Topological Data Analysis: Inference for spatially complex data

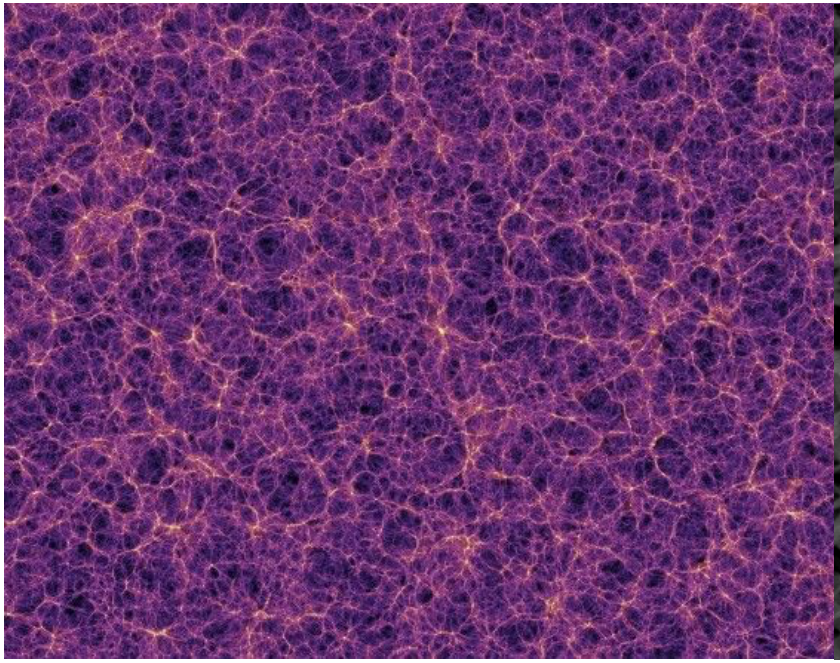
Jessi Cisewski-Kehe
Department of Statistics and Data Science
Yale University

Software Day
Department of Mathematics, College of Charleston

May 18, 2019

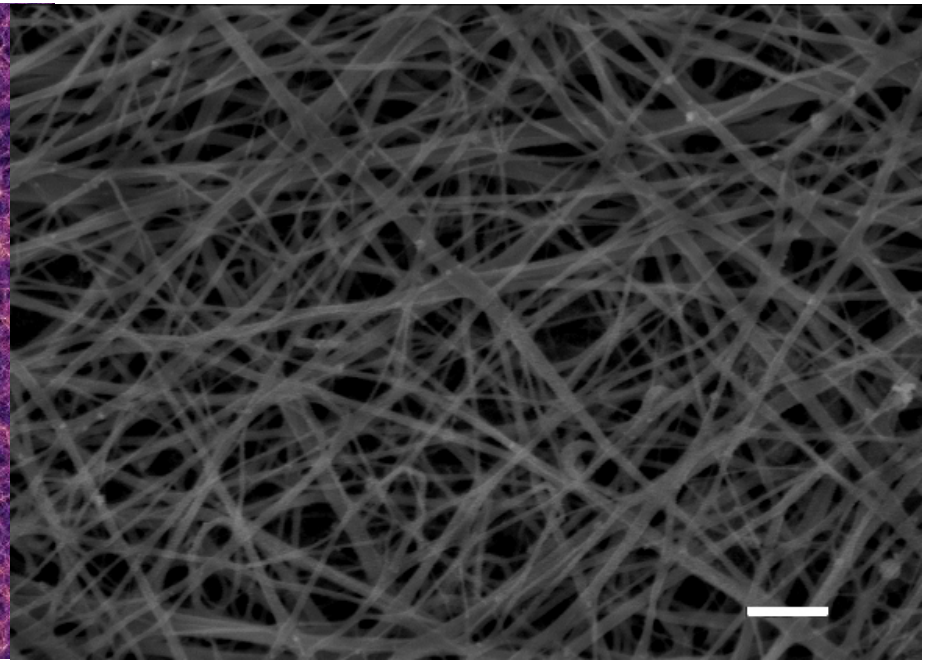
Spatially complex data

Universe



Millennium simulation from Springel et al. (2005)

Human fibrin network



Pretorius et al. (2009); the white scale bar is 1 μm

Outline

- Motivation: spatially complex data
- Quick overview of persistent homology using the `R TDA` package (see Jisu Kim's talk for more details)

  Rips filtration

  Function-based filtration
- Hypothesis testing with persistent homology

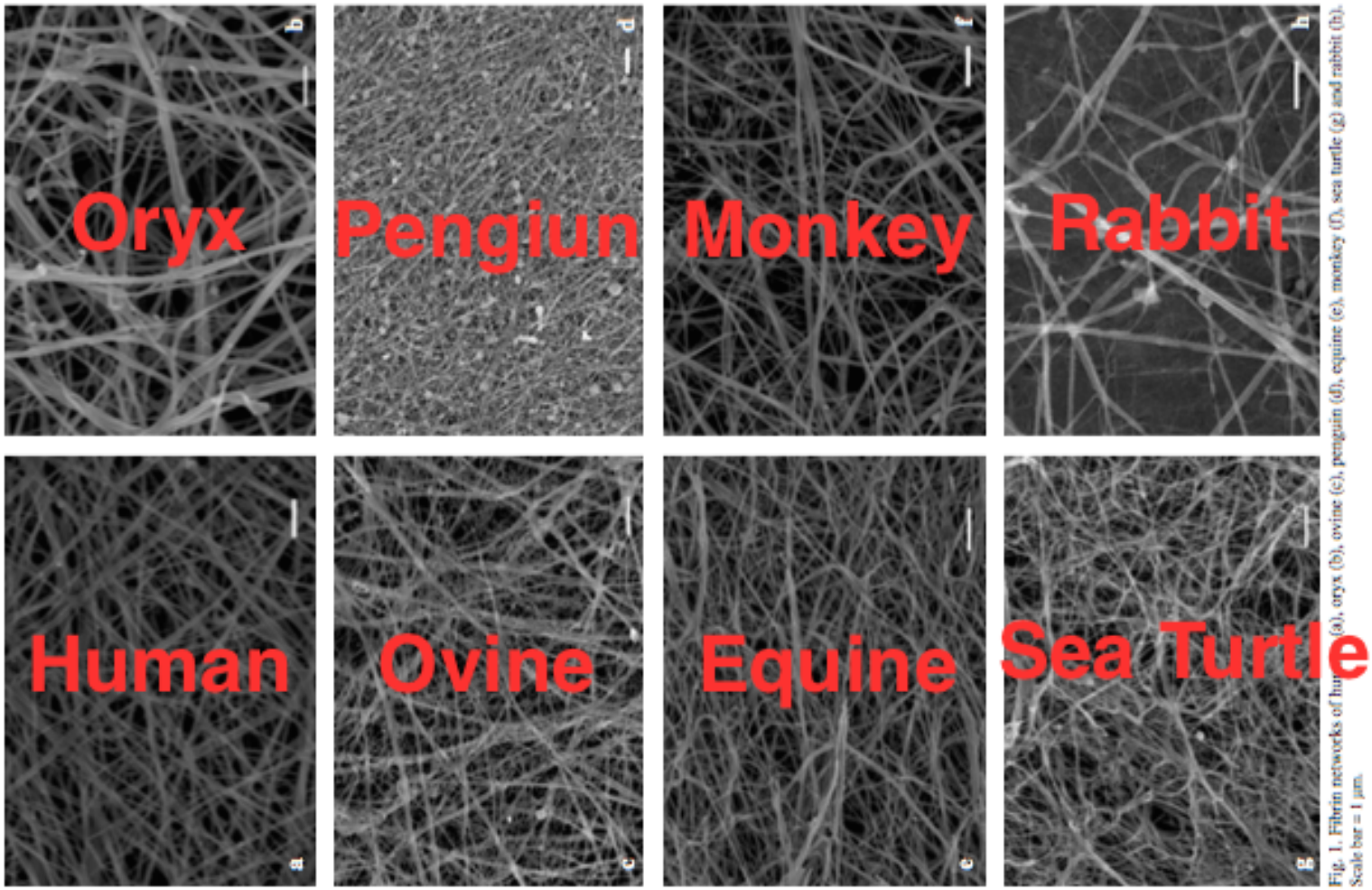
  Functional Summaries of Persistence Diagrams

