

# My presentation

Rui Dong

2022-10-14

## Contents

<b>1</b>	<b>How I can contribute to the SAPPHIRE project</b>	<b>1</b>
1.1	Update the <code>glottodist</code> with different metrics other than Gower's distance . . . . .	1
1.2	The Hausdorff distance of two datasets . . . . .	1
1.3	Topological data analysis (TDA) . . . . .	1
1.4	Spectral graph theory methods . . . . .	2
<b>2</b>	<b>Some codes</b>	<b>2</b>
2.1	Load the <code>wals</code> datasets . . . . .	2
2.2	The Hausdorff distance . . . . .	3
2.3	TDA . . . . .	4
2.4	Futher codes . . . . .	11

## 1 How I can contribute to the SAPPHIRE project

### 1.1 Update the `glottodist` with different metrics other than Gower's distance

We can implement other kinds of metrics like listed in [Boriah et al., 2008], Take Eskin's distance as an example:

$$S_k(X_k, Y_k) = \begin{cases} 1 & \text{if } X_k = Y_k \\ \frac{n_k^2}{n_k^2 + 2} & \text{otherwise} \end{cases}$$

### 1.2 The Hausdorff distance of two datasets

The **Hausdorff distance** is a quantity to measure the distance between two subsets of a metric space.

**Definition 1.1** (Hausdorff distance). Let  $X$  and  $Y$  be two subsets of a metric space  $(M, d)$ . The Hausdorff distance  $d_H(X, Y)$  is given by

$$d_H(X, Y) = \max \left\{ \sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y) \right\},$$

where  $d(x, Y) := \inf_{y \in Y} d(x, y)$ .

More than that, we can take the **Gromov-Hausdorff distance**, which measures the difference between two different datasets, into account.

### 1.3 Topological data analysis (TDA)

Using topological data analysis methods to analyze the `glottodata`

- Linguistic data analysis: [Port et al., 2018], [Port et al., 2022]
- Spatial data: [Feng et al., 2022]

## 1.4 Spectral graph theory methods

Construct a graph from the `glottodata` and analyse the graph, for example, analyse the spectrum of the Laplacian operator of the graph or apply heat kernel analysis methods [Ortegaray et al., 2021] to analyse the `glottodata`.

## 2 Some codes

### 2.1 Load the wals datasets

```
library(glottospace)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(TDAstats)
library(sf)

## Linking to GEOS 3.10.2, GDAL 3.4.1, PROJ 8.2.1; sf_use_s2() is TRUE

# load the dataset wals
wals <- glottoget("wals")

#select the data wrt South America
wals_sam <- subset(wals, continent == "South America")
wals_asia <- subset(wals, continent=="Asia")

# Define a function select.features to get all features of dataset that the percentage of
# NA values is less than the threshold a
count.na <- function (dataset, feature.name) {
  num.features <- length(st_drop_geometry(dataset))
  return(sum(is.na(dataset[[feature.name]])) / nrow(dataset))
}

select.features <- function(dataset, feature_names, a) {
  counts <- c()
  for (x in feature_names) {
    counts <- append(counts, count.na(wals_sam, x))
  }
  result <- t(as.matrix(counts[counts<0.5]))
  colnames(result) <- feature_names[which(counts<0.5)]

  return(result)
}

feature_names <- colnames(wals_sam)[-c(1, 194:208)]
select.features(wals_sam, feature_names, 0.5)
```

```
##           81A    82A    83A        86A    129A
## [1,] 0.4875 0.425 0.4125 0.4833333 0.4875

glottodata.wals <- function (continent_name) {
  structure <- glottocreate_structutable(rownames = c("81A", "82A", "83A", "86A", "129A"))
  structure$type <- rep("factor", 5)
  wals_data <- subset(wals, continent == continent_name)
  data <- select(wals_data, 'glottocode', '81A', '82A', '83A', '86A', '129A')
  data.df <- st_drop_geometry(data)

  data.df$`81A` <- as.factor(data.df$`81A`)
  data.df$`82A` <- as.factor(data.df$`82A`)
  data.df$`83A` <- as.factor(data.df$`83A`)
  data.df$`86A` <- as.factor(data.df$`86A`)
  data.df$`129A` <- as.factor(data.df$`129A`)

  data.unique <- match(unique(data.df$glottocode),
                        data.df$glottocode)
  data.df.unique <- data.df[data.unique, ]
  glottodata <- glottocreate_addtable(data.df.unique, structure,
                                     name="structure")

  return(glottodata)
}

glottodata_sam <- glottodata.wals("South America")
glottodata_asia <- glottodata.wals("Asia")
```

