ROBUST TOPOLOGICAL INFERENCE

Larry Wasserman
Carnegie Mellon University

Fred Chazal, Brittany Fasy, Jisu Kim, Fabrizio Lecci
Bertrand Michel, Alessandro Rinaldo

TOPSTAT: www.stat.cmu.edu/topstat

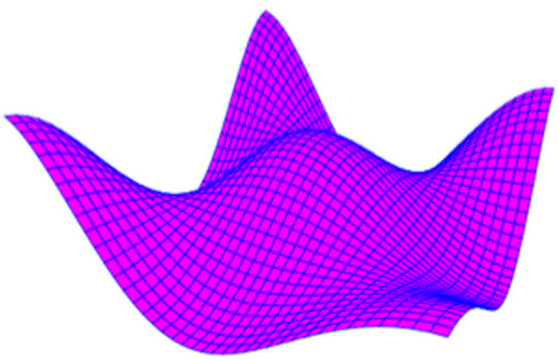
CMU TopStat

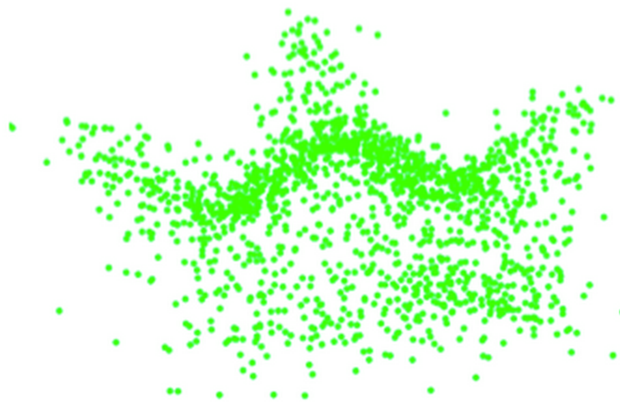
Homepage | Projects | Papers | Software | Talks | Private | Contact

People

- Sivaraman Balakrishnan
- Yen-Chi Chen
- Jessi Cisewski
- Brittany Fasy
- Christopher Genovese
- Brian Kent
- Jisu Kim
- Fabrizio Lecci
- Alessandro Rinaldo
- Aarti Singh
- Isa Verdinelli
- Larry Wasserman

The CMU Topological Statistics group is a research group at Carnegie Mellon University. The emphasis of our research is on statistical problems related to topological inference.
Visit the **Projects** page to see descriptions of our projects and relevant publications or preprints.





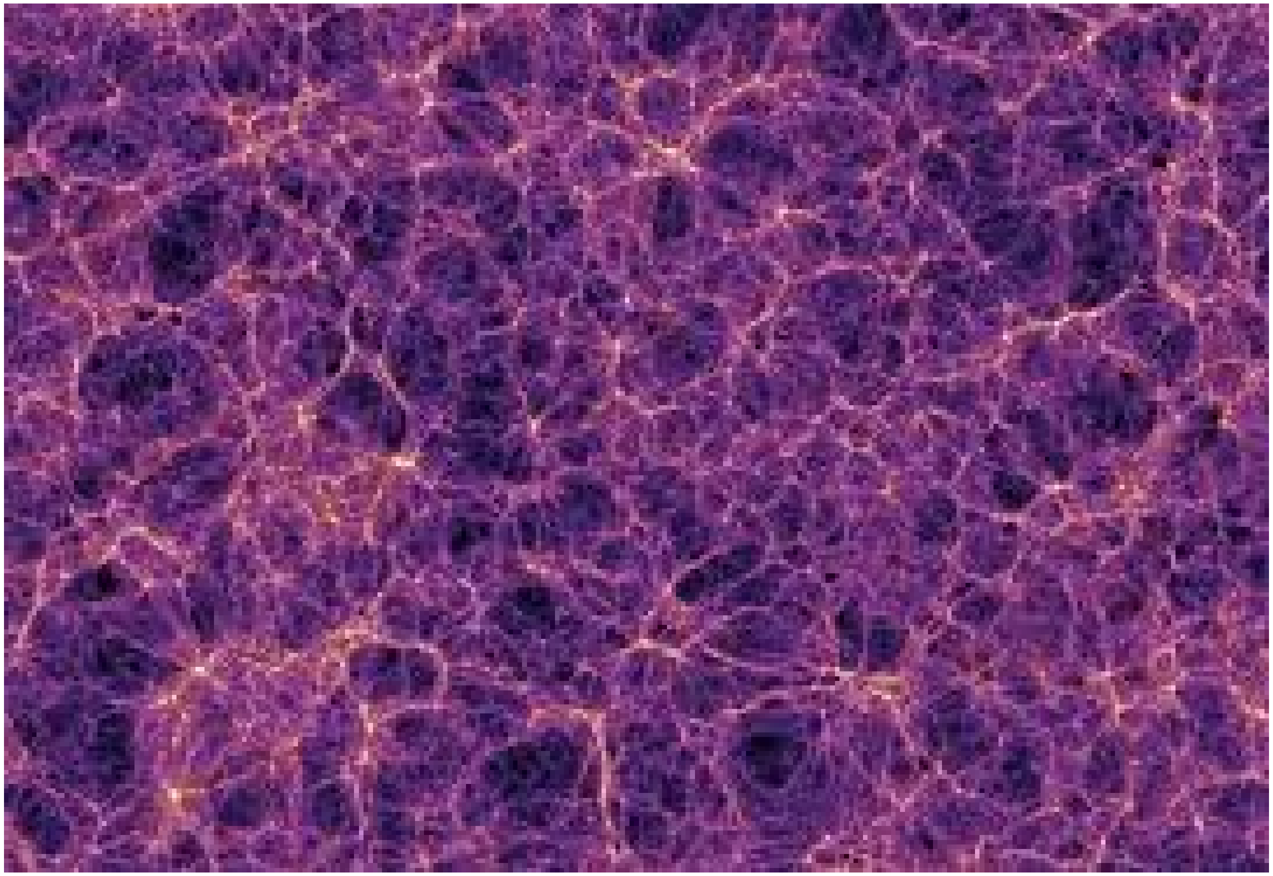
Topological Data Analysis (TDA): Extracting information from complex data.

The purpose of TDA is to extract features from complex point clouds (and images) for: summary, visualization, comparison, classification, inference.

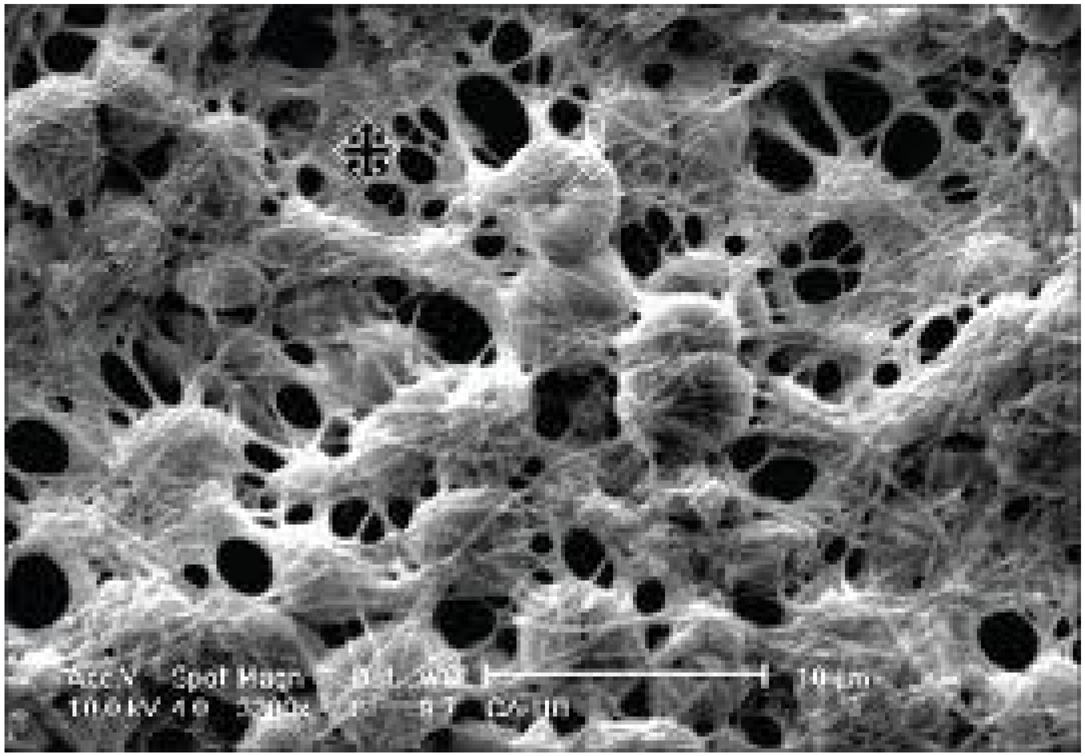
Many current methods are highly non-robust.

In this talk, I will describe a robust approach (distance-to-a-measure DTM) and I will discuss its statistical properties.

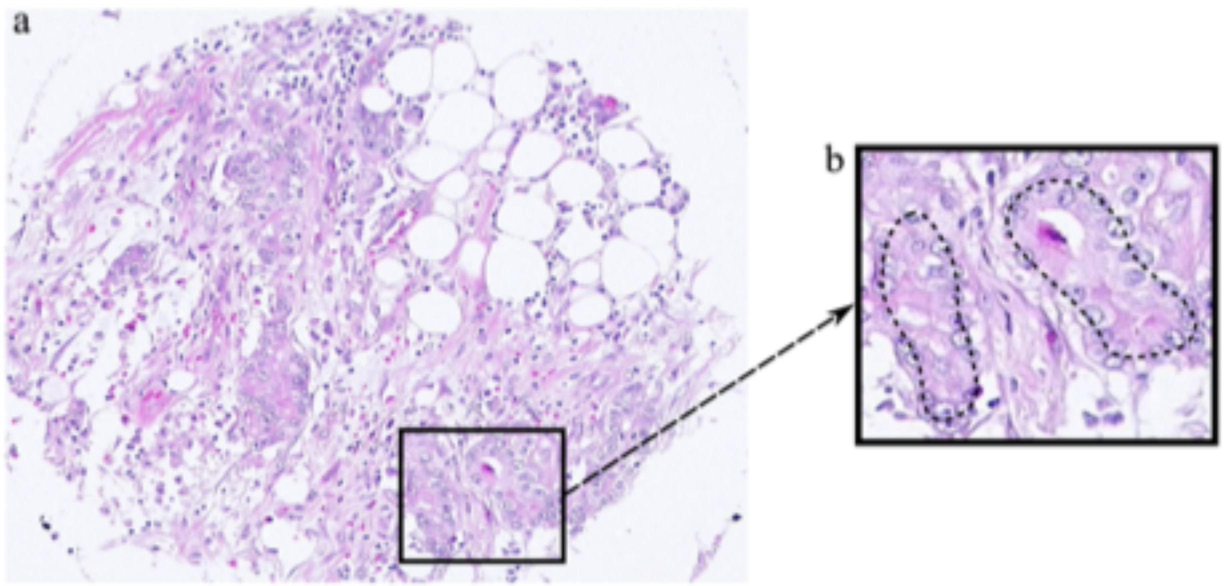
Before we start, here are some motivating examples:



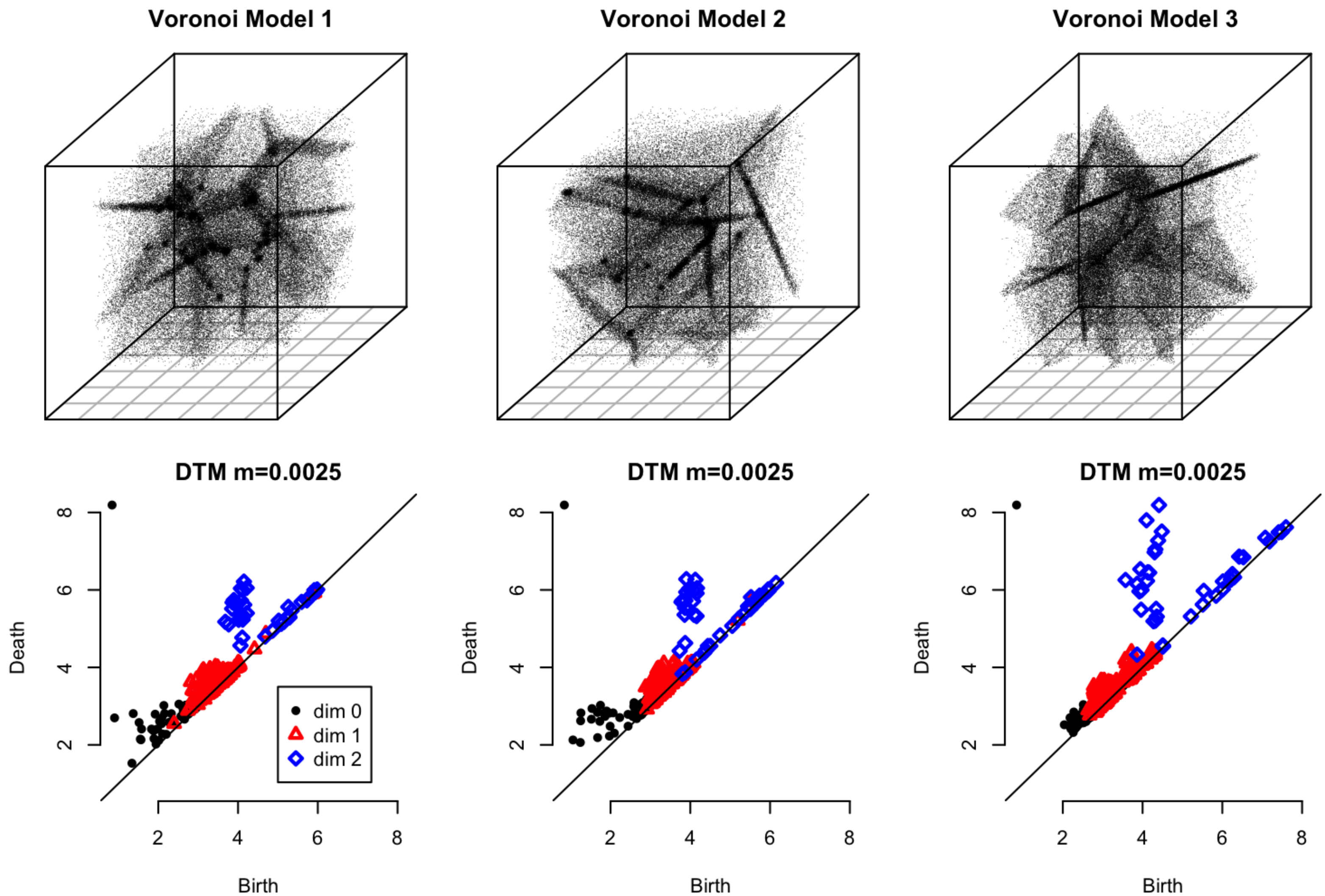
Cosmic web (source: Max Plank Institut; http://www.mpa-garching.mpg.de/galform/virgo/millennium/)



Fibrin network (source: Amiredly et al 2011).



Histological Images (source: Singh et al 2014)



Three Voronoi foam models. Which one is different?

NOTE ON SOFTWARE

All calculations were done with the R package: TDA
by Brittany T. Fasy, Jisu Kim, Fabrizio Lecci, Clement Maria.

Built on: Dionysus by Dmitriy Morozov, GUDHI by Clement Maria, PHAT by Ulrich Bauer, Michael Kerber, Jan Reininghaus.

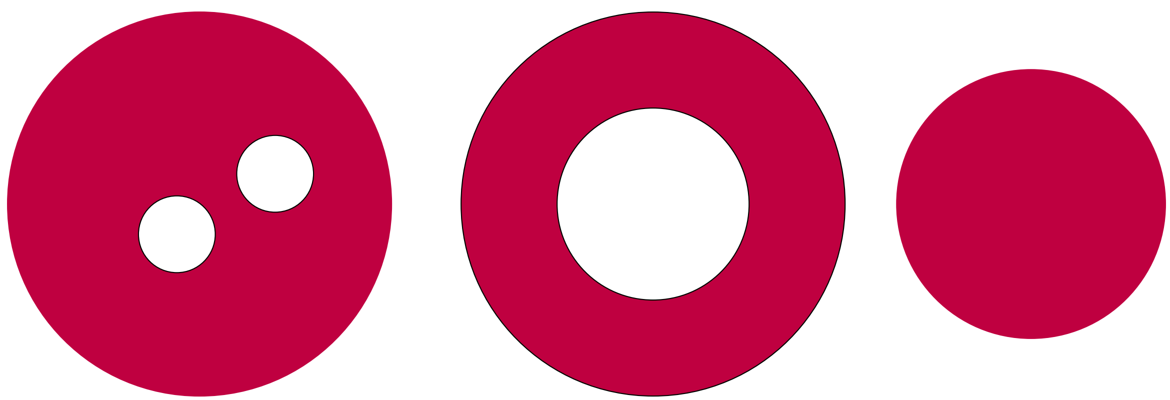
Download it from:

http://cran.us.r-project.org/web/packages/TDA/index.html

or

www.stat.cmu.edu/topstat

Algebraic Topology (Homology) in One Slide



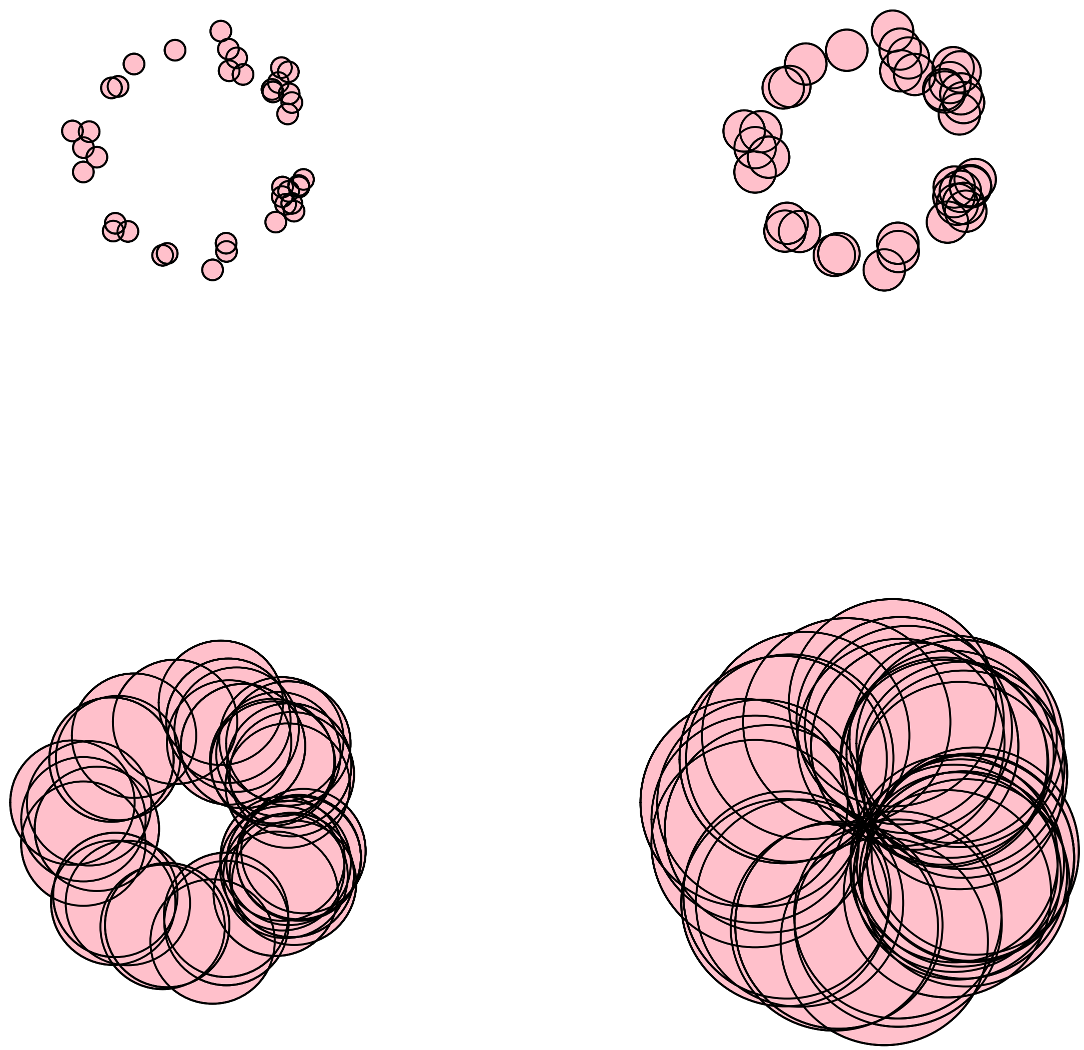
$\beta_0 = 3,\ \beta_1 = 3$

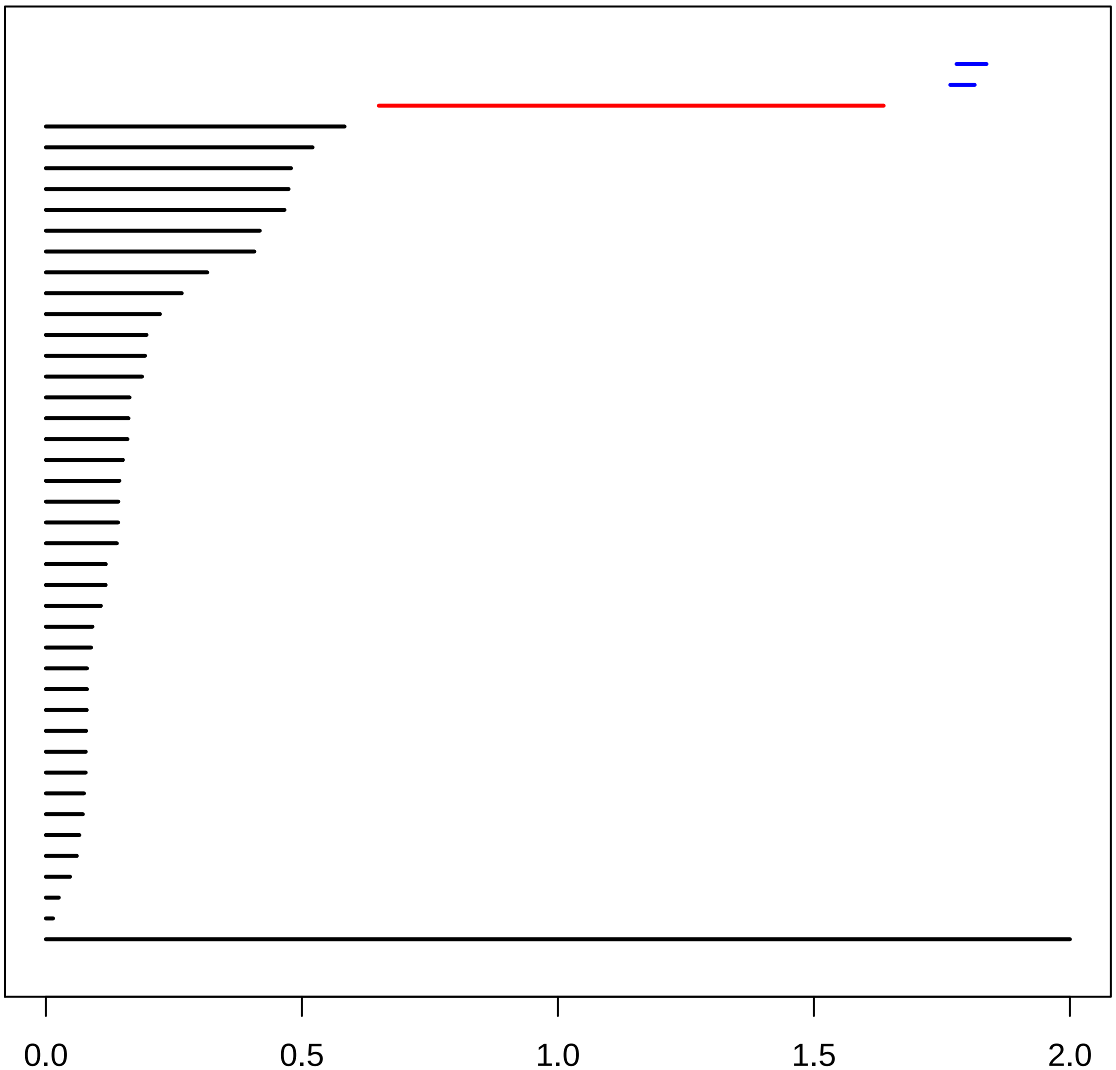
PERSISTENT HOMOLOGY

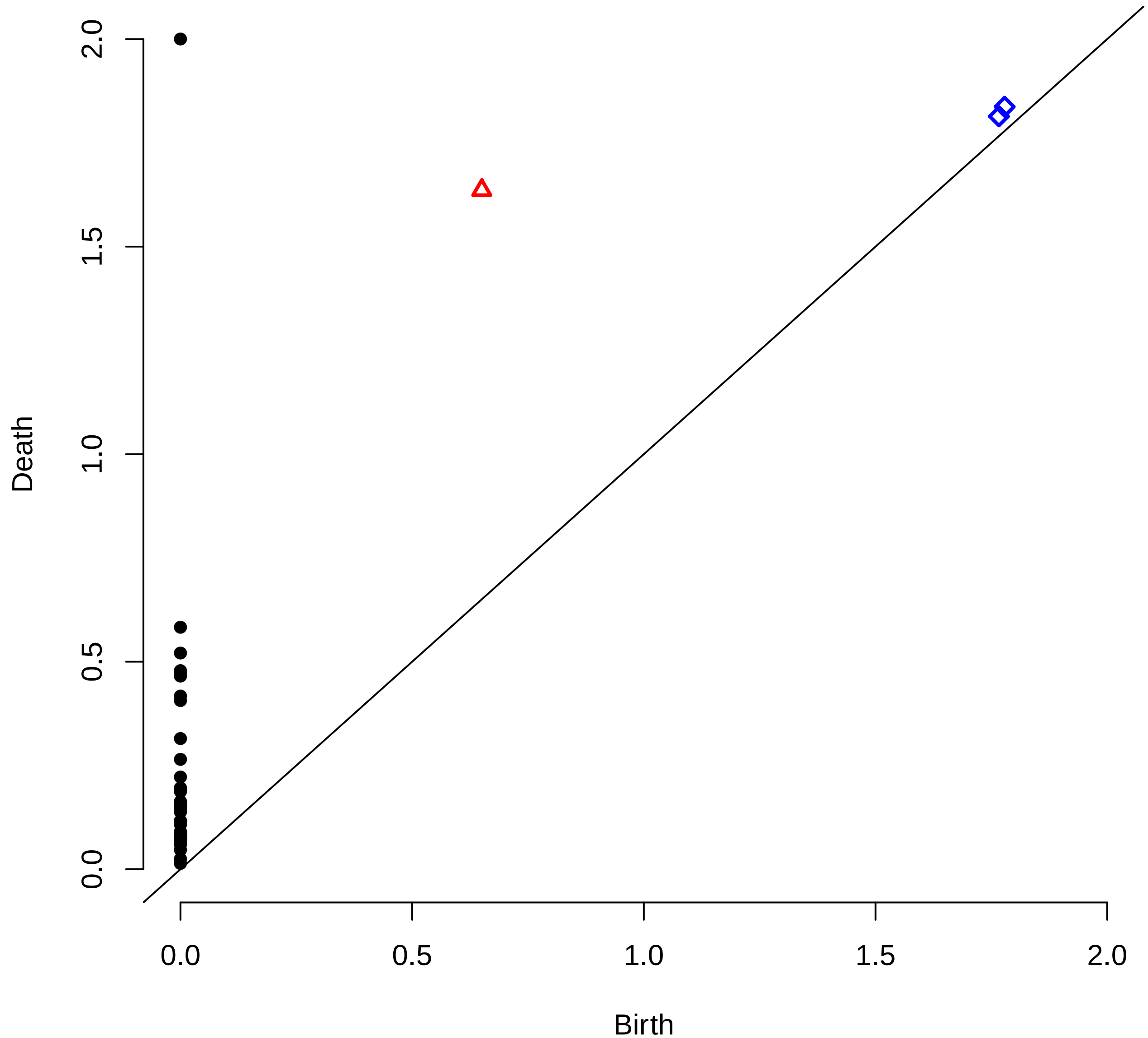
Persistent homology is a multiscale version of homology. (Edelsbrunner, Zomorodian, Harer, Carlsson, ...)

The idea is to find topological features (connected components, loops, voids etc) at different scales.

- First I will explain persistent homology using unions of balls.
- Then I will explain it using distance functions.
- Then we will robustify the distance function.




0.0
0.5
1.0
1.5
2.0


Death
Birth
0.0
0.5
1.0
1.5
2.0

IMPORTANT FACTS ABOUT THE PERSISTENCE DIAGRAM

- It is two dimensional, regardless of the dimension of the data.

- Points close to the diagonal are “small features.” (noise?)

- The diagram D includes the points plus all the points on the diagonal.

- There is a metric on the space of diagrams. The bottleneck distance.

Define $S_\epsilon = \bigcup_{i=1}^n B(X_i, \epsilon)$.

Persistent homology measures the evolution of features of

$$\left\{ S_\epsilon : \ \epsilon \geq 0 \right\}.$$

The diagram D is a collection of pairs (birth and death times) $\{(b_1, d_1), \ldots, (b_m, d_m)\}$. D includes all points on the diagonal.

Distance between two diagrams D_1, D_2:

$$\text{bottleneck}(D_1, D_2) = \min_{g: D_1 \to D_2} \sup_{z \in D_1} ||z - g(z)||_\infty.$$

COMPUTING HOMOLOGY

How do we actually find the connected components, holes, etc of S_ϵ?

We form a simplicial complex which is a set of simplices. This complex has the same topology as S_ϵ.

Computing the homology from the complex reduces to linear algebra (operations on matrices).

We won't discuss the details in this talk.

THE DISTANCE FUNCTION

Let S be a compact set.
Define $\Delta_S(x) = d(x, S) = \inf_{y \in S} ||x - y||$.

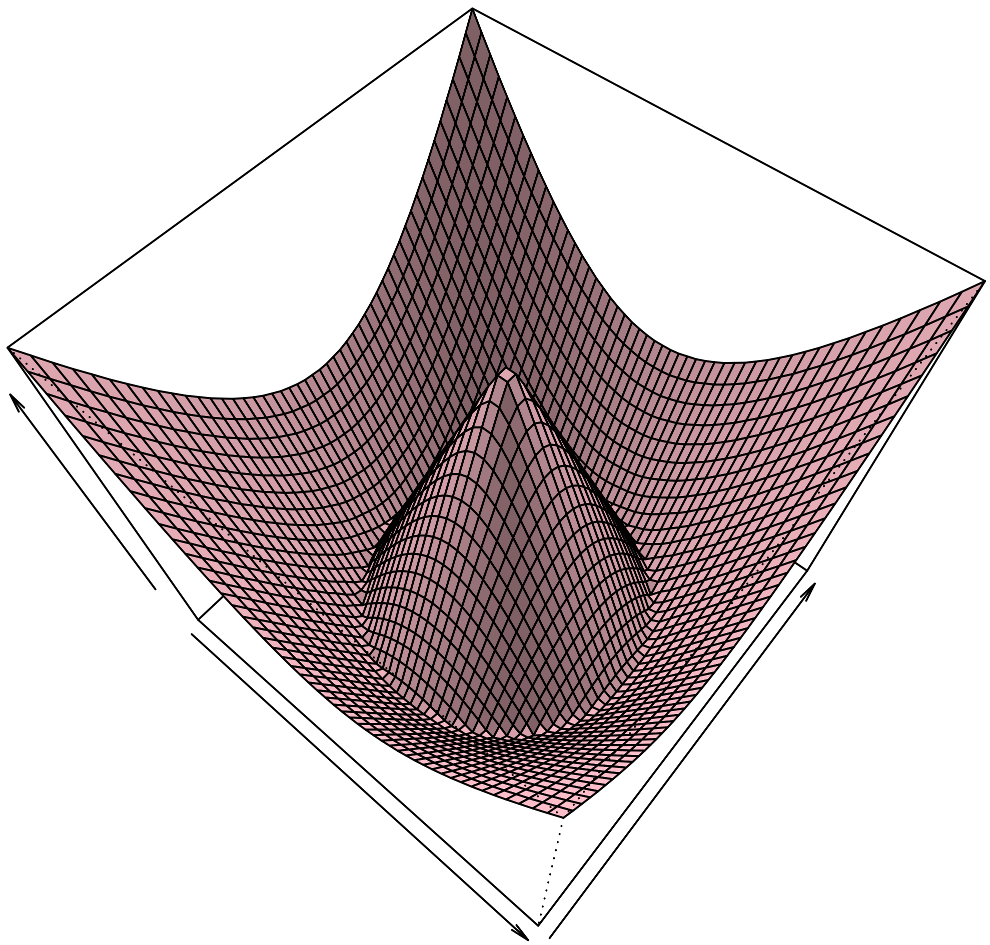
Let $L_t = \{x : \ \Delta_S(x) \leq t\}$ be a lower level set of the distance function.

The filtration $\{L_t : \ t \geq 0\}$ defines a persistent homology.

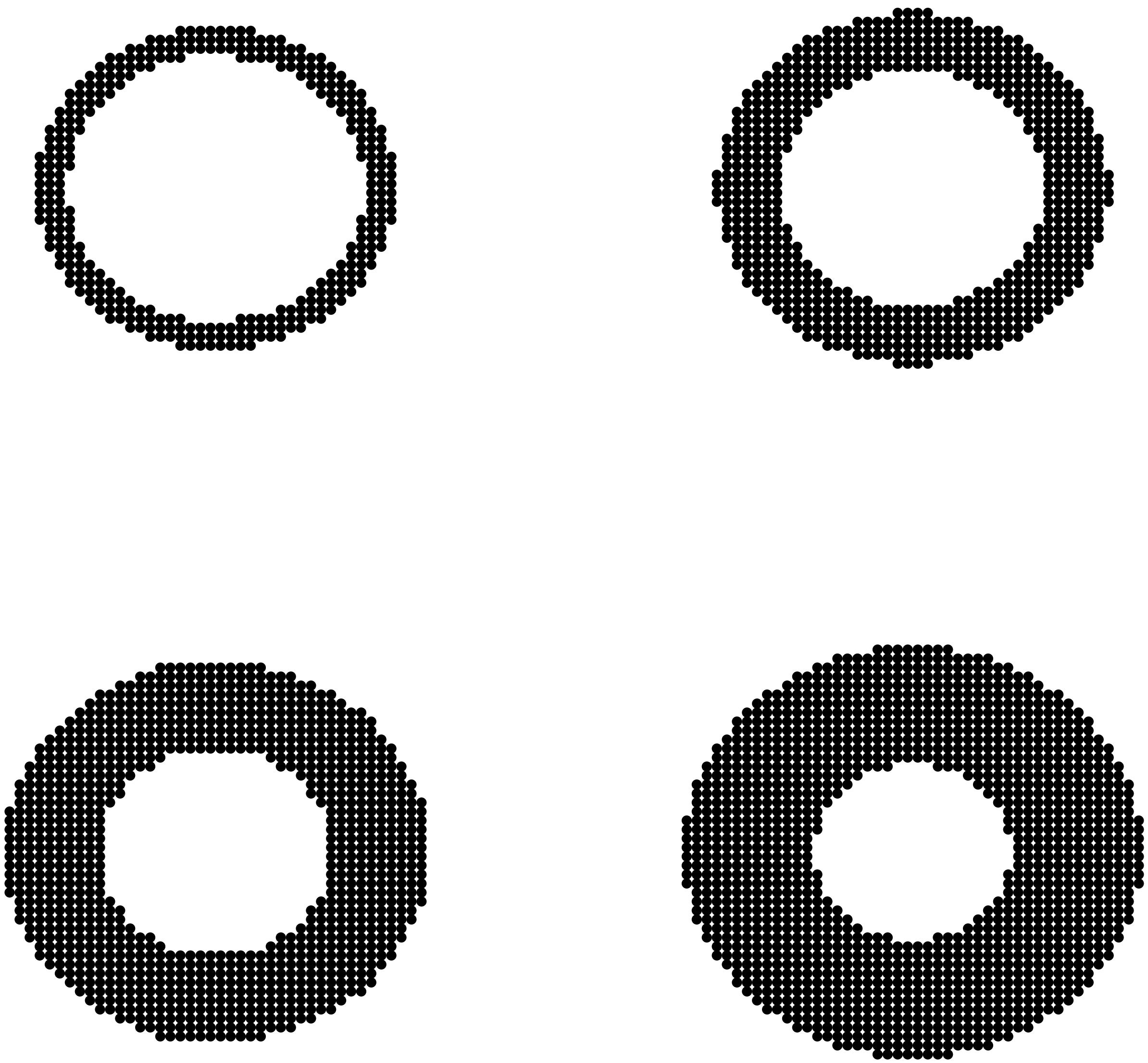
Cohen-Steiner, Edeslbrunner and Harer (2007) showed that:

$$\text{bottleneck}(D_1, D_2) \leq \sup_x ||\Delta_{S_1}(x) - \Delta_{S_2}(x)||.$$

Distance function for a circle in the plane.



Sublevel Sets



THE EMPIRICAL DISTANCE FUNCTION

Now let $S = \{X_1, \ldots, X_n\}$. Then

$$\Delta_S(x) = d(x, S) = \min_i ||x - X_i||.$$

Let

$$L_t = \{x : \ \Delta_S(x) \leq t\}.$$

Then

$$L_t = \bigcup_{i=1}^{n} B(X_i, t).$$

The union of balls is just the lower level sets of the empirical distance function.