TDA: persistent homology

Fibrin

Goal: Hypothesis Tests for spatially complex data

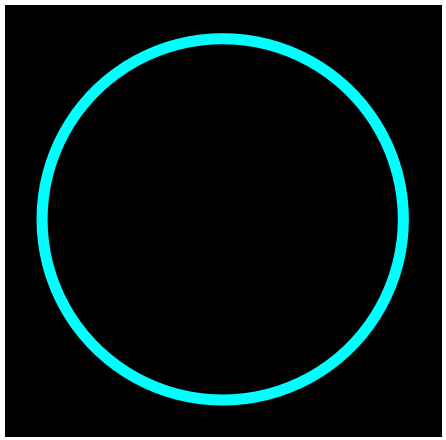
Human vs. Monkey fibrin



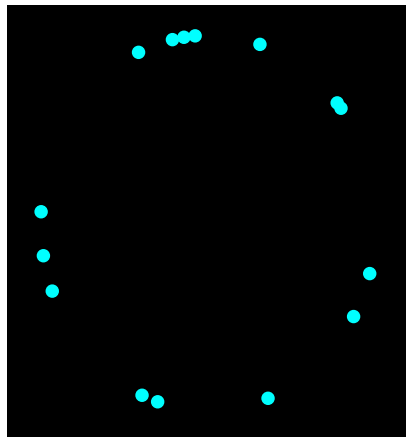
Pretorius et al. (2009)

Homology: considering data

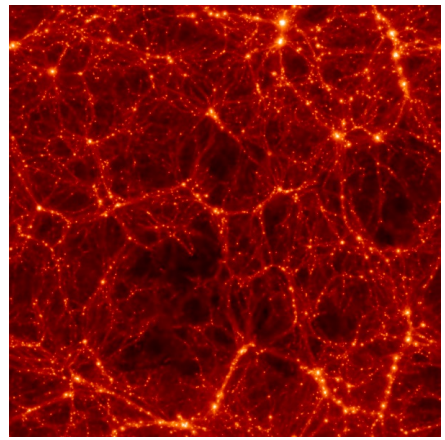
$\beta_0 = 1,\ \beta_1 = 1$



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



Universe



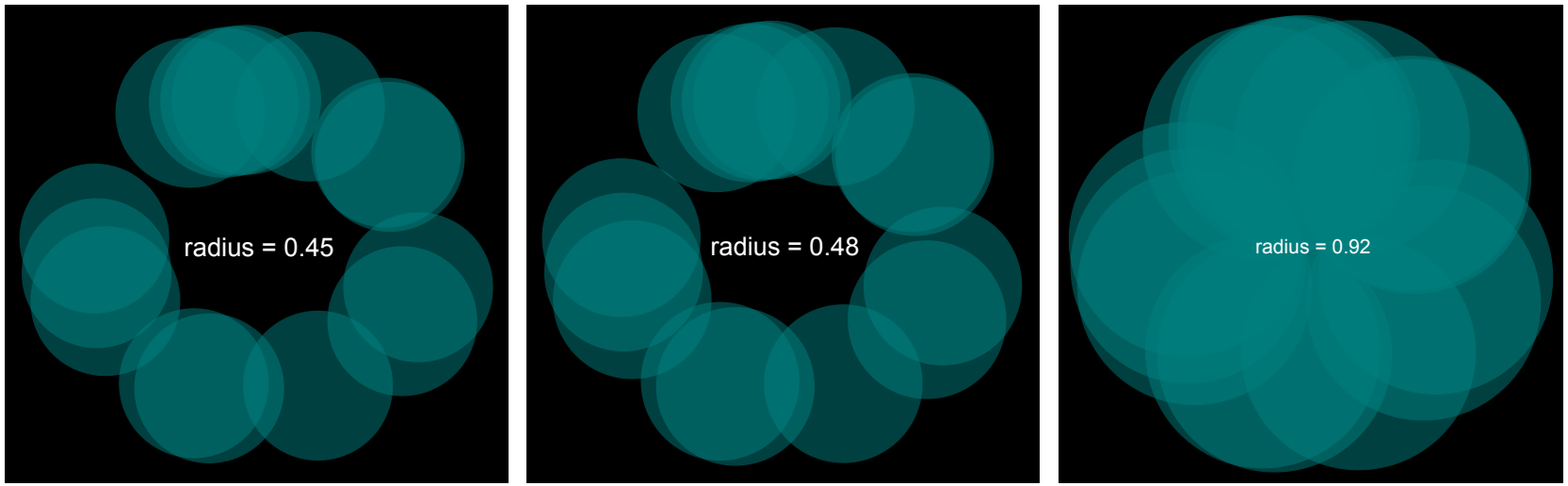
$\beta_0 = \#$ of connected components
$\beta_1 = \#$ of loops

Persistent homology is a multi-scale version of homology

(e.g., Edelsbrunner et al. 2002; Edelsbrunner and Harer 2008; Carlsson 2009)

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

Persistent homology: Rips filtration



Birth of loop: radius = 0.48

Death of loop: radius = 0.92

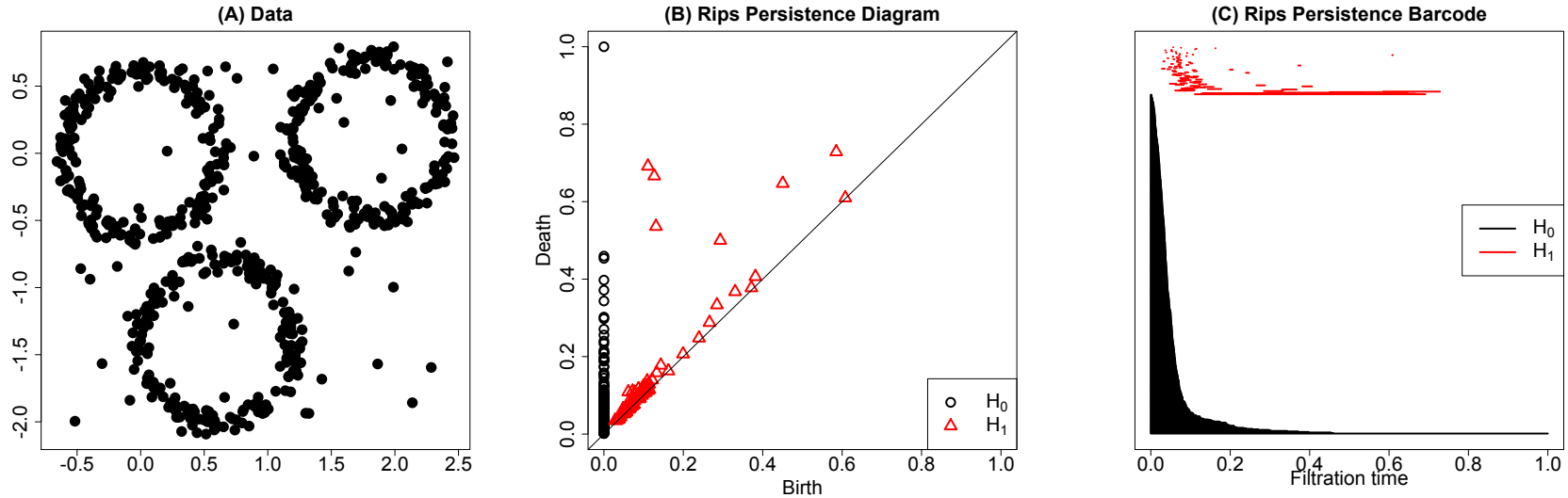
Persistence (or lifetime) of loop: 0.92 - 0.48 = 0.44

- Define $S_\epsilon = \cup_{i=1}^{n} B(Y_i, \epsilon)$ (union of balls with radius ϵ centered at observations $Y_1, \ldots, Y_n$)
- Persistent homology tracks the changing homology of S_ϵ across a range of ϵ's

TDA package R code: Rips filtration

```
library(TDA) #Load library
set.seed(123)  #Set random seed to reproduce results

#Generate three noisy circles
n <- 200
sig <- .5
data0 <- rbind(circleUnif(n, r = 6)+ matrix(rnorm(2*n,0,sig), ncol = 2), circleUnif(n, r = 6) +
                  cbind(rep(18,n), rep(1,n))+matrix(rnorm(2*n,0,sig), ncol = 2), circleUnif(n, r = 6) +
                   cbind(rep(6,n), rep(-14,n))+matrix(rnorm(2*n,0,sig), ncol = 2))

#Add some scatter
n <- 50
data1 <- rbind(data0, cbind(runif(n, min(data0[,1]), max(data0[,1])),
                   runif(n, min(data0[,2]), max(data0[,2]))))/10

#Plot the data (see next slide)
plot(data1, xlab = "", ylab = "", pch = 19)

#Get the persistence diagram and plot it (see next slide)
maxscale <- 1
maxdimension <- 1
diag1<- ripsDiag(data1, maxdimension, maxscale, printProgress = TRUE)$diagram
plot.diagram(diag1)

#Plot the barcode
plot.diagram(diag1, barcode = TRUE)
```

