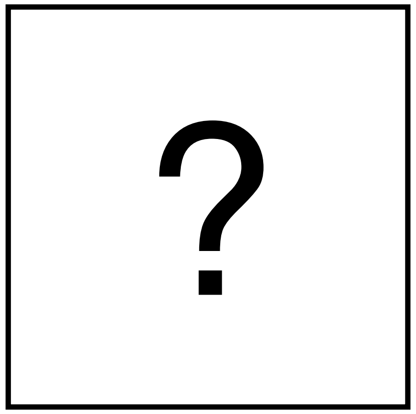
(joint work with S. Balakrishnan, B.T. Fasy, A. Rinaldo, A. Singh, L. Wasserman)

Inference and Topological Data Analysis

Data	→	Topological Summary	→	Inference

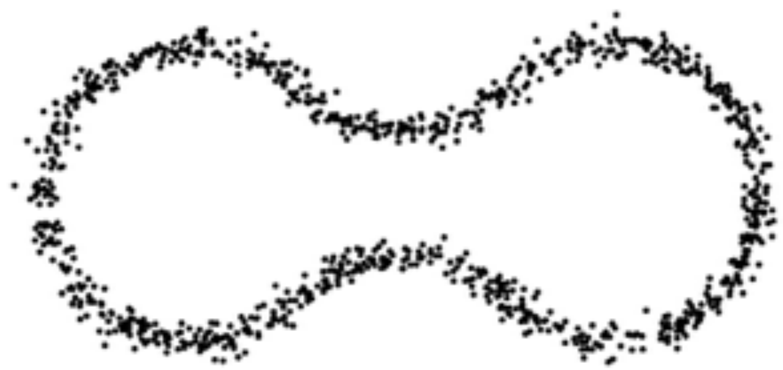




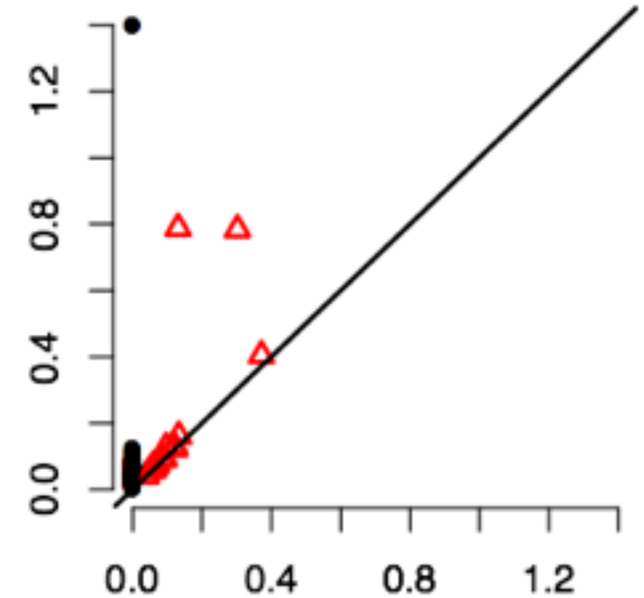


Objective: distinguish topological signal from topological noise

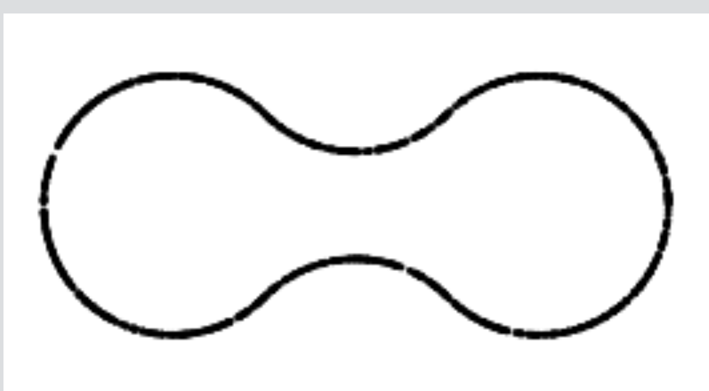
Topological Signal Vs Topological Noise



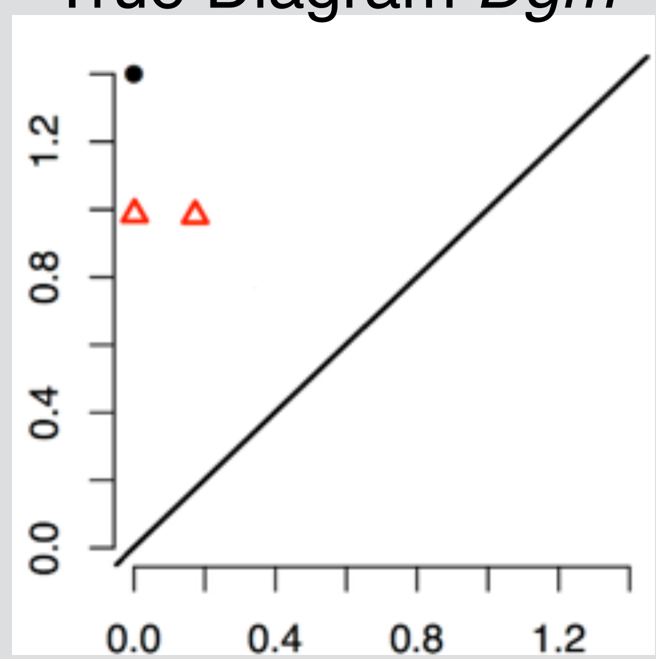
Empirical Diagram $\widehat{Dgm}_n$



Unobserved



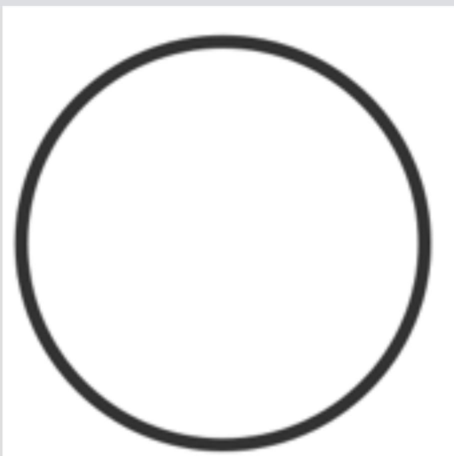
True Diagram *Dgm*



Persistent Homology of the distance function

Unobserved

manifold ***M***



$X_1, \dots, X_n \sim P$

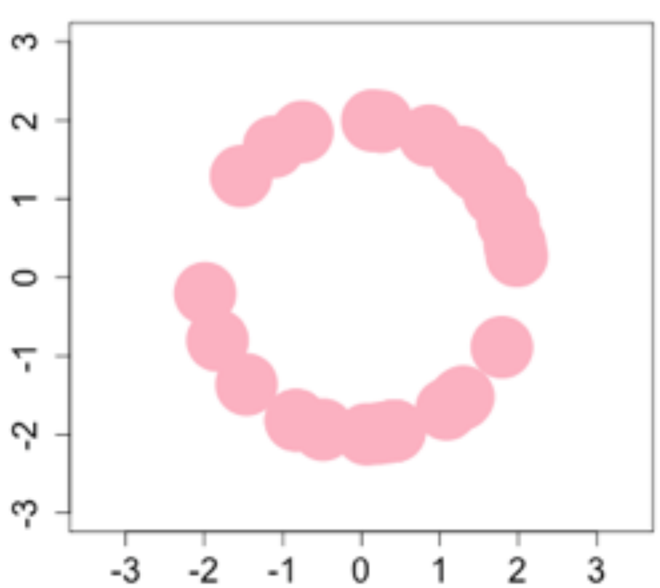
Distance function

$$d_S(x) = \inf_{y \text{ in Sample}} \| x - y \|$$

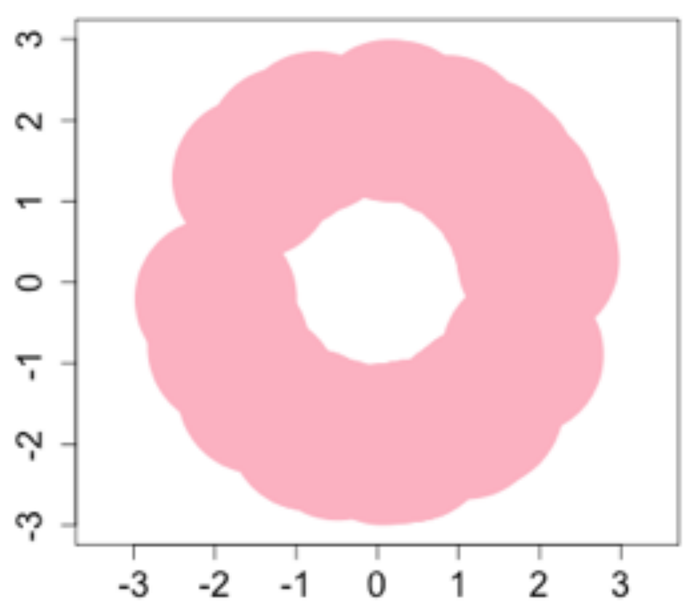
Sub-level set

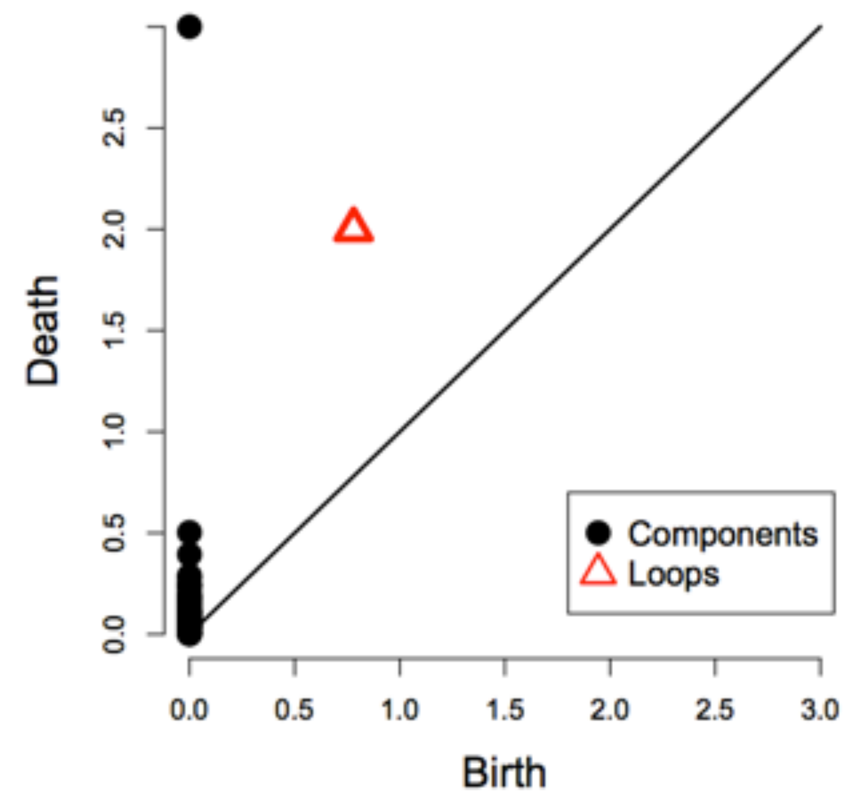
$$L_{\mathbf{t}} = \{x : d_S(x) < \mathbf{t}\}$$

