

Training in ade4 in R - Module I: Basic methods

Correspondence analysis and Principal Coordinates Analysis

Stéphane Dray

2023-12-03

Data structure



- One table with n rows and m columns
- Data are counts of individuals (sums by row and columns are meaningful)
- For instance
 - sites \times species

Objectives

- Identify what is the main information contained in the table
 - Identify the principal differences/similarities between row categories
 - Identify the principal differences/similarities between column categories
 - Identify the principal differences/similarities between row and column categories

Data

We consider the **meaudret** data set

```
library(ade4)
data(meaudret)
names(meaudret)
```

```
## [1] "env"      "design"    "spe"      "spe.names"
```

```
dim(meaudret$spe)
```

```
## [1] 20 13
```

```
head(meaudret$spe.names)
```

```
## [1] "Ephemera_danica" "Baetis_sp"      "Baetis_rhodani" "Baetis_niger"
```

The data set contains the abundances of 13 Ephemeroptera species in 20 samples. The measurements have been made in 6 sites at each season along a small French stream (see [?meaudret](#))

```
head(meaudret$design)
```

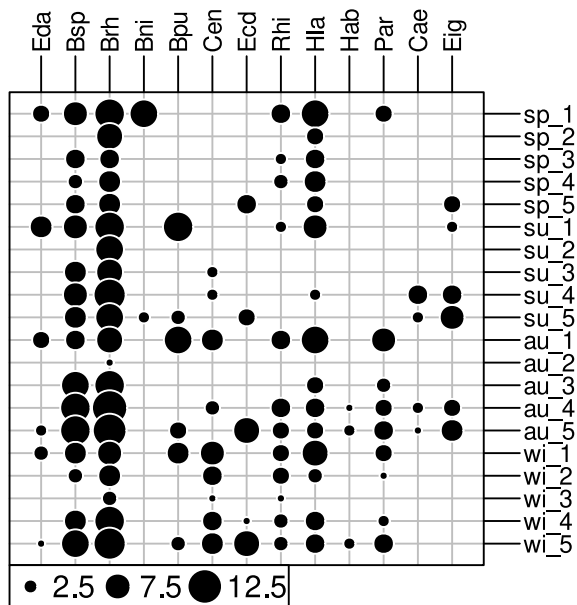
```
##      season site
## sp_1 spring  S1
## sp_2 spring  S2
## sp_3 spring  S3
## sp_4 spring  S4
## sp_5 spring  S5
## su_1 summer  S1
```

We want to know

- which species have similar distributions
- which sites have similar composition
- which species are mainly present in which sites

Contingency table

```
library(adegraphics)
table.value(meaudret$spe, symbol
```



```
head(rowSums(meaudret$spe))
```

```
## sp_1 sp_2 sp_3 sp_4 sp_5 su_1
##    48   12   17   18   24   44
```

```
head(colSums(meaudret$spe))
```

```
## Eda Bsp Brh Bni Bpu Cen
##   20 104 163  11  35  37
```

```
sum(meaudret$spe)
```

```
## [1] 595
```

Chi-square test

The χ^2 test allows to measure and evaluate the significance of the association between species and sites (the null hypothesis is the random distribution)

$$\chi_{obs}^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{\left(y_{ij} - \frac{y_{i.} y_{.j}}{y_{..}} \right)^2}{\frac{y_{i.} y_{.j}}{y_{..}}}$$

```
chisq.test(meaudret$spe)
```

```
## Warning in chisq.test(meaudret$spe): L'approximation du Chi-2 est peut-être
```

```
##
```

