Topological data analysis of the Cosmic Web

Jessi Cisewski-Kehe
Department of Statistics and Data Science
Yale University

Astro@Stats2017, University of Padova

September 8, 2017

Collaborators: Mike Wu (CS, Yale, UG 2016), Brittany Fasy (CS, Montana State), Wojciech Hellwing (Physics, Durham University), Mark R. Lovell (Physics, University of Amsterdam), Alessandro Rinaldo (Stat, CMU), Larry Wasserman (Stat, CMU)

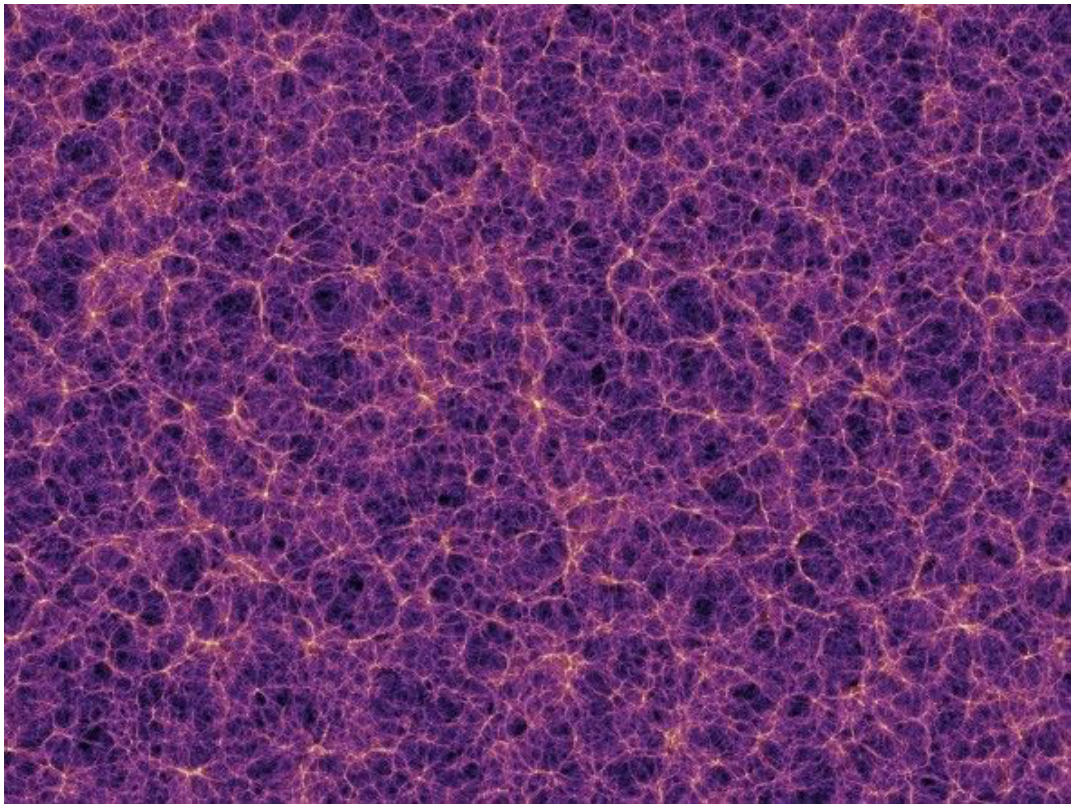
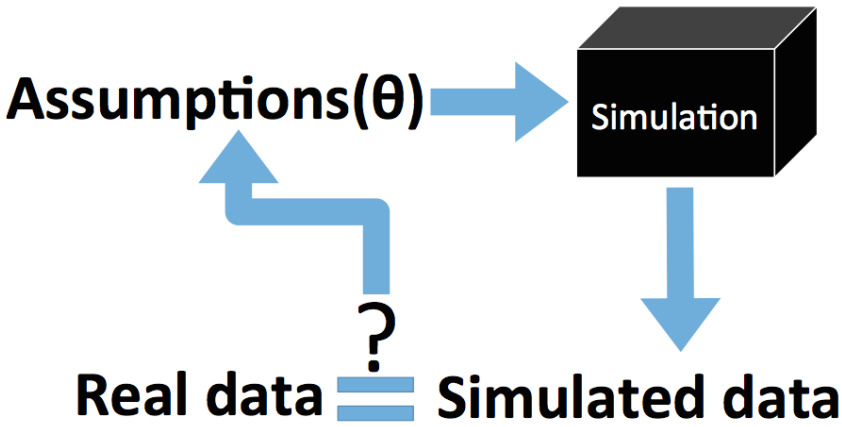


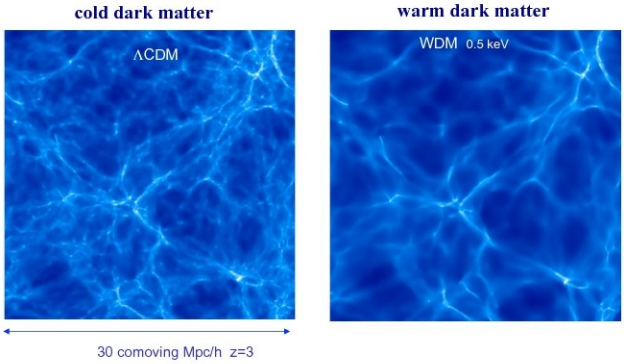
Image: Millennium simulation from Springel et al. (2005)



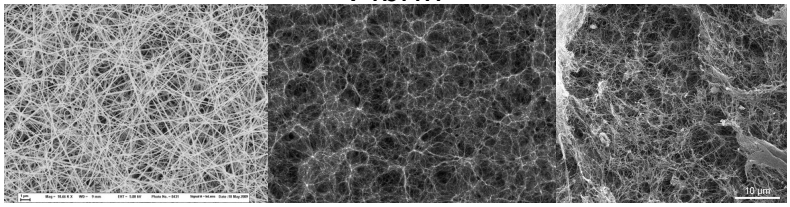
General idea: use topological summaries to visualize and compare complex data

Goal: Hypothesis Tests for spatially complex data

Large-scale structure of the Universe



Fibrin



Source: http://www.kicc.cam.ac.uk, http://www.med.unc.edu, NASA, Klumpp et al. (2011)