$$L_{\mathbf{0.5}} = \{x : d_S(x) < \mathbf{0.5}\}$$



$$L_{\mathbf{1}} = \{x : d_S(x) < \mathbf{1}\}$$

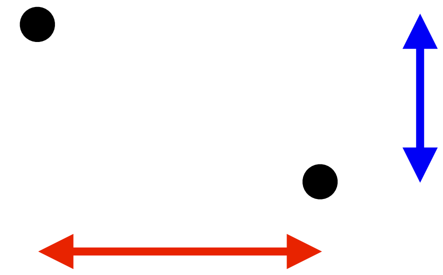




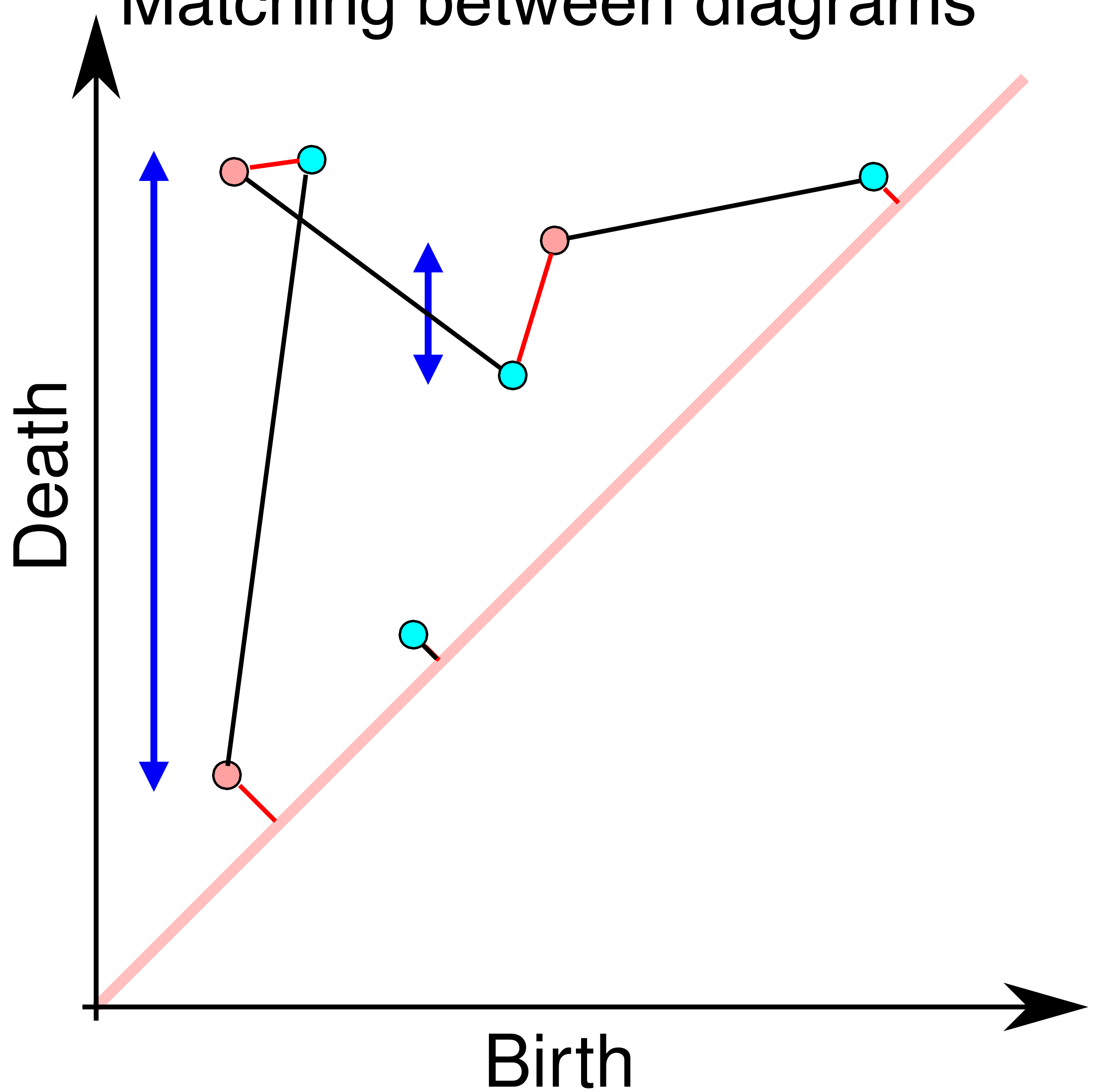
Distance between Persistence Diagrams

Distance between points

$$d_\infty(a,b) = \max\{|a_x - b_x|, |a_y - b_y|\}$$



Matching between diagrams



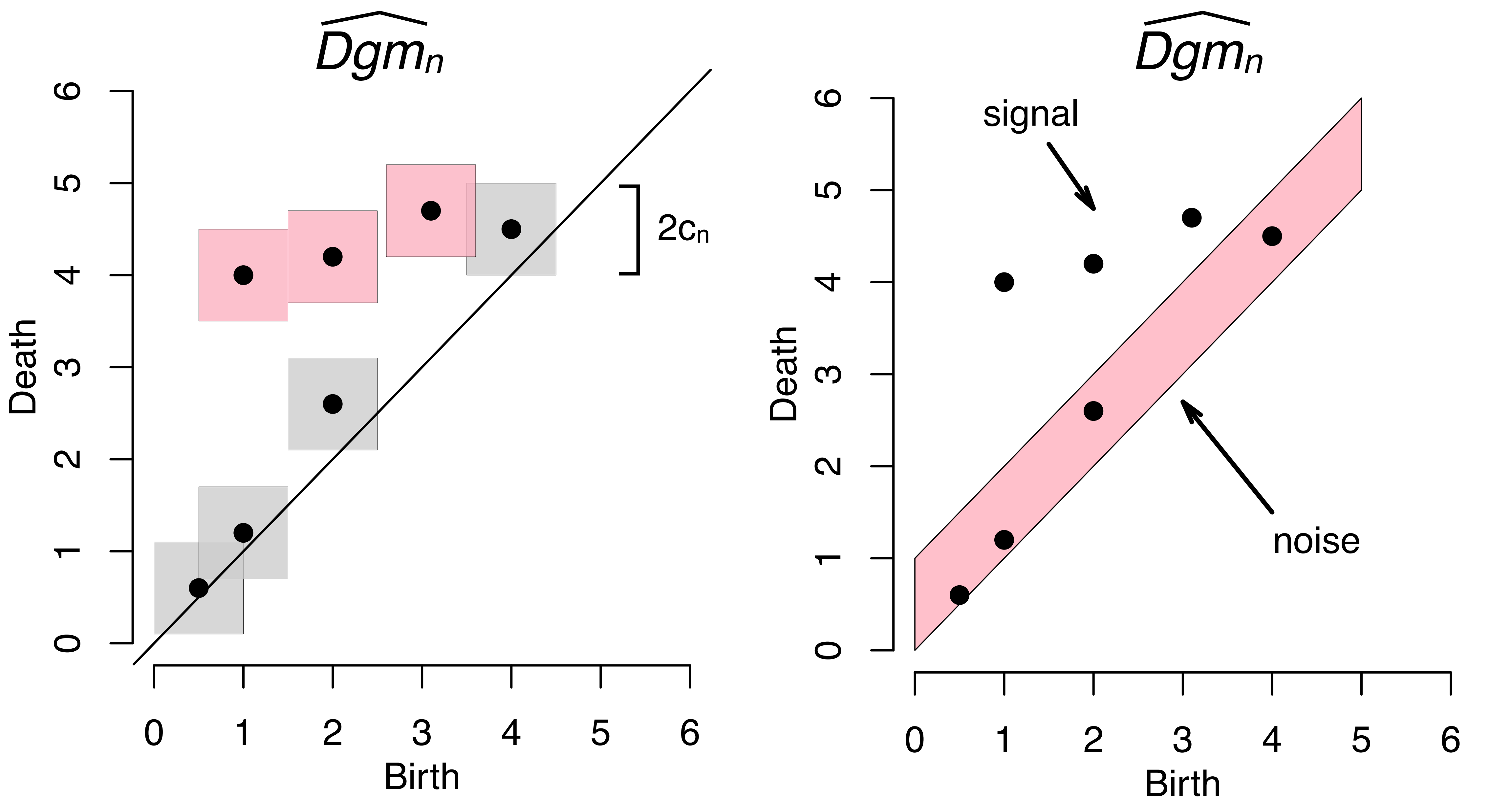
Bottleneck distance

$$W_\infty(D_1, D_2) = \min_{M \in \mathcal{M}(D_1, D_2)} \max_{(a,b)\in M} d_\infty(a,b)$$

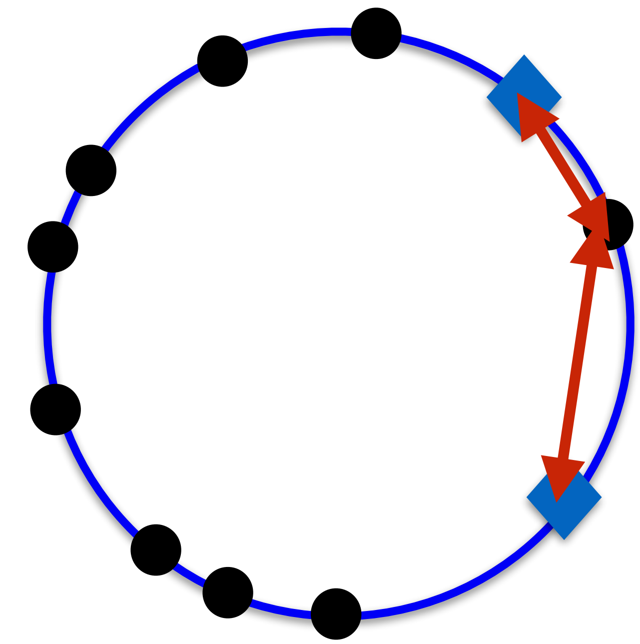
Objective: Confidence sets for Persistence Diagrams

A 1-α confidence set for *Dgm* is an interval $[0,\ c_n]$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(W_\infty(\widehat{Dgm}_n, Dgm) \in [0, c_n]\right) \geq 1-\alpha$$



Confidence sets for Persistence Diagrams: general strategy



Hausdorff distance

$$\mathcal{H}(\mathbf{M}\text{anifold}, \mathbf{S}\text{ample}) = \max_{x \in \mathbf{M}} \min_{y \in \mathbf{S}} \|x - y\|$$

Stability Theorem [Cohen-Steiner et al. 2005]

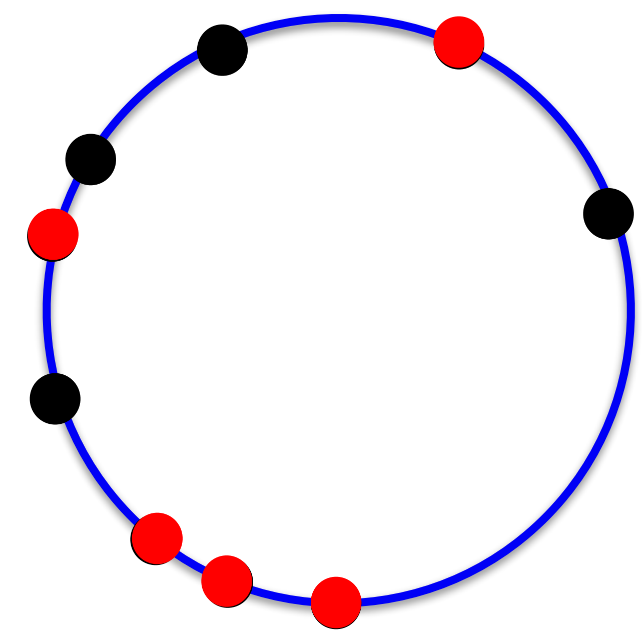
$$W_\infty(Dgm, \widehat{Dgm}_n) \leq \mathcal{H}(\mathbf{M}\text{anifold}, \mathbf{S}\text{ample})$$

