Methods for distance

Principal Coordinates Analysis and related methods

Stéphane Dray

2024-01-28

Introduction

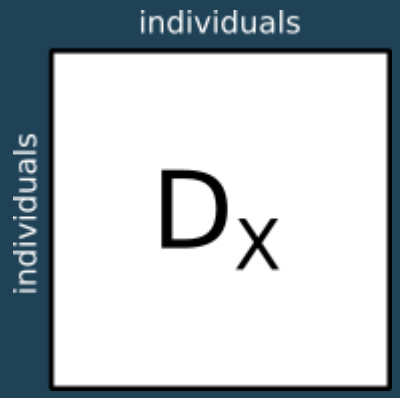
```
library(ade4)
library(adegraphics)
adegpar(paxes.draw = TRUE, pbackground.col = "lightgrey",
    pgrid.col = "white")
data(yanomama)
str(yanomama, max.level = 1)
```

```
## List of 3
##  $ geo: num [1:19, 1:19] 0 9 28 152 149 169 172 253 244 82 ...
##  $ gen: num [1:19, 1:19] 0 35 44 47 52 57 65 69 34 47 ...
##  $ ant: num [1:19, 1:19] 0 96 147 295 284 253 289 507 488 203 ...
```

Differences among 19 Yanomama Indian villages, 3 distance matrices:

- Geographic
- Genetic
- Anthropometric

Introduction

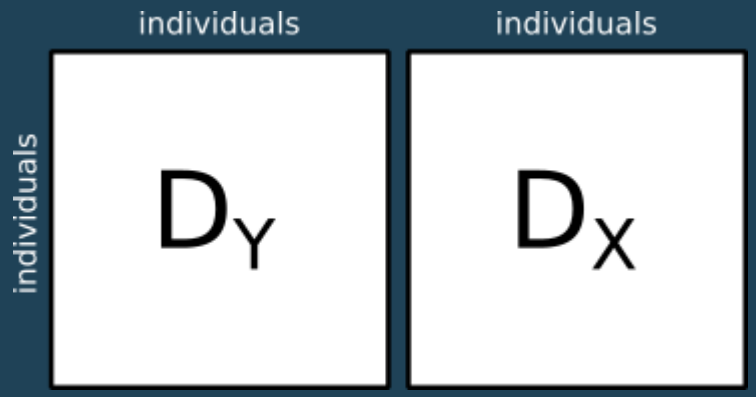


One distance matrix measured on *n* individuals

Describe the information contained in the table:

- Identify the differences/similarities between individuals

Introduction



Two distance matrices measured on *n* individuals

Describe the information contained in the two table:

- Identify the differences/similarities between individuals in each table
- Indentify relationships between both matrices

Why distances?

- Biological or ecological hypothesis can be defined using distances
- From spatial, genetic, phylogenetic, ecological data, etc.
- Distances can be directly measurued
- Distances can be computed from raw data
    - binary
    - numeric
    - percentage
    - factor
    - etc.
- Own definition of differences between individuals.

Dissimilarity and distances

Let E be a set of n individuals. Let x, y, z be 3 elements of E

A **dissimilarity** is an application $d : E \times E \to \mathbb{R}^+$ satisfying

(1) : $\mathrm{d}(x, y) \geq 0$

(2) : $\mathrm{d}(x, y) = 0 \Rightarrow x = y$

(3) : $\mathrm{d}(x, y) = d(y, x)$

The dissimilarity is **metric** (a.k.a **distance**) if the following condition is satisfied

(4) : $\mathrm{d}(x, y) \leq \mathrm{d}(x, z) + \mathrm{d}(z, y)$

It is **ultrametric** if

(4) : $\mathrm{d}(x, y) \leq max(\mathrm{d}(x, z), \mathrm{d}(z, y))$

It is euclidean if we can define n points in an Euclidean space so that distances computed are exactly the dissimilarities.

Methods based on distance matrices

in R, objects of class `dist`

- ultrametric $\longrightarrow$ clustering / classification
- euclidean $\longrightarrow$ Principal Coordinates Analysis (PCoA)
- non-euclidean $\longrightarrow$ Non-Metric Multidimensional Scaling (NMDS)

Principal coordinates analysis

- PCA, CA methods induce implicitly a way to compute distances
- Several other distances have been proposed (e.g., genetic, presence-absence)
- PCoA takes a distance matrix as input and returns coordinates in a low dimensional space that best preserve the original distances.