INTERLUDE: THE STATISTICAL PERSPECTIVE

We are focusing on the following situation:

The data: $X_1, \ldots, X_n \sim P$.

We are interested in some function $T(P)$ (population quantity).

Example: $T(P) =$ persistent homology of the support of P.

Anything we compute from the data should be viewed as an estimate of population quantity.

Success means:
-consistency (get correct answer as $n \to \infty$)
-some measurement of uncertainty (bootstrap confidence sets)
-robustness (don't require fragile conditions on P)

BOOTSTRAP INFERENCE IN ONE SLIDE

$X_1, \ldots, X_n \sim P$.

$P_n =$ empirical measure (mass $1/n$ at each data point).

Estimate $\theta = T(P)$ with $\widehat{\theta} = T(P_n)$.

Bootstrap: Draw $X_1^*, \ldots, X_n^* \sim P_n$. Compute $\widehat{\theta}^* = T(P_n^*)$. Repeat. Find $\widehat{c}$ such that

$$\mathbb{P}(\sqrt{n}|\widehat{\theta}^* - \widehat{\theta}| > \widehat{c} \,|\, X_1, \ldots, X_n) = \alpha.$$

$C_n = \widehat{\theta} \pm \widehat{c}/\sqrt{n}$.

Then $\mathbb{P}(\theta \in C_n) = 1 - \alpha + O_P\left(\frac{1}{\sqrt{n}}\right)$.

Let $X_1, \ldots, X_n \sim P$ where P has support S.

True distance function Δ_S with persistence diagram D.
Empirical distance function:

$$\Delta_n(x) = \min_i ||x - X_i||$$

with diagram $\widehat{D}$.

Under not so weak conditions, (see our paper, Annals to appear),

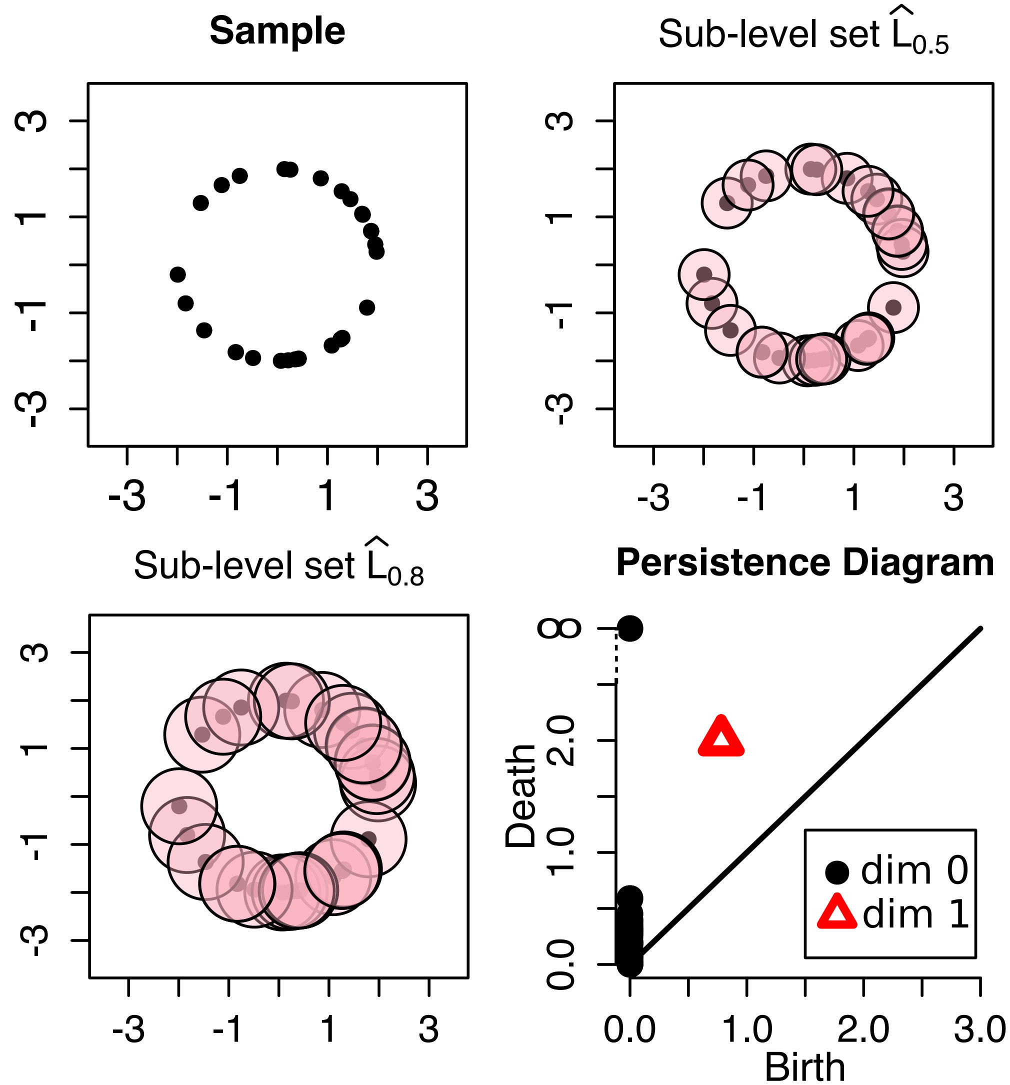
$$\sup_x ||\Delta_n(x) - \Delta_S(x)|| \stackrel{P}{\to} 0$$

and this implies that

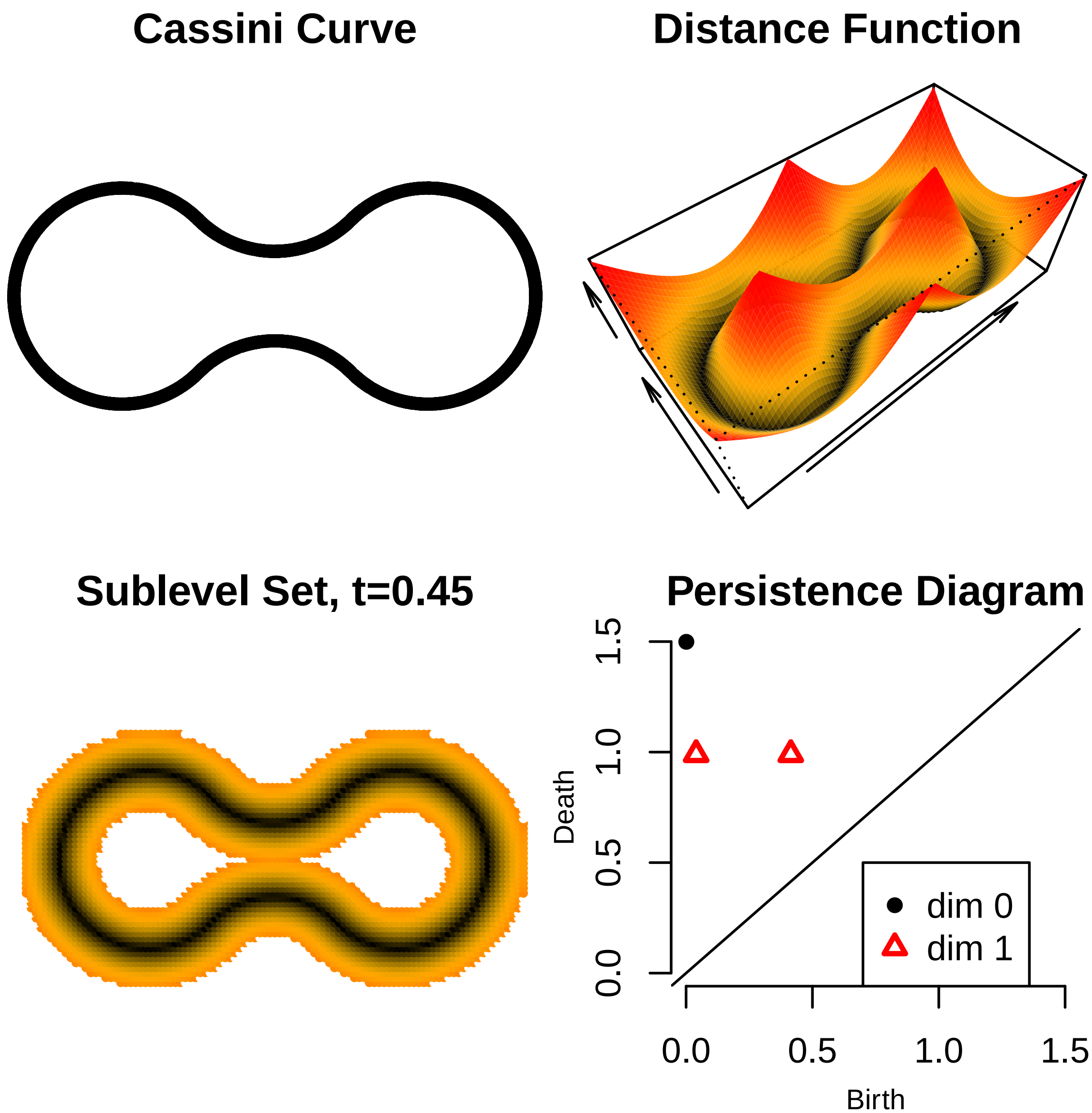
$$\text{bottleneck}(\widehat{D}, D) \stackrel{P}{\to} 0.$$

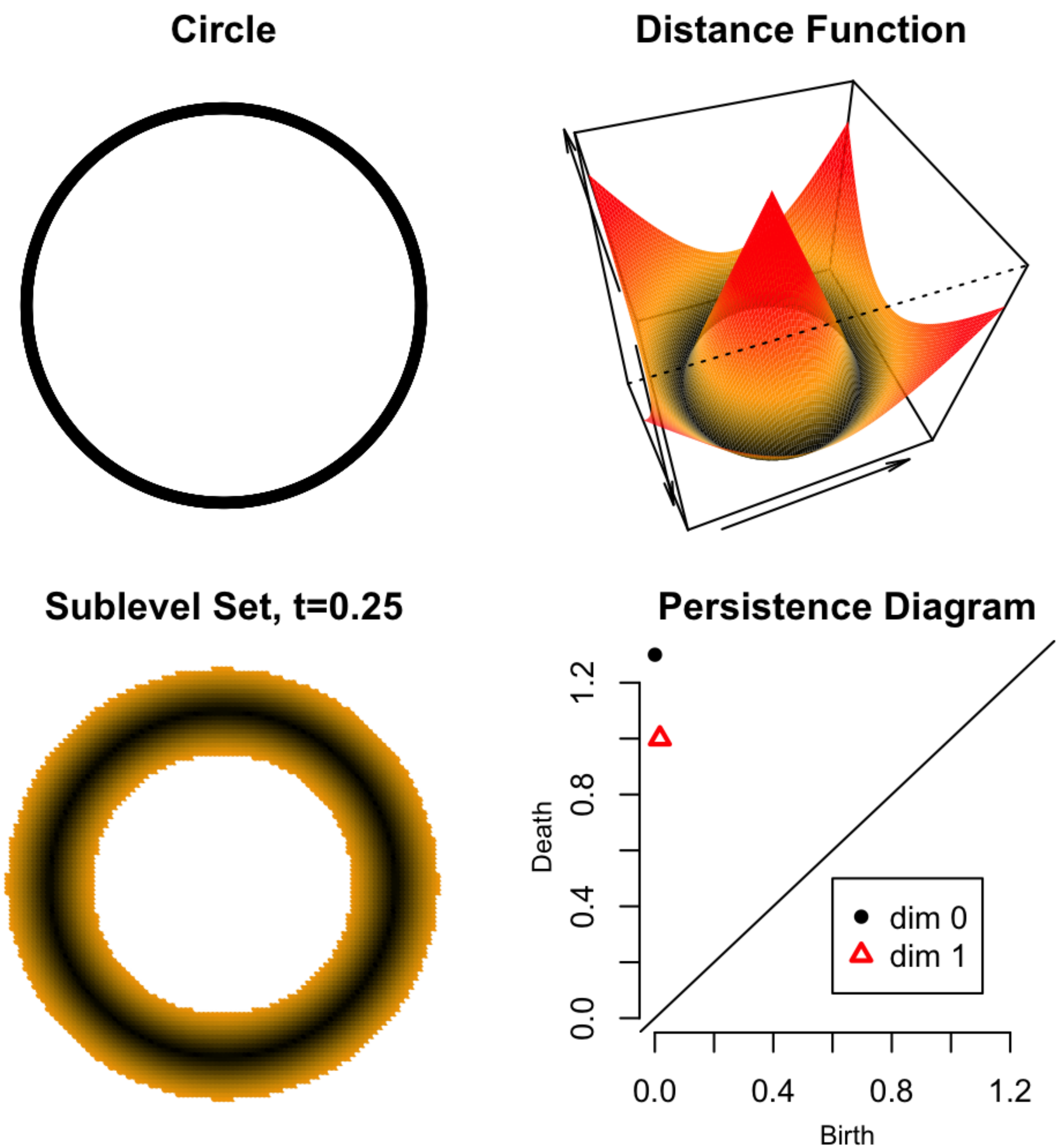
But: if there is any noise or outliers, $\Delta_n(x)$ is a disaster!