Idea: construct a confidence interval for $\mathcal{H}(\mathbf{M}, \mathbf{S})$ and use it for $W_\infty(Dgm, \widehat{Dgm}_n)$

$$\mathbb{P}\left(W_\infty(Dgm, \widehat{Dgm}_n) > c_n\right) \leq \mathbb{P}\left(\mathcal{H}(\mathbf{M}, \mathbf{S}) > c_n\right) \leq \alpha$$

Subsampling Algorithm

- Sample n points from the manifold
- Draw N subsamples of size b: $S_b^1, \ldots, S_b^N$
- Let $T_j = \mathcal{H}(\mathbf{S}, \mathbf{S_b^j}) \quad \text{for} \quad j = 1, \ldots, N$
- Define $L(t) = \frac{1}{N}\sum_{j=1}^{N} I(T_j > t)$
- Let $c_n = 2L^{-1}(\alpha)$



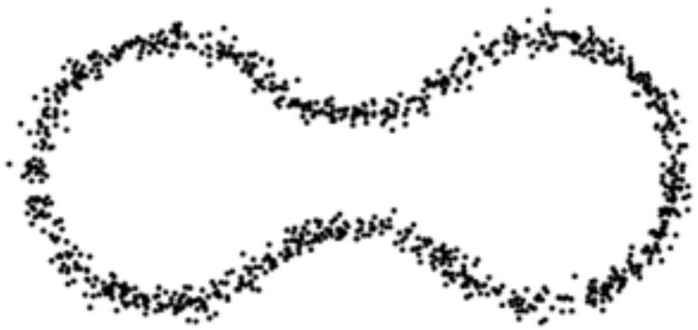
Theorem [Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, Singh, 2013]

$$\mathbb{P}\left(W_\infty(Dgm, \widehat{Dgm}_n) > c_n\right) \leq \mathbb{P}\left(\mathcal{H}(\mathbf{M},\mathbf{S}) > c_n\right) \leq \alpha + O\left(\frac{b}{n}\right)^{1/4}$$