- 😄 it allows to choose a particular distance measure between sites (or species).

- 😢 it focuses either on individuals or variables, not both.

- Useful if distances are directly recorded or computed from raw data tables

PCoA algorithm

PCoA is based on the diagonalization of a bicentered matrix of squared distances.

$$\mathbf{H} = \left[-\frac{1}{2} d_{ij}^2 - m_i - m_j + m \right]$$

$$\mathbf{H} = \mathbf{U}\Lambda\mathbf{U}^t = \mathbf{U}\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}\mathbf{U}^t = \mathbf{X}\mathbf{X}^t$$

What about non-euclidean distances

```
geo <- as.dist(yanomama$geo)
is.euclid(geo)
```

```
## [1] FALSE
```

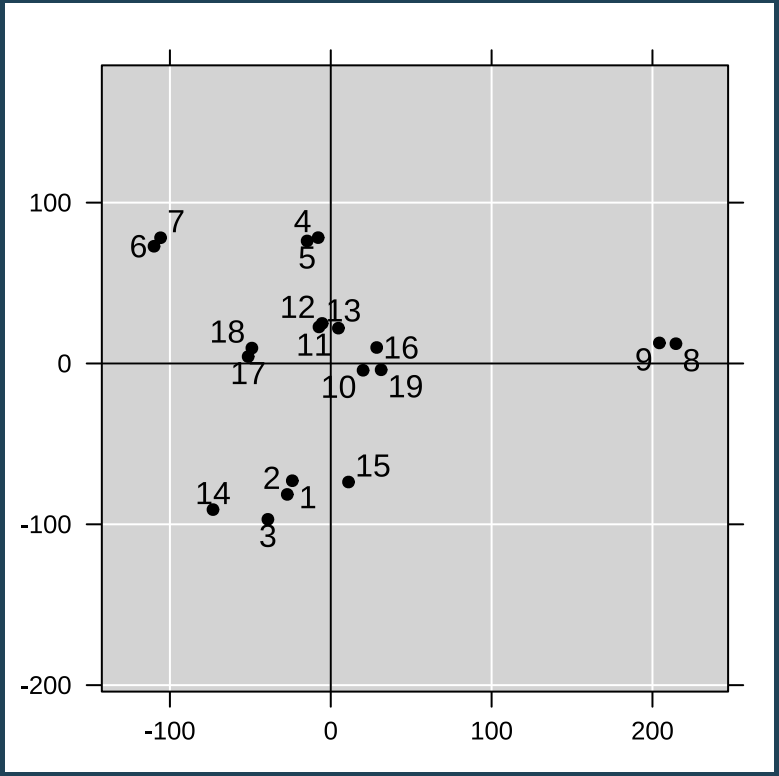
In this case, PCoA would returns negative eigenvalues as no euclidean representation is possible to preserves exactly the distances.

Some solutions:

- `cailliez` / `lingoes` : add constant values to distances
- `quasieuclid` : modify distances for quasi-euclidean distances
- NMDS : aims to preserve orders of distances (`vegan::metaMDS`)