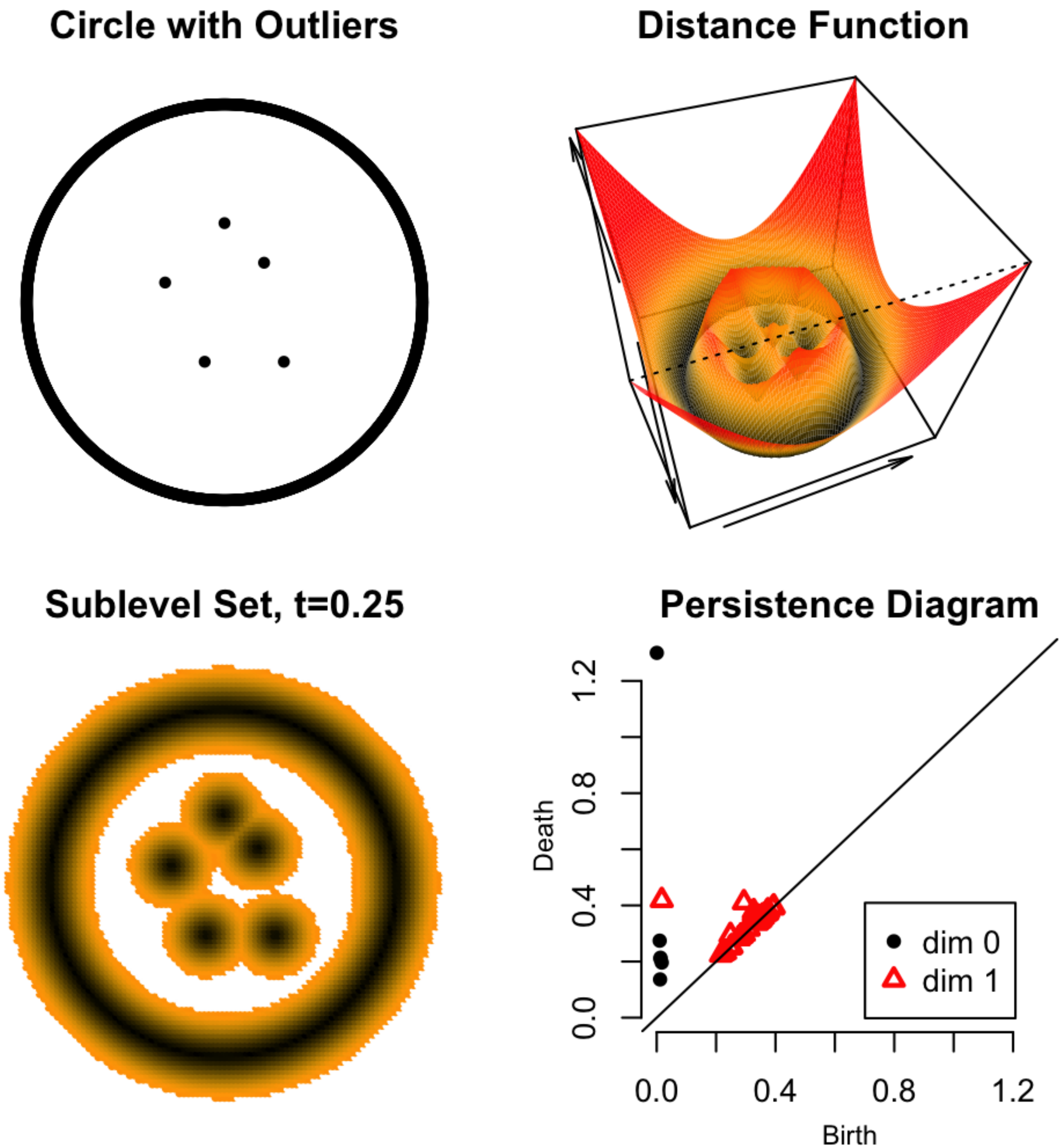
Circle, no outliers:



Cassini curve, no outliers:




Circle
Distance Function
Sublevel Set, t=0.25
Persistence Diagram
Death
Birth
0.0
0.4
0.8
1.2
dim 0
dim 1


Circle with Outliers
Distance Function
Sublevel Set, t=0.25
Persistence Diagram
Death
Birth
0.0
0.4
0.8
1.2
dim 0
dim 1

ROBUST TDA

Suppose that

$$X_1, \ldots, X_n \sim P = \pi R + (1 - \pi)(Q \star \Phi_\sigma)$$

where R is a smooth distribution over $\mathbb{R}^d$ (outliers), Φ is noise ($N(0, \sigma^2 I)$) and Q is supported on a "small set" S. We want to estimate the homology of S or the persistent homology of S.

Two robust approaches:

(1) DTM (distance to a measure); described on next slide.

(2) Upper level sets of density p.

First we focus on DTM.

DTM

Distance-to-a-measure (DTM) invented by: Chazal, Cohen-Steiner and Merigot (2011).

For each x, let

$$G_x(t) = P(||X - x|| \leq t).$$

Given $0 < m < 1$, the DTM is

$$\delta^2(x) = \frac{1}{m}\int_0^m [G_x^{-1}(u)]^2 du = \mathbb{E}\left[||X - x||^2 I(||X - x|| \leq G_x^{-1}(m))\right].$$

The sublevel sets of δ define a persistence diagram D.

Stability Theorem (Chazal, Cohen-Steiner and Merigot, 2011)

Let P_1 have DTM δ_1 with diagram D_1 and P_2 have DTM δ_2 with diagram D_2.

Then,

$$\text{bottleneck}(D_1, D_2) \leq ||\delta_1 - \delta_2||_\infty.$$

This will help us with statistical inference.

Suppose that

$$P = \pi R + (1 - \pi)(Q \star \Phi_\sigma)$$

and Q is supported on S and satisfies (a,b)-condition:

$$Q(B(x, \epsilon)) \geq a\epsilon^b.$$

Let D be the diagram from δ and let D_S be the diagram for the distance function of S. Then

$$\text{bottleneck}(D, D_S) \leq a^{-1/b} m^{1/b} + \frac{c\sqrt{\pi} + \sigma(1 + \pi)}{\sqrt{m}}.$$

So, when π, σ, m are small, $D \approx D_S$.

ESTIMATION AND INFERENCE

The DTM $\delta(x) = \delta_P(x)$ is a function of P. If we insert the empirical measure

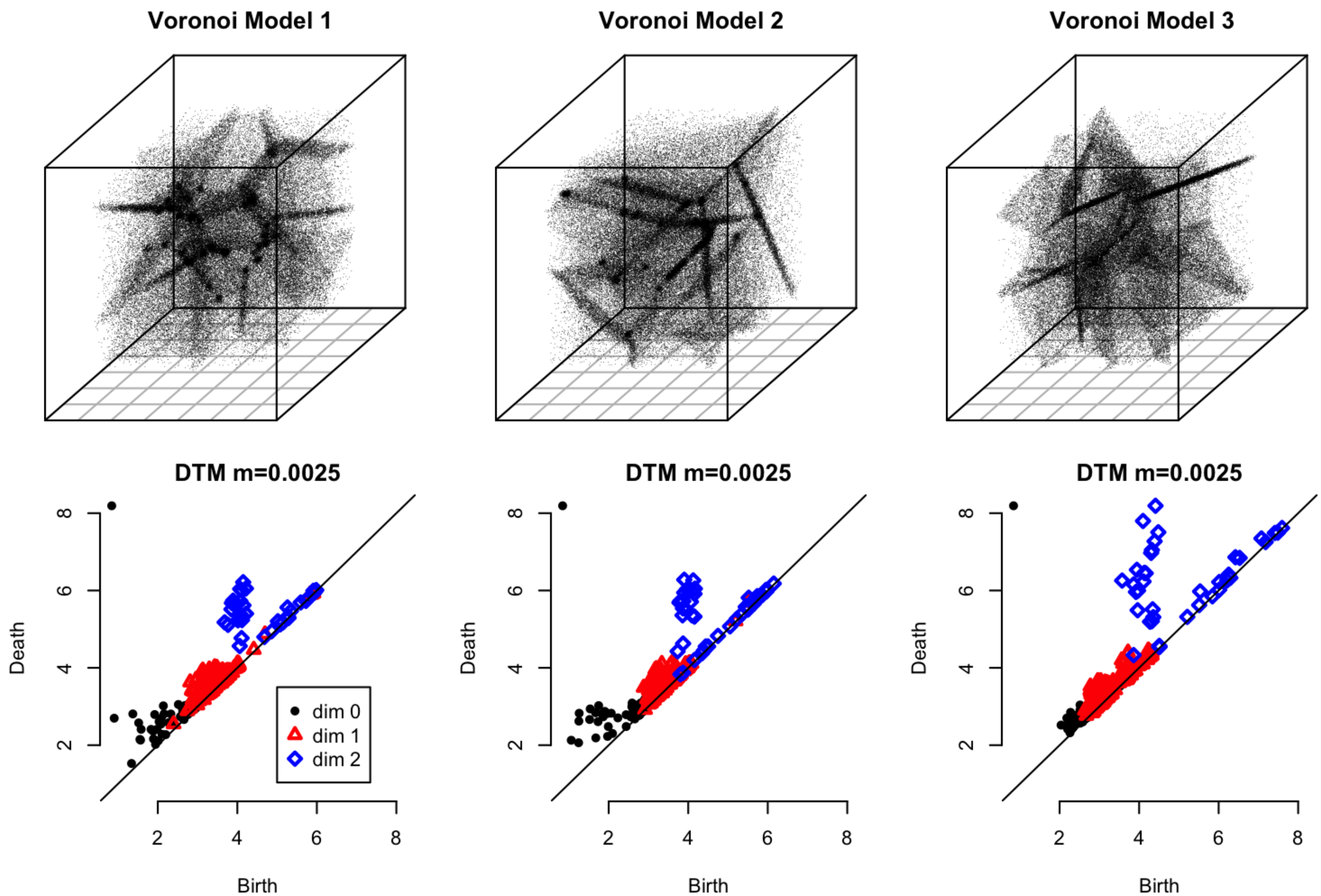
$$P_n = \frac{1}{n} \sum_{i=1}^{n} \theta_{X_i}$$