Persistent homology summaries



- Persistence diagram D is a collection of birth (b_j) and death (d_j) times of homology group generators of a particular rank (r_j):

$$D = \{(r_j, b_j, d_j) : j = 1, \ldots, l\}$$

where l represents the number of homology group generators off the diagonal

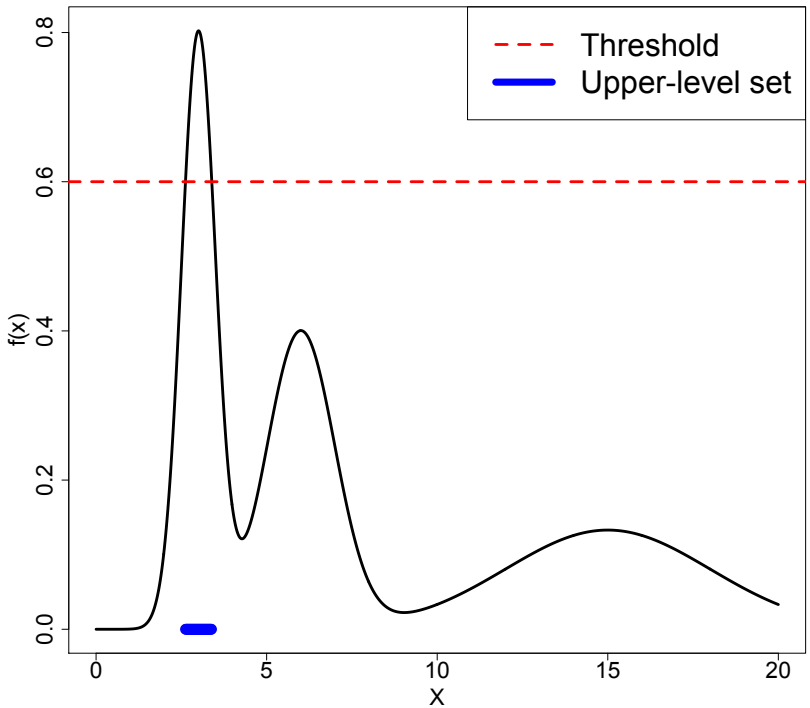
Function-based persistent homology

- Rather than defining the filtration using a Rips Complex over the data points, a function can be used for persistent homology
- Kernel density estimates (e.g. Fasy et al. 2014) or Distance-to-Measure (DTM) functions (e.g. Chazal et al. 2011) are popular approaches in TDA for turning a point-cloud of data into a function

Function-based persistent homology

Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$. An **upper level set**, relative to a threshold $\lambda \in \mathbb{R}$ is the set of points $x \in \mathbb{R}^d$ defined by $E_\lambda = \{x \in \mathbb{R}^d : f(x) \geq \lambda\}$

Similarly, **lower level set**: $E_\lambda = \{x \in \mathbb{R}^d : f(x) < \lambda\}$



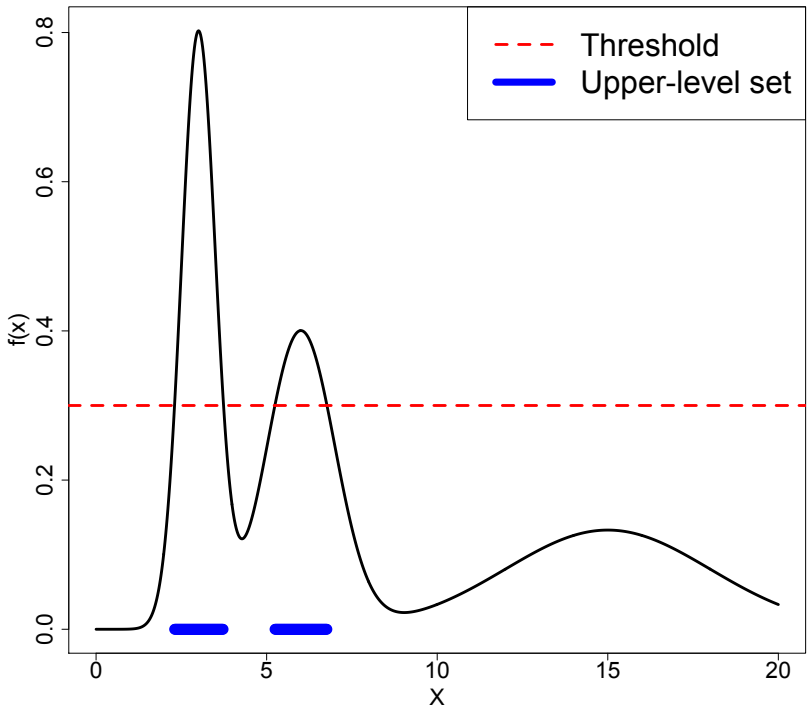
⋆Construct simplicial complexes on the **upper level sets**

⋆Birth and death of separate components of the upper level set is related to the birth and death of maxima and minima

Function-based persistent homology

Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$. An **upper level set**, relative to a threshold $\lambda \in \mathbb{R}$ is the set of points $x \in \mathbb{R}^d$ defined by $E_\lambda = \{x \in \mathbb{R}^d : f(x) \geq \lambda\}$

Similarly, **lower level set**: $E_\lambda = \{x \in \mathbb{R}^d : f(x) < \lambda\}$



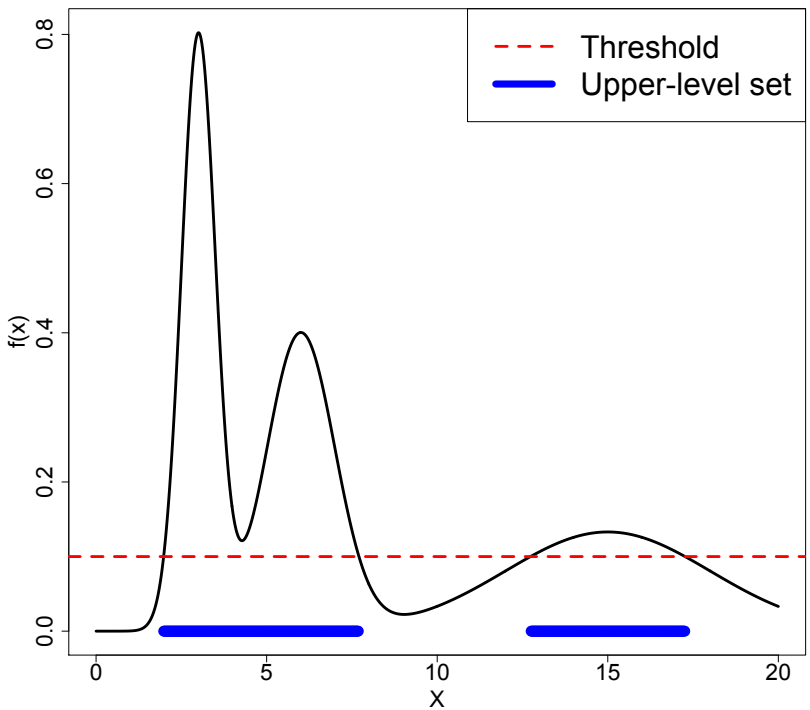
⋆Construct simplicial complexes on the **upper level sets**

⋆Birth and death of separate components of the upper level set is related to the birth and death of maxima and minima

Function-based persistent homology

Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$. An **upper level set**, relative to a threshold $\lambda \in \mathbb{R}$ is the set of points $x \in \mathbb{R}^d$ defined by $E_\lambda = \{x \in \mathbb{R}^d : f(x) \geq \lambda\}$

Similarly, **lower level set**: $E_\lambda = \{x \in \mathbb{R}^d : f(x) < \lambda\}$



*Construct simplicial complexes on the **upper level sets**

*Birth and death of separate components of the upper level set is related to the birth and death of maxima and minima

Distance-to-a-Measure (DTM) Function

- The DTM function can be defined for a probability measure P with support $Y \subset \mathbb{R}^d$ and point $y \in \mathbb{R}^d$ as

$$d_{m_0}(y) = \sqrt{\frac{1}{m_0} \int_0^{m_0} [G_y^{-1}(u)]^2 du},$$

where $G_y(t) = P(\|Y - y\| \leq t)$ and tuning parameter $0 \leq m_0 \leq 1$.

- Given observations $y_1, y_2, \ldots, y_n$, $d_{m_0}(y)$ can be **estimated** using

$$\hat{d}_{m_0}(y) = \sqrt{\frac{1}{k} \sum_{y_i \in N_k(y)} \|y_i - y\|^2},$$

$0 < m_0 < 1$ is a tuning parameter, $k = \lfloor n m_0 \rfloor$, and $N_k(y) = k$ nearest neighbors of $y_1, y_2, \ldots, y_n$ to y.