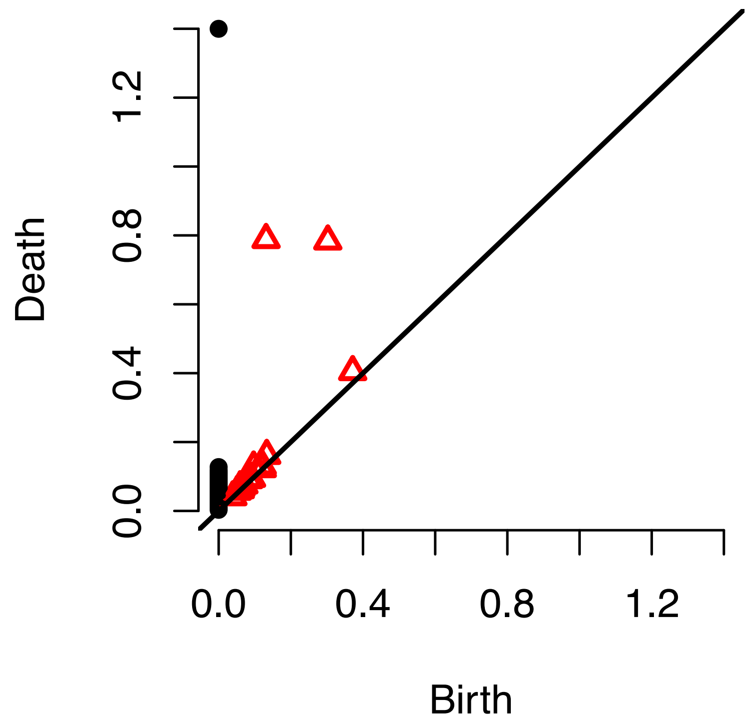
Stability theorem

Example: the Cassini Curve

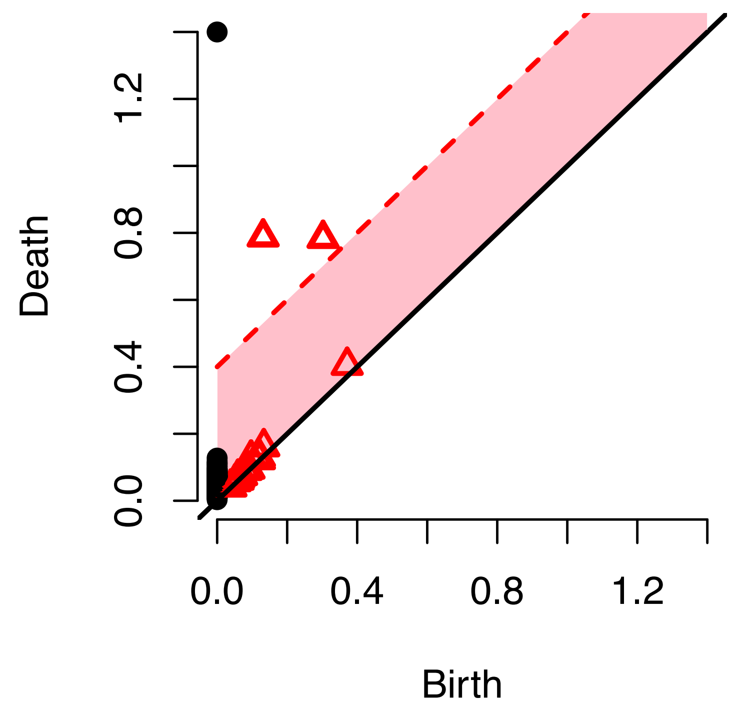
Sample



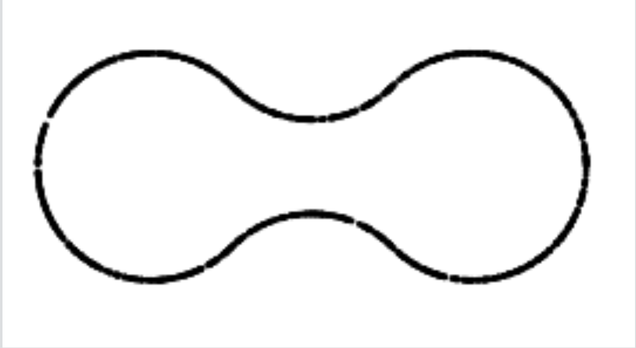
Empirical diagram $\widehat{Dgm_n}$



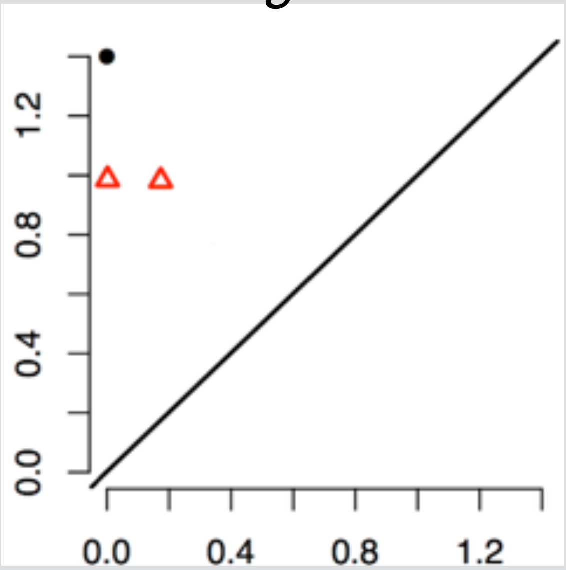
95% Confidence band



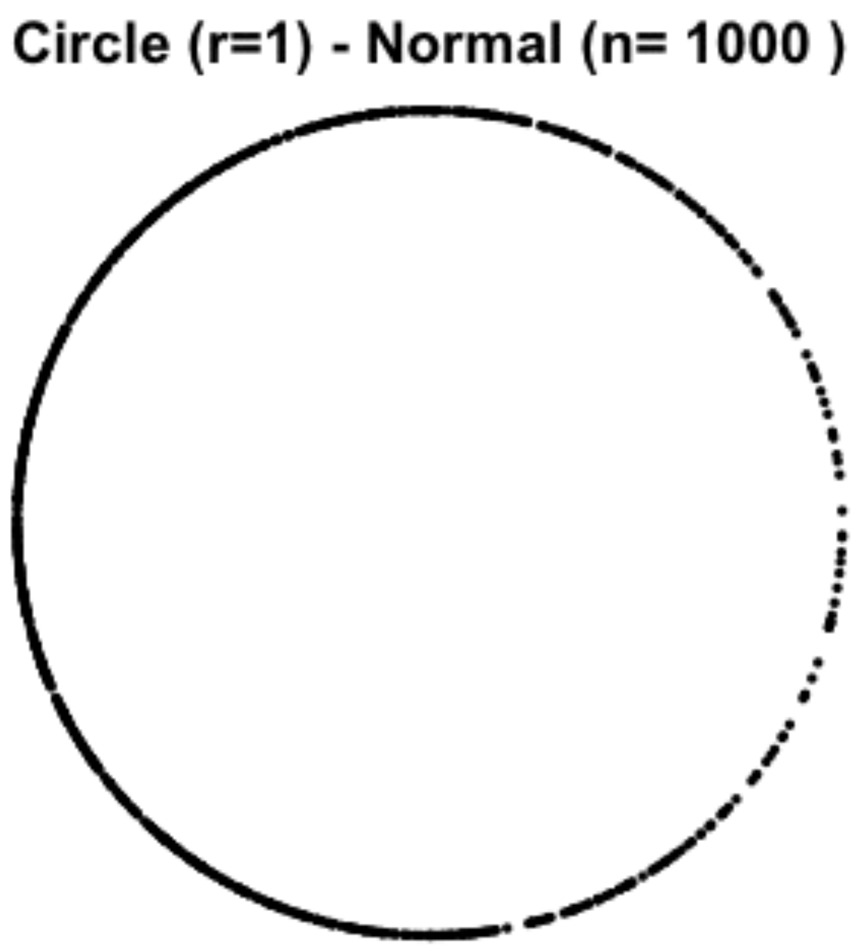
Unobserved

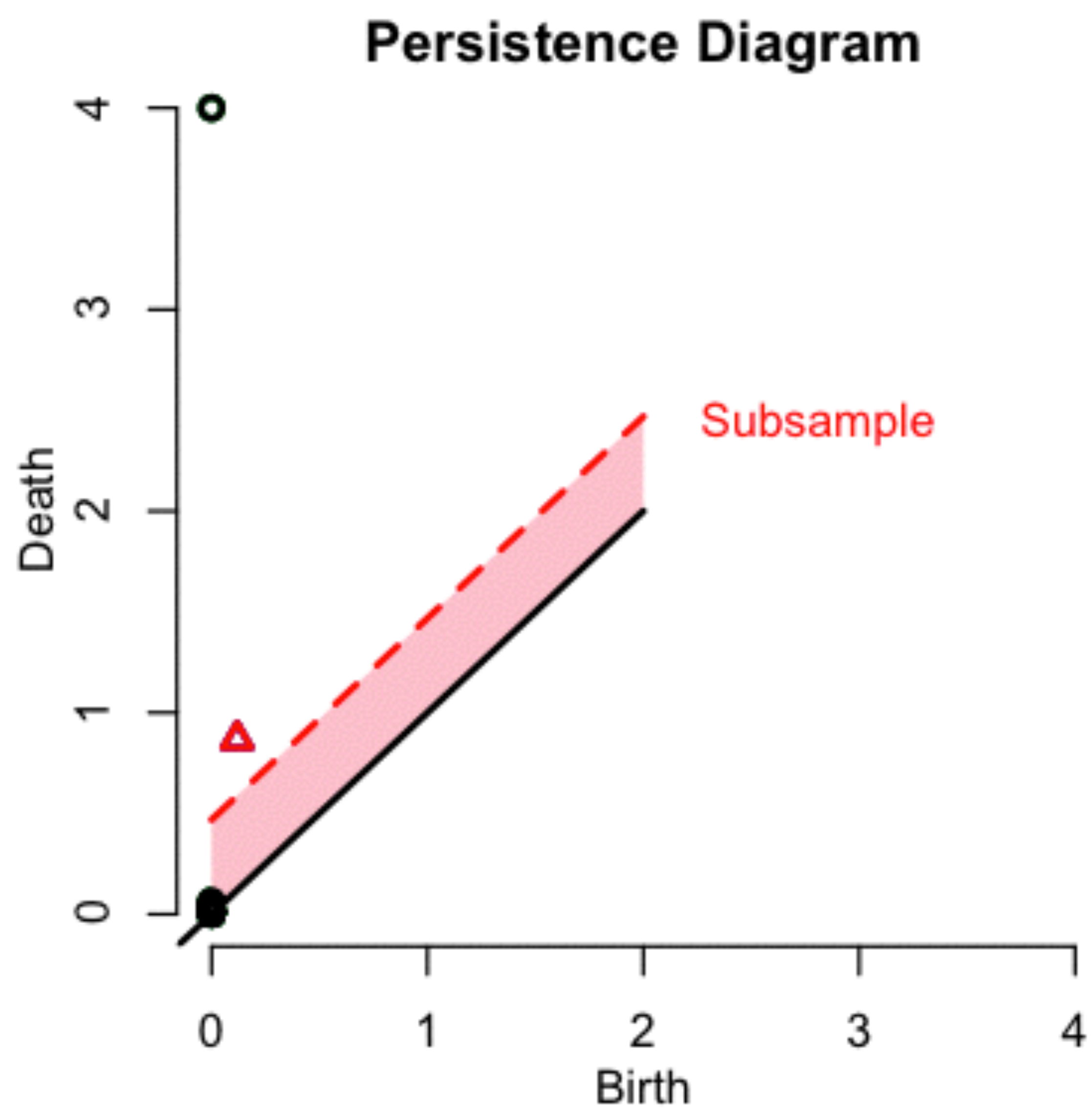


Dgm



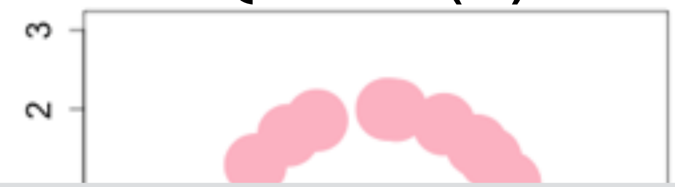
Example: the Circle





A different approach: Density Persistence Diagrams

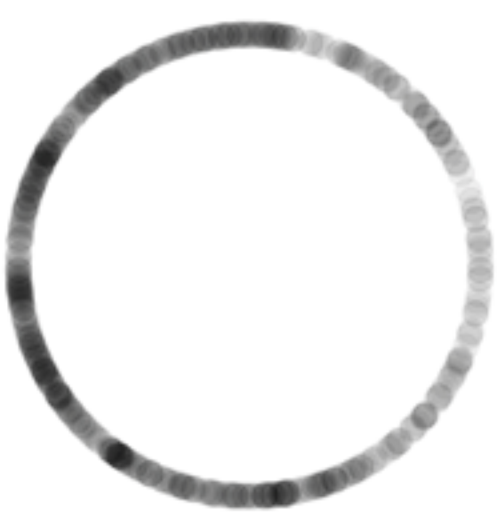
$L_t = \{x : d_S(x) < t\}$



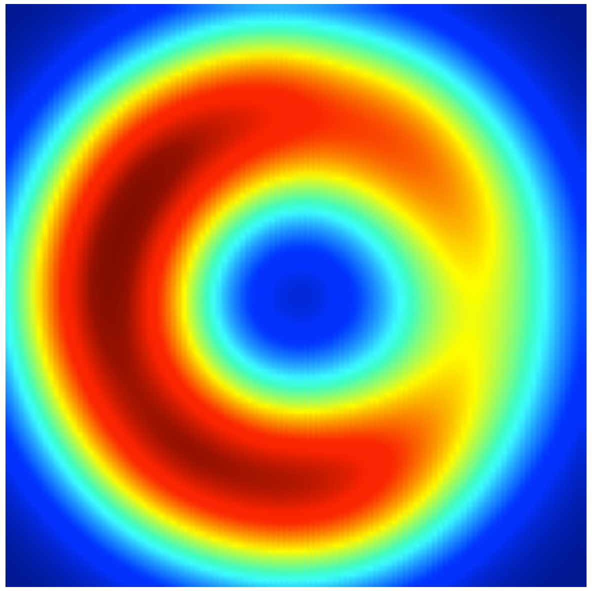
Instead of the distance fun

$$\hat{p}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^D} K\left(\frac{\|x - X_i\|}{h} \right)$$

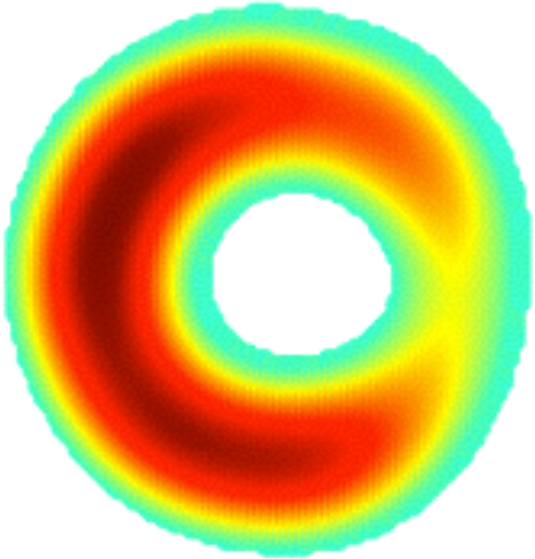
.. we can consider a density estimator:



Kernel Density estimator $\hat{p}_h(x)$



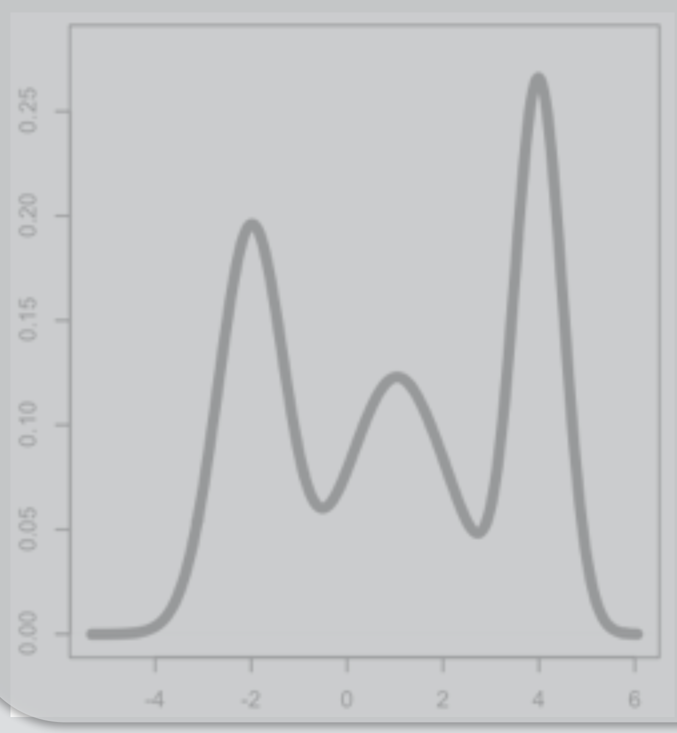
$U_{0.7} = \{x : \hat{p}_h(x) > 0.7\}$



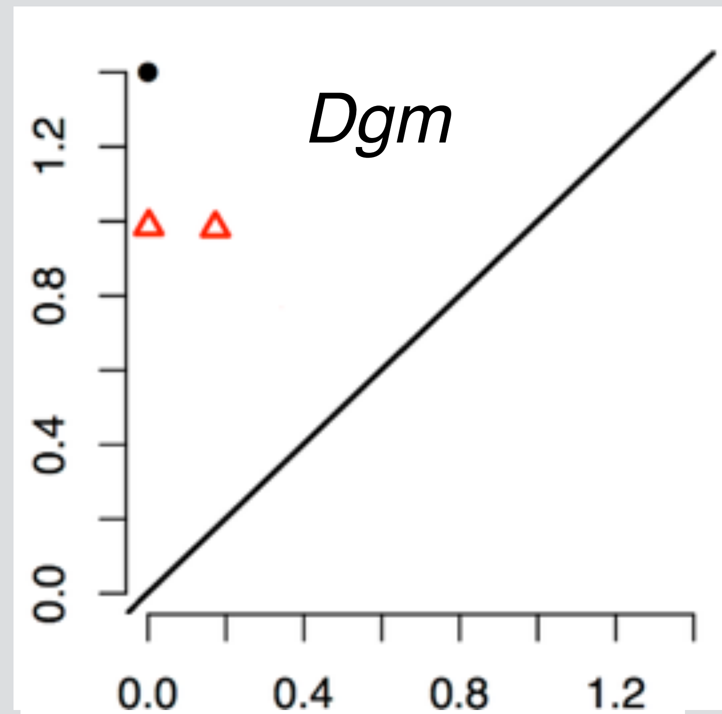
Demo

Confidence sets for Diagrams: general strategy

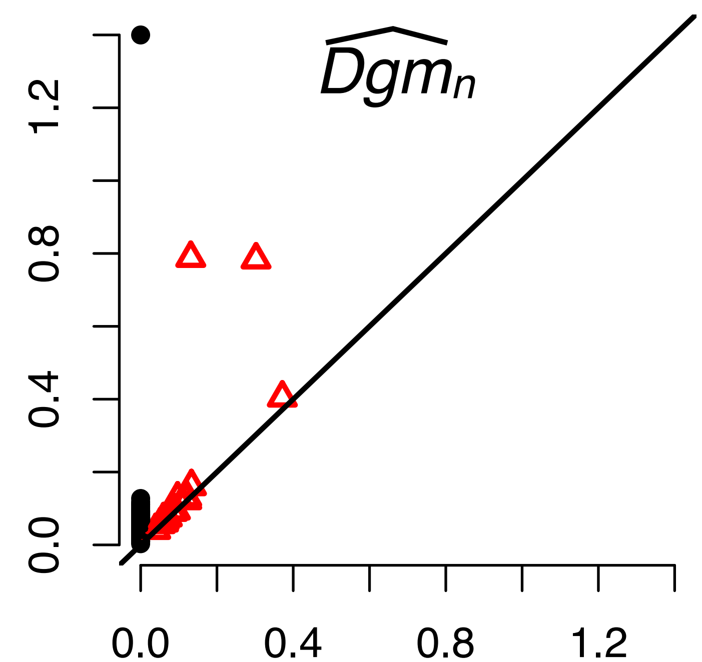
p



$p_h = E[\hat{p}_h]$



$\hat{p}_h$



Stability Theorem (Cohen-Steiner et al. 2005)