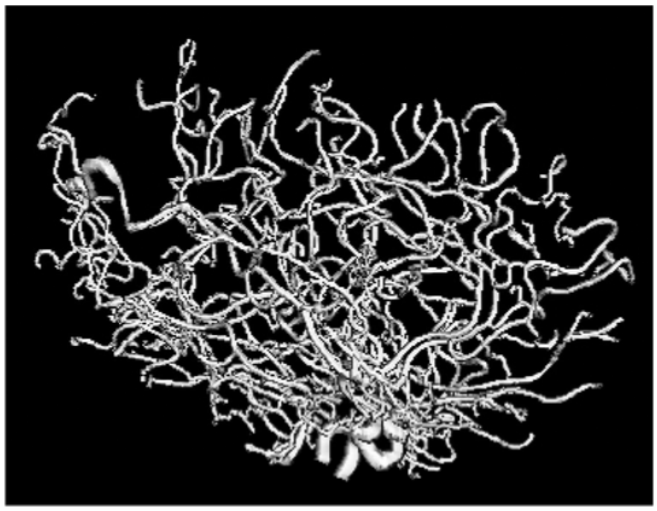


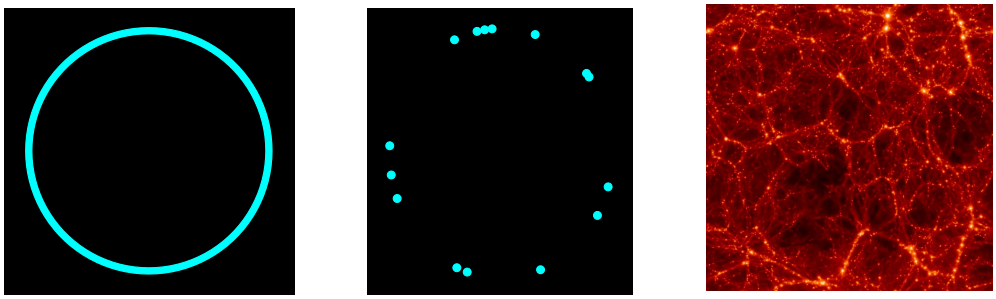
FIG. 1. *Tree of arteries from the brain of one person, showing one data object. Thickest arteries appear near the bottom. Arteries bend, twist and branch through three dimensions, which results in meaningful aspects of the data being captured by persistent homology representations. The resolution is* $0.5 \times 0.5 \times 0.8\ mm^3$.

- Bendich et al. (2016)
- 98 brain artery trees, subject ages from 18 to 72
- Looking for correlations in brain artery trees between sex and age

Homology

Looking for holes of order 0 (connected components), 1 (loops), 2 (voids), ...

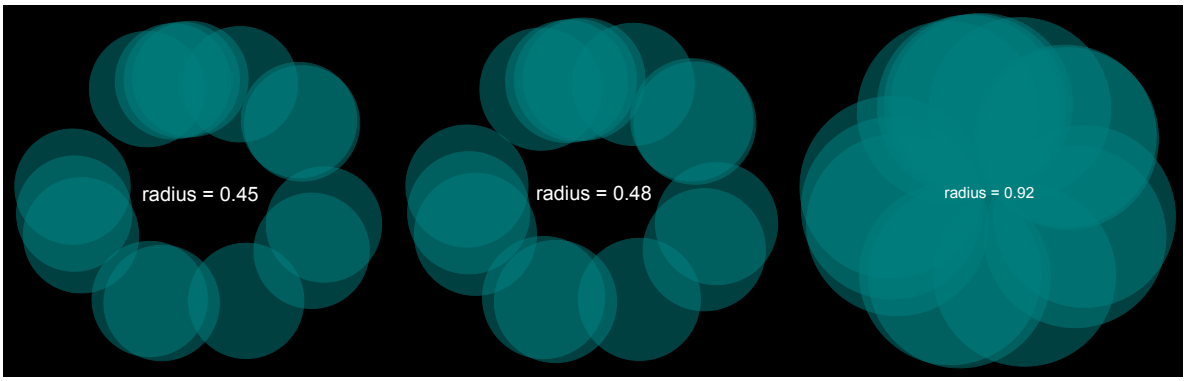
$\beta_0 = 1,\ \beta_1 = 1$ $\beta_0 = 15,\ \beta_1 = 0$ Universe