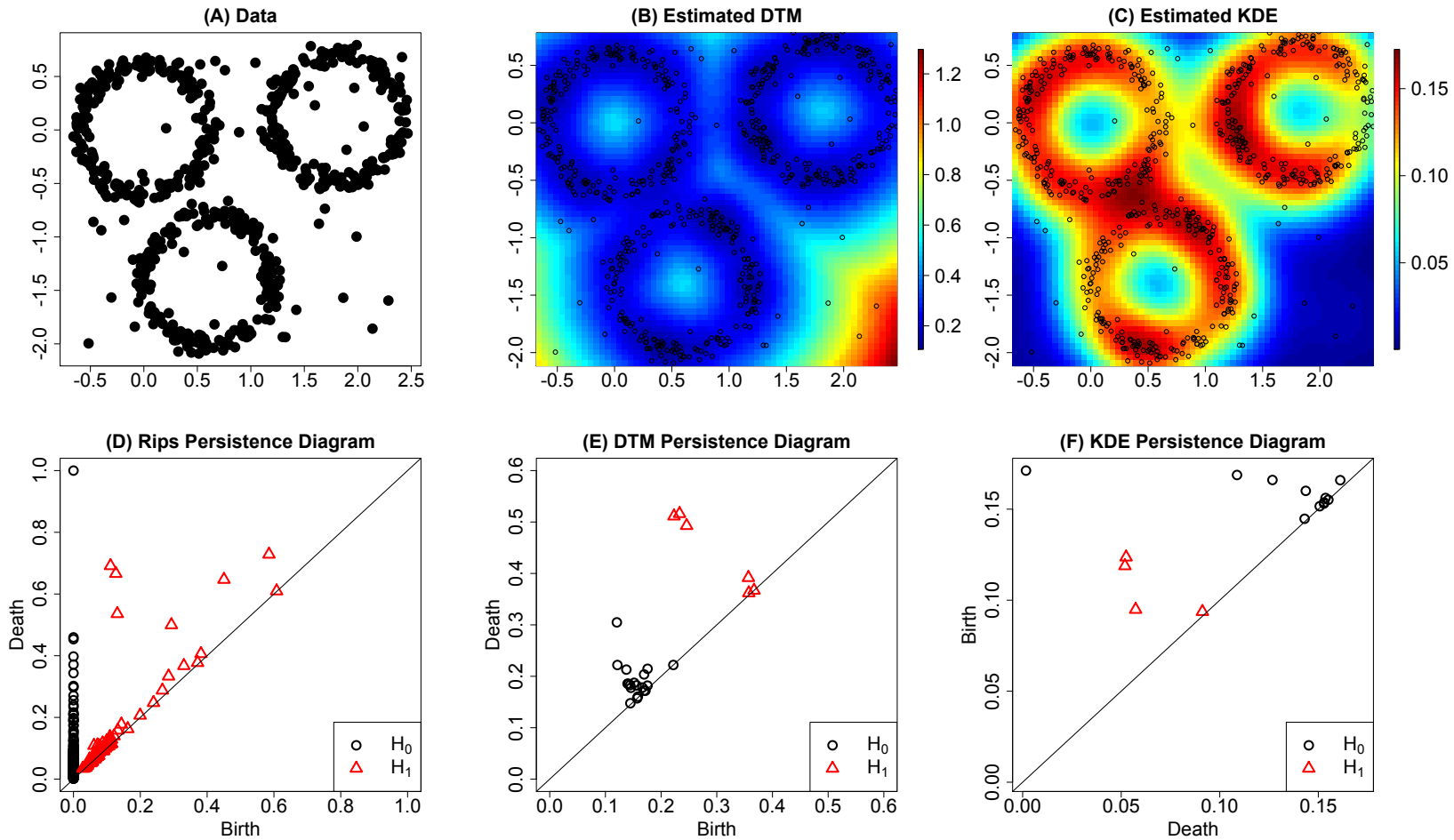
References: Chazal et al. (2011, 2016)

TDA package R code: KDE and DTM filtrations

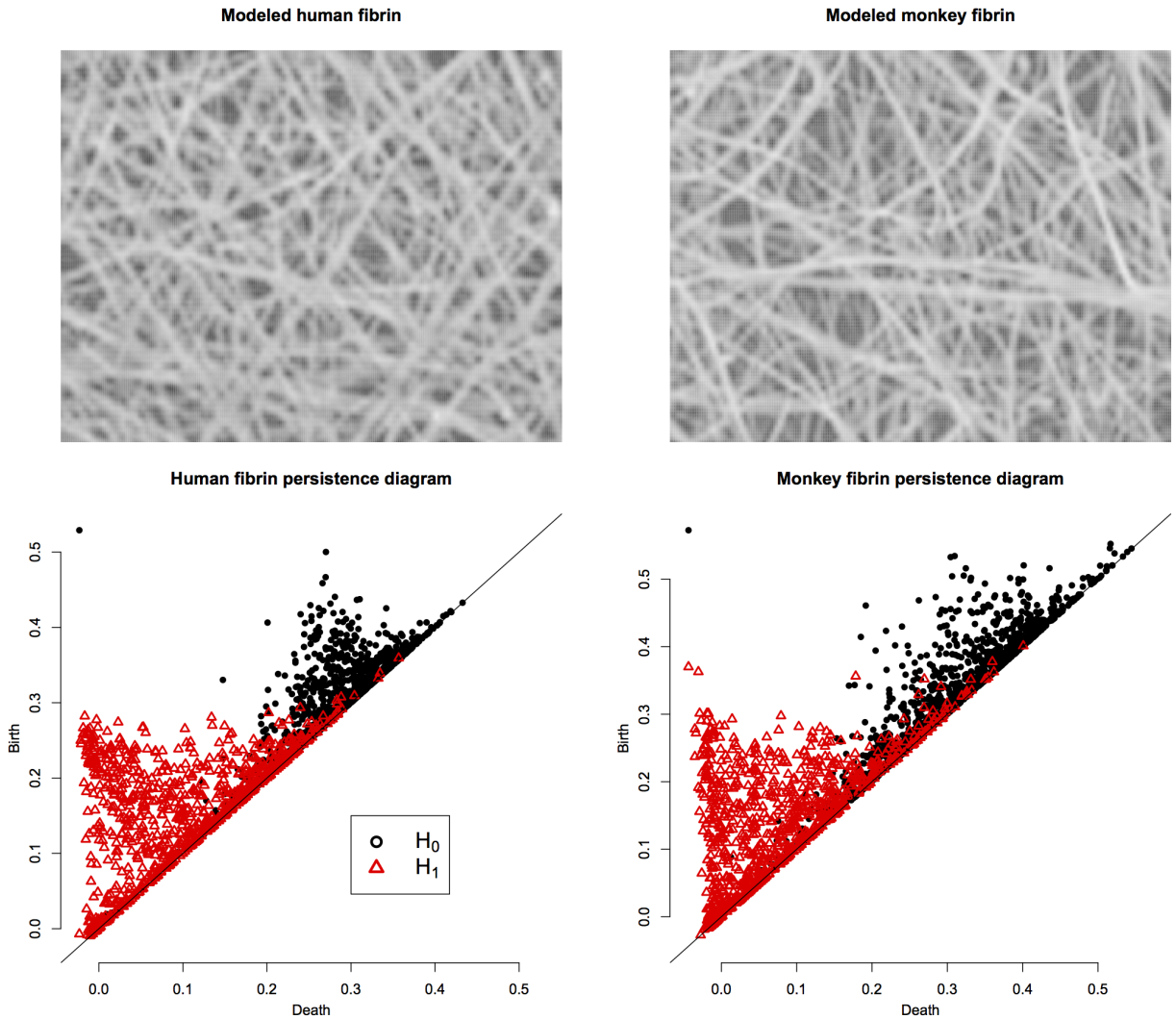
Same data1 as used previously.

```
#Construct a grid of points over which we evaluate the functions
by <- 0.05
Xseq <- seq(min(data1[,1]), max(data1[,1]), by = by)
Yseq <- seq(min(data1[,2]), max(data1[,2]), by = by)
Grid <- expand.grid(Xseq, Yseq)

#DTM
m0 <- 0.05
data1.dtm <- matrix(dtm(data1, Grid, m0), nrow = length(Xseq), ncol = length(Yseq)) #calculate DTM
image(data1.dtm) #Plot image of DTM
diag1.dtm <- gridDiag(FUNvalues = data1.dtm, sublevel = TRUE, location = FALSE,
                      printProgress = TRUE, maxdimension = 1)$diagram

plot.diagram(diag1.dtm) #Plot diagram

#KDE
h <- .25
data1.kde <- kde(data1, Grid, h, kertype = "Gaussian", weight = 1, printProgress = FALSE) #calculate KDE
kde_matrix <- matrix(data1.kde,nrow=length(Xseq), ncol=length(Yseq)) #format as matrix
image(Xseq, Yseq, kde_matrix) #Plot image of KDE
diag1.kde <- gridDiag(FUNvalues = kde_matrix, sublevel = FALSE,location = FALSE,
                      printProgress = TRUE, maxdimension = 1)$diagram

plot.diagram(diag1.kde) #Plot diagram
```

Illustration of different filtrations



Two-sample hypothesis tests



- Modeled human fibrin network (left) and monkey fibrin network (right); original images are from (Pretorius et al., 2009).