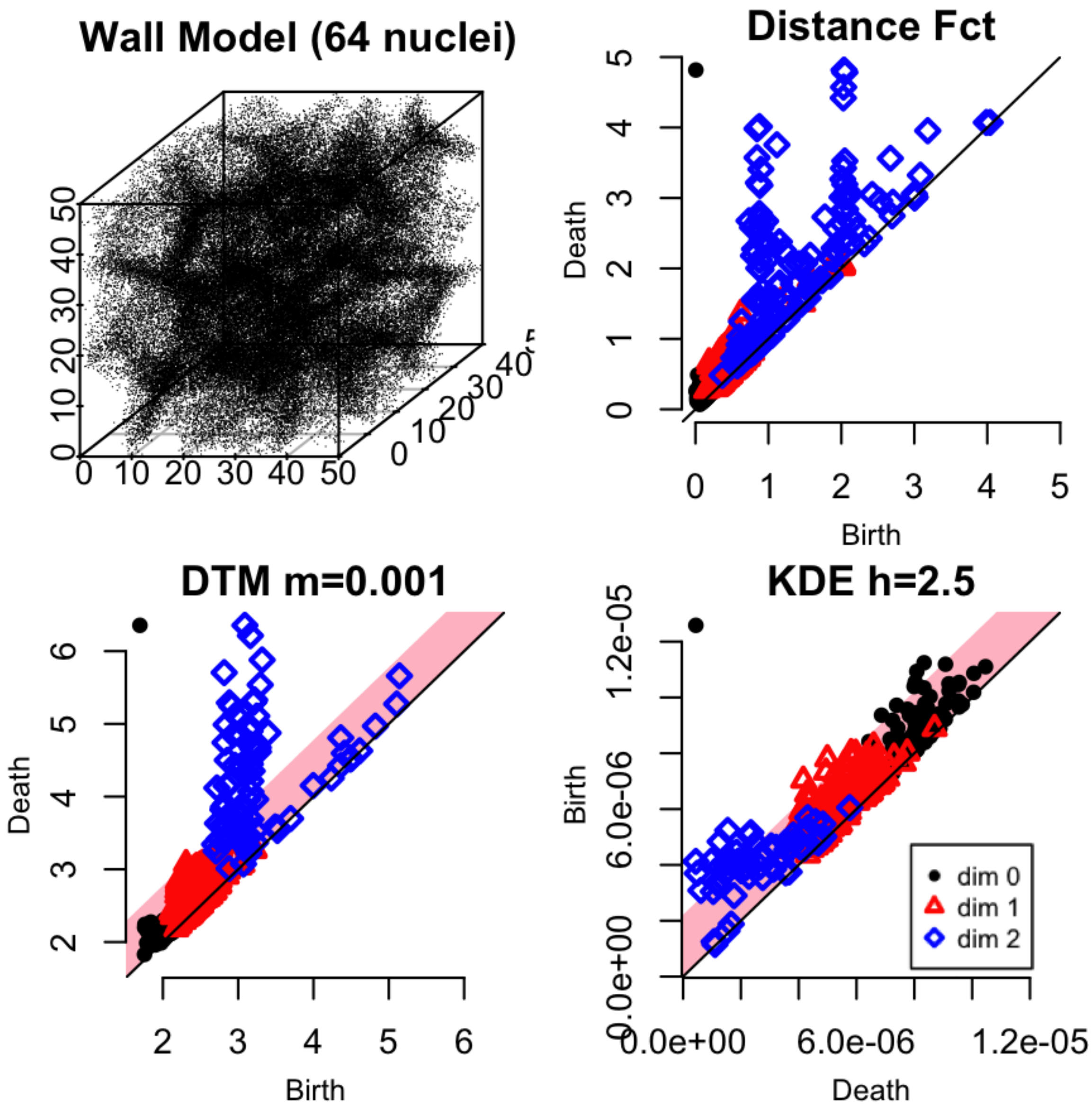
we get the plug-in estimator

$$\widehat{\delta}^2(x) = \left(\frac{1}{k_n}\right) \sum_{i=1}^{k_n} ||x - X_{(i)}||^2$$

where $k_n = mn$ and $||X_{(1)} - x|| \geq ||X_{(2)} - x|| \geq \cdots$

Voronoi foams (astronomical models): the first two have have similar topological features, the third has more voids:





Wall Model (64 nuclei)
Distance Fct
DTM m=0.001
KDE h=2.5
Birth
Death
dim 0
dim 1
dim 2

THEOREM

Under regularity conditions,

$$\sqrt{n}(\widehat{\delta}^2(x) - \delta^2(x)) \rightsquigarrow \mathbb{B}(x)$$

where $\mathbb{B}$ is a centered Gaussian process with covariance kernel

$$\kappa(x,y) = \frac{1}{m^2}\int_0^{F_x^{-1}(m)}\int_0^{F_y^{-1}(m)}\left(\mathbb{P}\left[B(x,\sqrt{t}) \cap B(y,\sqrt{s})\right] - F_x(t)F_y(s)\right) ds\, dt$$

and $F_x(t) = \mathbb{P}(||X - x||^2 \le t)$.

Recall the stability theorem:

$$\text{bottleneck}(\widehat{D}, D) \le \sup_x ||\widehat{\delta}(x) - \delta(x)||.$$

BOOTSTRAP CONFIDENCE BAND FOR δ

Draw: $X_1^*, \ldots, X_n^* \sim P_n$. Compute $\widehat{\delta}^*$. Repeat.

THEOREM: The map δ taking probability measures to DTM's is Hadamard differentiable. Hence, if we define $\widehat{c}_\alpha$ by

$$\mathbb{P}(\sqrt{n}||\widehat{\delta}^* - \widehat{\delta}||_\infty > \widehat{c}_\alpha \,|\, X_1, \ldots, X_n) = \alpha.$$

Then

$$\mathbb{P}\left(||\delta - \widehat{\delta}||_\infty \leq \frac{\widehat{c}_\alpha}{\sqrt{n}}\right) \to 1 - \alpha.$$

SIGNIFICANCE OF TOPOLOGICAL FEATURES

Confidence set for true diagram D:

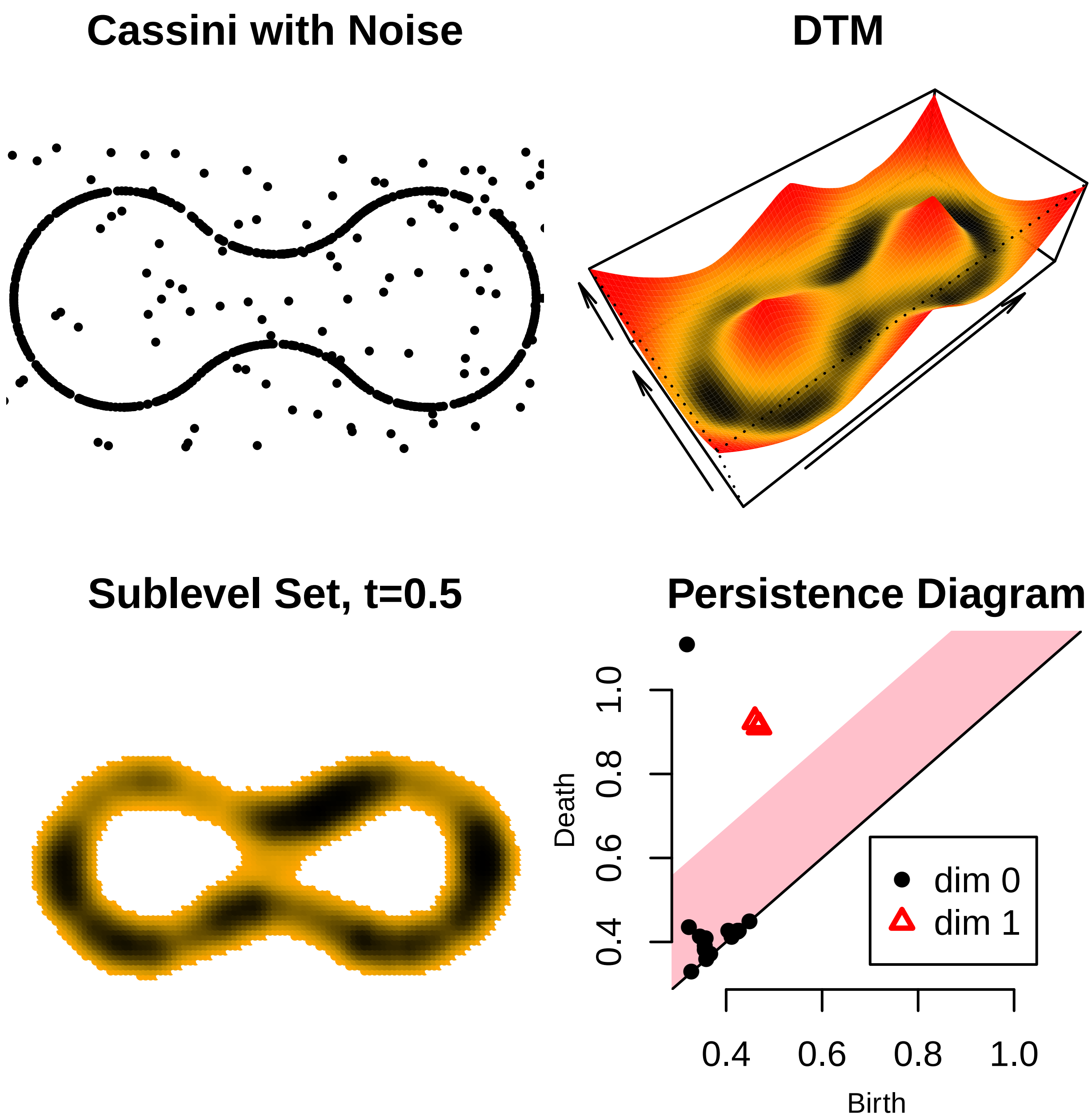
$$\mathcal{D} = \left\{ D :\ \text{bottleneck}(D, \widehat{D}) \le \frac{\widehat{c}_\alpha}{\sqrt{n}} \right\}.$$

How to display this?

Consider a feature (a point on the diagram) with birth and death time (b, d). A feature is significant if it is not matched to the diagonal for any diagram in $\mathcal{D}$ i.e. if

$$d - b > \frac{\widehat{c}_\alpha}{\sqrt{n}}.$$

We can display this by adding a “noise band” on the diagram.


Cassini with Noise
DTM
Sublevel Set, t=0.5
Persistence Diagram
Death
0.4
0.6
0.8
1.0
Birth
0.4
0.6
0.8
1.0
dim 0
dim 1

ANOTHER APPROACH: DENSITIES

$$\text{KDE}: \quad \widehat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{||x - X_i||}{h}\right)$$

which estimates $p_h(x) = \mathbb{E}[\widehat{p}_h(x)]$. The upper-level sets $\{\widehat{p}_h(x) > t\}$ define a persistence diagram $\widehat{D}$. In TDA we do not let $h > 0$. This means that the rates are $O_P(1/\sqrt{n})$.

The diagram $\widehat{D}$ of $\{\widehat{p}_h > t\}$ estimates the diagram D of $\{p_h > t\}$. Then

$$\text{bottleneck}(\widehat{D}, D) = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Alternative view:
Kernel Distances (Phillips, Wang and Zheng 2014):

$$
\begin{aligned}
D^2(P,Q) = &\int\int K_h(u,v)dP(u)dP(v) + \int\int K_h(u,v)dQ(u)dQ(v) \\
&- 2\int\int K_h(u,v)dP(u)dQ(v).
\end{aligned}
$$

Let θ_x be a point mass at x. Define

$$
\begin{aligned}
D^2(x) &\equiv D^2(P,\theta_x) \\
&= \int\int K_h(u,v)dP(u)dP(v) + K_h(x,x) - 2\int K_h(x,u)dP(u)
\end{aligned}
$$

PLUG-IN ESTIMATOR

$$\widehat{D}^2(x) = \frac{1}{n^2}\sum_i \sum_j K_h(X_i, X_j) + K_h(x,x) - \frac{2}{n}\sum_i K_h(x, X_i).$$

The lower-level sets of $\widehat{D}$ are (essentially) the same as the upper level sets of $\widehat{p}_h$.

Now we proceed as with the DTM: get diagram, bootstrap etc. (Similar limiting theorems apply.)

Technical note: $\widehat{\delta}$ estimates the persistent homology of S. $\widehat{p}$ really estimates the homology of S.

INFERENCE

The inferences are based on the stability theorem:

$$\text{bottleneck}(\widehat{D}, D) \leq ||\widehat{p}_h - p_h||_\infty.$$

Now we can construct estimate, confidence band, etc.

But: sometimes $\text{bottleneck}(\widehat{D}, D) < ||\widehat{p}_h - p_h||_\infty$.

A SHARPER LIMIT THEOREM

If we make slightly stronger assumptions, we get a better limiting result.

THEOREM:

$$\sqrt{n}\ \text{bottleneck}(\widehat{D}, D) \rightsquigarrow ||Z||_\infty$$

where, $Z \in \mathbb{R}^k$, $Z \sim N(0, \Sigma)$, and Σ is a function of the gradient and Hessian of p_h.

This sidesteps the stability theorem. It is directly about the bottleneck distance.