## 2.2 The Hausdorff distance

```
overlap.dist <- function (x, y){
  sum(x != y) / length(x)
}

overlap.x.Y.dist <- function(x, Y) {
  result <- c()
  for (i in 1:nrow(Y)) {
    result <- append(result, overlap.dist(x, Y[i,]))
  }
  return(min(result))
}

overlap.supX.Y.dist <- function(X, Y) {
  result <- c()
  for (i in 1:nrow(X)) {
    result <- append(result, overlap.x.Y.dist(X[i, ], Y))
  }
  return(max(result))
}

overlap.hausdorff.dist <- function(X, Y) {
  X <- glottoclean(X)
  structure <- X[["structure"]]
  X <- glottosimplify(X)
  X <- tibble::column_to_rownames(X, "glottocode")
```

```

X <- data.frame(lapply(X, as.character),
                stringsAsFactors = FALSE)
X[is.na(X)] <- "unknown"

Y <- glottoclean(Y)
Y <- glottosimplify(Y)
Y <- tibble::column_to_rownames(Y, "glottocode")
Y <- data.frame(lapply(Y, as.character),
                stringsAsFactors = FALSE)
Y[is.na(Y)] <- "unknown"

result <- max(overlap.supX.Y.dist(X, Y), overlap.supX.Y.dist(Y,X))
return(result)
}

```

Compute the Hausdorff distance between South America and Asia:

```
overlap.hausdorff.dist(glottodata_asia, glottodata_sam)
```

```

## Missing values recoded to NA
##
## Missing values recoded to NA
## [1] 0.6

```

## 2.3 TDA

### 2.3.1 The linguistic syntactic data of South America and Asia

```
glottodist_sam <- glottodist(glottodata_sam)
```

```

## Missing values recoded to NA
## All variables have two or more levels (excluding NA)
glottodist_sam[is.na(glottodist_sam)] <- 0

```

```
glottodist_asia <- glottodist(glottodata_asia)
```

```

## Missing values recoded to NA
## All variables have two or more levels (excluding NA)
glottodist_asia[is.na(glottodist_asia)] <- 0

```

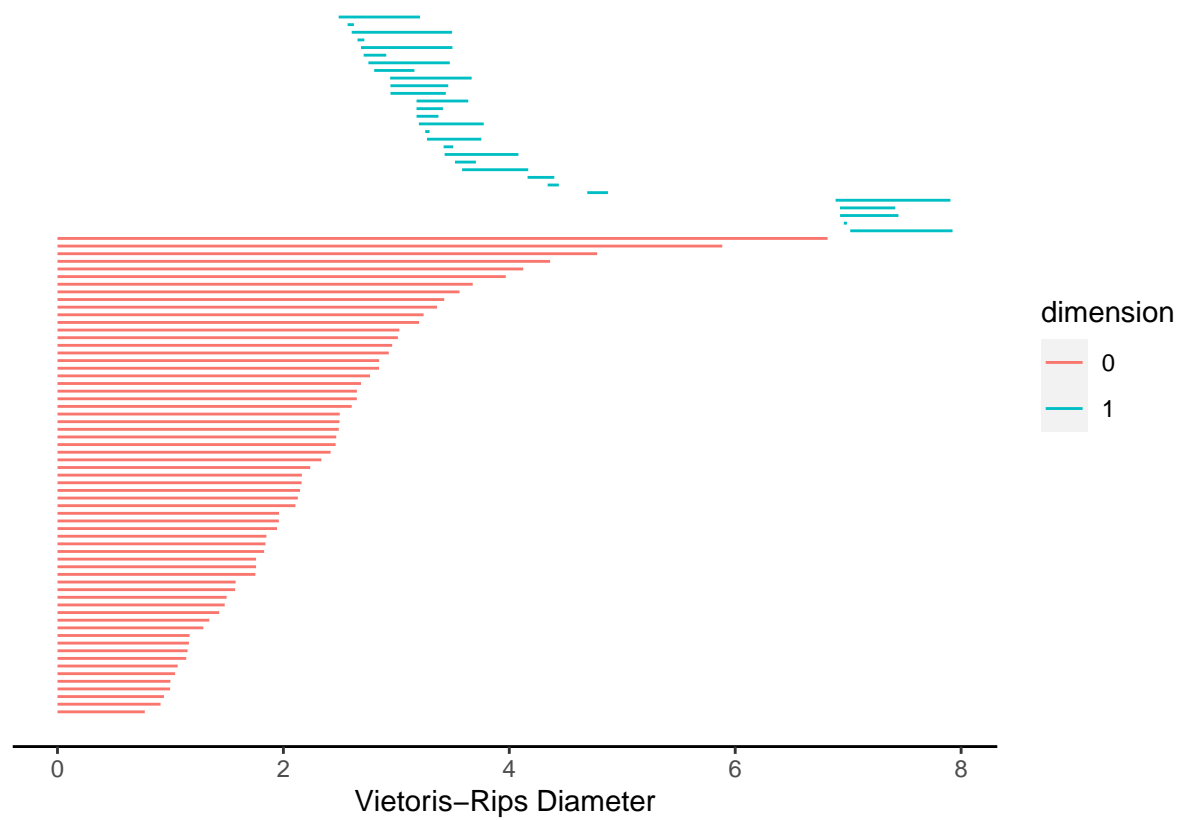
The persistence barcode and persistence diagram of South America language syntactic structures

```
sam.phom <- calculate_homology(glottodist_sam)
```