Two-sample hypothesis testing: overview

- Setting: samples from two, potentially different, populations

  Human vs. monkey fibrin

  Or maybe there is population, $P^{(1)}$, such that a random draw produces data on a noisy circle, and another population, $P^{(2)}$, that produces random noise (but in advance you do not know there is such a difference)

- In a two-sample hypothesis testing framework, you might have the following hypotheses

  Null hypothesis: There is no difference between $P^{(1)}$ and $P^{(2)}$. Differences in the samples would just be due to chance.

  Alternative hypothesis: There is a difference between $P^{(1)}$ and $P^{(2)}$.

  - General goal is to have evidence *against* the null hypothesis in favor of the alternative hypothesis
  - Two possible conclusions: (i) reject the null hypothesis, or (ii) do not reject the null hypothesis. (In this setting, we do not *accept* the null hypothesis.)

Test statistics and p-values

- Consider a simple example of comparing the means of the two populations:

  Null hypothesis: $\mu_1 = \mu_2$

  Alternative hypothesis: $\mu_1 \neq \mu_2$

- General idea: assume the null hypothesis is true, and find a test statistic, T, to check the compatibility between the null hypothesis and the data

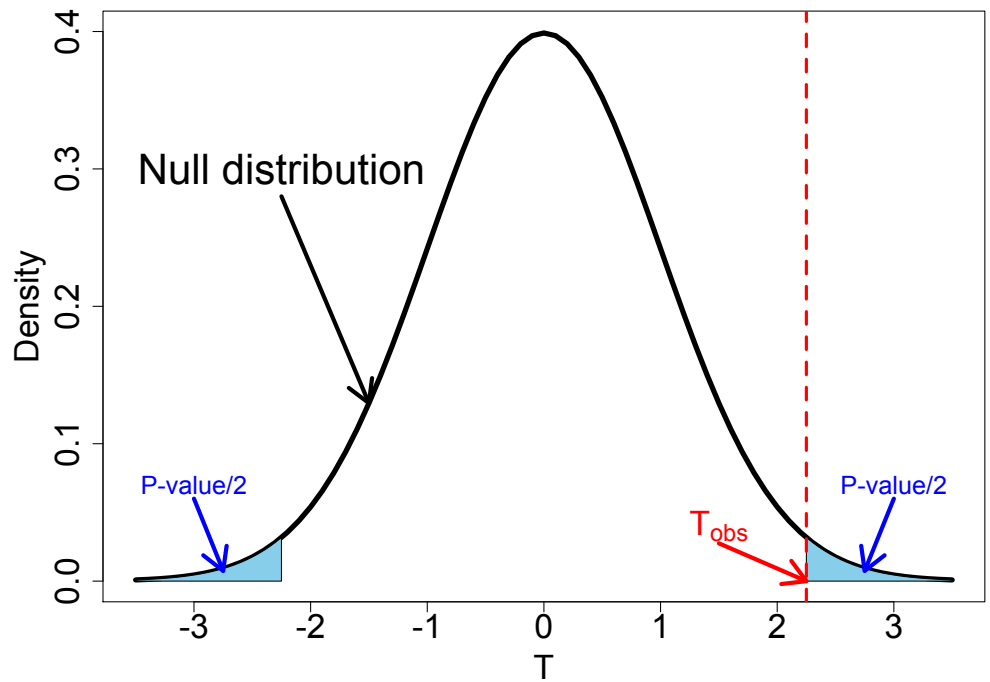
  Example: $T = (\bar{x}_1 - \bar{x}_2)/\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ (where $\bar{x}_l$ is the sample mean for sample drawn from population $l = 1, 2$ with sample size n_l and [known] population standard deviations σ_l).

  $\longrightarrow$ very positive or very negative values of T would be evidence *against* the null hypothesis

- How positive or negative depends on the distribution of the test statistic
- In this simple example, it turns out we know the distribution of the test statistic follows a normal distribution with mean 0 and variance 1

  $\longrightarrow$ p-value $= 2P(\mathbf{T} > |T_{obs}|)$, where $\mathbf{T}$ is a random variable representing the test statistic and T_{obs} is the observed test statistic

  $\longrightarrow$ small p-values ($< .05$, .01, etc) would be evidence *against* the null hypothesis



Two-sample hypothesis tests: TDA

Back to the TDA setting...

- Given two sets of persistence diagrams, $D_1^{(1)}, \ldots, D_{n_1}^{(1)} \sim \mathcal{P}^{(1)}$ and $D_1^{(2)}, \ldots, D_{n_2}^{(2)} \sim \mathcal{P}^{(2)}$
  where $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are the true underlying distributions of persistence diagrams for group 1 and 2, respectively.
  (existence of distributions established in Mileyko et al. (2011))

- $H_0 : \mathcal{P}^{(1)} = \mathcal{P}^{(2)}$ vs. $H_1 : \mathcal{P}^{(1)} \neq \mathcal{P}^{(2)}$

- What to use for the test statistic?

  Persistence diagrams are difficult objects to work with $\longrightarrow$ consider functional summaries of persistence diagrams

- Several functional summaries have been proposed (e.g., Chazal et al. 2014; Adams et al. 2015; Bubenik 2015; Chen et al. 2015)

- In Berry, Chen, Cisewski-Kehe, and Fasy (2018), we develop a unified framework for univariate functional summaries of persistence diagrams then prove some basic functional convergence theorems using tools from functional data analysis

- Given two sets of persistence diagrams, $D_1^{(1)}, \ldots, D_{n_1}^{(1)} \sim \mathcal{P}^{(1)}$ and $D_1^{(2)}, \ldots, D_{n_2}^{(2)} \sim \mathcal{P}^{(2)}$.
- $H_0 : \mathcal{P}^{(1)} = \mathcal{P}^{(2)}$ vs. $H_1 : \mathcal{P}^{(1)} \neq \mathcal{P}^{(2)}$

Let $F_{l,i} = F(D_i^{(l)})$ be the functional summary for diagram i of set $l = 1, 2$, and $\hat{F}_l(t) = \sum_{i=1}^{n_l} F_{l,i}(t)$

$\hat{F}_l(t)$ is a consistent estimator of the population mean functional summary, $\mathbb{E}F_l(t)$ (Berry, Chen, Cisewski-Kehe, and Fasy, 2018)

Then use test statistic $T = d(\hat{F}_1(t), \hat{F}_2(t))$ for some metric $d(\cdot, \cdot)$

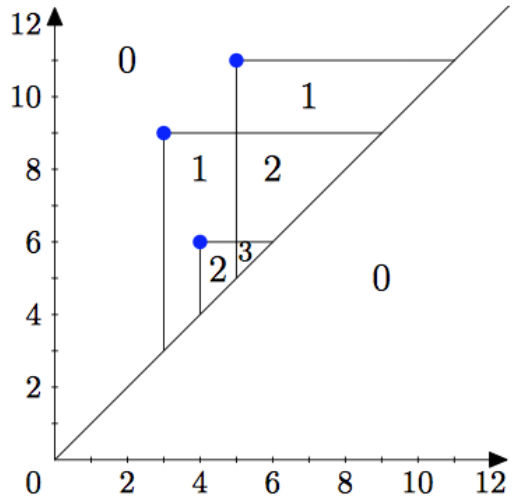
Landscape functions

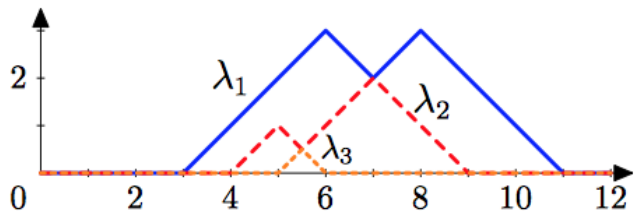
Landscape functions are the collection of functions $\mathbb{F}_k : \mathcal{D} \to \mathcal{F}$ s.t. for each $k \in \mathbb{N}$

$$\mathbb{F}_k(D; t) = \underset{i=1,\ldots,l}{\text{kmax}} \Lambda_i(t)$$

for $t \in [t_{\text{min}}, t_{\text{max}}]$, kmax selects the kth largest value

$$\Lambda_i(t) = \begin{cases} t - b_i & t \in [b_i, \frac{d_i+b_i}{2}] \\ d_i - t & t \in [\frac{d_i+b_i}{2}, d_i] \\ 0 & \text{otherwise} \end{cases}$$





Bubenik (2015)