PCoA in practice

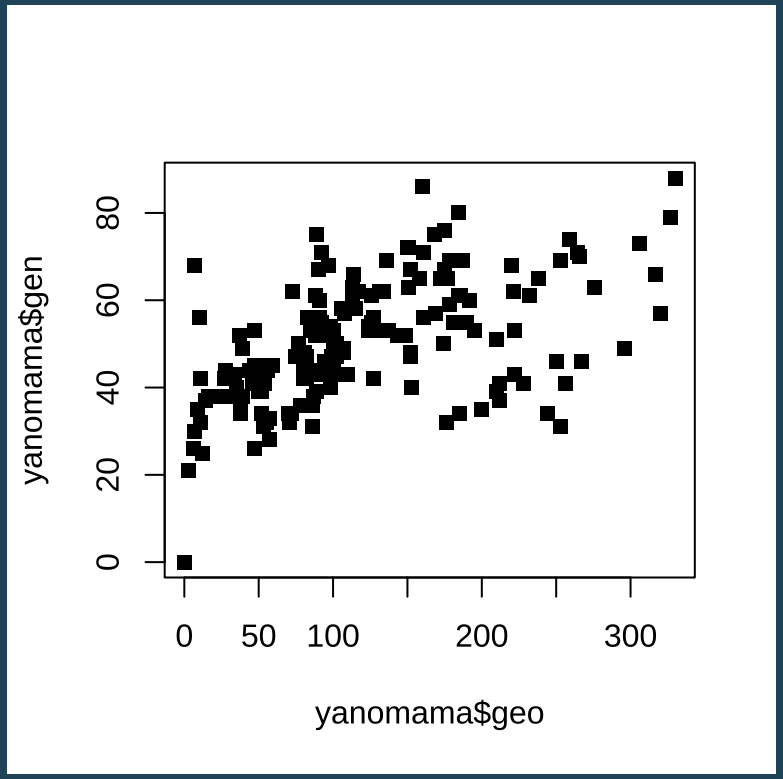
```
geo <- quasieuclid(geo)
pco.geo <- dudi.pco(geo, scan = F)
s.label(pco.geo$li, plabel.optim = TRUE)
```



Relationships among distance matrices

Is there a link between spatial distances and allelic distances ?

```
plot(yanomama$gen ~ yanomama$geo, pch = 15)
```



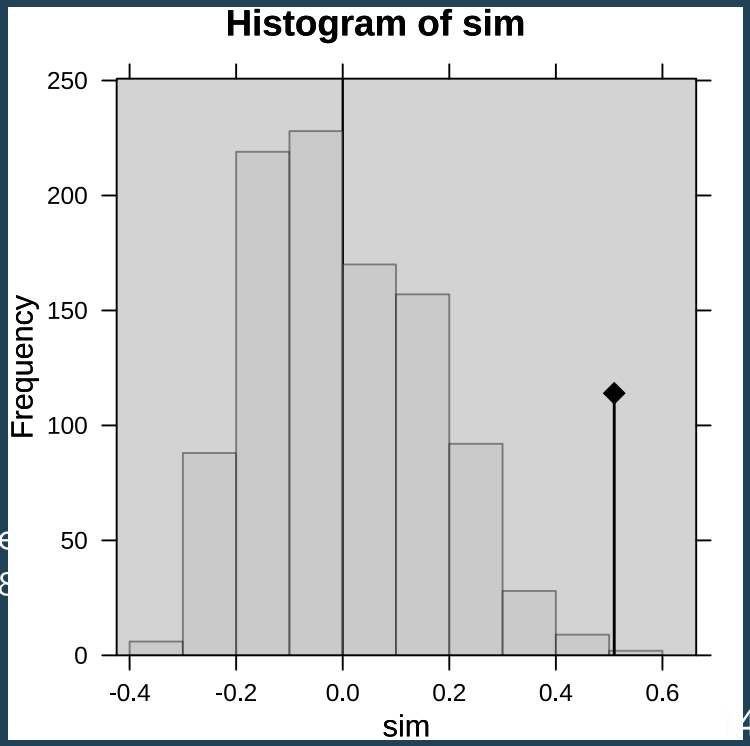
Mantel test

Permutation-based testing procedure based on linear correlation (Pearson's). Can be monotonic (not only linear) if Spearman's correlation is used:

```
gen <- quasieuclid(as.dist(yanoma
m1 <- mantel.randtest(gen, geo)
m1
```

```
## Monte-Carlo test
## Call: mantel.randtest(m1 = gen, m2
##
## Observation: 0.5095199
##
## Based on 999 replicates
## Simulated p-value: 0.003
## Alternative hypothesis: greater
##
##     Std.Obs Expectation    Variance
## 3.106564135 0.001964964 0.026693498
```

