R Package TDA for Statistical Inference on Topological Data Analysis

Jisu Kim

Carnegie Mellon University

2016-06-16

Introduction

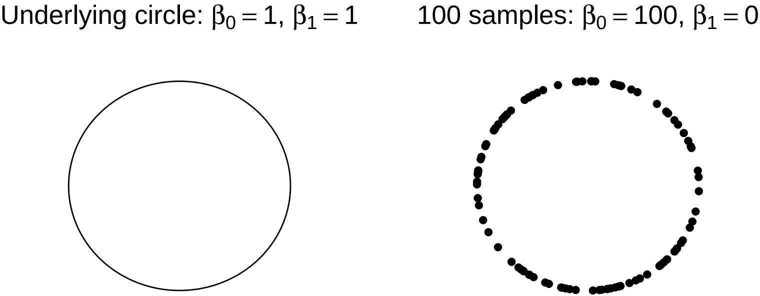
Preliminary functions in R Package TDA

Persistent Homology

Statistical Inference on Persistence Homology

Reference

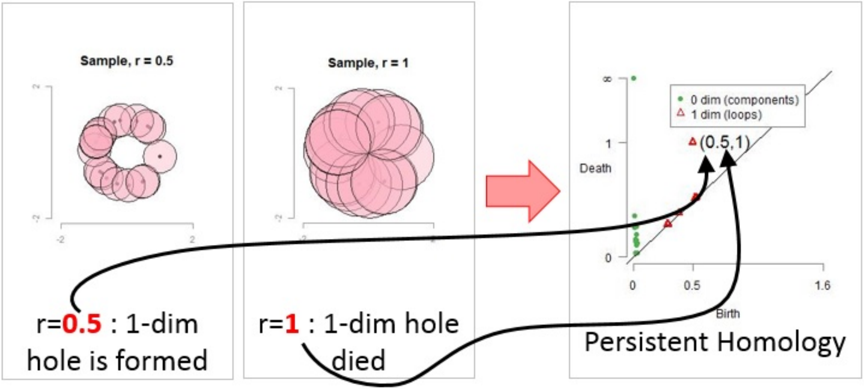
Homology of finite sample is different from homology of underlying manifold, hence it cannot be directly used for inference.



Underlying circle: $\beta_0 = 1$, $\beta_1 = 1$

100 samples: $\beta_0 = 100$, $\beta_1 = 0$

Persistence homology computes homologies on collection of sets, and tracks when topological features are born and when they die.

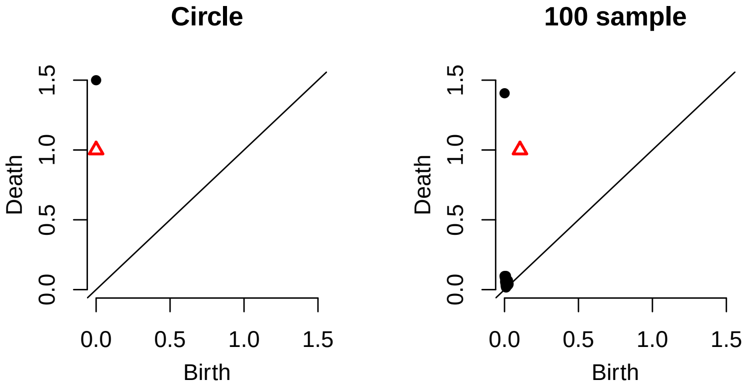


r=**0.5** : 1-dim hole is formed

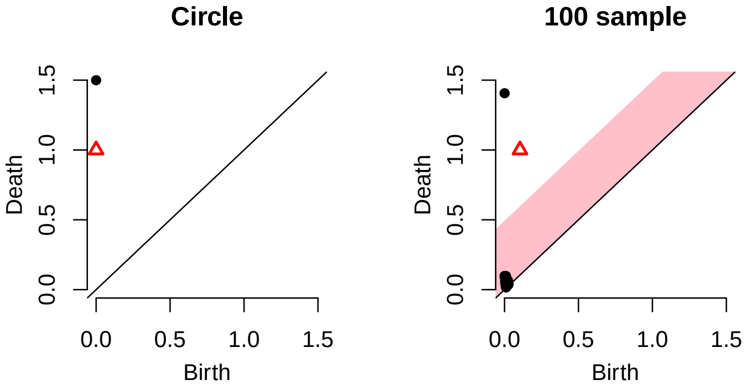
r=**1** : 1-dim hole died

Persistent Homology

Persistent homology of underlying manifold can be inferred from persistent homology of finite samples.