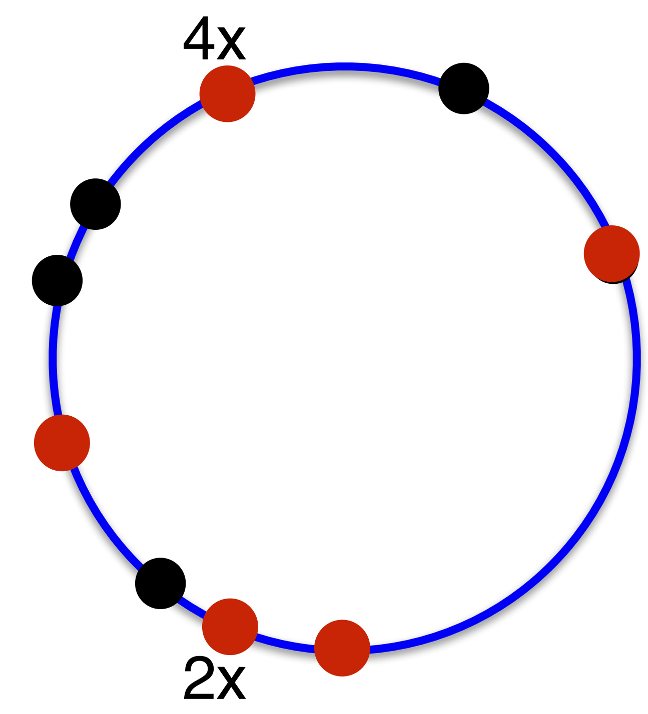
$$W_\infty(Dgm, \widehat{Dgm}_n) \leq \|p_h - \hat{p}_h\|_\infty$$

$$\mathbb{P}\left(W_\infty(Dgm, \widehat{Dgm}_n) > c_n\right) \leq \mathbb{P}\left(\|p_h - \hat{p}_h\|_\infty > c_n\right) = \alpha$$

Confidence Intervals for $\|p_h - \hat{p}_h\|_\infty$ using bootstrap

Bootstrap

- Sample n points from the manifold
- Draw B subsamples of size n: $S_n^1, \ldots, S_n^B$
- Let $T_j = \sqrt{nh^D}\|\hat{p}_h - \hat{p}_h^j\|_\infty$
- Define $Z_\alpha = \inf\left\{z : \frac{1}{B}\sum_{j=1}^{B} I(T_j > z) \leq \alpha\right\}$

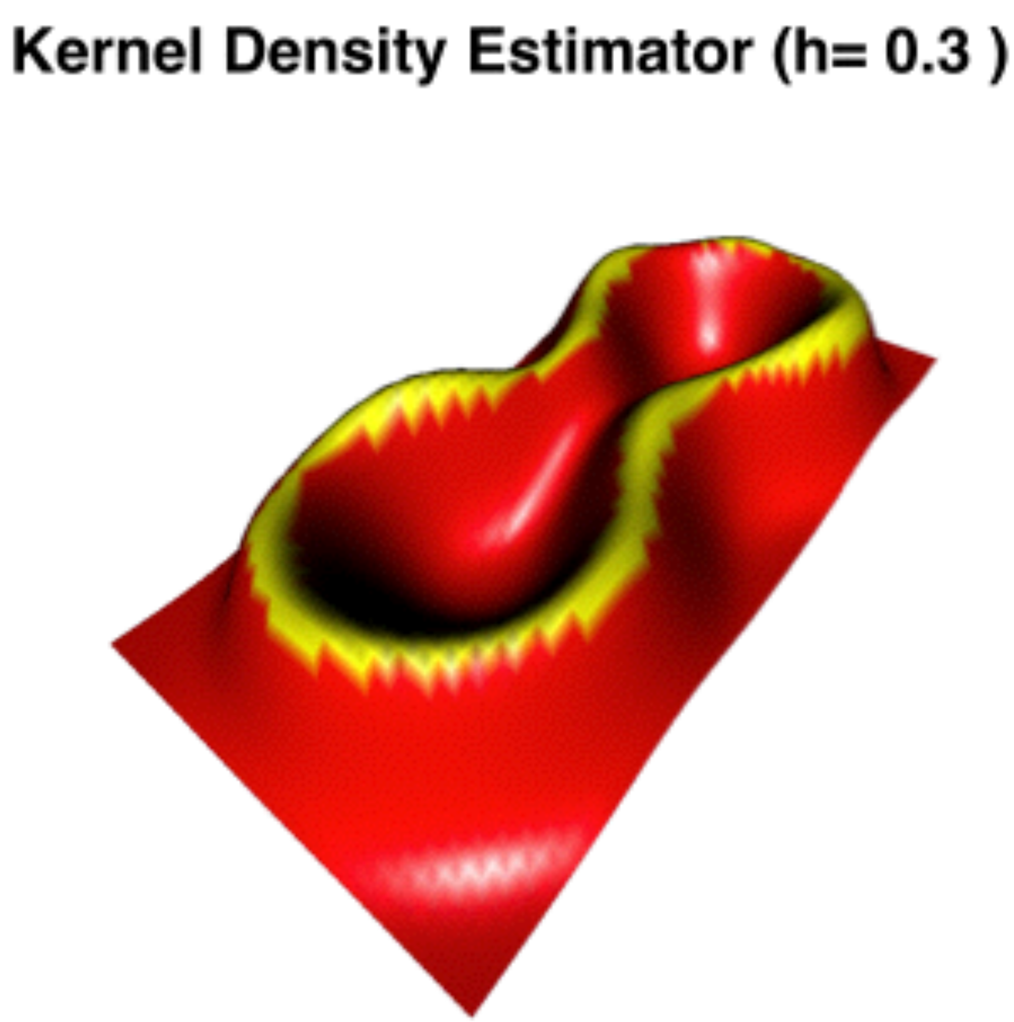


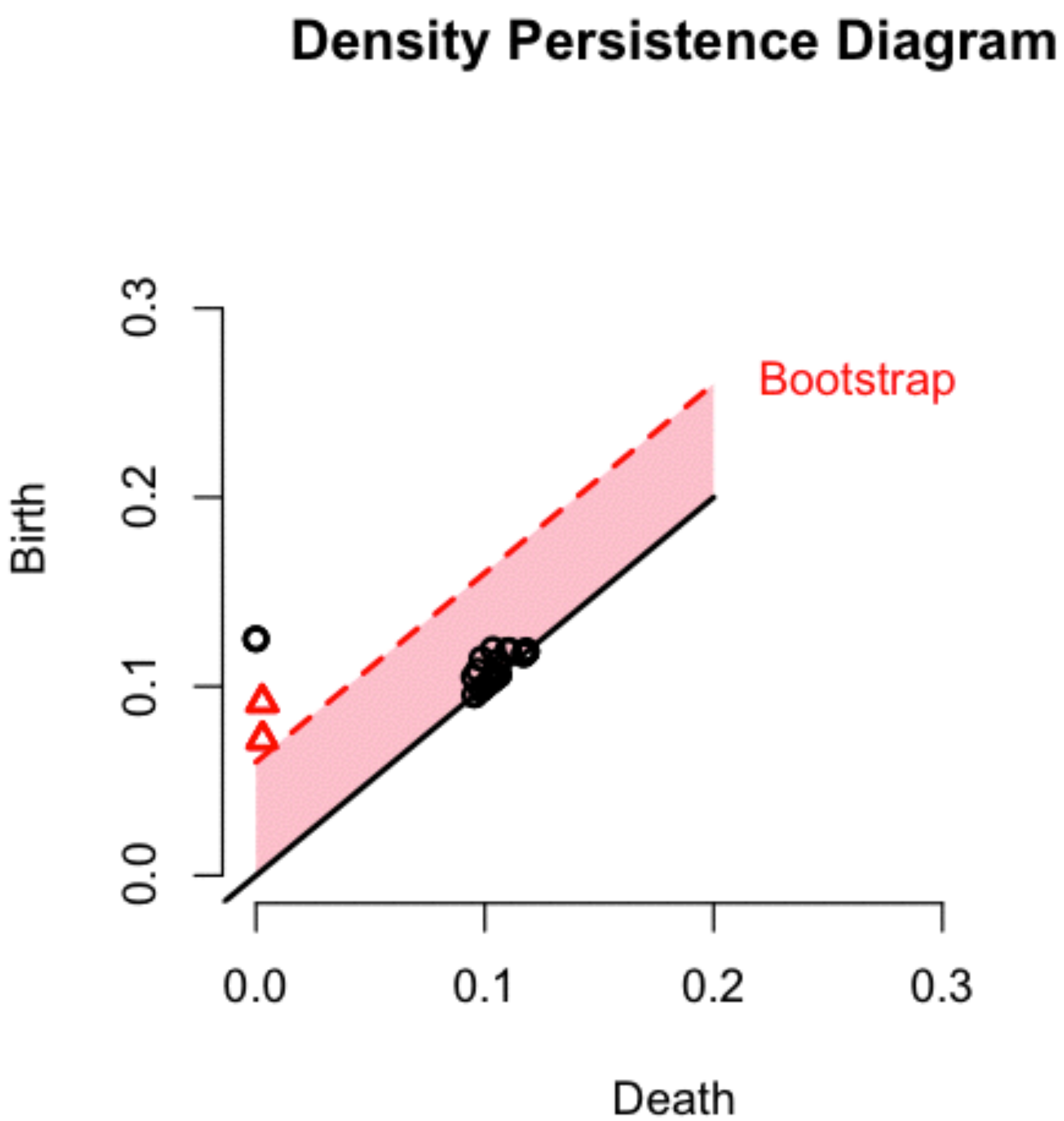
Theorem [Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, Singh, 2013]

$$\mathbb{P}\left(W_\infty(Dgm, \widehat{Dgm}_n) > \frac{Z_\alpha}{\sqrt{nh^D}}\right) \leq \mathbb{P}\left(\sqrt{nh^D}\|p_h - \hat{p}_h\|_\infty > Z_\alpha\right) = \alpha + O\left(\sqrt{\frac{1}{n}}\right)$$