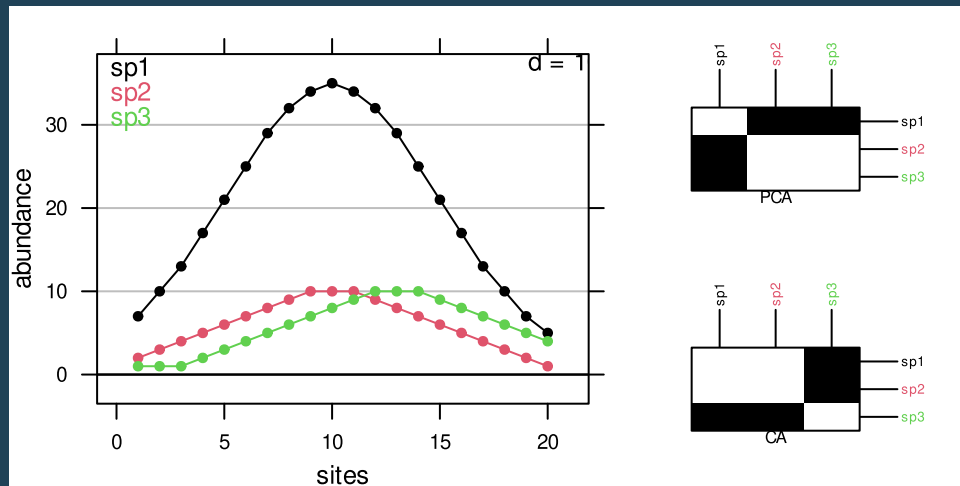
```
##      Pearson's Chi-squared test
```

```
##
```

```
## data:  meaudret$spe
```

```
## X-squared = 534.52, df = 228, p-value < 2.2e-16
```

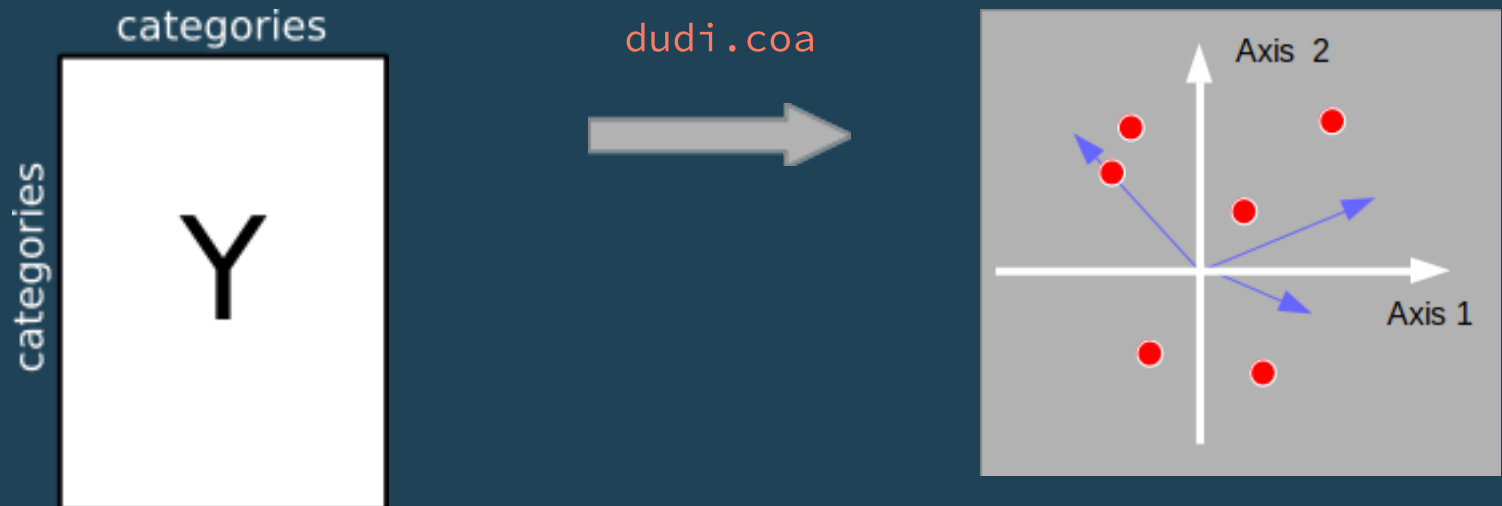

Absolute and relative frequencies



- In PCA (Euclidean distances), species 2 and 3 are closer
- In COA (χ^2 distances), species 1 and 2 are closer (null distance)