$\beta_0 = \#$ of connected components, $\beta_1 = \#$ of loops

Image: `http://astro.berkeley.edu`

Persistent homology

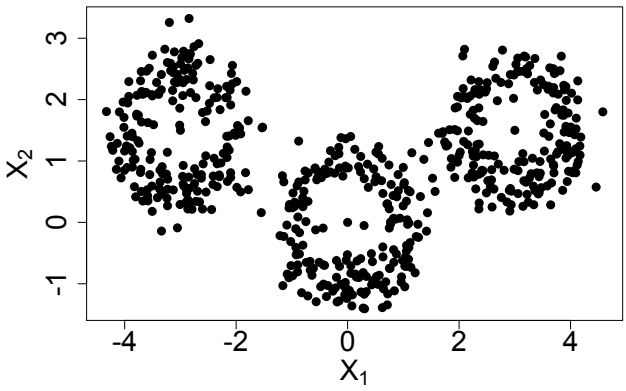


Birth of loop: radius = 0.48

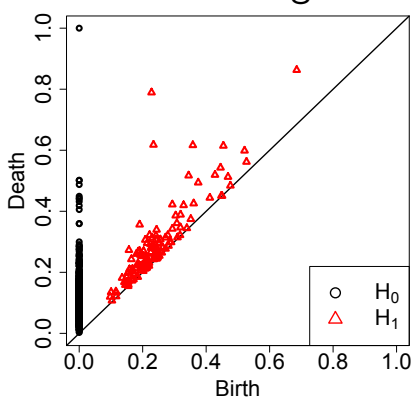
Death of loop: radius = 0.92

Persistence (or lifetime) of loop: 0.92 - 0.48 = 0.44

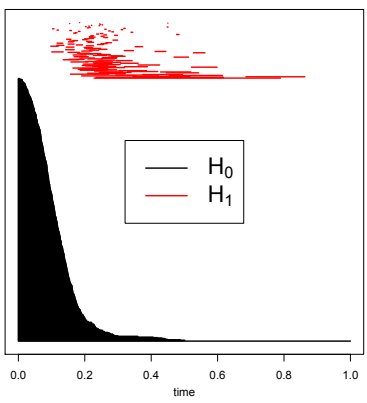
Persistent homology summaries

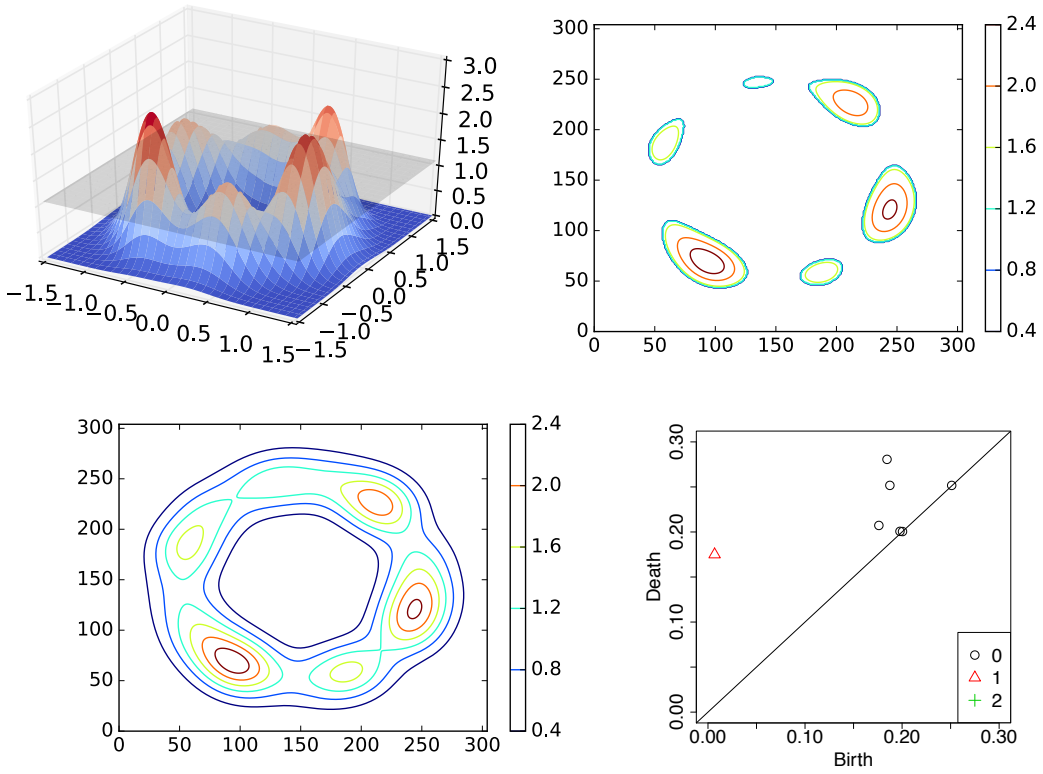


Persistence diagram

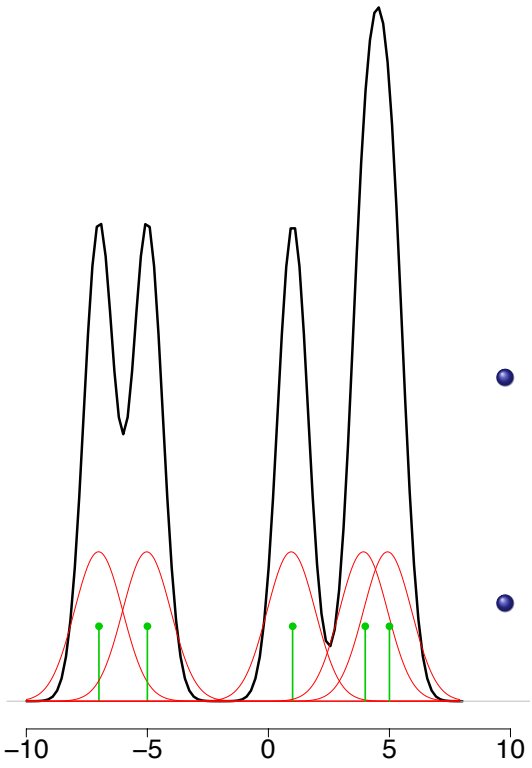


Persistence barcode plot





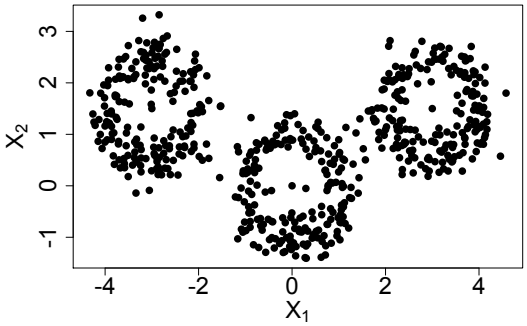
Kernel density estimation

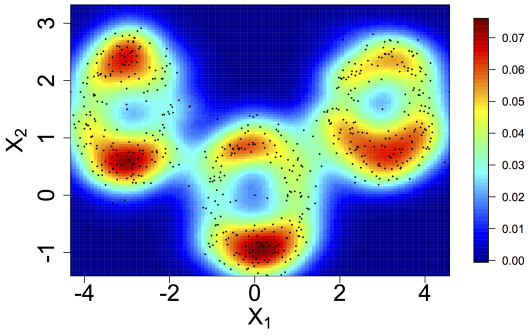


$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

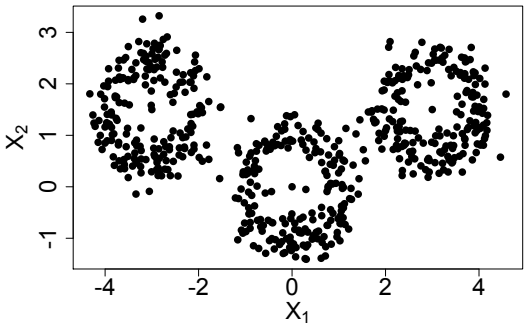
- $K(\cdot)$ is a kernel function (e.g. Gaussian density) ← Red curves
- Parameter h controls the amount of smoothing ← Width of red curves

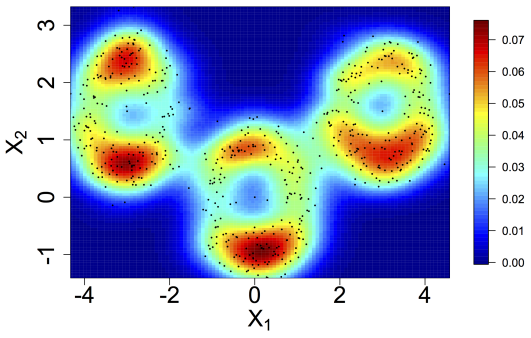
Persistent homology summaries



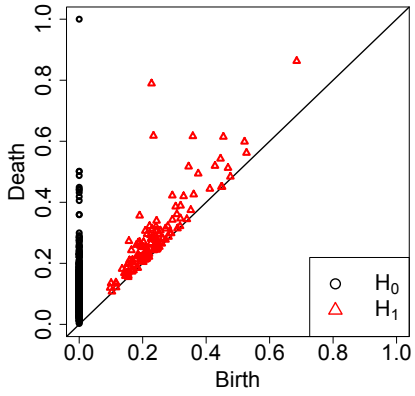


Persistent homology summaries

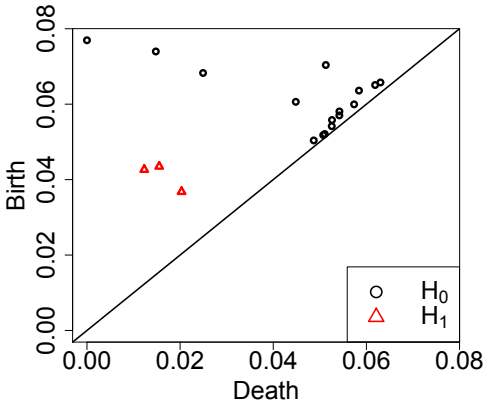




Persistence diagram (Rips)

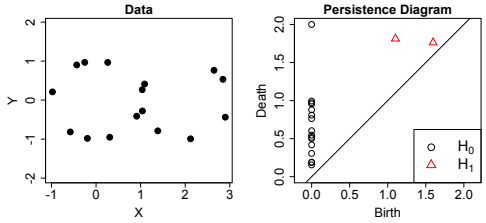


Persistence diagram (KDE)