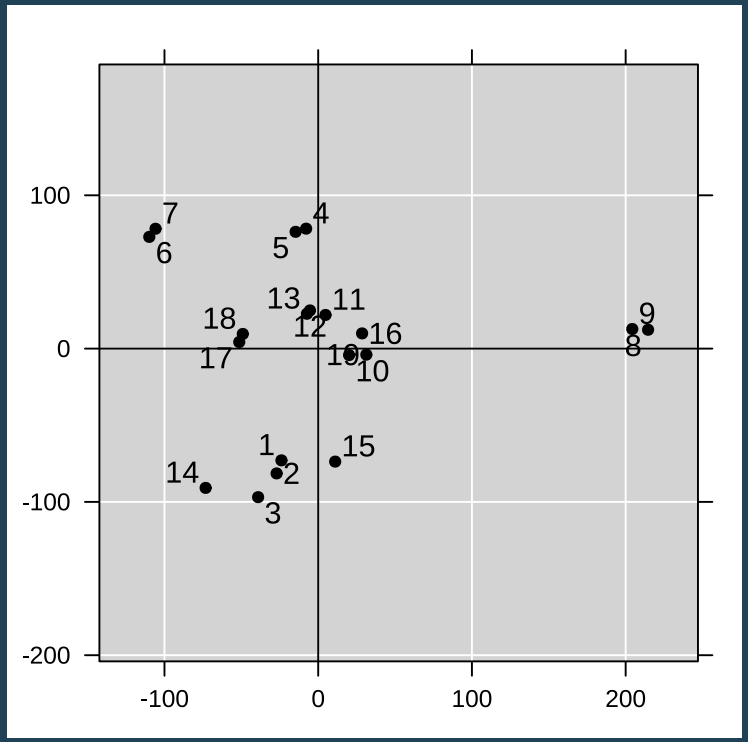
```
plot(m1)
```



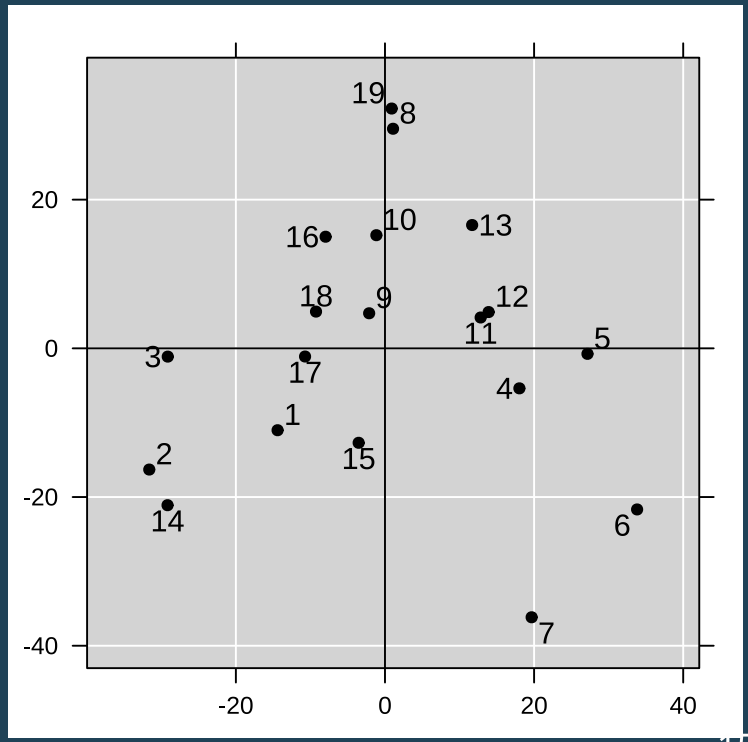
Comparing two configurations

```
pco.gen <- dudi.pco(gen, scan = F)
```

```
s.label(pco.geo$li, plabel.optim
```