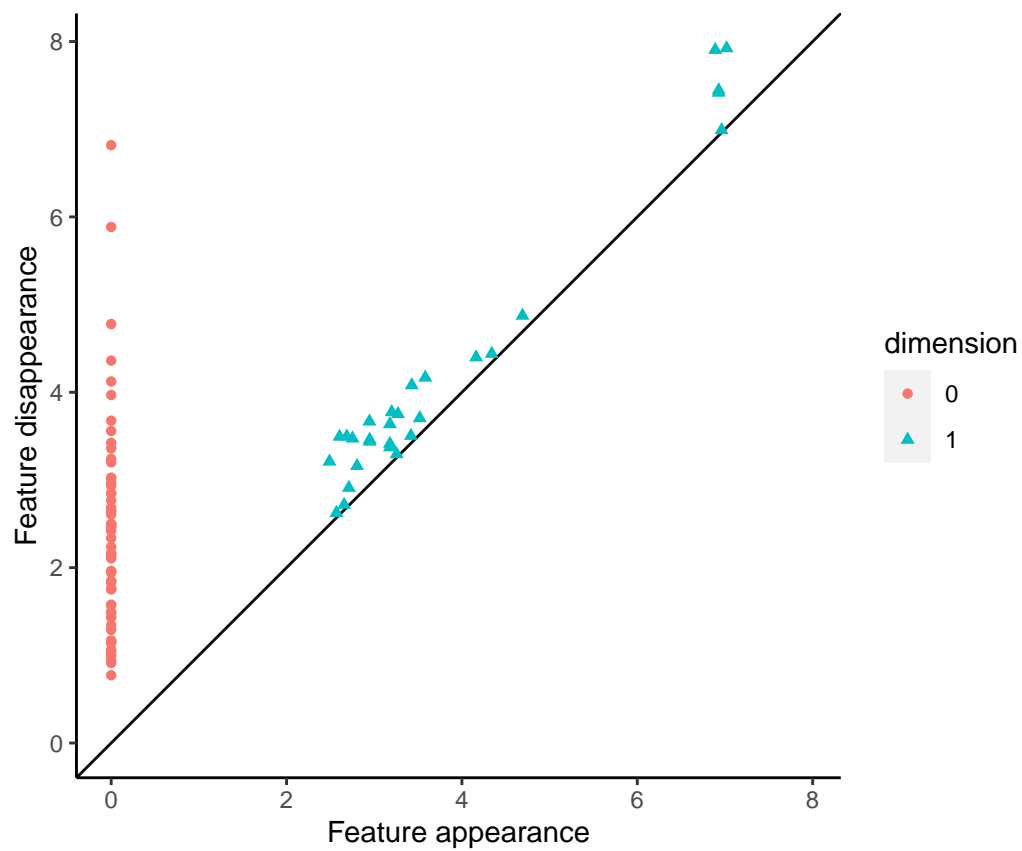
```

par(mfrow=c(1,2))
plot_barcode(sam.phom)