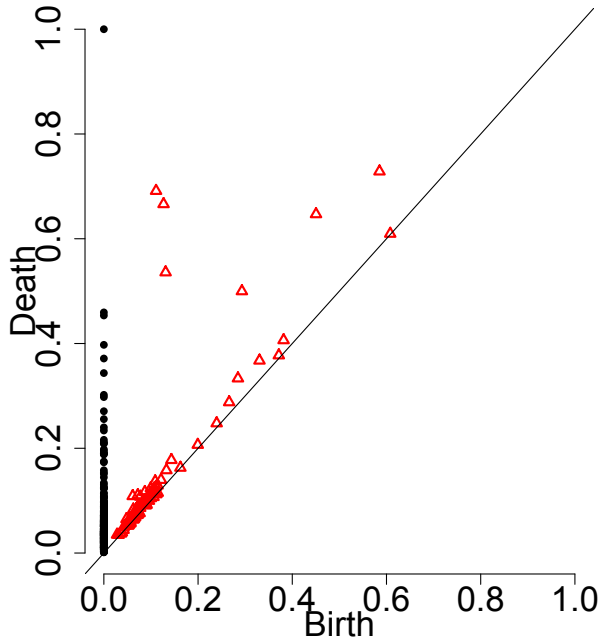
Generalized landscapes Berry, Chen, Cisewski-Kehe, and Fasy (2018):

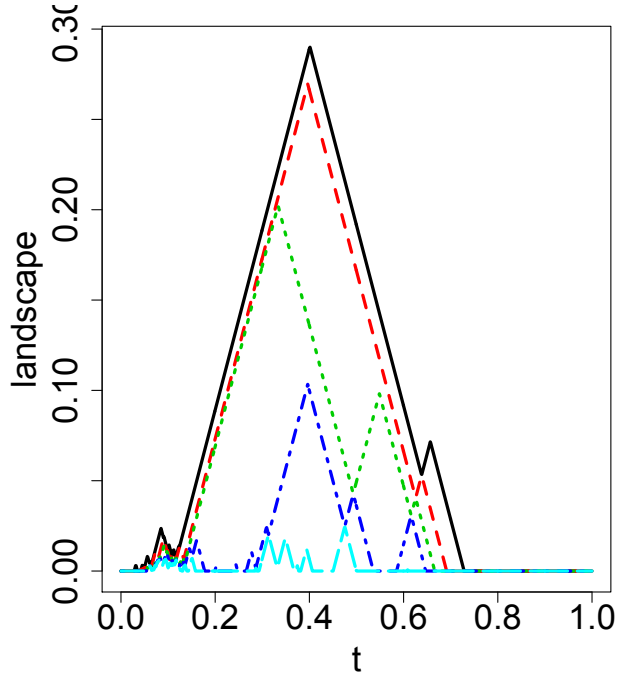
R code available at `https://github.com/JessiCisewskiKehe/generalized_landscapes`

TDA package R code: landscape functions

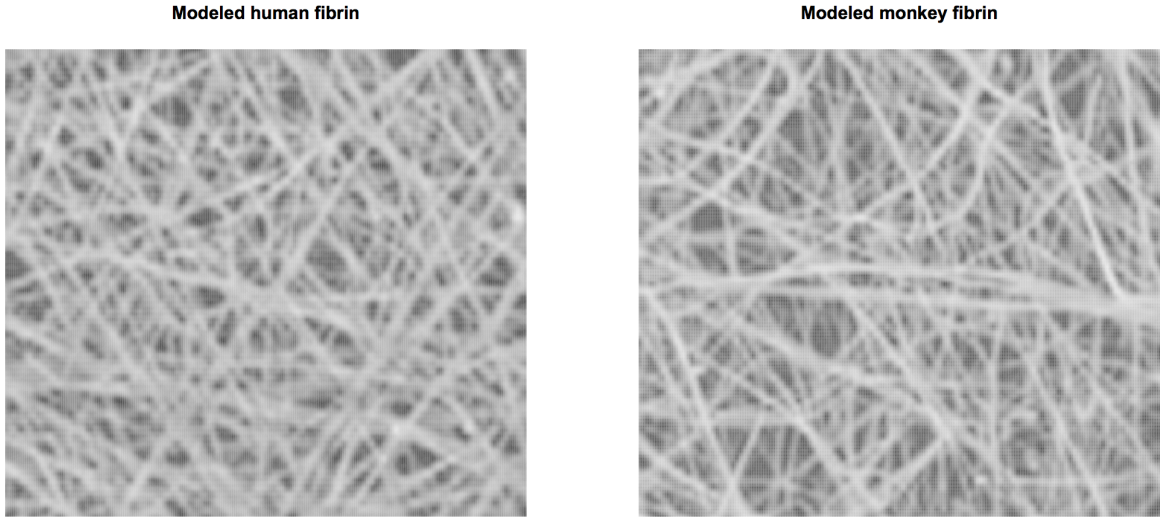
Same data1 as used previously and Rips persistence diagram diag1

```
#set sequence for function
tseq <- seq(min(diag1[,2:3]),max(diag1[,2:3]), length = 1000)

#get landscapes 1 to 5
land1 <- landscape(diag1, dimension = 1, KK = 1:5, tseq)

#plot first landscape
plot(tseq, land1[,1], type = "l", xlab = "t", ylab = "landscape")
```





Fibrin data



- Suppose we had a sample of Monkey fibrin images and of Human fibrin images, and then two sets of persistence diagrams: $D_1^{(1)}, \ldots, D_{n_1}^{(1)} \sim \mathcal{P}^{(1)}$ and $D_1^{(2)}, \ldots, D_{n_2}^{(2)} \sim \mathcal{P}^{(2)}$.
- $H_0 : \mathcal{P}^{(1)} = \mathcal{P}^{(2)}$ vs. $H_1 : \mathcal{P}^{(1)} \neq \mathcal{P}^{(2)}$

  Let $F_{l,i} = F(D_i^{(l)})$, be the first landscape function for diagram i of set $l = 1, 2$

  Calculate average landscape for each group: $\hat{F}_l(t) = \sum_{i=1}^{n_l} F_{l,i}(t)$

  Then use test statistic such as $T = \int |\hat{F}_1(t) - \hat{F}_2(t)| dt$
- But what is the distribution of T? Needed to compute a p-value...

Permutation tests

Null hypothesis: There is no difference between $P^{(1)}$ and $P^{(2)}$. Differences in the samples would just be due to chance.

Alternative hypothesis: There is a difference between $P^{(1)}$ and $P^{(2)}$.

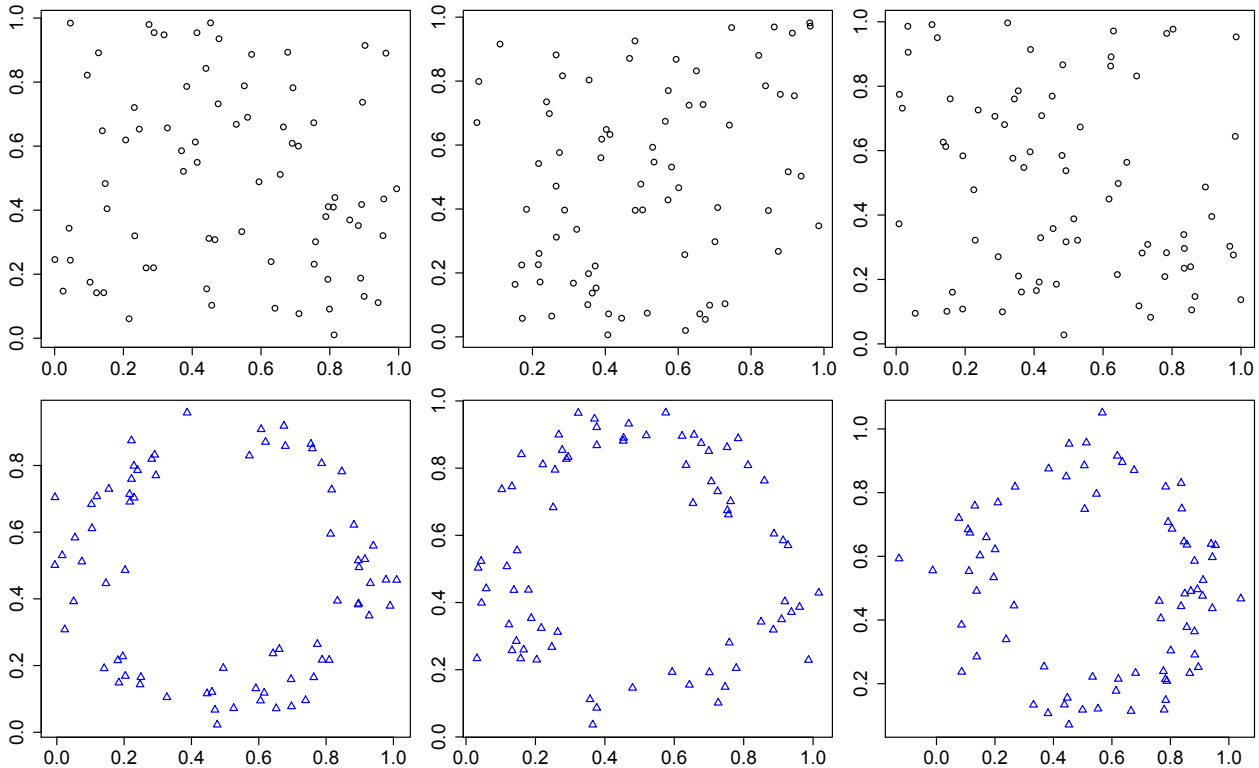
$\longrightarrow$ Can estimate the null distribution of the test statistic, T, by randomly mixing (i.e., permuting) the labels of 1 or 2 a bunch of times to get many realizations of T under the null hypothesis

$\longrightarrow$ See where the *observed* T falls on the null distribution to calculate a permutation p-value

Permutation test: Example

Null hypothesis: There is no difference between $P^{(1)}$ and $P^{(2)}$. Differences in the samples would just be due to chance.