BOTTLENECK BOOTSTRAP

Let

$$F_n(t) = \mathbb{P}(\sqrt{n}\ \text{bottleneck}(\widehat{D}, D) \le t).$$

Let $X_1^*, \ldots, X_n^* \sim P_n$ where P_n is the empirical distribution. Let $\widehat{D}^*$ be the diagram from $\widehat{p}_h^*$ and let

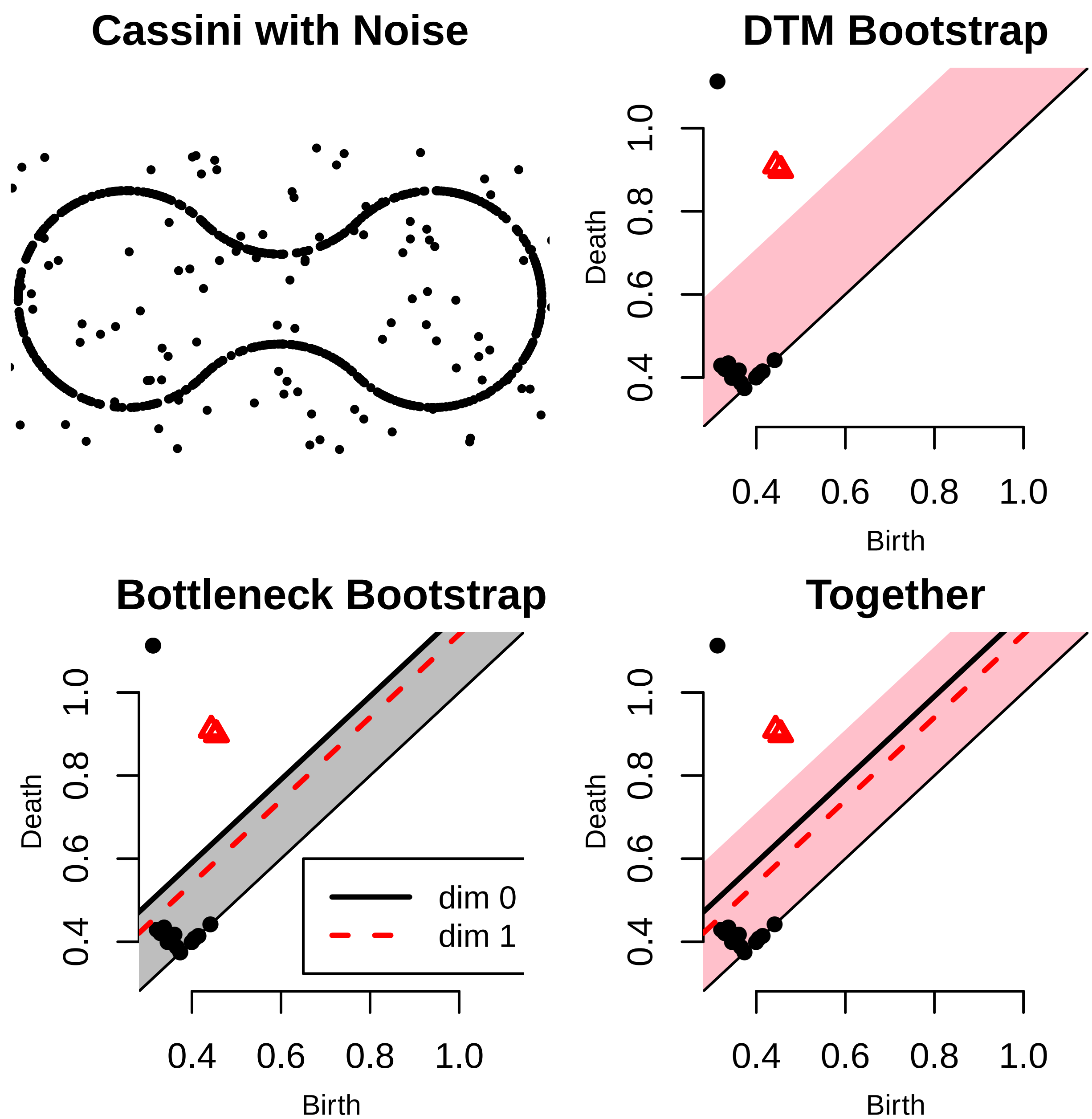
$$\widehat{F}_n(t) = \mathbb{P}(\sqrt{n}\ \text{bottleneck}(\widehat{D}^*, \widehat{D}) \mid X_1, \ldots, X_n) \le t)$$

be the bootstrap approximation to F_n.

THEOREM:

$$\sup_t |\widehat{F}_n(t) - F_n(t)| \xrightarrow{P} 0.$$

So we can use $\widehat{c}_\alpha = \widehat{F}_n(1-\alpha)/\sqrt{n}$.


Cassini with Noise
DTM Bootstrap
Death
Birth
0.4
0.6
0.8
1.0
Bottleneck Bootstrap
dim 0
dim 1
Together

TUNING PARAMETERS

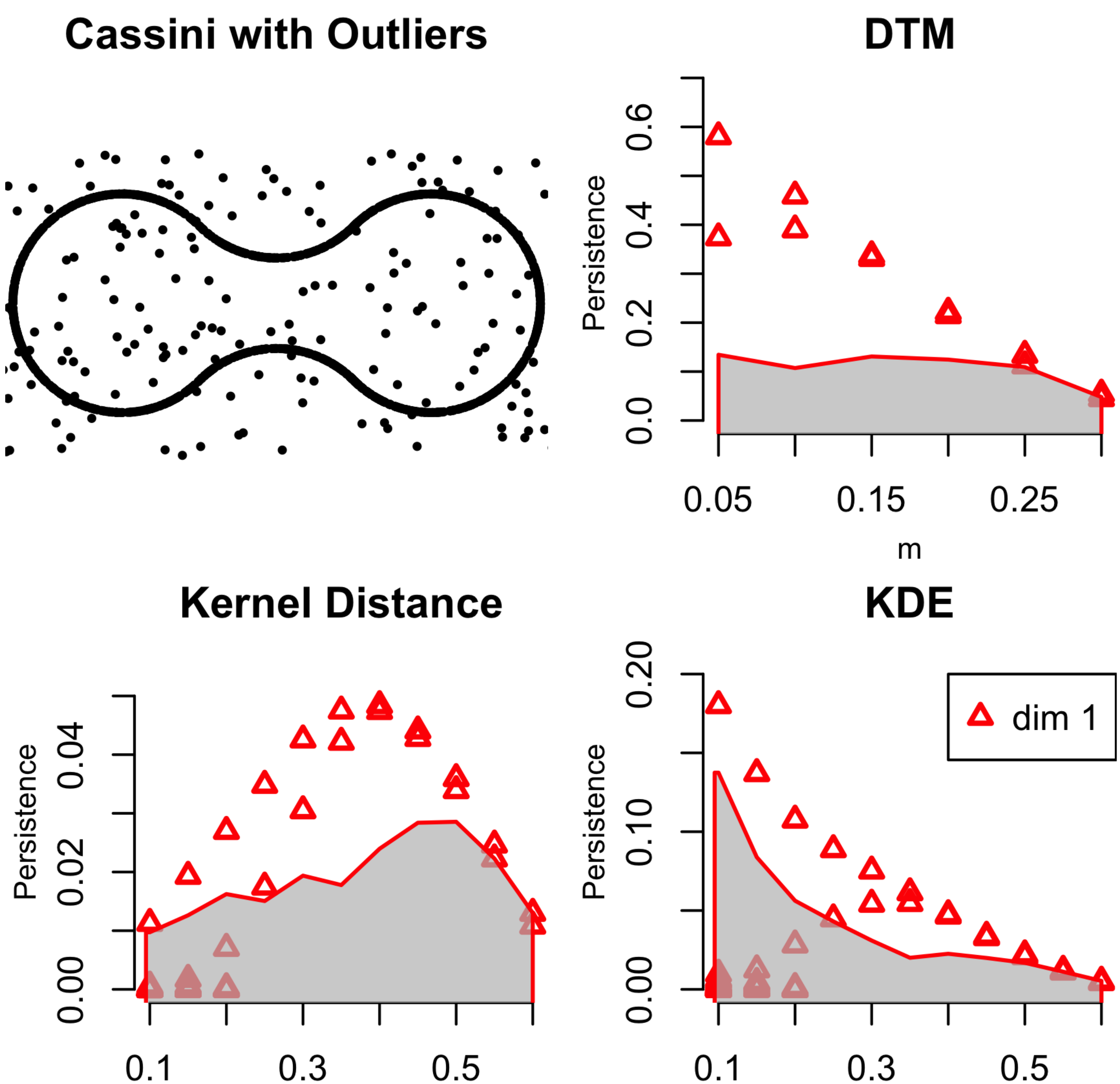
How to choose the tuning parameter: m for DTM, and h for kernels?

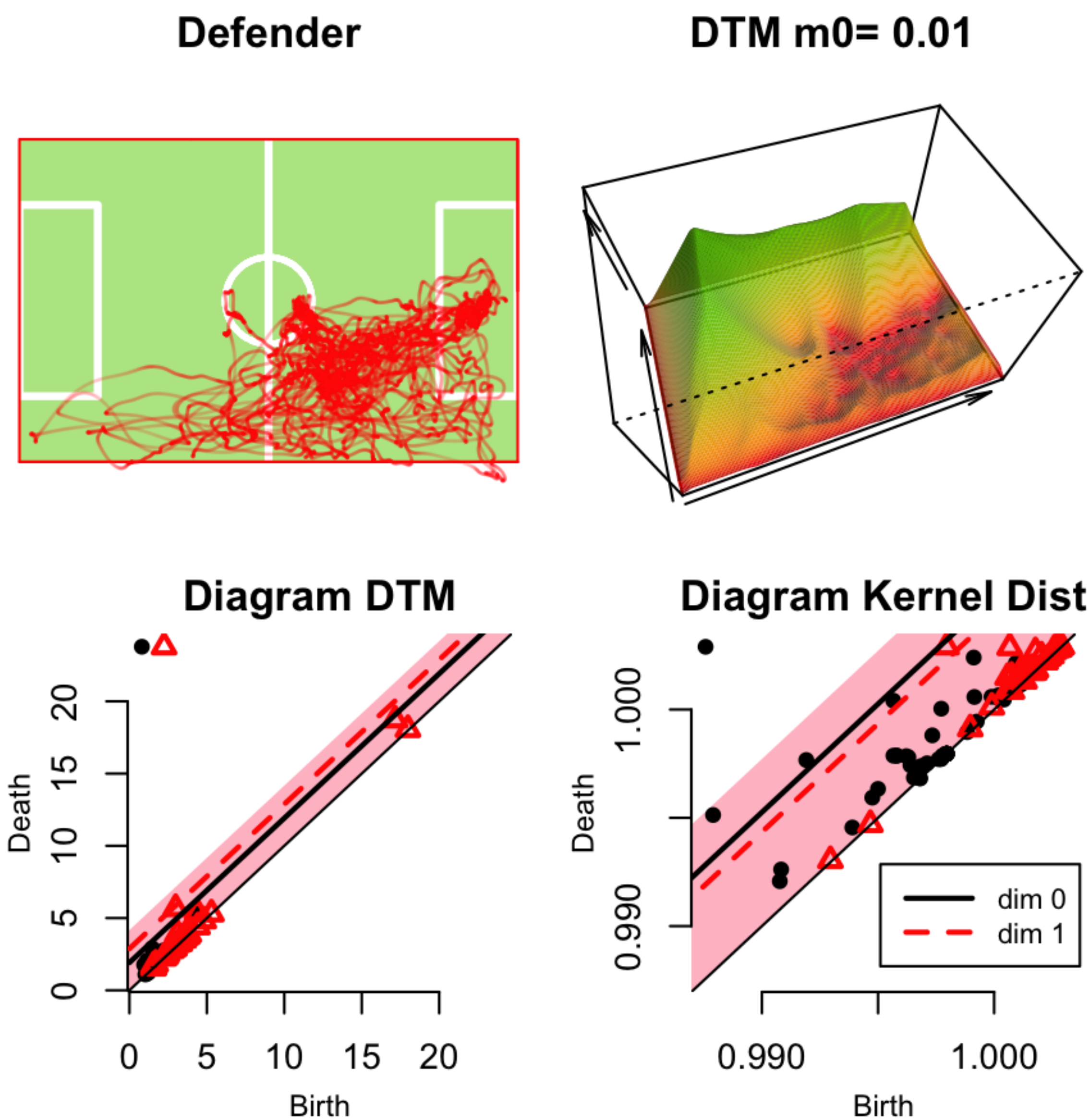
Births and deaths: $\{(b_1, d_1), \ldots, (b_k, d_m)\}$.