```



```
plot_persist(sam.phom)
```

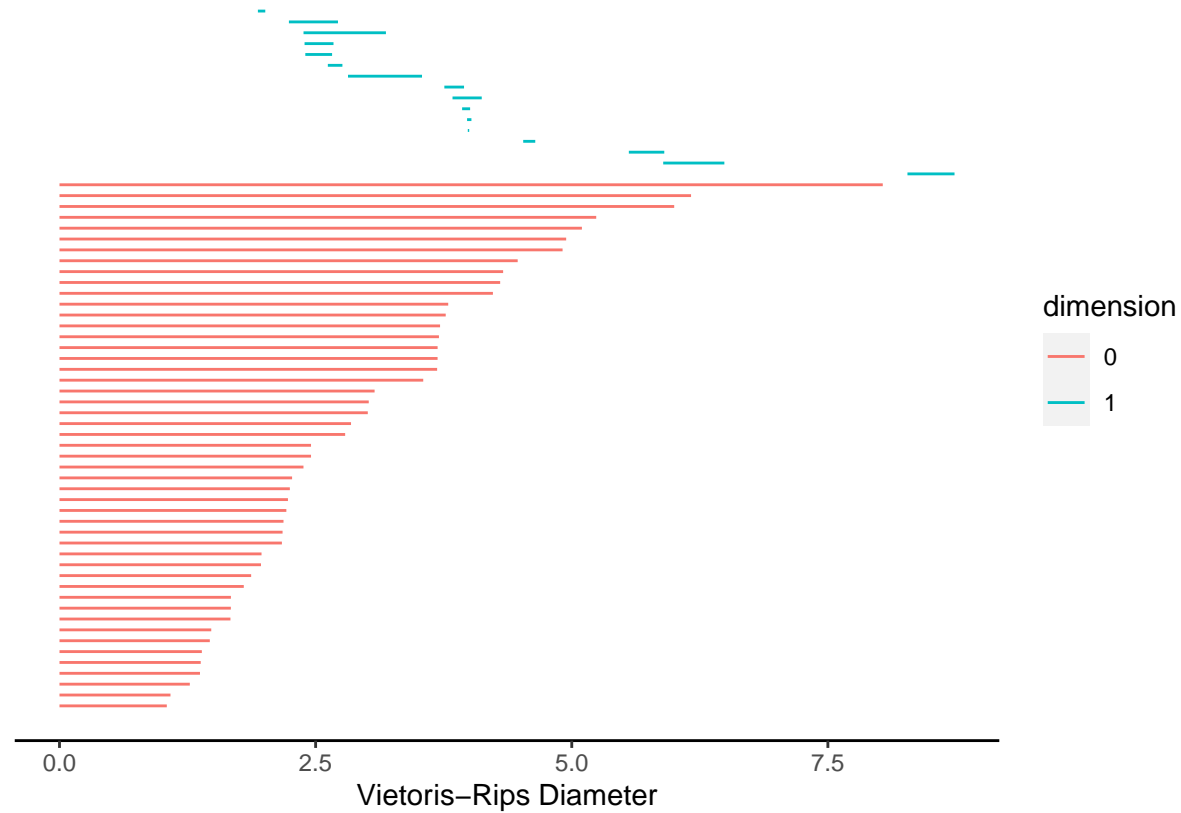


The persistence barcode and persistence diagram of Asia language syntactic structures

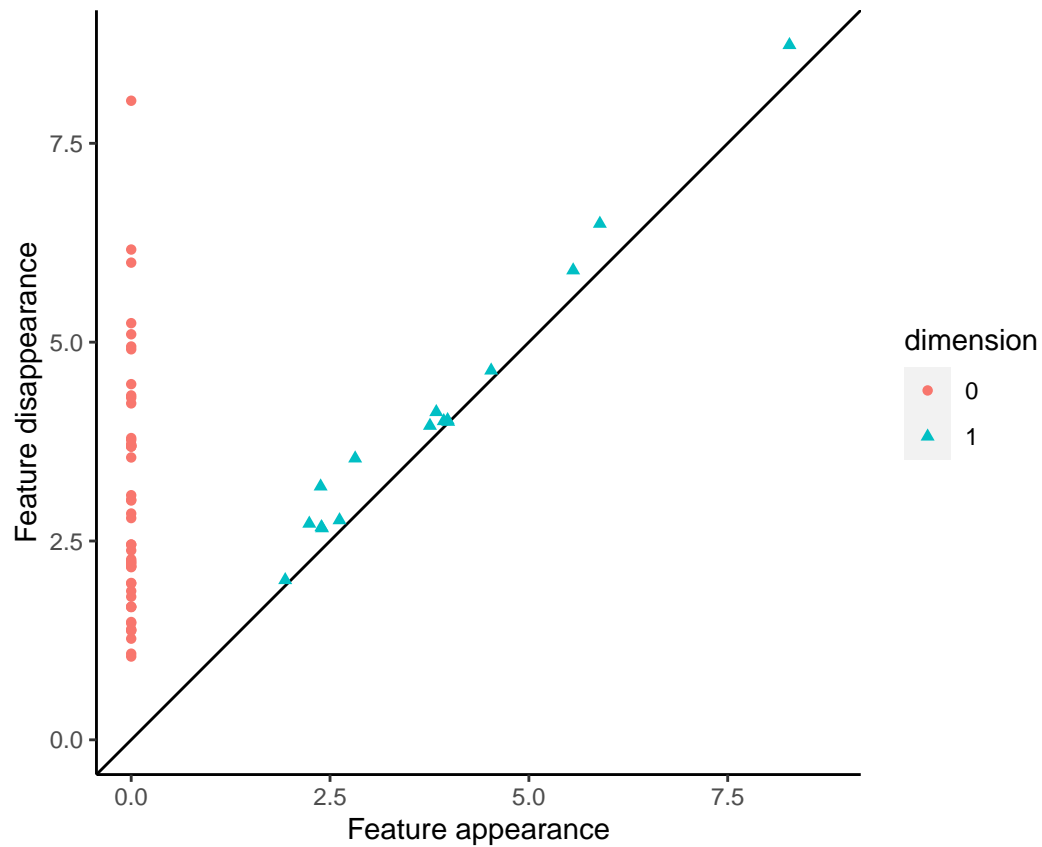
```
asia.phom <- calculate_homology(glottodist_asia)
```

```
par(mfrow=c(1,2))
```

```
plot_barcode(asia.phom)
```



```
plot_persist(asia.phom)
```