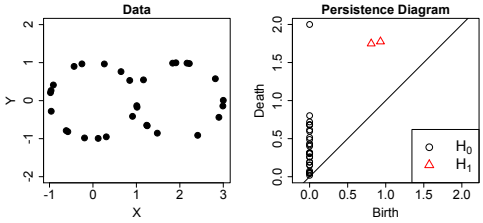


Topological randomness - I

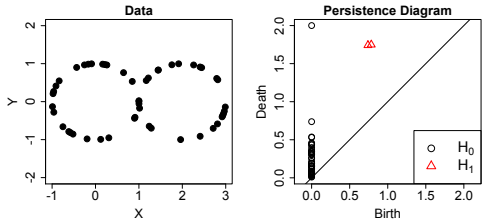
Sample size = 16



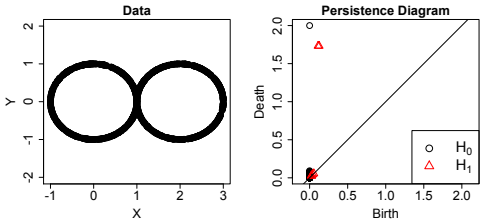
Sample size = 30



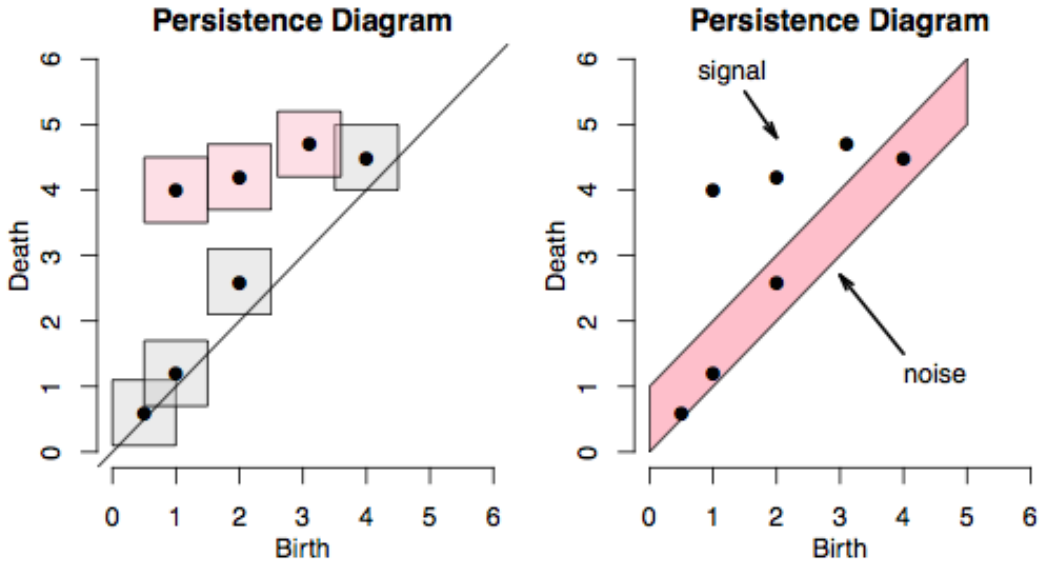
Sample size = 50



Sample size = 800



Confidence sets for persistent homology



Fasy et al. (2014)

Topological randomness - II

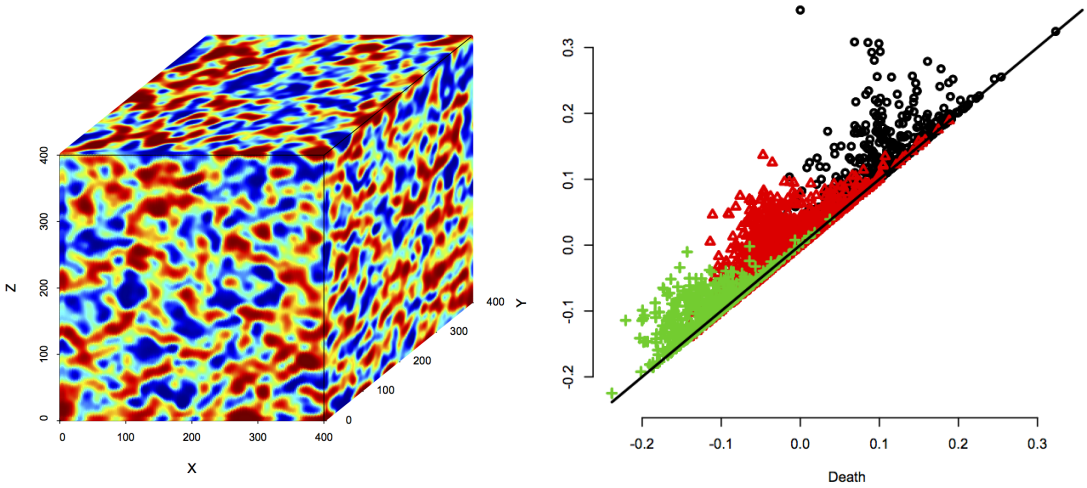
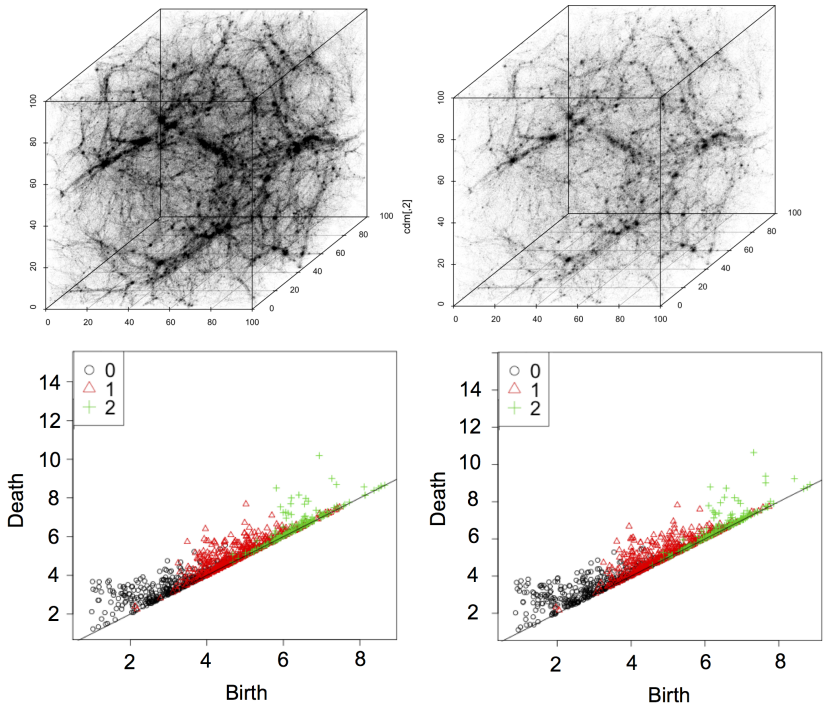


Figure: Cisewski et al. (2014)

Hypothesis testing

Cold Dark Matter (left) vs. Warm Dark Matter (right)