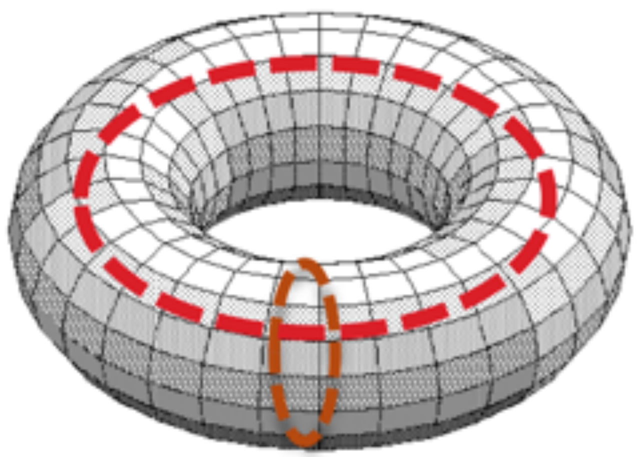
Stability theorem

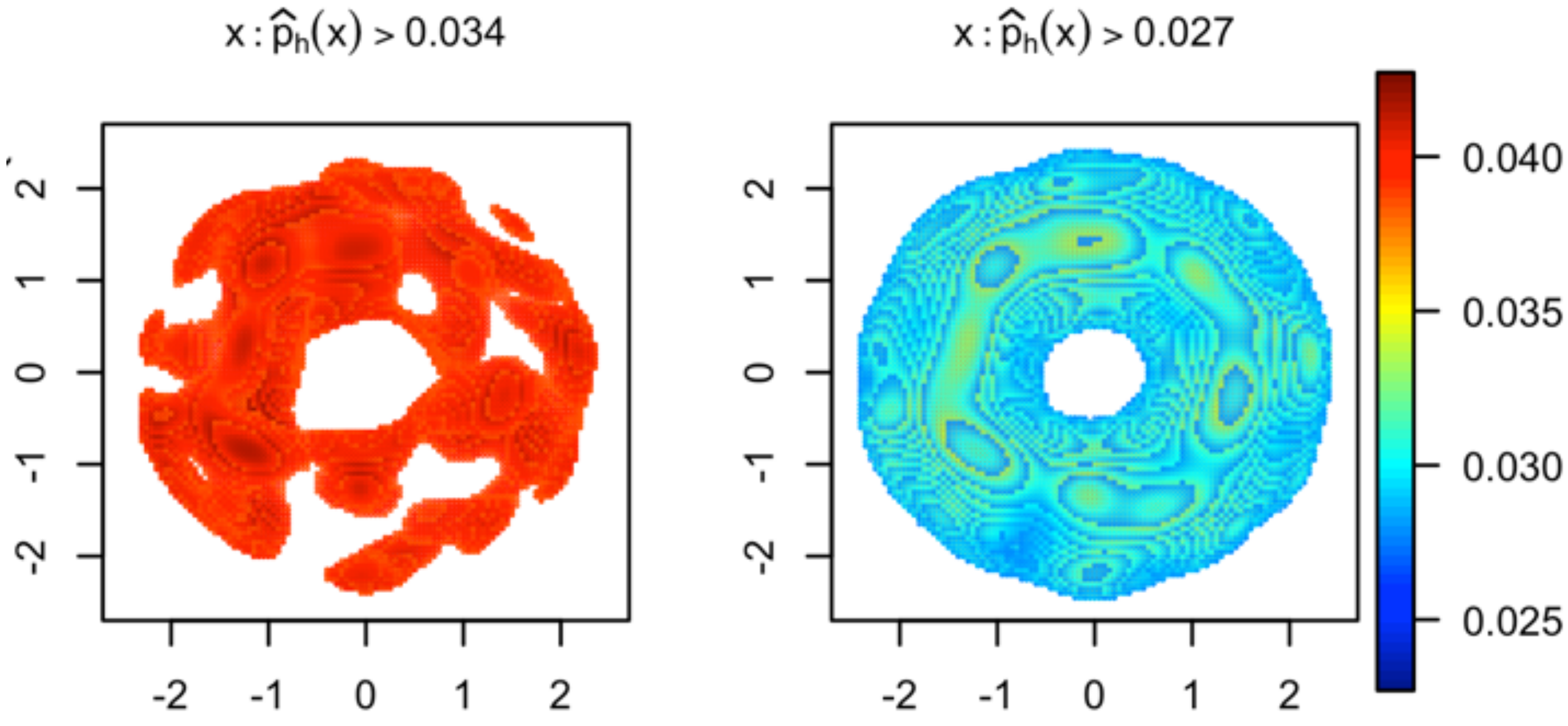
Example: Persistent Homology of the uniform density over the Cassini Curve

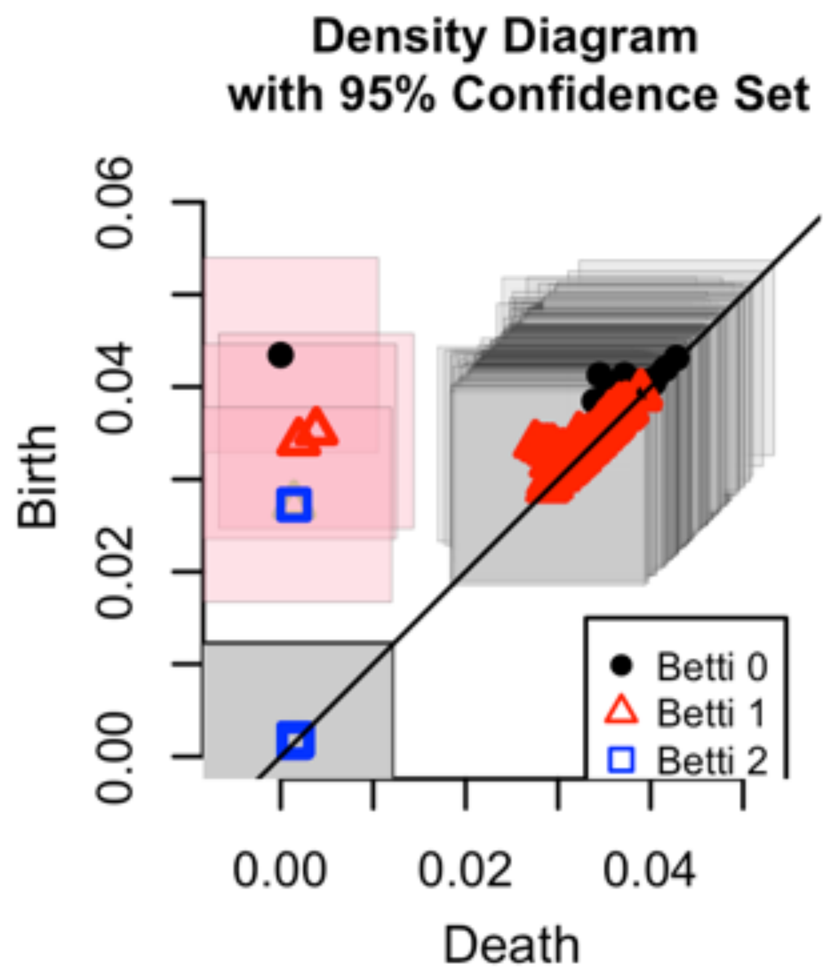




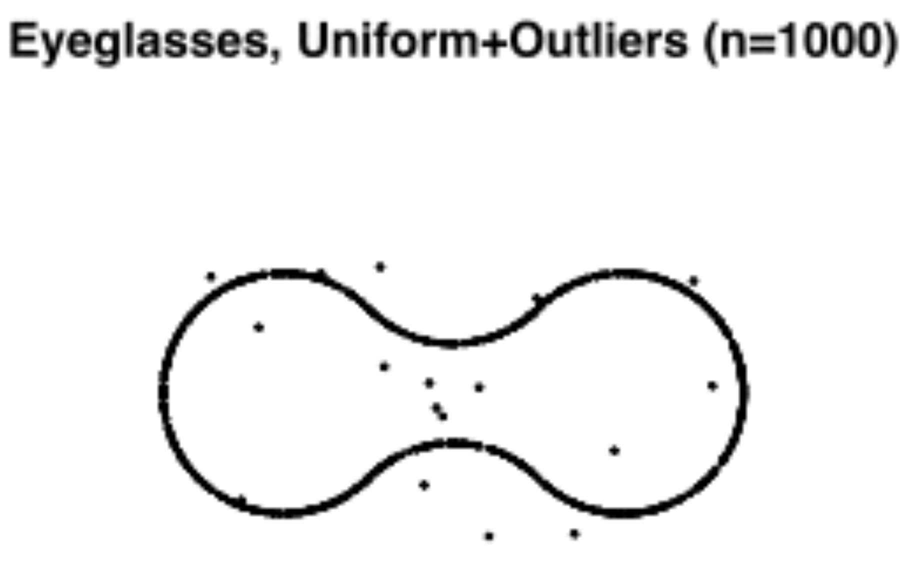
Example: uniform distribution over the Torus