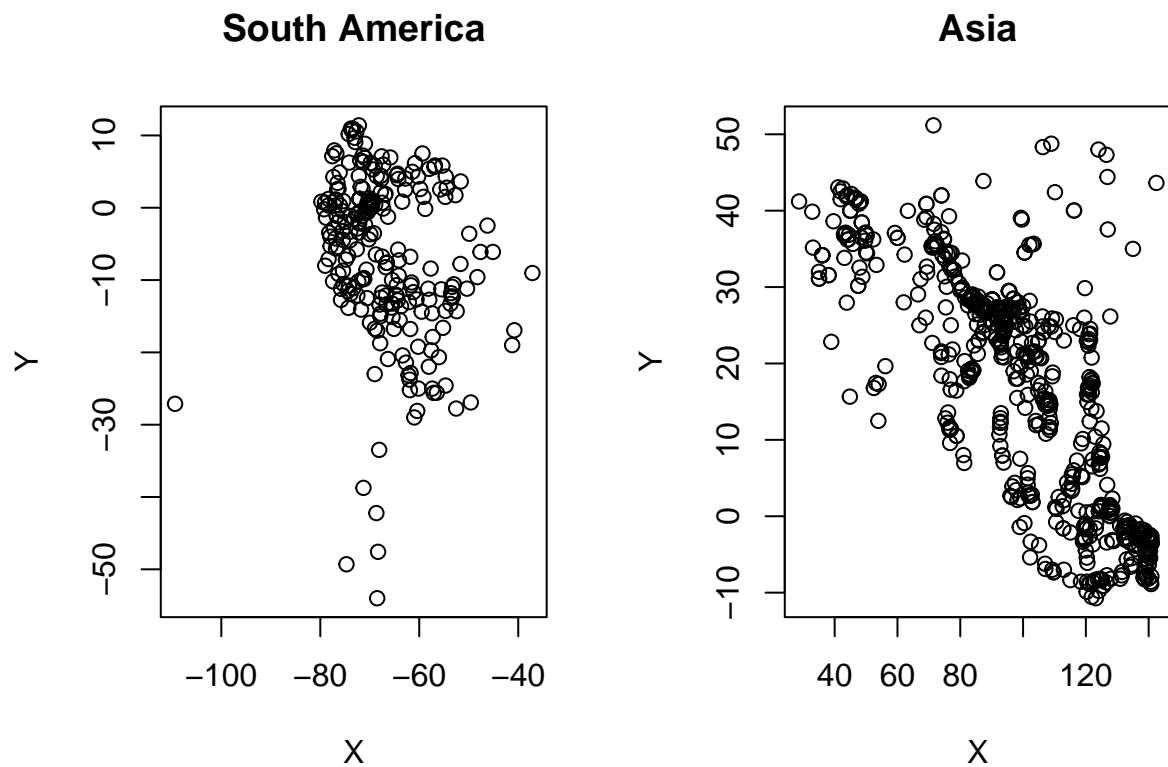


### 2.3.2 The spatial data of South America and Asia

```
sam_coordinates <- st_coordinates(wals_sam)
asia_coordinates <- st_coordinates(wals_asia)

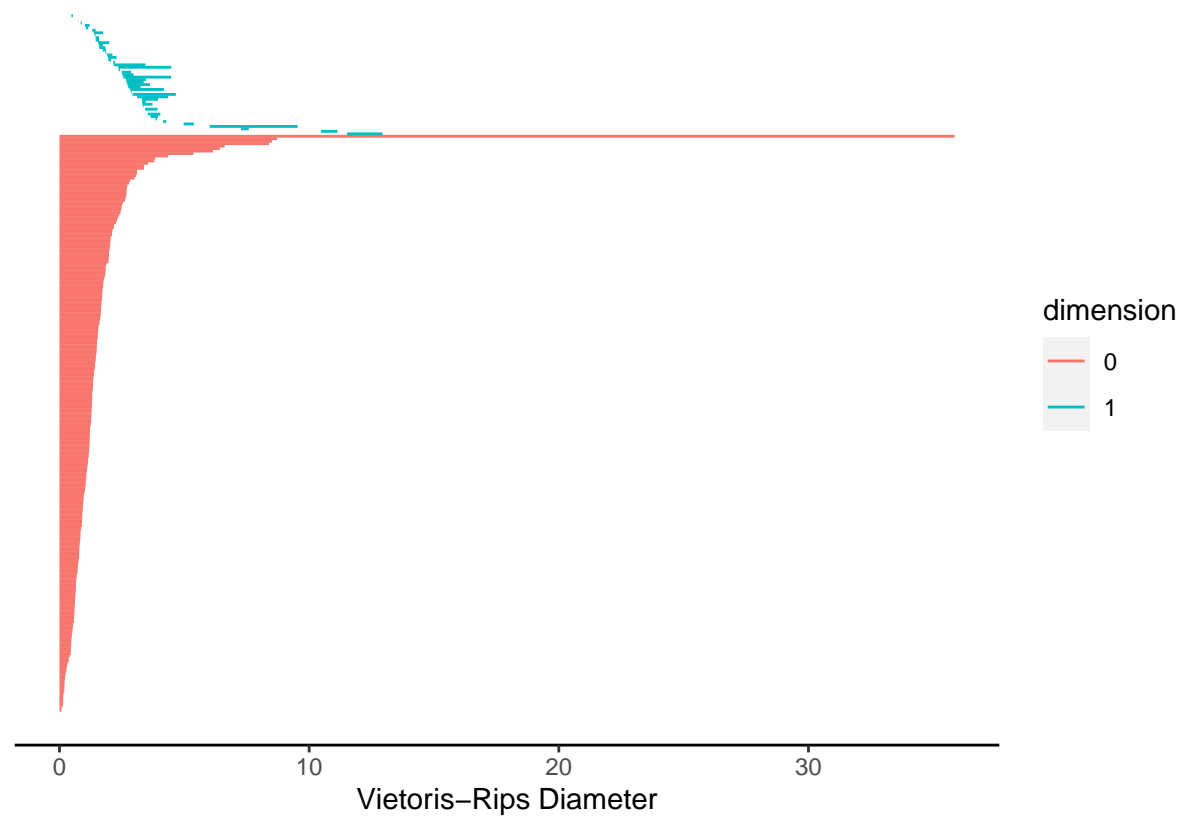
sam_geo_phm <- calculate_homology(sam_coordinates)
asia_geo_phm <- calculate_homology(asia_coordinates)

par(mfrow=c(1,2))
plot(sam_coordinates, main="South America")
plot(asia_coordinates, main="Asia")
```