How can we distinguish statistically significant homological features from noisy homological features?



R Package TDA bridges between R and C++ library GUDHI/Dionysus/PHAT.

- website:
  `https://cran.r-project.org/web/packages/TDA/index.html`
- Author: Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, Clément Maria, and Vincent Rouvreau.
- R is a programming language for statistical computing and graphics.
- R has short development time, while C/C++ has short execution time.
- R package TDA provides an R interface for C++ library GUDHI[Maria, 2014]/Dionysus[Morozov, 2007]/PHAT[Bauer et al., 2012].

Introduction

Preliminary functions in R Package TDA

Persistent Homology

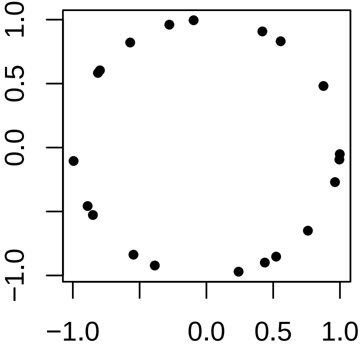
Statistical Inference on Persistence Homology

Reference

R Package TDA provides a function to sample on a circle.

The function circleUnif() generates n sample from the uniform distribution on the circle in $\mathbb{R}^2$ with radius r.

```r
circleSample <- circleUnif(n = 20, r = 1)
plot(circleSample, xlab = "", ylab = "", pch = 20)
```



R Package TDA provides distance functions over a grid.

Suppose *n* points are generated from the unit circle, and grid of points are generated.

```r
X <- circleUnif(n = 400, r = 1)

lim <- c(-1.7, 1.7)
by <- 0.05
margin <- seq(from = lim[1], to = lim[2], by = by)
Grid <- expand.grid(margin, margin)
```

R Package TDA provides distance functions over a grid.

The distance function $\Delta : \mathbb{R}^d \to [0, \infty)$ is defined as

$$\Delta(y) = \inf_{x \in X} \|x - y\|_2.$$

The function distFct() computes the distance function Δ on a grid of points.

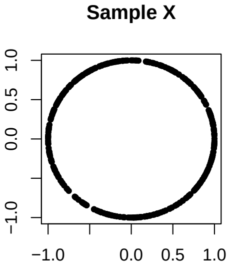
```r
distance <- distFct(X = X, Grid = Grid)

par(mfrow = c(1,2))
plot(X, xlab = "", ylab = "", main = "Sample X", pch = 20)
persp(x = margin, y = margin,
  z = matrix(distance, nrow = length(margin), ncol = length(margin)),
  xlab = "", ylab = "", zlab = "", theta = -20, phi = 35, scale = FALSE,
  expand = 3, col = "red", border = NA, ltheta = 50, shade = 0.5,
  main = "Distance Function")
```

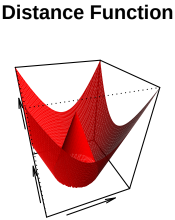
R Package TDA provides distance functions over a grid.

The distance function $\Delta : \mathbb{R}^d \to [0, \infty)$ is defined as

$$\Delta(y) = \inf_{x \in X} \|x - y\|_2.$$

The function distFct() computes the distance function Δ on a grid of points.


Sample X


Distance Function

R Package TDA provides density functions over a grid.

The Gaussian Kernel Density Estimator (KDE) $\hat{p}_h : \mathbb{R}^d \to [0, \infty)$ is defined as

$$\hat{p}_h(y) = \frac{1}{n(\sqrt{2\pi}h)^d} \sum_{i=1}^{n} \exp\left(\frac{-\|y - x_i\|_2^2}{2h^2} \right),$$

where h is a smoothing parameter.
The function kde() computes the KDE function $\hat{p}_h$ on a grid of points.

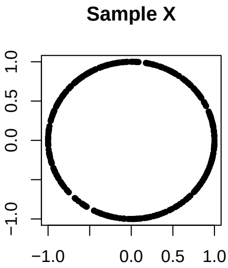
```r
h <- 0.3
KDE <- kde(X = X, Grid = Grid, h = h)

par(mfrow = c(1,2))
plot(X, xlab = "", ylab = "", main = "Sample X", pch = 20)
persp(x = margin, y = margin,
  z = matrix(KDE, nrow = length(margin), ncol = length(margin)),
  xlab = "", ylab = "", zlab = "", theta = -20, phi = 35, scale = FALSE,
  expand = 3, col = "red", border = NA, ltheta = 50, shade = 0.5,
  main = "KDE")
```

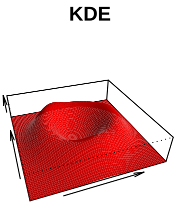
R Package TDA provides density functions over a grid.