Correspondence analysis

- $\mathbf{X} = \mathbf{D}_n^{-1} \mathbf{P} \mathbf{D}_m^{-1} - \mathbf{1}_n \mathbf{1}_m^\top$ is the transformed and centred table of relative frequencies with $\mathbf{P} = [y_{ij}/y_{..}]$
- $\mathbf{Q} = \mathbf{D}_m$ where $\mathbf{D}_m = \text{diag}(\mathbf{P}^\top \mathbf{1}_n)$ contains the column category frequencies
- $\mathbf{D} = \mathbf{D}_n$ where $\mathbf{D}_n = \text{diag}(\mathbf{P} \mathbf{1}_m)$ contains the row category frequencies



Maximized criteria

- For rows

$$Q(\mathbf{a}) = \|\mathbf{D}_n^{-1} \mathbf{P}_0 \mathbf{D}_m^{-1} \mathbf{D}_m \mathbf{a}\|_{\mathbf{D}_n}^2 = \|\mathbf{D}_n^{-1} \mathbf{P}_0 \mathbf{a}\|_{\mathbf{D}_n}^2$$

In this viewpoint, columns have a unit-variance score \mathbf{a} that maximises the variance between the row barycenters.

- For columns

$$\|\mathbf{D}_m^{-1} \mathbf{P}_0^\top \mathbf{D}_n^{-1} \mathbf{D}_n \mathbf{b}\|_{\mathbf{D}_m}^2 = \|\mathbf{D}_m^{-1} \mathbf{P}_0^\top \mathbf{b}\|_{\mathbf{D}_m}^2$$

In this viewpoint, rows have a unit-variance score \mathbf{b} that maximises the variance between the column barycenters.

The `dudi.coa` function

Arguments

```
args(dudi.coa)
```

```
## function (df, scannf = TRUE, nf = 2)  
## NULL
```

- `df` is a `data.frame` with the positive values (counts)
- `scannf` and `nf` allow to set the number of dimensions to interpret

```
coa.meau <- dudi.coa(meaudret$spe, scannf = FALSE)
```

Returned values

```
names(coa.meau)
```

```
## [1] "tab" "cw" "lw" "eig" "rank" "nf" "c1" "li" "co" "l1"
```

It returns an object of class **dudi** containing:

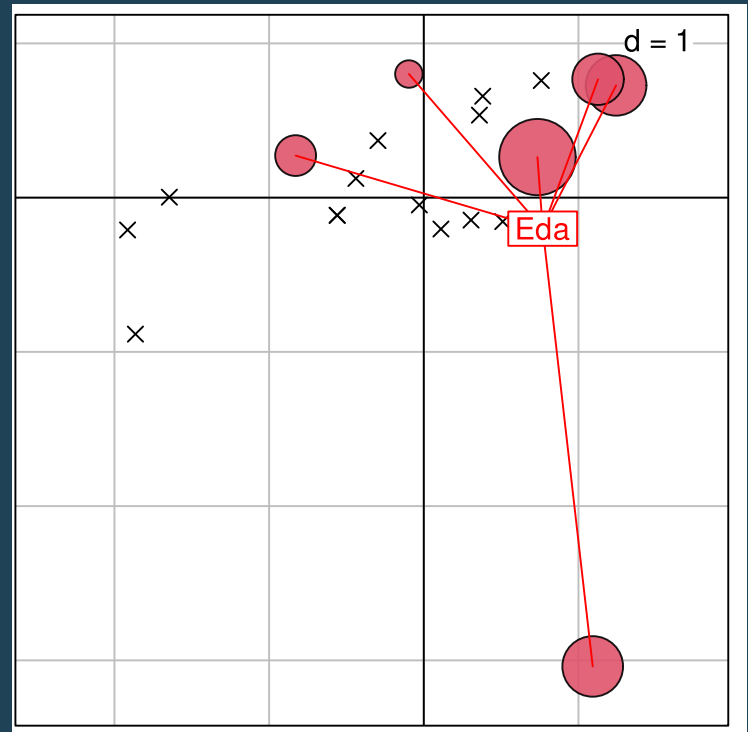
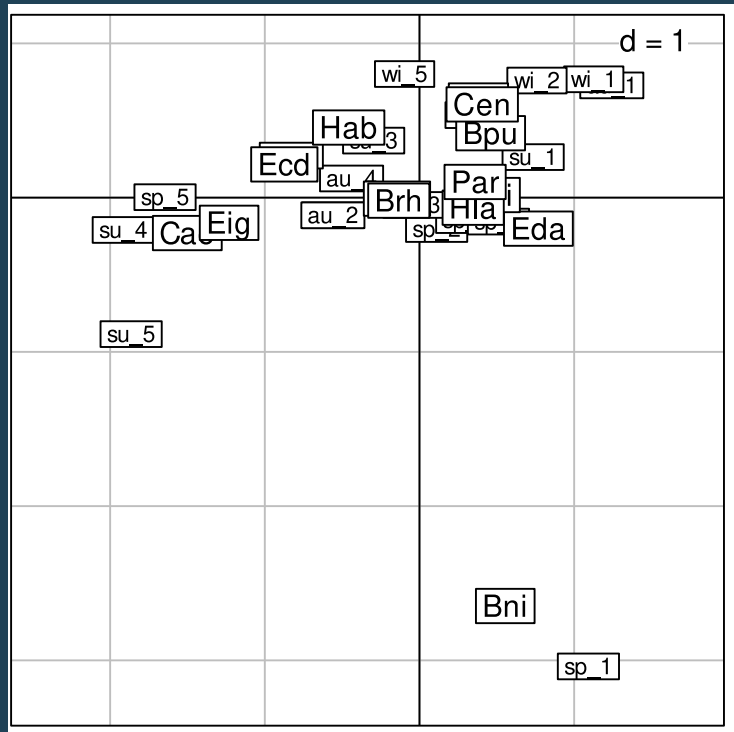
- **\$eig**: eigenvalues (Λ)
- **\$cw**: column weights (\mathbf{D}_m)
- **\$lw**: row weights (\mathbf{D}_n)
- **\$tab**: centred relative frequencies table ($\mathbf{D}_n^{-1} \mathbf{P}_0 \mathbf{D}_m^{-1}$)
- **\$c1**: unit-variance column scores (\mathbf{A})
- **\$li**: row scores as weighted averages ($\mathbf{L} = \mathbf{D}_n^{-1} \mathbf{P}_0 \mathbf{A}$)
- **\$l1**: unit-variance row scores (\mathbf{B})
- **\$co**: column scores as weighted averages ($\mathbf{C} = \mathbf{D}_m^{-1} \mathbf{P}_0^\top \mathbf{B}$)
- **\$N**: total sum ($y_{..}$)

Graphical representations

Biplot can be produced for CA using the `biplot` function. Three types of biplots can be produced using the argument `method`