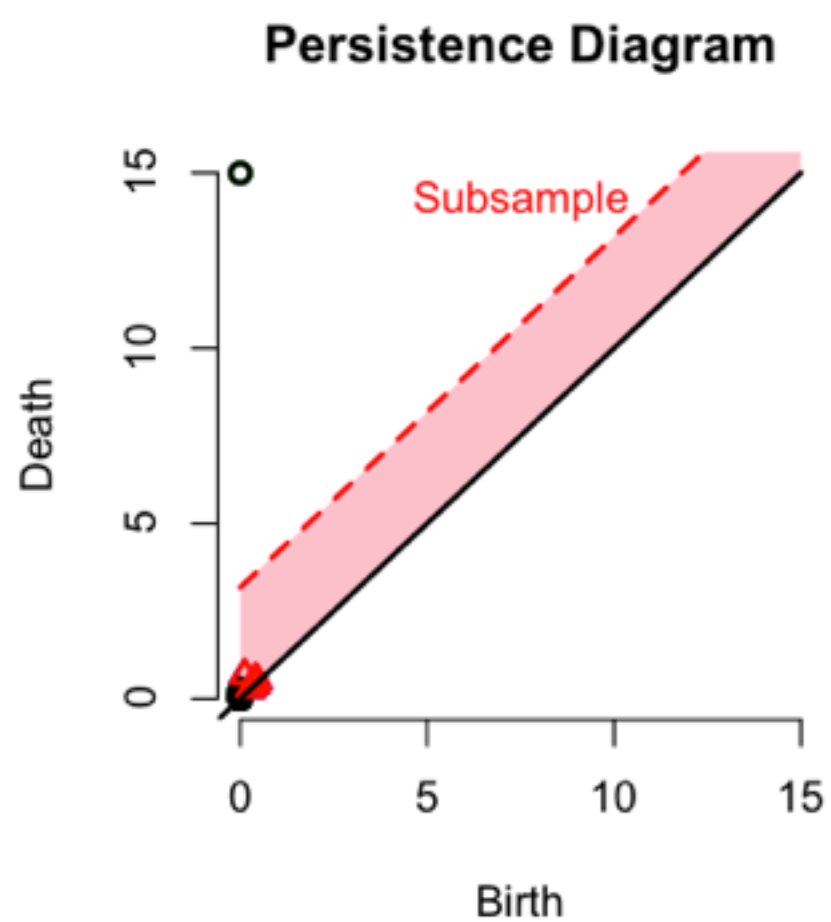


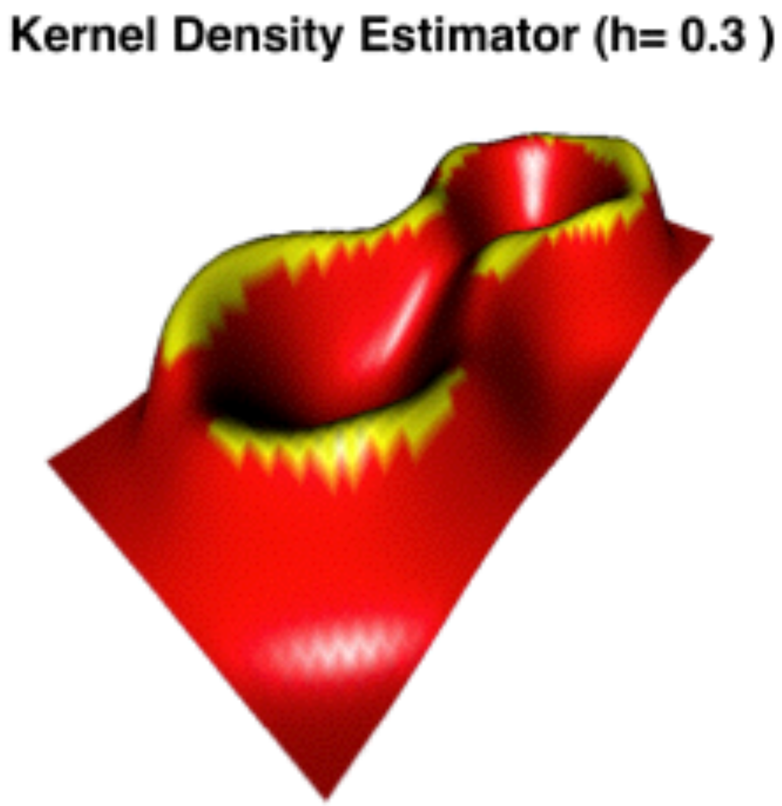


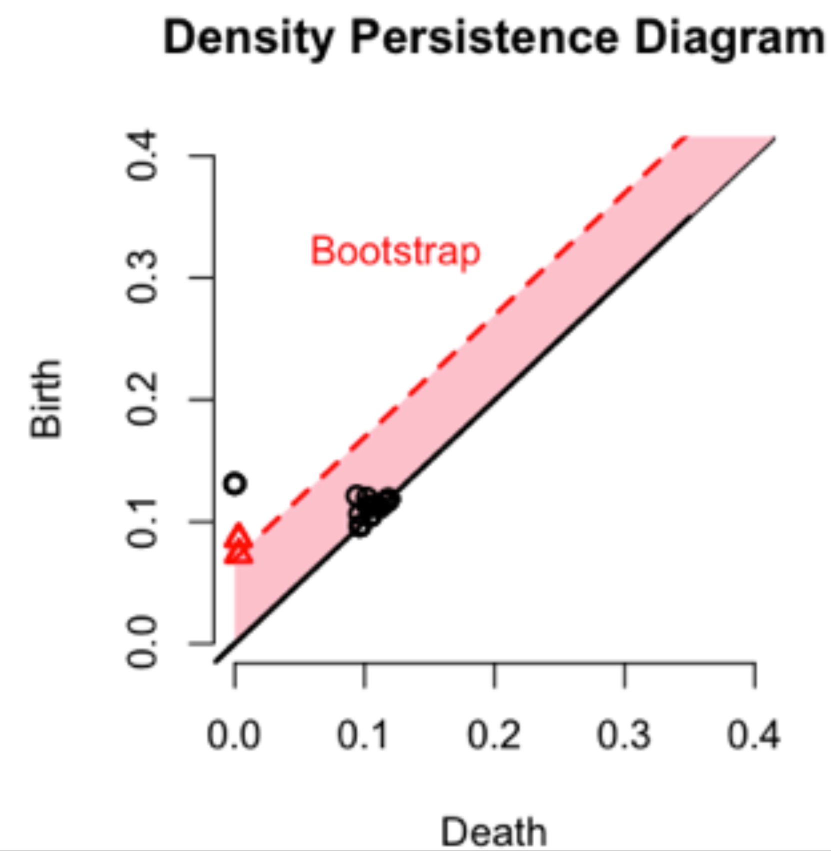


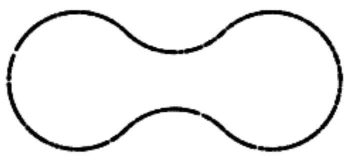
The Density Persistence Diagram is insensitive to outliers











Varying alpha and n

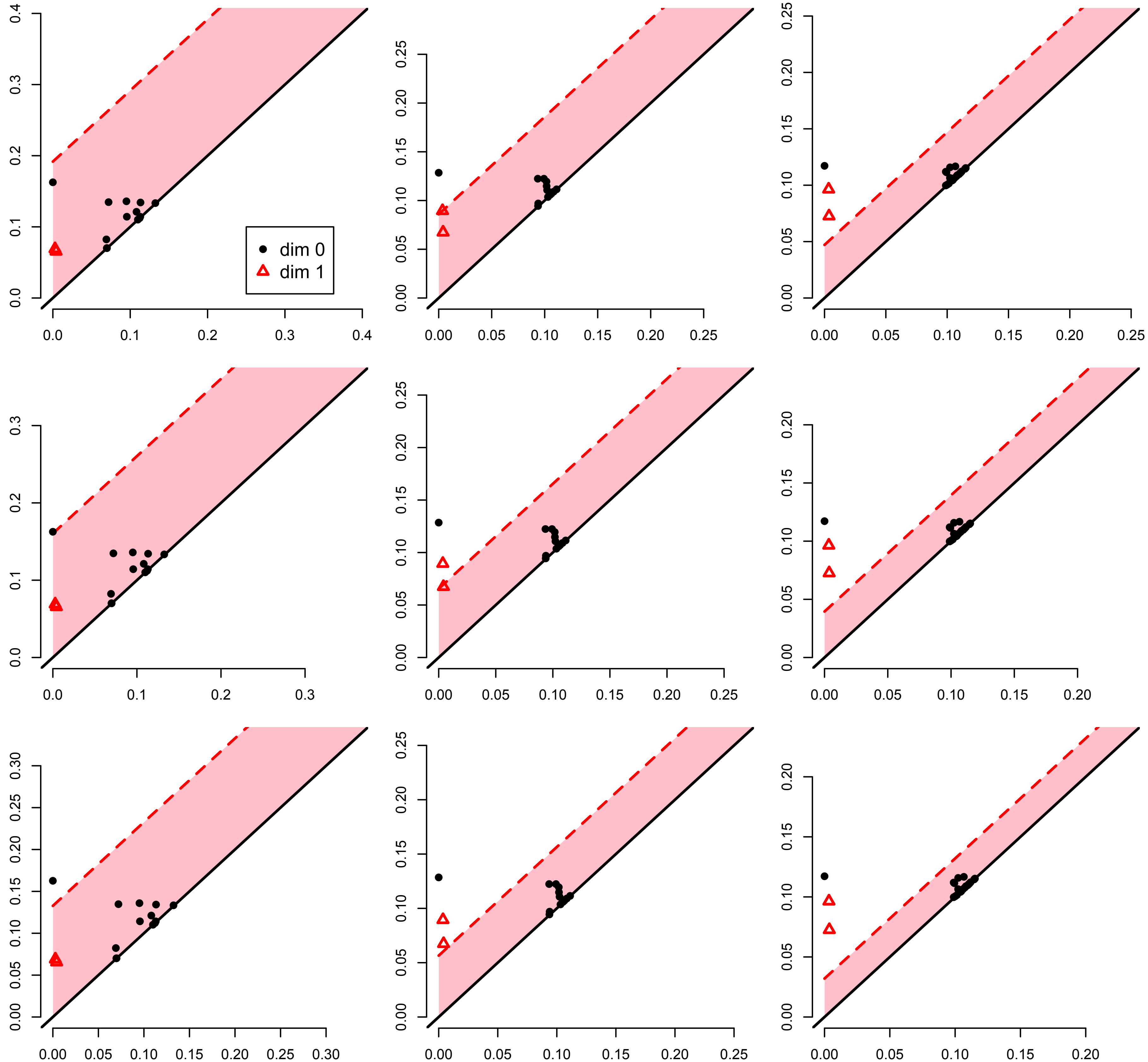
Fixed
h=0.3

n=200 n=1000 n=3000

α=0.01

α=0.05

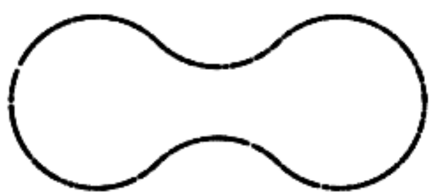
α=0.1



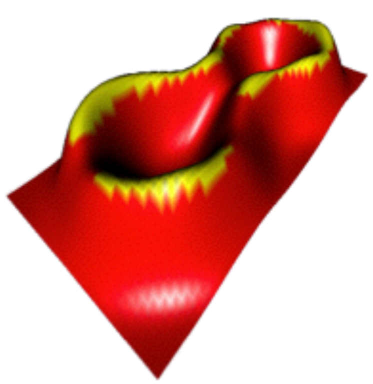
Parameter selection

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^D}K\left(\frac{\|x - X_i\|}{h}\right)$$

From Larry's Talk



h=0.1



h=0.3

h=0.5

FAILURE OF CROSS-VALIDATION FOR TDA

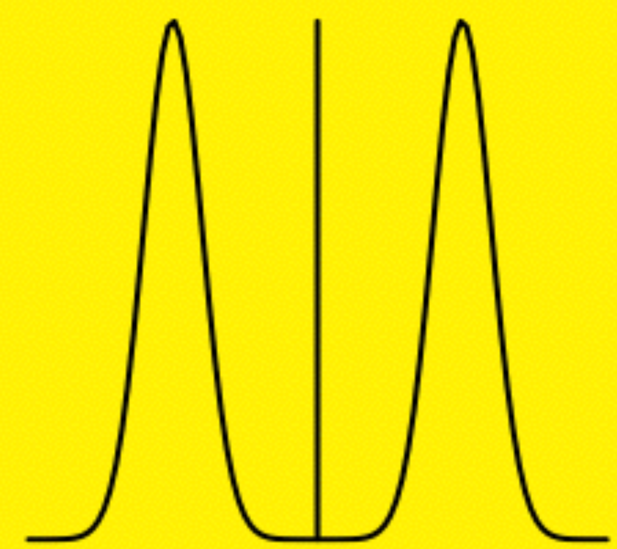
But in TDA, p might be singular or nearly singular. Consider

$$P = \frac{1}{3}N(-5,1) + \frac{1}{3}\delta_0 + \frac{1}{3}N(5,1)$$

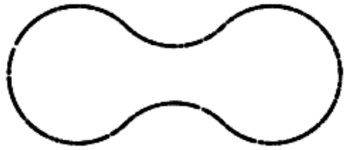
where δ_0 is a point mass at 0.