The Gaussian Kernel Density Estimator (KDE) $\hat{p}_h : \mathbb{R}^d \to [0, \infty)$ is defined as

$$\hat{p}_h(y) = \frac{1}{n(\sqrt{2\pi}h)^d} \sum_{i=1}^{n} \exp\left(\frac{-\|y - x_i\|_2^2}{2h^2} \right),$$

where h is a smoothing parameter.
The function kde() computes the KDE function $\hat{p}_h$ on a grid of points.


Sample X


KDE

Introduction

Preliminary functions in R Package TDA

Persistent Homology

Statistical Inference on Persistence Homology

Reference

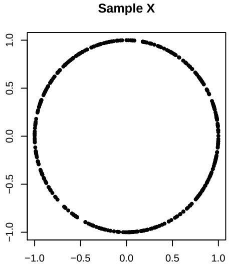
R Package TDA computes Persistent Homology over a grid.

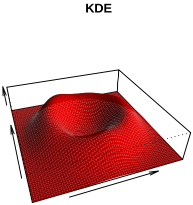
- The function gridDiag() computes the persistent homology of sublevel (and superlevel) sets of the input function.
  - gridDiag() evaluates the real valued input function over a grid.
  - gridDiag() constructs a filtration of simplices using the values of the input function.
  - gridDiag() computes the persistent homology of the filtration.
- The user can choose to compute persistent homology using either Dionysus library or PHAT library.

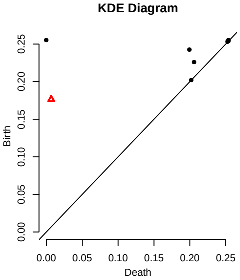
```r
Diag <- gridDiag(X = X, FUN = kde, lim = cbind(lim, lim), by = by,
  sublevel = FALSE, library = "Dionysus", printProgress = FALSE, h = 0.3)

par(mfrow = c(1,3))
plot(X, xlab = "", ylab = "", main = "Sample X", pch = 20)
persp(x = margin, y = margin,
  z = matrix(KDE, nrow = length(margin), ncol = length(margin)),
  xlab = "", ylab = "", zlab = "", theta = -20, phi = 35, scale = FALSE,
  expand = 3, col = "red", border = NA, ltheta = 50, shade = 0.9,
  main = "KDE")
plot(x = Diag[["diagram"]], main = "KDE Diagram")
```

R Package TDA computes Persistent Homology over a grid.

- The function gridDiag() computes the persistent homology of sublevel (and superlevel) sets of the input function.
    - gridDiag() evaluates the real valued input function over a grid.
    - gridDiag() constructs a filtration of simplices using the values of the input function.
    - gridDiag() computes the persistent homology of the filtration.
- The user can choose to compute persistent homology using either Dionysus library or PHAT library.


Sample X


KDE


KDE Diagram

Introduction

Preliminary functions in R Package TDA

Persistent Homology

Statistical Inference on Persistence Homology

Reference

Bottleneck distance gives a metric on the space of Persistent Homology.

Definition

Let $D_1,\ D_2$ be multiset of points. Bottleneck distance is defined as

$$W_\infty(D_1,\ D_2) = \inf_\gamma \sup_{x \in D_\mathbf{1}} \|x - \gamma(x)\|_\infty,$$

where γ ranges over all bijections from D_1 to D_2.

Bottleneck distance can be controlled by the corresponding distance on functions: Stability Theorem.

Theorem

[Edelsbrunner and Harer, 2010][Chazal, de Silva, Glisse, and Oudot, 2012] Let $\mathbb{X}$ *be finitely triangulable space and* $f,\ g : \mathbb{X} \to \mathbb{R}$ *be two continuous functions. Then for each dimension* p*,*

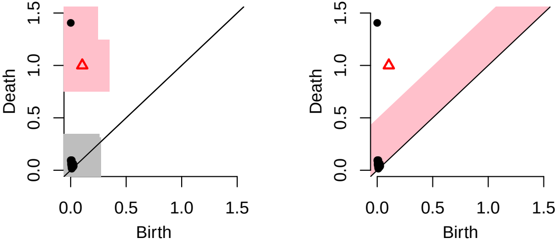
$$W_\infty(Dgm_p(f), Dgm_p(g)) \leq \|f - g\|_\infty.$$

Confidence band for Persistent Diagram can be obtained by the corresponding confidence band for functions.