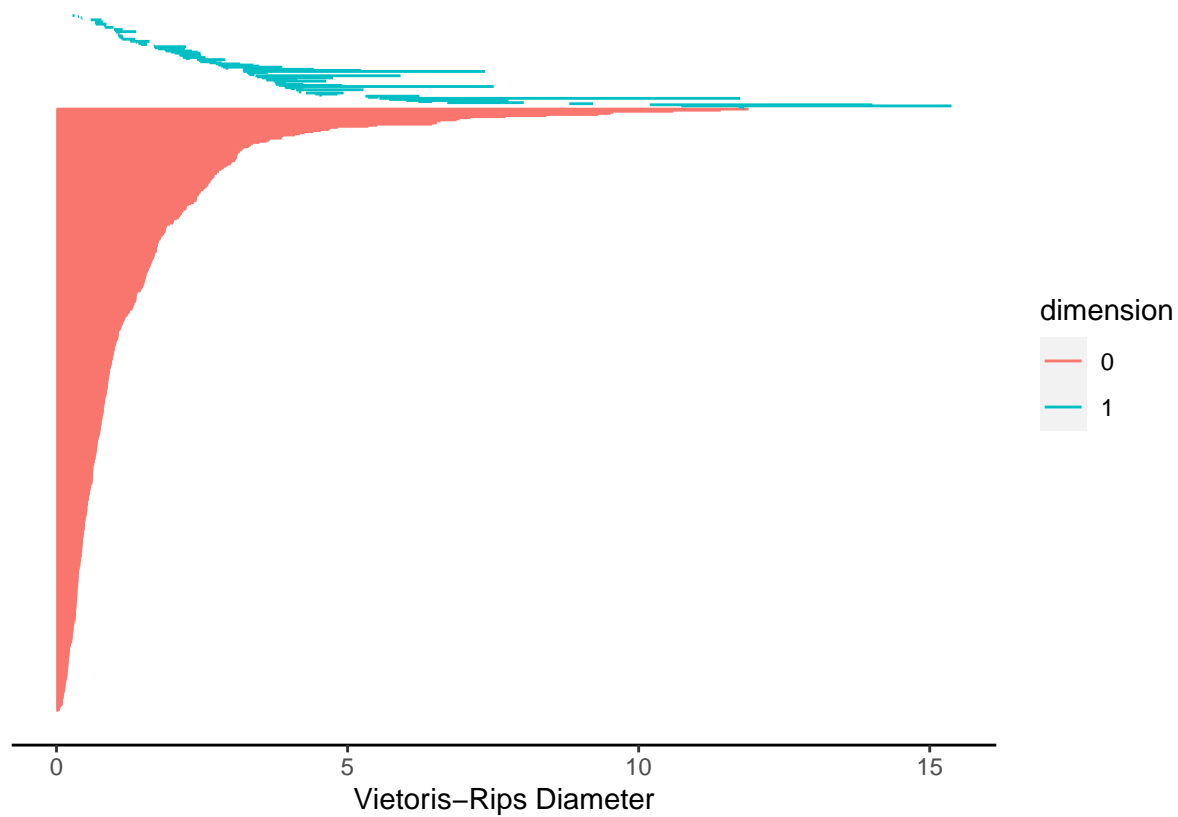
The persistence barcodes of spatial data of South America and Asia:

```
par(mfrow=c(1,2))
plot_barcode(sam_geo_phm)
```