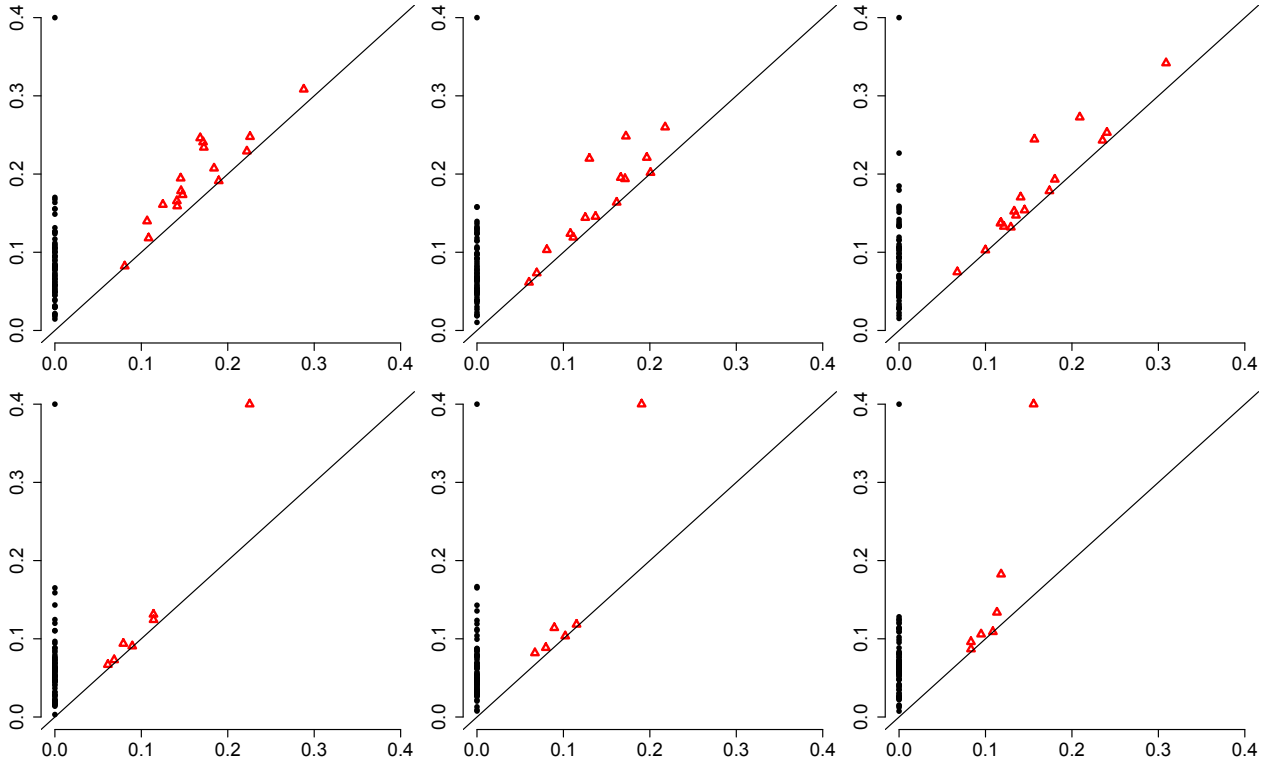
Alternative hypothesis: There is a difference between $P^{(1)}$ and $P^{(2)}$.



R code: get samples

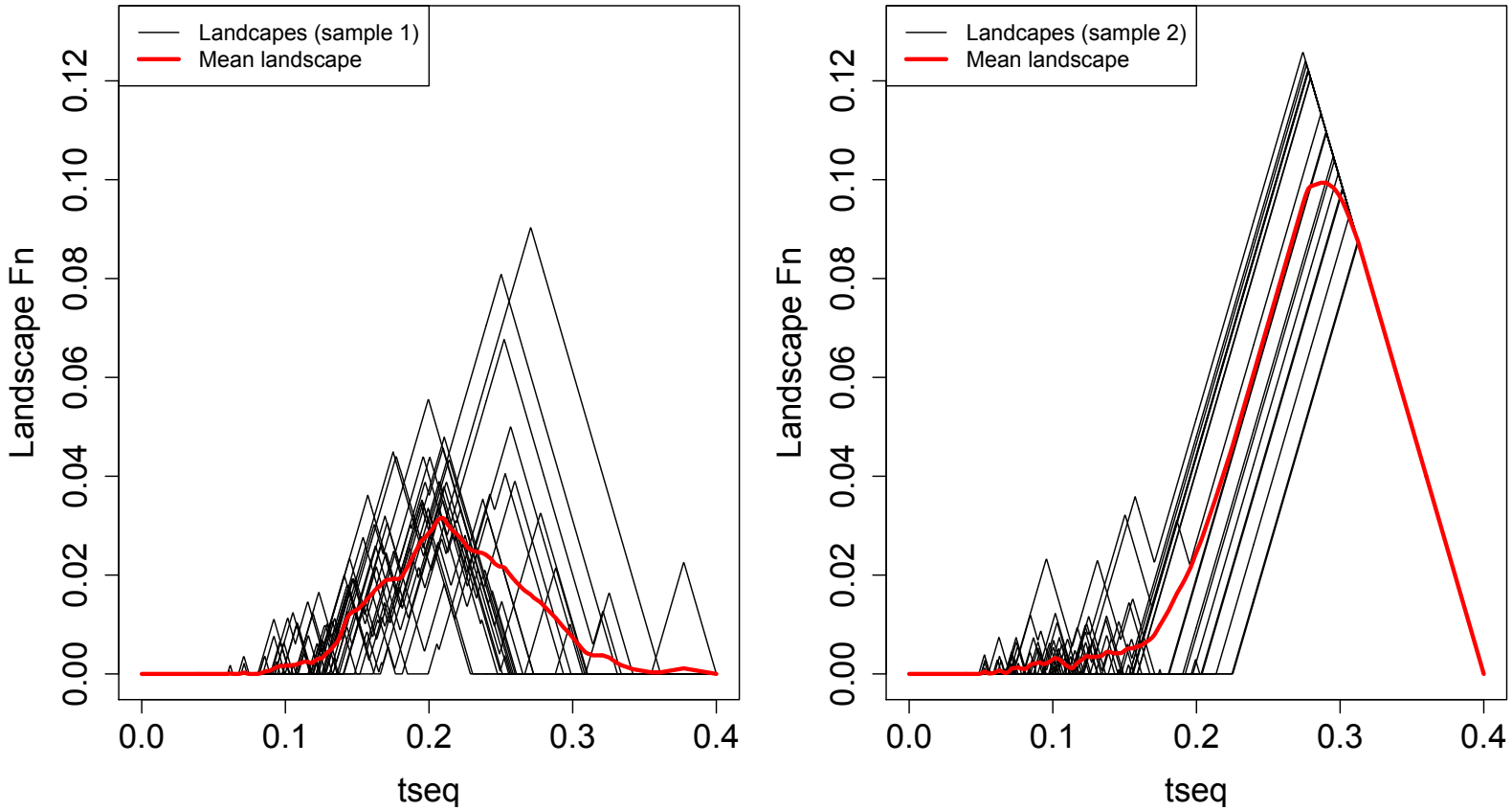
```
library(TDA)

set.seed(123)
pop1 <- function(n){
return(matrix(runif(2*n),ncol=2))
}

pop2 <- function(n,sig,rad){
              data0 <- circleUnif(n, r = rad)+
                    matrix(rnorm(2*n,0,sig), ncol = 2)+c(.5,.5)
              return(data0)
}

n_samples <- 20
n1 <- 75
n2 <- 75

sample1 <- lapply(1:n_samples, function(ii) pop1(n1))
sample2 <- lapply(1:n_samples, function(ii) pop2(n2,.06,.4))
```