P doesn't have a density but p_h does, where

$$p_h(x) = \mathbb{E}[\widehat{p}_h(x)] = \frac{d}{dx}(P \star K_h).$$

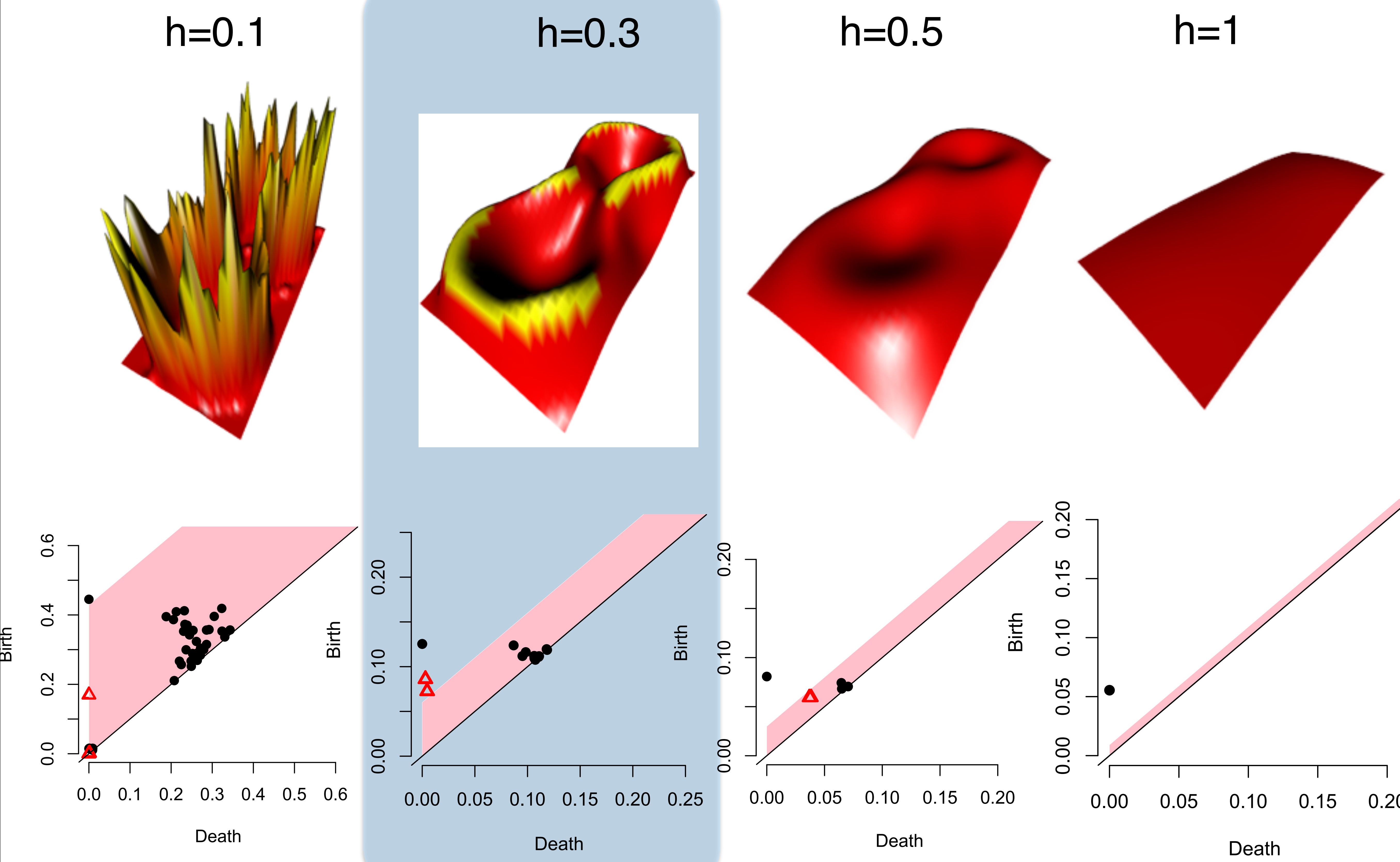


Cross-validation gives $h = 0$ which is useless.

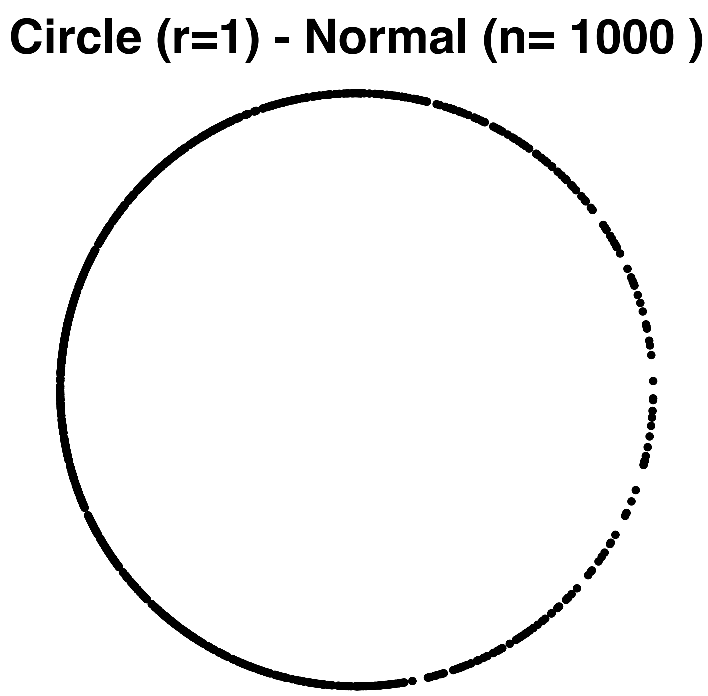


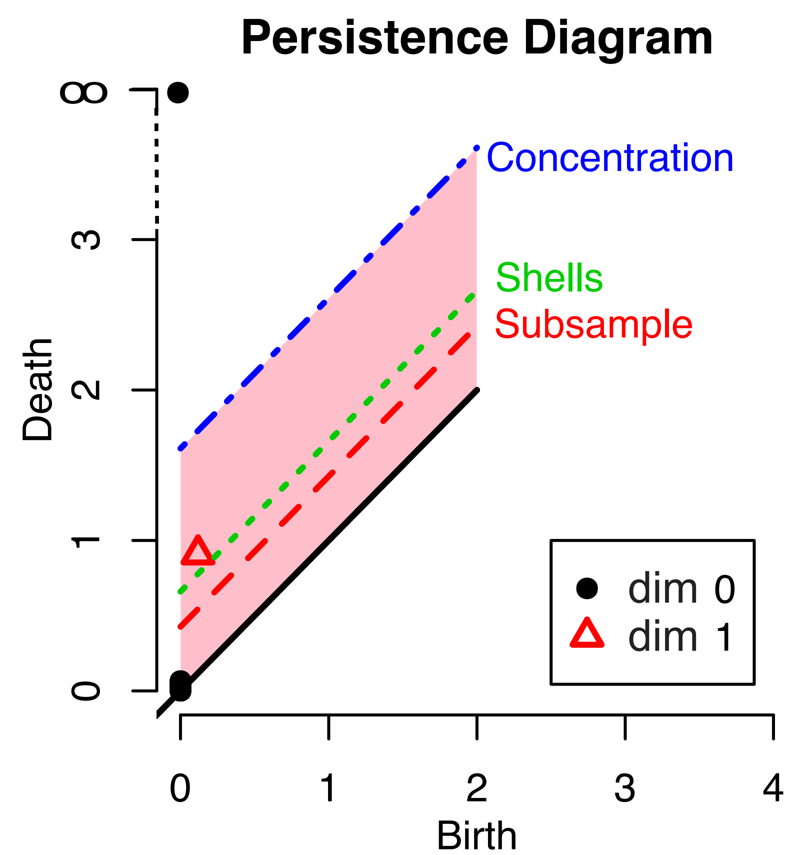
Maximum Significant Topological Signal Strength

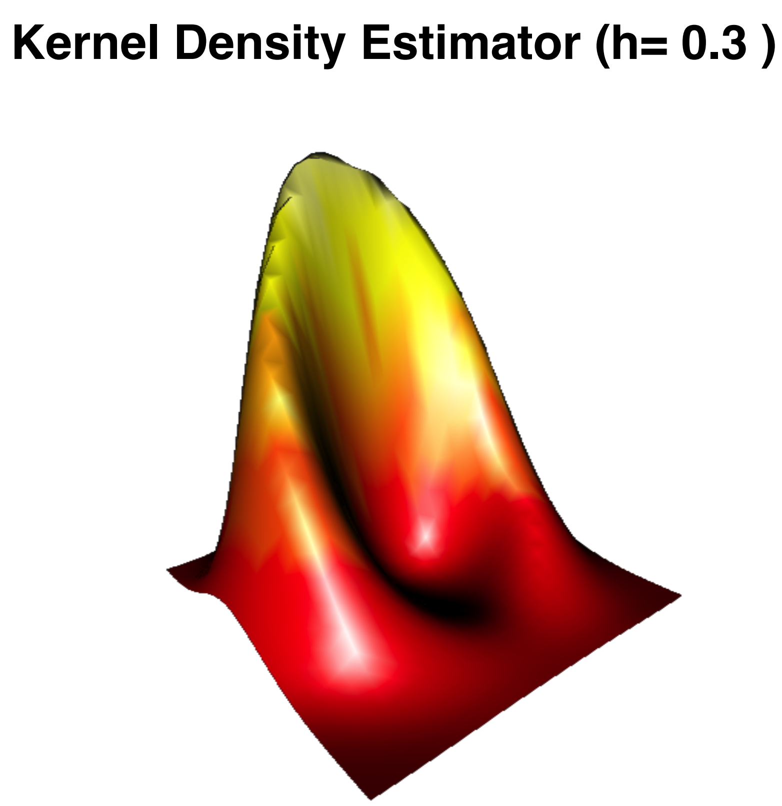
Fixed
n=1000, $\alpha = 0.05$

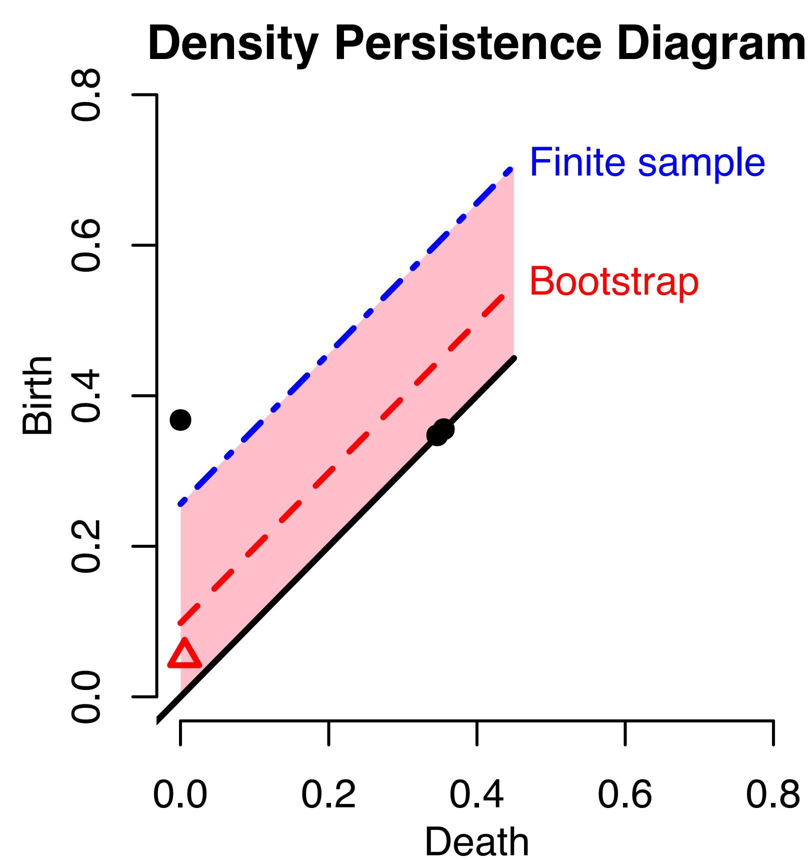


More methods: [Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, Singh, 2013]





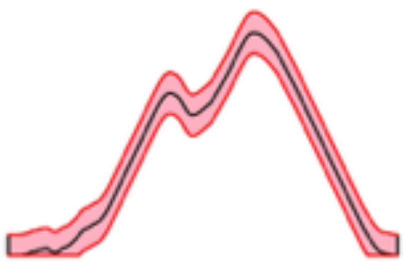


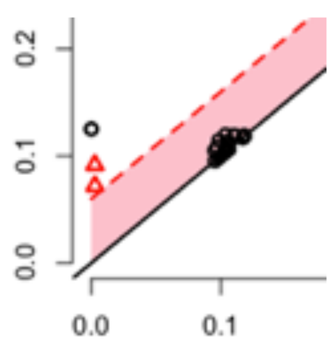


Open Questions

- More on tuning parameters
  (e.g. locally adaptive bandwidth)
- Confidence Sets for Diagrams of other distances
  (e.g. distance to measure)
- Hypothesis Tests
  see e.g. [Bubenik, 2012], [Robinson, Turner, 2013]

Summary





- Monday: Larry Wasserman showed confidence bands for landscapes
- Today: I’ve showed confidence sets for diagrams
- Tomorrow, 10 AM: Jessi Cisewski will show applications of these methods to real data.

Thank you

lecci@cmu.edu
www.stat.cmu.edu/~flecci www.stat.cmu.edu/topstat