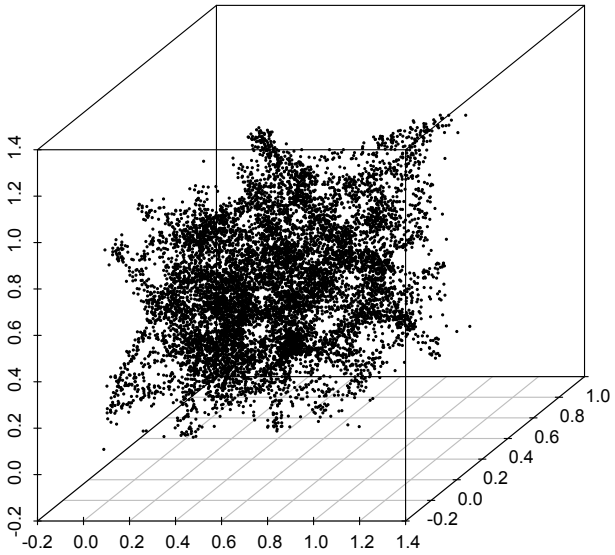
Two-sample hypothesis tests

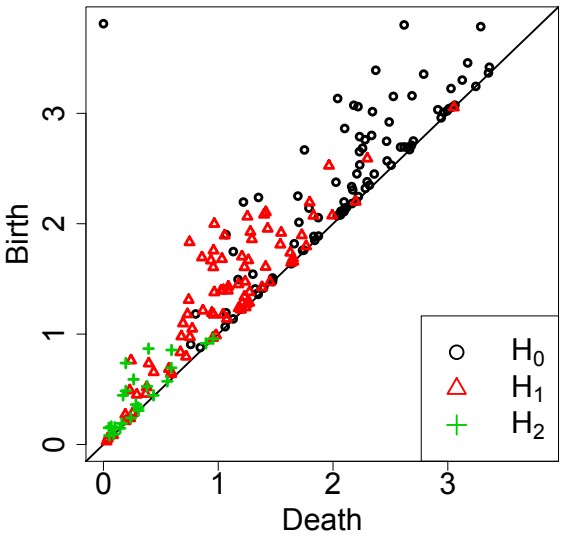
- Given two sets of persistence diagrams, $\{\mathcal{P}_1^{(1)}, \ldots, \mathcal{P}_n^{(1)}\}$ and $\{\mathcal{P}_1^{(2)}, \ldots, \mathcal{P}_m^{(2)}\}$.
- $H_0 : \mathcal{P}^{(1)} = \mathcal{P}^{(2)}$ vs. $H_1 : \mathcal{P}^{(1)} \neq \mathcal{P}^{(2)}$
  where $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are the true underlying distributions of persistence diagrams for group 1 and 2, respectively.

Two-sample hypothesis tests

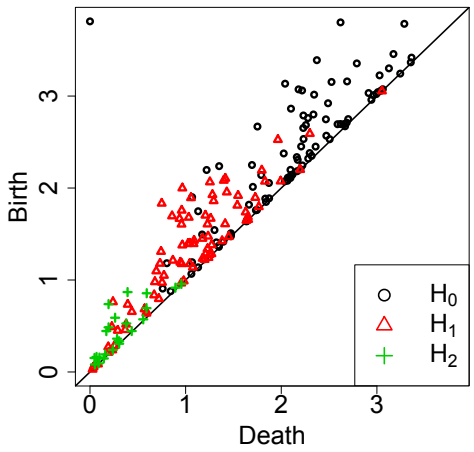
- Given two sets of persistence diagrams, $\{\mathcal{P}_1^{(1)}, \ldots, \mathcal{P}_n^{(1)}\}$ and $\{\mathcal{P}_1^{(2)}, \ldots, \mathcal{P}_m^{(2)}\}$.
- $H_0 : \mathcal{P}^{(1)} = \mathcal{P}^{(2)}$ vs. $H_1 : \mathcal{P}^{(1)} \neq \mathcal{P}^{(2)}$
  where $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are the true underlying distributions of persistence diagrams for group 1 and 2, respectively.
- What to use for the test statistic?

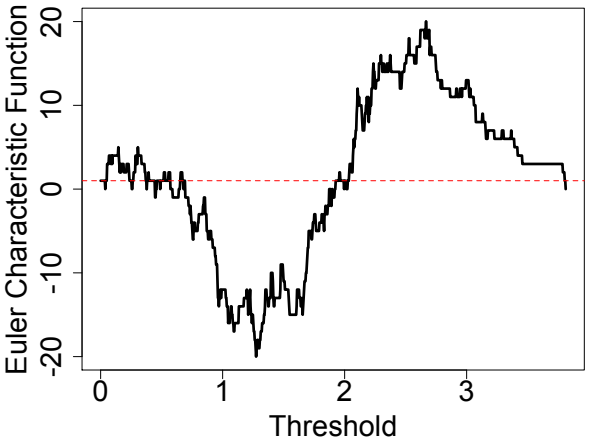
Proposed summary statistics





Proposed summary: Euler Characteristic Fn





$$\chi(t) = \beta_0(t) - \beta_1(t) + \beta_2(t)$$

for threshold t.

(alternating sum of the rank of the homology groups)

Landscape functions

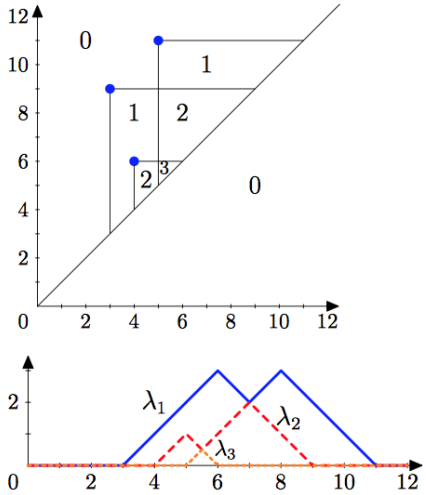
For $h = 0, 1, 2, \ldots$ (homology dim) and birth and deaths $\{b_{hi}, d_{hi}\}_{i=1}^{n}$

$$\Lambda_{p_{hi}}(t) = \begin{cases} t - b_{hi} & t \in [b_{hi}, \frac{d_{hi}+b_{hi}}{2}] \\ d_{hi} - t & t \in [\frac{d_{hi}+b_{hi}}{2}, d_{hi}] \\ 0 & \text{otherwise} \end{cases}$$

Landscape functions are the following collection of functions:

$$\lambda_{D_h}(k, t) = \underset{p_{hi} \in D_h}{\text{kmax}} \Lambda_{p_{hi}}(t),$$

$$t \in [t_{\min}, t_{\max}], k = 1, \ldots, n_h$$



Bubenik (2015)

Silhouette functions

Rather than working with each k of $\lambda_{D_h}(k, t)$ (landscapes) individually, silhouettes provide a way of combining the triangle functions:

$$\phi_h(t) = \frac{\sum_{i=1}^{m} w_{hi} \Lambda_{hi}(t)}{\sum_{i=1}^{m} w_{hi}}$$

with weights w_i. We use the recommend $w_{hi} = |d_{hi} - b_{hi}|^p$, ($p$ is a tuning parameter)