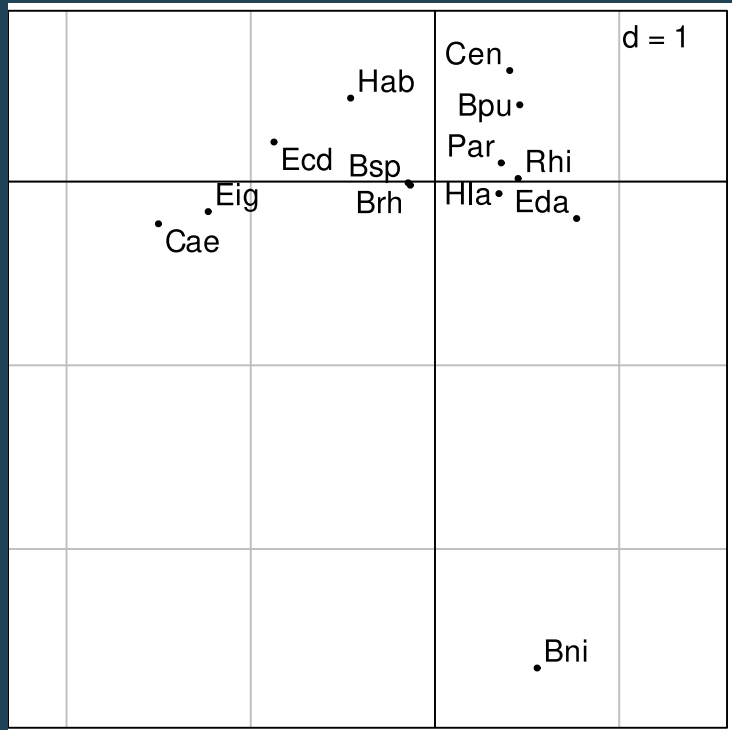
- If `method = 2`, species are positioned by a unit-variance score (`$c1`) and sites by weighted averaging (`li`).
- If `method = 3`, sites are positioned by a unit-variance score (`$l1`) and species by weighted averaging (`$co`).
- By default, `method = 1` corresponds to a compromise between these two representations (`$li` and `$co`).