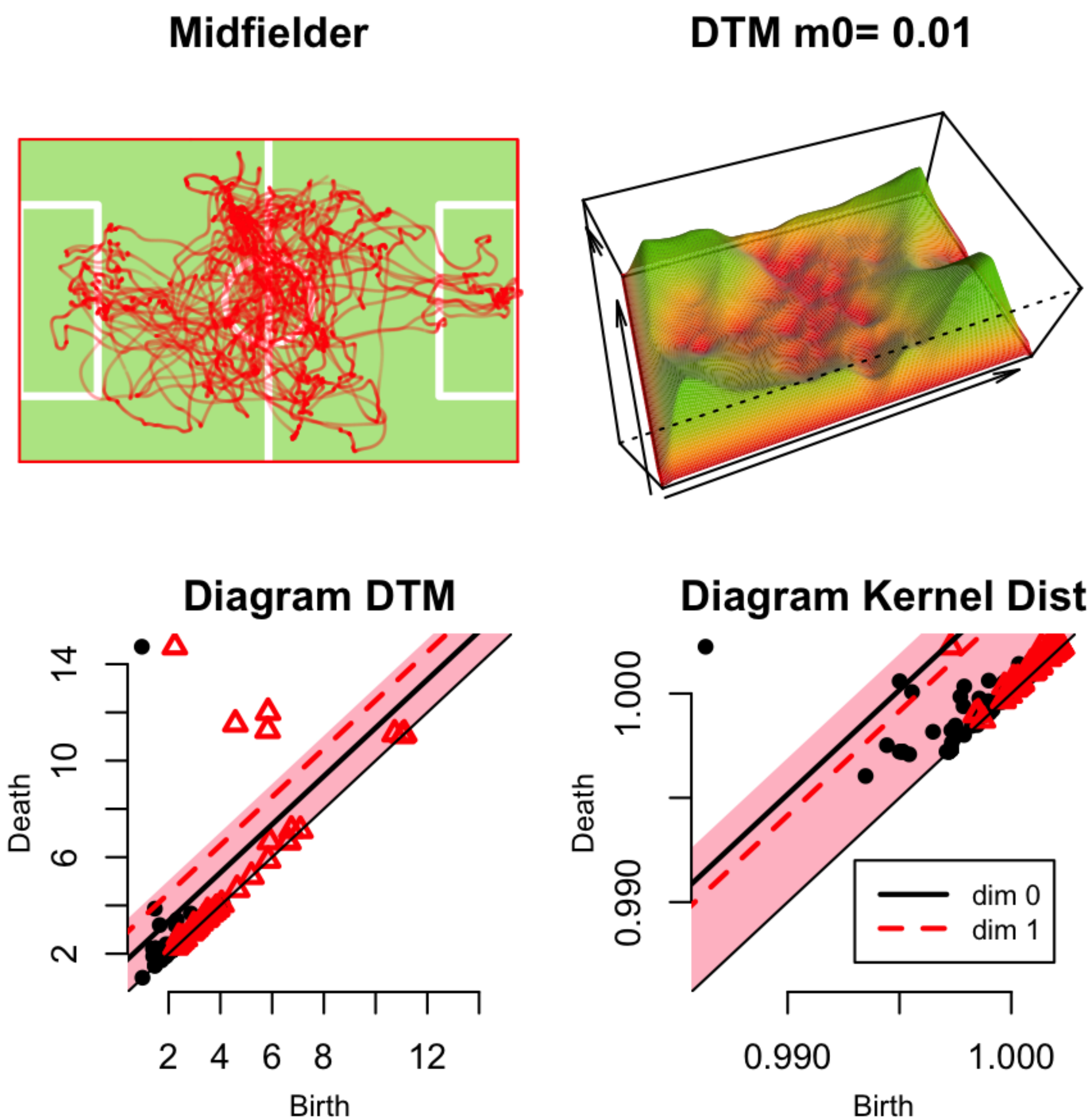
Choose parameter to maximize the number of significant features:

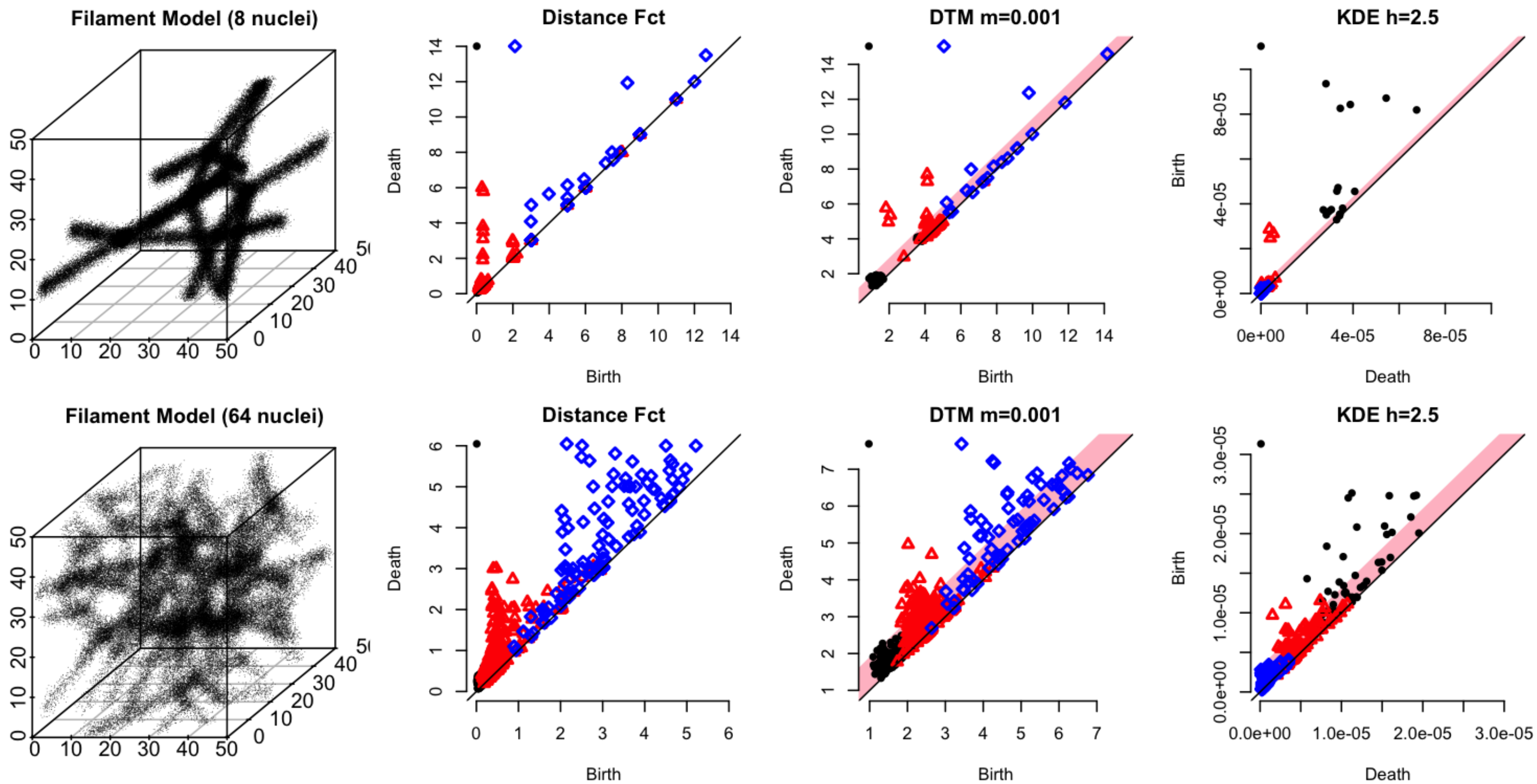
$$d_i - b_i > \frac{\widehat{c}_\alpha}{\sqrt{n}}$$

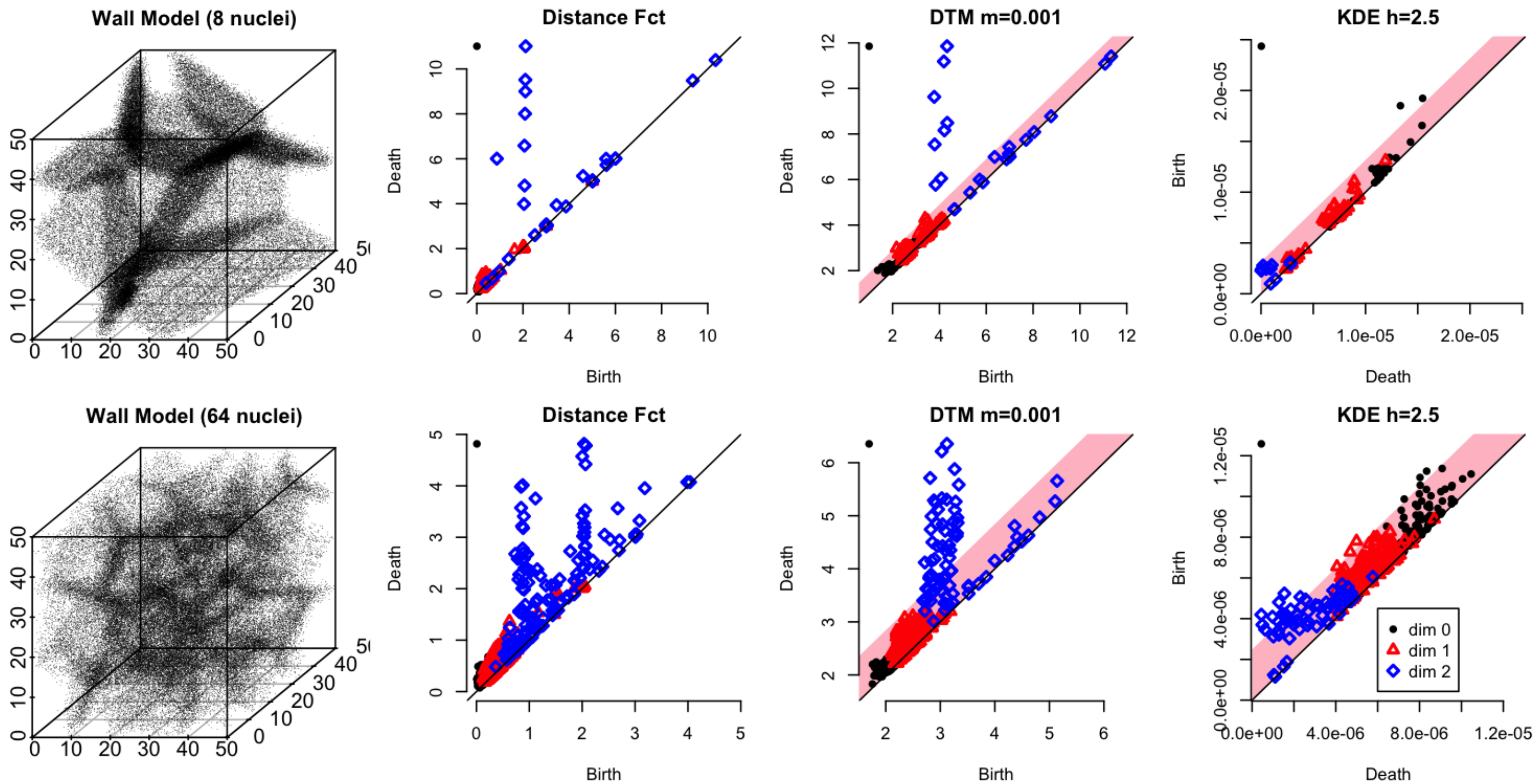
(First suggested informally in Guibas, Morozov and Merigot, 2013, without the notion of statistical significance).


Cassini with Outliers
DTM
Persistence
0.0
0.2
0.4
0.6
0.05
0.15
0.25
m
Kernel Distance
Persistence
0.00
0.02
0.04
0.1
0.3
0.5
KDE
Persistence
0.00
0.10
0.20
0.1
0.3
0.5
dim 1


Defender
DTM m0= 0.01
Diagram DTM
Death
Birth
0
5
10
15
20
Diagram Kernel Dist
Death
Birth
0.990
1.000
dim 0
dim 1


Midfielder
DTM m0= 0.01
Diagram DTM
Death
Birth
2
6
10
14
2
4
6
8
12
Diagram Kernel Dist
Death
Birth
0.990
1.000
0.990
1.000
dim 0
dim 1


Filament Model (8 nuclei)
Distance Fct
DTM m=0.001
KDE h=2.5
Filament Model (64 nuclei)
Distance Fct
DTM m=0.001
KDE h=2.5
Birth
Death


Wall Model (8 nuclei)
Distance Fct
DTM m=0.001
KDE h=2.5
Wall Model (64 nuclei)
Distance Fct
DTM m=0.001
KDE h=2.5
Birth
Death
dim 0
dim 1
dim 2

CONCLUDING REMARKS: (2 Slides)

- Boundary bias: padding or reflection.

- Two sample testing: in progress

- Applications to astronomy:
Nonparametric 3D map of the IGM using the Lyman-alpha forest. Cisewski, Croft, Freeman, Genovese, Khandai, Ozbek and Larry Wasserman. arXiv:1401.1867.
Other applications in progress.

- Computation is slow (high memory demands).
Subsampling Methods for Persistent Homology. Chazal, Fasy, Lecci, Michel, Rinaldo and Wasserman. arXiv:1406.1901

CONCLUDING REMARKS

- There are other useful "topological features."

-Peter Bubenik's landscapes.
Stochastic Convergence of Persistence Landscapes and Silhouettes.
Chazal, Fasy, Lecci, Rinaldo, Wasserman. arXiv:1312.0308

-Low dimensional, high density structures (ridges)
Nonparametric ridge estimation. Genovese, Perone-Pacifico, Verdinelli, Wasserman. (Annals 2014). arXiv:1212.5156.

- Papers can be found at www.stat.cmu.edu/topstat

- Software: also at: www.stat.cmu.edu/topstat

THANKS