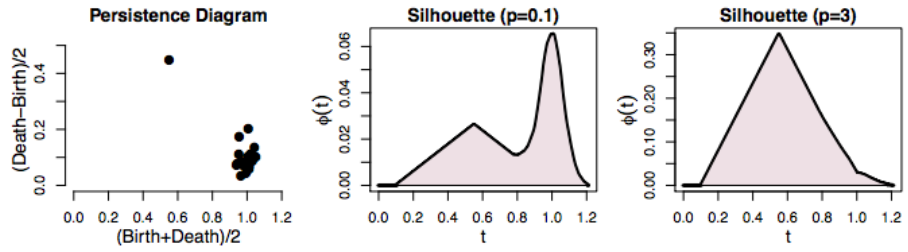
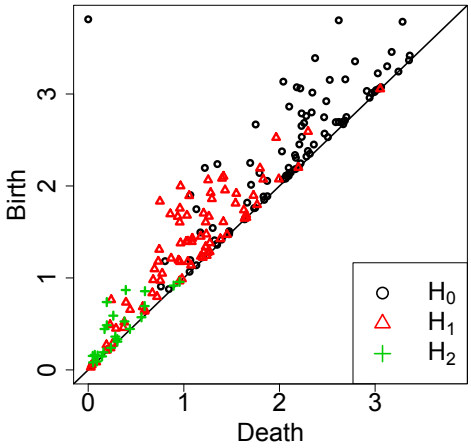
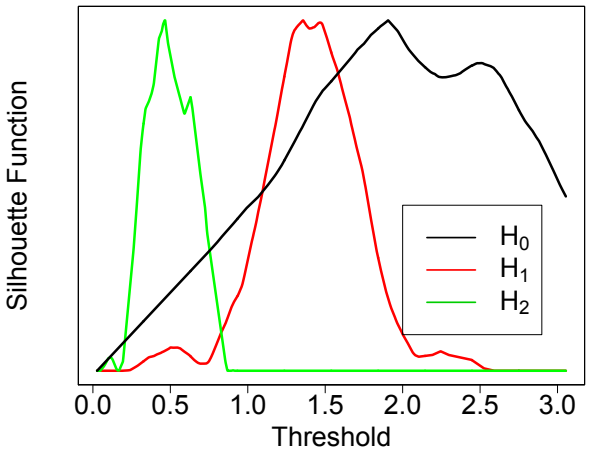


Figure 3: An example of power-weighted silhouettes for different choices of p. Note that the axes are on different scales. The weighted silhouette is one-Lipschitz.

Chazal et al. (2014)

Proposed summary: Silhouette functions





Proposed summary: Intensity function

- For each persistence diagram, estimate corresponding intensity function (Chen et al., 2015):

$$\hat{\kappa}_\tau = \sum_j (b_j - d_j) \frac{1}{\tau^2} K\left(\frac{x - b_j}{\tau}\right) K\left(\frac{x - d_j}{\tau}\right)$$

with symmetric kernel K and smoothing parameter τ

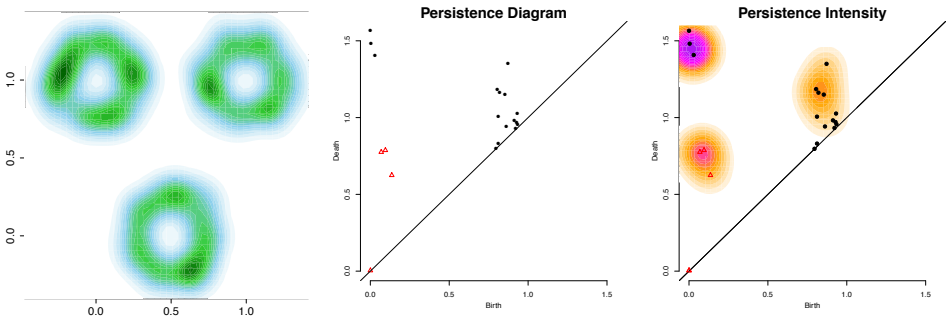
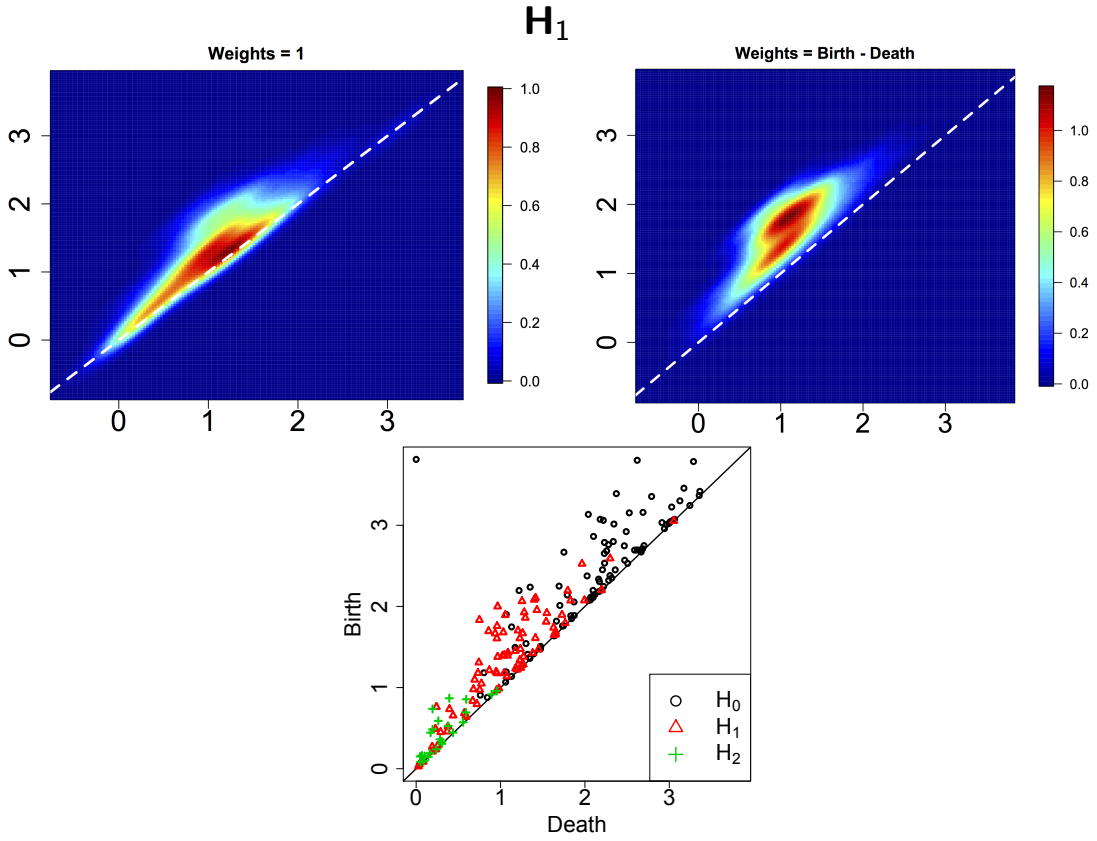


Figure 1: An example of a persistence diagram and the smoothed persistence intensity estimator constructed from a density estimator. Left: the density estimator. Middle: the persistence diagram. Each black dot is a 0-dimensional topological feature and each red triangle is a 1-dimensional topological feature. Right: the smoothed persistence intensity estimator. Note that in this case we only use the topological feature of dimension 0 to compute the intensities. i.e. the connected components.