Weighted averaging

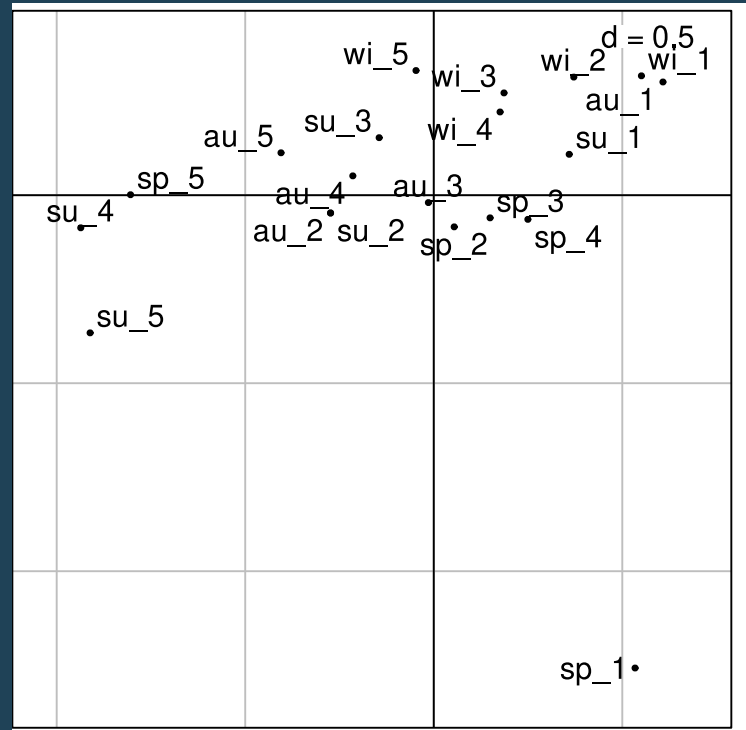


Separate representations

```
s.label(coa.meau$co, plabels.optf
```

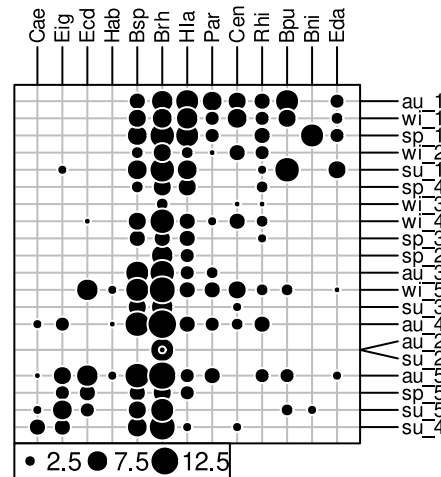
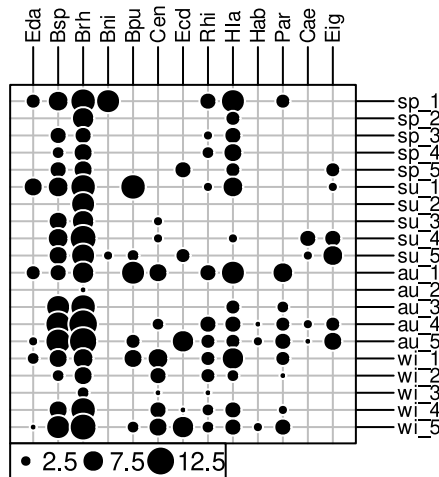


```
s.label(coa.meau$li, plabels.optf
```



Reordering of a table

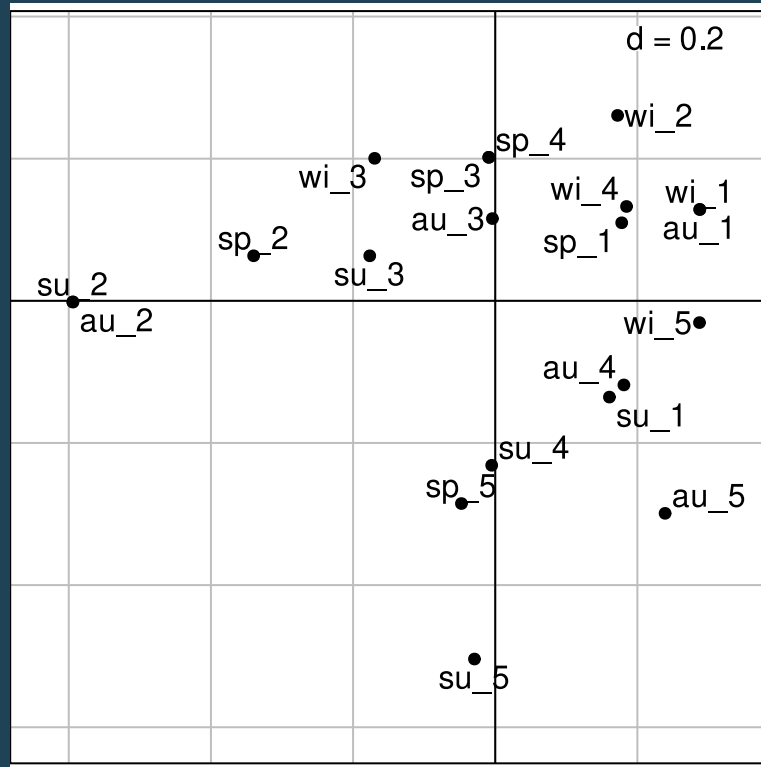
```
g1 <- table.value(meaudret$spe, ppoint.cex = 0.5, symbol = "circle",  
  plot = FALSE)  
g2 <- table.value(meaudret$spe, ppoint.cex = 0.5, symbol = "circle",  
  coordsx = rank(coa.meau$co[, 1]), coordsy = rank(coa.meau$li[,  
    1]), plot = FALSE)  
cbindADEg(g1, g2, plot = TRUE)
```



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

```
dJ <- dist.binary(meaudret$spe, method = 1) # Jaccard
pcoJ <- dudi.pco(dJ, scannf = FALSE)
s.label(pcoJ$li, plabels.optim = TRUE)
```



COA in practice

[Go to practical 4](#)