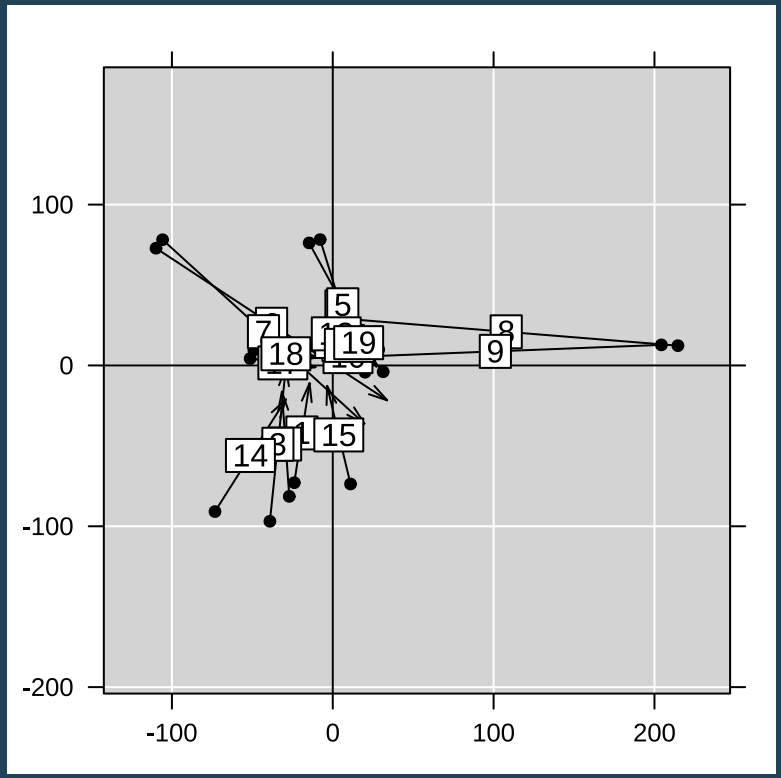



```
s.label(pco.gen$li, plabel.optim
```



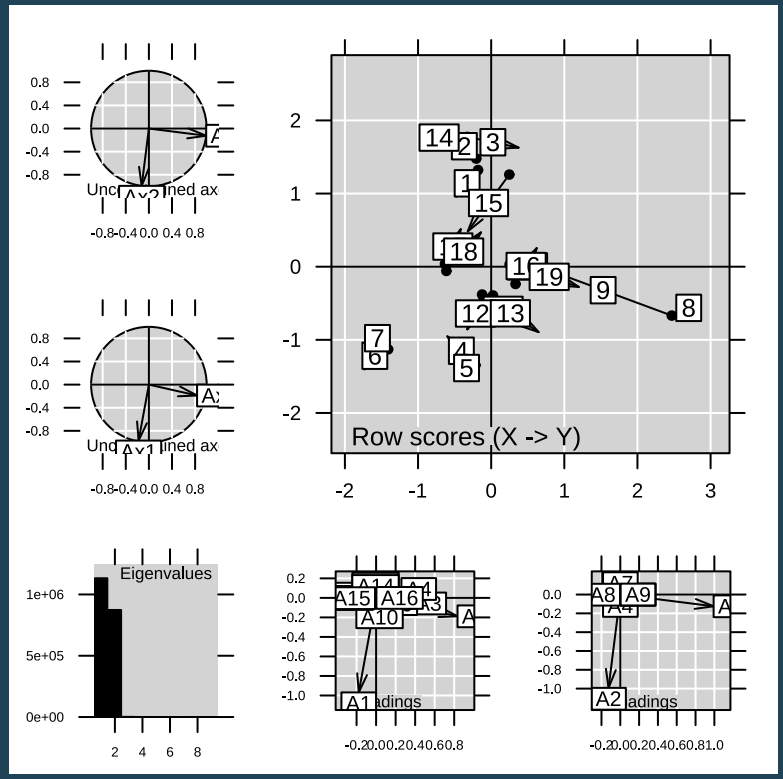
Fit the configurations of points

```
s.match(pco.geo$li, pco.gen$li, plabel.optim = TRUE)
```



Rotate, reflect

```
coi <- coinertia(pco.geo, pco.gen, scannf = FALSE)
plot(coi)
```



RV coefficient

```
RV.randtest(pco.geo$li, pco.gen$li)
```

```
## Monte-Carlo test
## Call: RV.randtest(df1 = pco.geo$li, df2 = pco.gen$li)
##
## Observation: 0.5128779
##
## Based on 999 replicates
## Simulated p-value: 0.001
## Alternative hypothesis: greater
##
##     Std.Obs Expectation    Variance
## 5.728076144 0.106374766 0.005036284
```