Let M be a compact manifold, and $X = \{X_1, \cdots, X_n\}$ be n samples whose support is M. Let f_M and f_X be corresponding functions whose persistent homology is of interest.

Given the significance level $\alpha \in (0, 1)$, $(1 - \alpha)$ confidence band $c_n = c_n(X_1, \cdots, X_n)$ is a random variable satisfying

$$\mathbb{P}\left(W_\infty(Dgm_p(f_M),\, Dgm_p(f_X)) > c_n\right) \leq \alpha.$$



Confidence band for Persistent Diagram can be obtained by the corresponding confidence band for functions.

From Stability Theorem, $\mathbb{P}\left(||f_M - f_X|| > c_n\right) \leq \alpha$ implies

$$\mathbb{P}\left(W_\infty(Dgm_p(f_M), \, Dgm_p(f_X)) > c_n\right) \leq \mathbb{P}\left(||f_M - f_X|| > c_n\right) \leq \alpha,$$

so the confidence band of corresponding functions f_M can be used for confidene band of persistence diagrams $Dgm_p(f_M)$.

Confidence band can be computed using the bootstrap algorithm.

1. Given a sample $X = \{x_1, \ldots, x_n\}$, compute the kernel density estimator $\hat{p}_h$.
2. Draw $X^* = \{x_1^*, \ldots, x_n^*\}$ from $X = \{x_1, \ldots, x_n\}$ (with replacement), and compute $\theta^* = \sqrt{n}||\hat{p}_h^*(x) - \hat{p}_h(x)||_\infty$, where $\hat{p}_h^*$ is the density estimator computed using X^*
3. Repeat the previous step B times to obtain $\theta_1^*, \ldots, \theta_B^*$
4. Compute $q_\alpha = \inf \left\{ q : \frac{1}{B} \sum_{j=1}^{B} I(\theta_j^* \geq q) \leq \alpha \right\}$
5. The $(1-\alpha)$ confidence band for $\mathbb{E}[\hat{p}_h]$ is $\left[\hat{p}_h - \frac{q_\alpha}{\sqrt{n}} \, , \, \hat{p}_h + \frac{q_\alpha}{\sqrt{n}} \right]$.

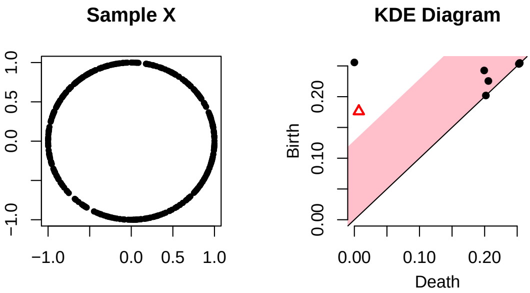
Confidence band can be computed using the bootstrap algorithm.

- The validity of the bootstrap algorithm is proved and used in the framework of persistent homology.
  - [Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, and Singh, 2014] proved for kernel density estimator,
  - [Chazal, Fasy, Lecci, Michel, Rinaldo, and Wasserman, 2014] proved for distance to measure and kernel distance.

R Package TDA computes the bootstrap confidence band.

The function bootstrapBand() computes $(1-\alpha)$ bootstrap confidence band.

```
bandFun <- bootstrapBand(X = X, FUN = kde, Grid = Grid, B = 20,
                         parallel = FALSE, alpha = 0.1, h = h)

par(mfrow = c(1,2))
plot(X, xlab = "", ylab = "", main = "Sample X", pch = 20)
plot(x = Diag[["diagram"]], band = 2 * bandFun[["width"]],
     main = "KDE Diagram")
```



Introduction

Preliminary functions in R Package TDA

Persistent Homology

Statistical Inference on Persistence Homology

Reference

Reference

Ulrich Bauer, Michael Kerber, and Jan Reininghaus. PHAT, a software library for persistent homology, 2012. `https://bitbucket.org/phat-code/phat`.

Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance-to-a-measure and kernel distance. *Technical Report*, 2014.

Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Society, 2010.

Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 2014.

Clément Maria. GUDHI, simplicial complexes and persistent homology packages, 2014. `https://project.inria.fr/gudhi/software/`.

Dmitriy Morozov. Dionysus, a c++ library for computing persistent homology, 2007. `http://www.mrzv.org/software/dionysus/`.

Thank you!