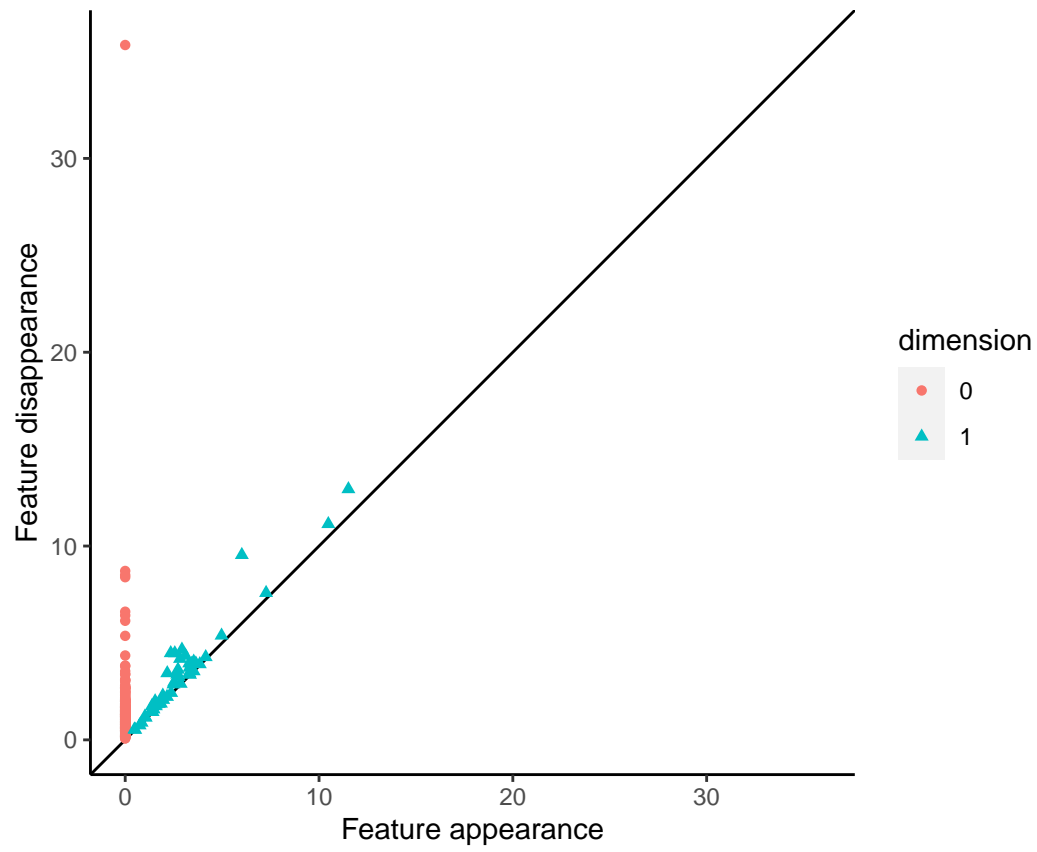


```
plot_barcode(asia_geo_phm)
```

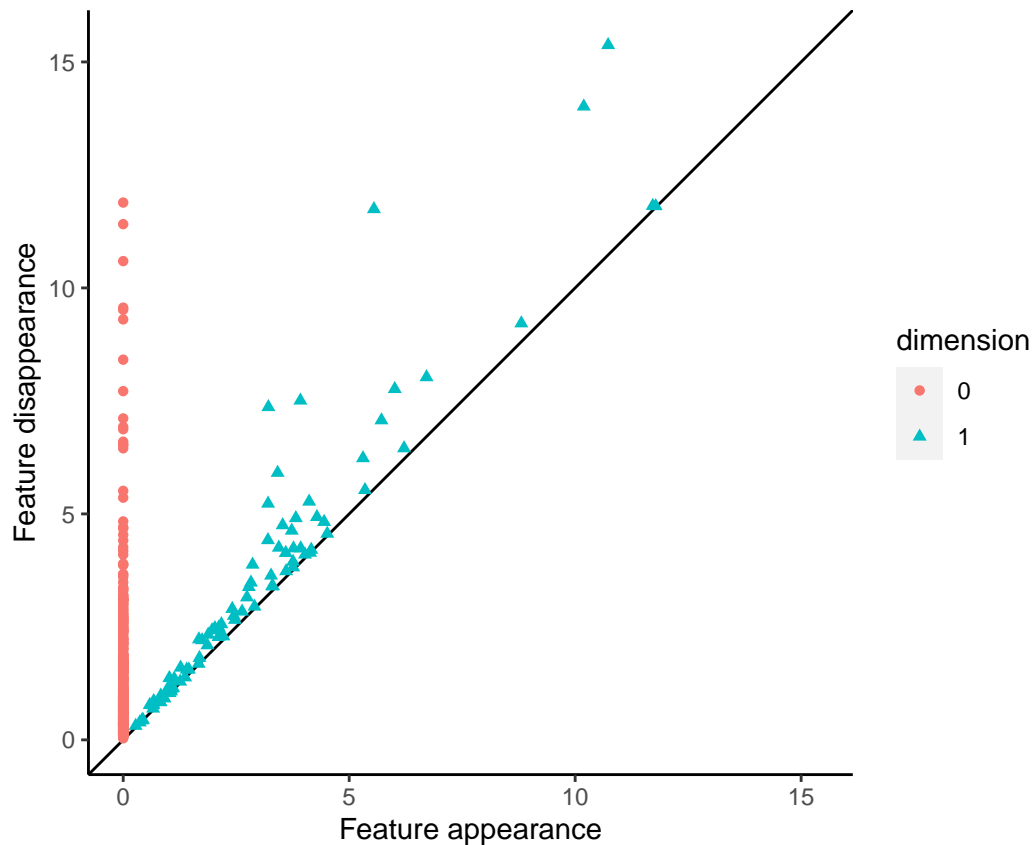


The persistence diagrams of spatial data of South America and Asia:

```
par(mfrow=c(1,2))  
plot_persist(sam_geo_phm)
```



```
plot_persist(asia_geo_phm)
```



## 2.4 Futher codes

Codes about spectral graph theory methods and heat kernel analysis methods like described in [Ortegaray et al., 2021] in the future?

## References

- [Boriah et al., 2008] Boriah, S., Chandola, V., and Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In *SDM*.
- [Feng et al., 2022] Feng, M., Hickok, A., and Porter, M. A. ([2022] ©2022). Topological data analysis of spatial systems. In *Higher-order systems*, Underst. Complex Syst., pages 389–399. Springer, Cham.
- [Ortegaray et al., 2021] Ortegaray, A., Berwick, R. C., and Marcolli, M. (2021). Heat kernel analysis of syntactic structures. *Math. Comput. Sci.*, 15(4):643–660.
- [Port et al., 2018] Port, A., Gheorghita, I., Guth, D., Clark, J. M., Liang, C., Dasu, S., and Marcolli, M. (2018). Persistent topology of syntax. *Math. Comput. Sci.*, 12(1):33–50.
- [Port et al., 2022] Port, A., Karidi, T., and Marcolli, M. (2022). Topological analysis of syntactic structures. *Math. Comput. Sci.*, 16(1):Paper No. 2, 68.