- Related idea: *Persistent images* (Adams et al., 2015)

Proposed summary: Intensity functions



- Perform a (permutation) kernel test (Gretton et al., 2012)
  For intensity functions $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$

$$T = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_h(X_i, X_j) - \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} K_h(X_i, Y_j) + \frac{1}{m^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_h(Y_i, Y_j),$$

  $K_h(X, Y)$ can be thought of as a similarity measure between intensity functions X and Y

  E.g. Gaussian kernel $K_h(X, Y) = exp(-\frac{||X-Y||^2}{h^2})$

- Perform a (permutation) kernel test (Gretton et al., 2012)
  For intensity functions $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$

$$T = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_h(X_i, X_j) - \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} K_h(X_i, Y_j) + \frac{1}{m^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_h(Y_i, Y_j),$$

  $K_h(X, Y)$ can be thought of as a similarity measure between intensity functions X and Y

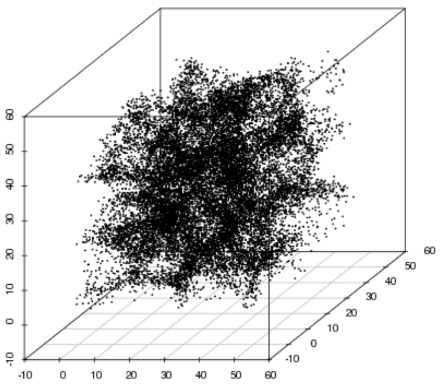
  E.g. Gaussian kernel $K_h(X, Y) = exp(-\frac{||X-Y||^2}{h^2})$

- For the tests based on functional summaries, $\phi_i(t)$, for each persistence diagram i, we estimate

$$\widehat{x_i} = \int_{t_{\min}}^{t_{\max}} |\phi_i(t)| \, \mathrm{d}t.$$

  Then do a two-sample t-test.

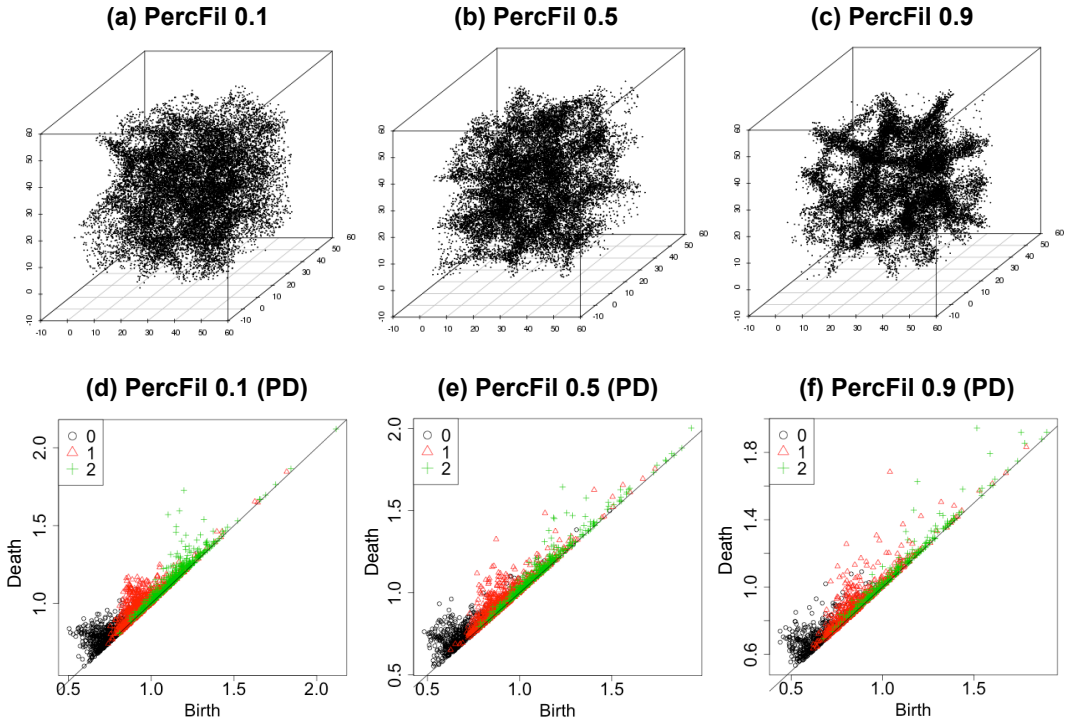
Large-scale structure (LSS) simulation model*



Abbrev	Definition	Value
percWall	Percentage of particles on the walls	$0.98 - p_f$
percFil	Percentage of particles on the filaments	p_f
percClust	Percentage of particles in the clusters	0.02

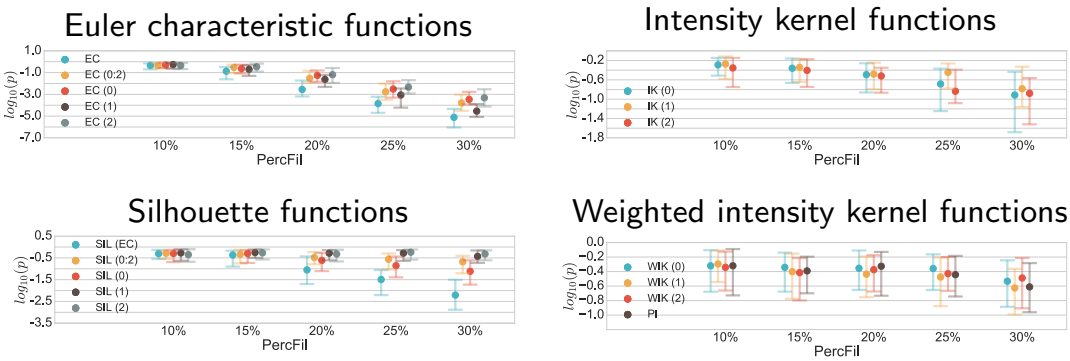
Table: Parameters of LSS model. For the simulation study, p_f will vary from 0.1 to 0.9.

*Variation of "Voronoi Foam Model" of Icke and van de Weygaert (1991)

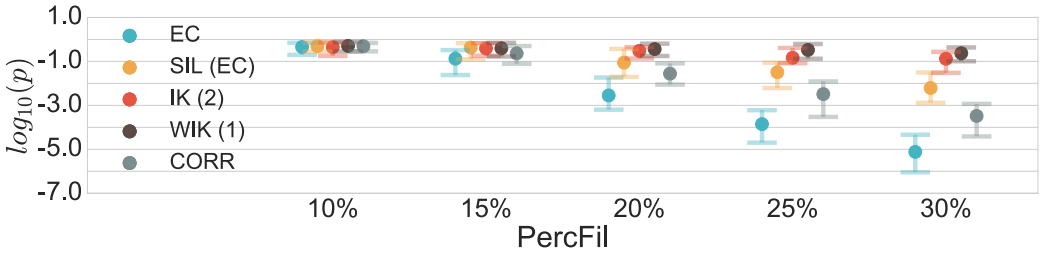


Simulation study details

- Vary p_f (% of points on filaments) from 0.1 to 0.9; $p_f = 0.1$ is the null model
- 15 datasets in each sample
- 1000 repetitions