R code: get persistence diagrams




```
maxscale <- .4
maxdimension <- 1
diag1 <- lapply(1:n_samples, function(ii) ripsDiag(sample1[[ii]],
                    maxdimension, maxscale)$diagram)
diag2 <- lapply(1:n_samples, function(ii) ripsDiag(sample2[[ii]],
                    maxdimension, maxscale)$diagram)
```

R code: get landscapes



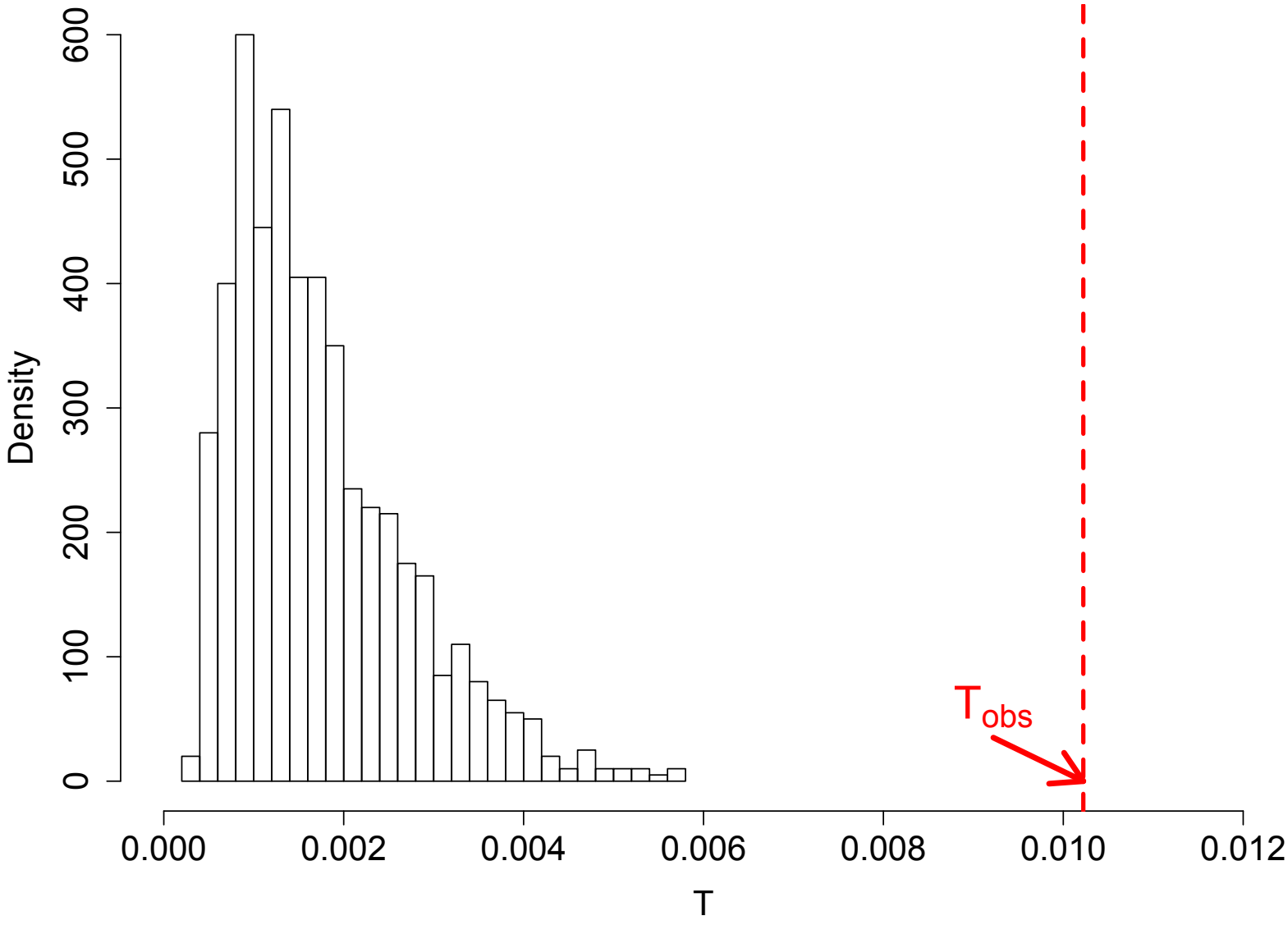

```
tseq <- seq(0,.4, length = 1000)
land1 <- sapply(1:n_samples, function(ii)
                  landscape(diag1[[ii]], dimension = 1, KK = 1, tseq))
land2 <- sapply(1:n_samples, function(ii)
                  landscape(diag2[[ii]], dimension = 1, KK = 1, tseq))
```

R code: permutation tests

```
library(sfsmisc) #for integrate.xy

n_perm <- 1000
tseq <- seq(0,.4, length = 1000)
land_mean1 <- apply(land1,1,mean)
land_mean2 <- apply(land2,1,mean)
T_obs <- integrate.xy(tseq,abs(land_mean1 - land_mean2))

landscapes_all <- t(cbind(land1,land2))
T_stat <- c()
for(i in 1:n_perm){
     which_landscapes <- sample(1:nrow(landscapes_all),
                        nrow(landscapes_all)/2, replace = FALSE)
     mean1 <- apply(landscapes_all[which_landscapes,],2,mean)
     mean2 <- apply(landscapes_all[-which_landscapes,],2,mean)
     T_stat[i] <- integrate.xy(tseq,abs(mean1 - mean2))
}
```

R code: permutation tests

Approximate distribution of **T** under the null hypothesis

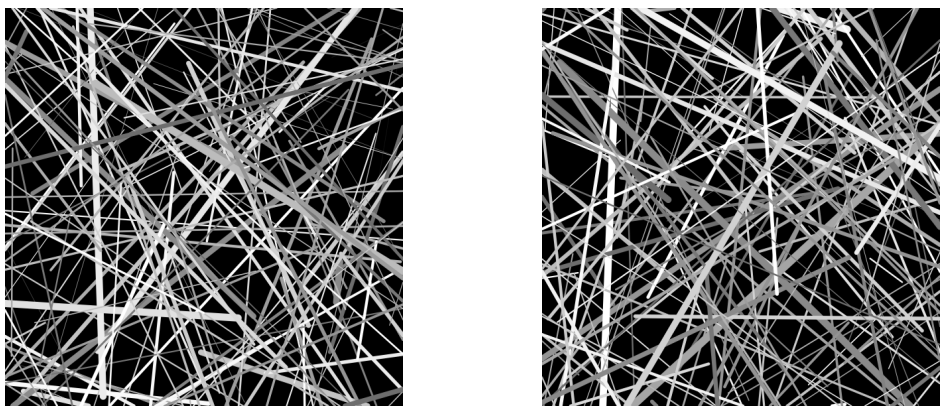


Pickup Sticks Simulator (STIX)

Another dataset we can consider in the coding sprints:

To generate an image with n segments, or sticks

1. Two sets of n points are randomly sampled from a Uniform distribution: $\{u_{i1}, u_{i2}\}_{i=1}^{n}$
2. Segments drawn between points in the same position of the two lists of random numbers (i.e. between u_{i1} and u_{i2})
3. The thickness of each segment is randomly drawn from a χ^2 distribution with thickness $= t$ degrees of freedom.



Realizations of the Pick-up Sticks Simulation Data (STIX) with average thicknesses of (left) 5 and (right) 6

Concluding remarks

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

  However, analyzing these data is not always straightforward

- Hypothesis testing using persistent homology

  Functional summaries of persistence diagrams can be used as test statistics

Concluding remarks

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

  However, analyzing these data is not always straightforward

- Hypothesis testing using persistent homology

  Functional summaries of persistence diagrams can be used as test statistics

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*.

Berry, E., Chen, Y.-C., Cisewski-Kehe, J., and Fasy, B. T. (2018), "Functional Summaries of Persistence Diagrams," ArXiv preprint arXiv: 1804.01618.

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

Carlsson, G. (2009), "Topology and Data," *Bulletin of the American Mathematical Society*, 46, 255 – 308.

Chazal, F., Cohen-Steiner, D., and Mérigot, Q. (2011), "Geometric inference for probability measures," *Foundations of Computational Mathematics*, 11, 733–751.

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.

Chazal, F., Massart, P., Michel, B., et al. (2016), "Rates of convergence for robust geometric inference," *Electronic journal of statistics*, 10, 2243–2286.

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

Edelsbrunner, H. and Harer, J. (2008), "Persistent homology - a survey," *Contemporary mathematics*, 453, 257 – 282.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002), "Topological persistence and simplification," *Discrete and Computational Geometry*, 28, 511–533.

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.

Mileyko, Y., Mukherjee, S., and Harer, J. (2011), "Probability measures on the space of persistence diagrams," *Inverse Problems*, 27, 124007.

Pretorius, E., Vieira, W., Oberholzer, H., and Auer, R. (2009), "Comparative scanning electron microscopy of platelets and fibrin networks of humans and different animals," *International Journal of Morphology*, 27, 69–76.