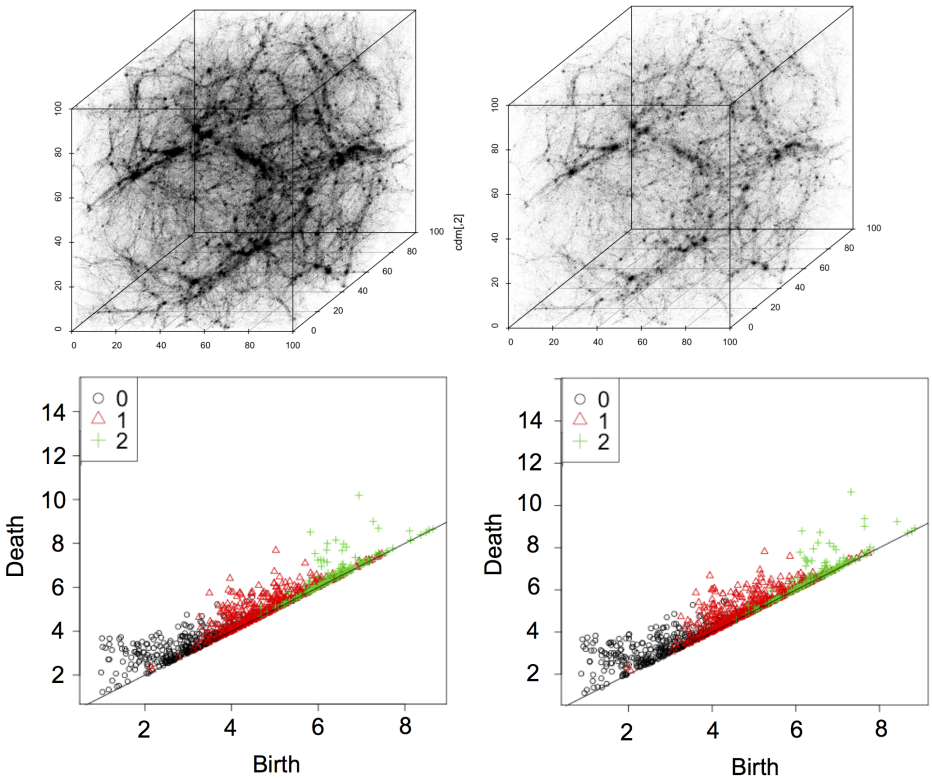
Comparison of the best performers within each group



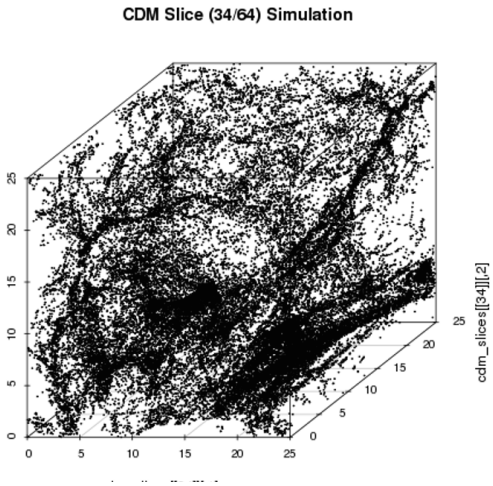
- Median $\log_{10}$ p-value $\pm$ 25th and 75th percentiles of the 1000 iterations
- The horizontal axis represents the true PercFil (10%, 15%, 20%, 25%, 30%), compared to the null PercFil of 10%
- The vertical axis shows the $\log_{10}$(p-values)

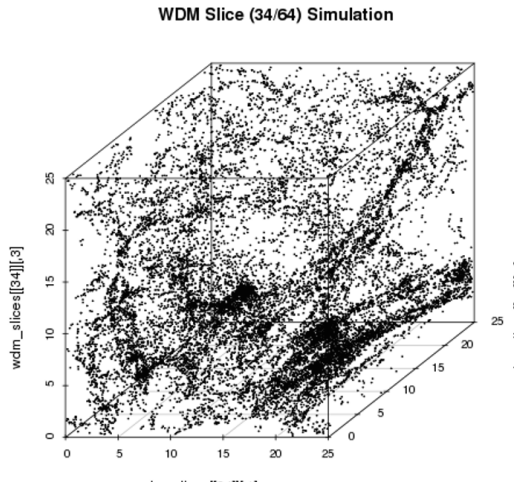
Dark Matter Simulations

Cold DM (left) vs. Warm DM (right)



- Only have one realization of each cosmological simulation
  - we divide each cube into 2^3 sub-cubes and 2^4 sub-cubes
  - do matched-pairs t-test





Test	2^3 Sub-cubes	4^3 Sub-cubes
EC	1.2e-06	7.4e-26
EC(0:2)	2.1e-05	7.4e-28
EC(0)	3.3e-08	3.8e-30
EC(1)	1.8e-05	8.2e-21
EC(2)	0.340	0.088
Sil(EC)	7.7e-08	2.5e-20
Sil(0:2)	1.9e-06	1.1e-33
Sil(0)	3.0e-08	1.5e-34
Sil(1)	1.2e-05	2.9e-23
Sil(2)	0.925	0.035
CORR	6.7e-04	7.4e-16

Summary

- Spatially complex data is becoming more common in science (e.g. Cosmic Web, fibrin, brain artery trees)

  However, analyzing these data is not straightforward

- We explore several summary statistics derived from persistence diagrams to carryout two sample hypothesis tests

- The test based on the Euler Characteristic function performs the best in the simulation study with large-scale structure

Summary

- Spatially complex data is becoming more common in science (e.g. Cosmic Web, fibrin, brain artery trees)

  However, analyzing these data is not straightforward

- We explore several summary statistics derived from persistence diagrams to carryout two sample hypothesis tests

- The test based on the Euler Characteristic function performs the best in the simulation study with large-scale structure

Thank you!

Bibliography I

Adams, H., Chepushtanova, S., Emerson, T., Hanson, E., Kirby, M., Motta, F., Neville, R., Peterson, C., Shipman, P., and Ziegelmeier, L. (2015), "Persistent images: A stable vector representation of persistent homology," *arXiv preprint arXiv:1507.06217*.

Bendich, P., Marron, J. S., Miller, E., Pieloch, A., Skwerer, S., et al. (2016), "Persistent homology analysis of brain artery trees," *The Annals of Applied Statistics*, 10, 198–218.

Bubenik, P. (2015), "Statistical topological data analysis using persistence landscapes," *Journal of Machine Learning Research*, 16, 77–102.

Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., and Wasserman, L. (2014), "Stochastic convergence of persistence landscapes and silhouettes," in *Proceedings of the thirtieth annual symposium on Computational geometry*, ACM, p. 474.

Chen, Y.-C., Wang, D., Rinaldo, A., and Wasserman, L. (2015), "Statistical analysis of persistence intensity functions," *arXiv preprint arXiv:1510.02502*.

Cisewski, J., Croft, R. A. C., Freeman, P. E., Genovese, C. R., Khandai, N., Ozbek, M., and Wasserman, L. (2014), "Non-parametric 3D map of the intergalactic medium using the Lyman-alpha forest," *Monthly Notices of the Royal Astronomical Society*, 440, 2599–2609.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., Singh, A., et al. (2014), "Confidence sets for persistence diagrams," *The Annals of Statistics*, 42, 2301–2339.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012), "A kernel two-sample test," *The Journal of Machine Learning Research*, 13, 723–773.

Icke, V. and van de Weygaert, R. (1991), "The galaxy distribution as a Voronoi foam," *Quarterly Journal of the Royal Astronomical Society*, 32, 85–112.

Robinson, A. and Turner, K. (2013), "Hypothesis Testing for Topological Data Analysis," *arXiv preprint arXiv:1310.7467*.