Training in ade4 in R - Module II: Advanced methods

Introduction

Stéphane Dray

2022-02-24

Material

The content of the course is available at

https://github.com/sdray/LausanneAdvanced/

In R, you can download the course by

```
usethis::use_course("sdray/LausanneAdvanced", destdir = "~/")
```

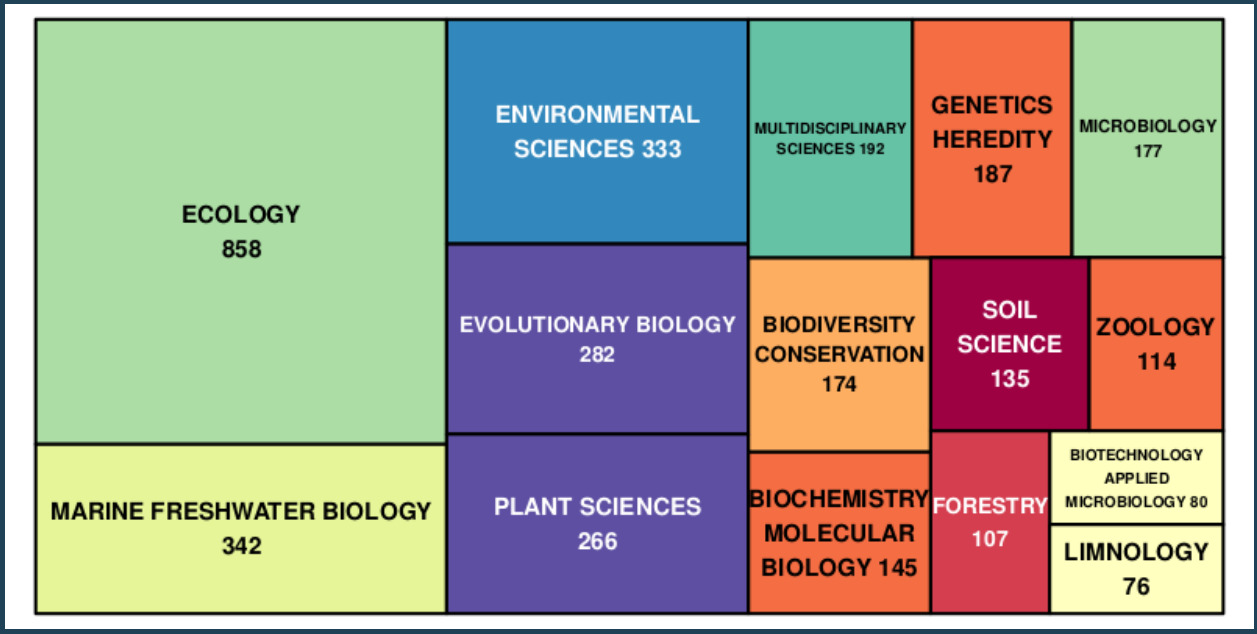
Online version at https://sdray.github.io/LausanneAdvanced

Required packages

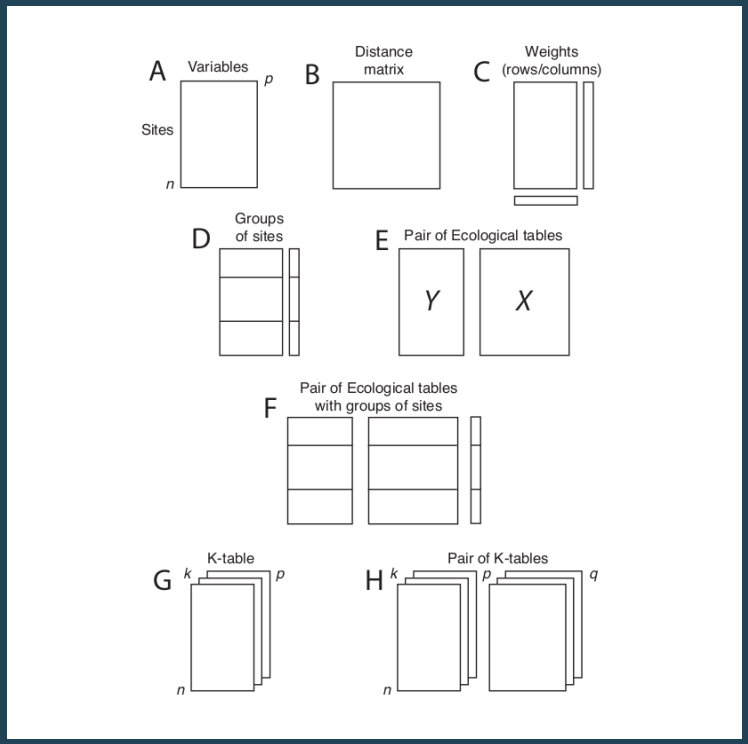
- `ade4` to run the analyses
- `adegraphics` to represent results
- `adespatial` and `spdep` for spatial analysis
- `rgl` to understand multivariate methods in interactive 3D

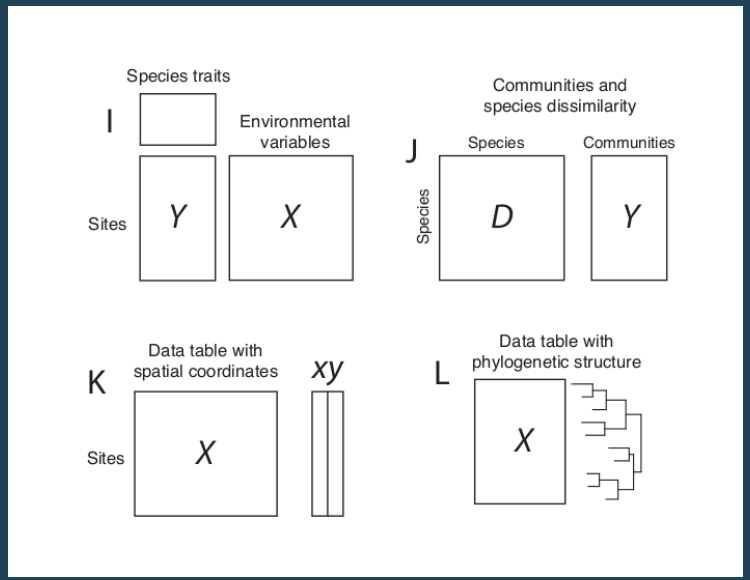
ade4

- R package since 2002
- Exploratory analysis of ecological data
  - Multivariate methods
  - Graphical functions

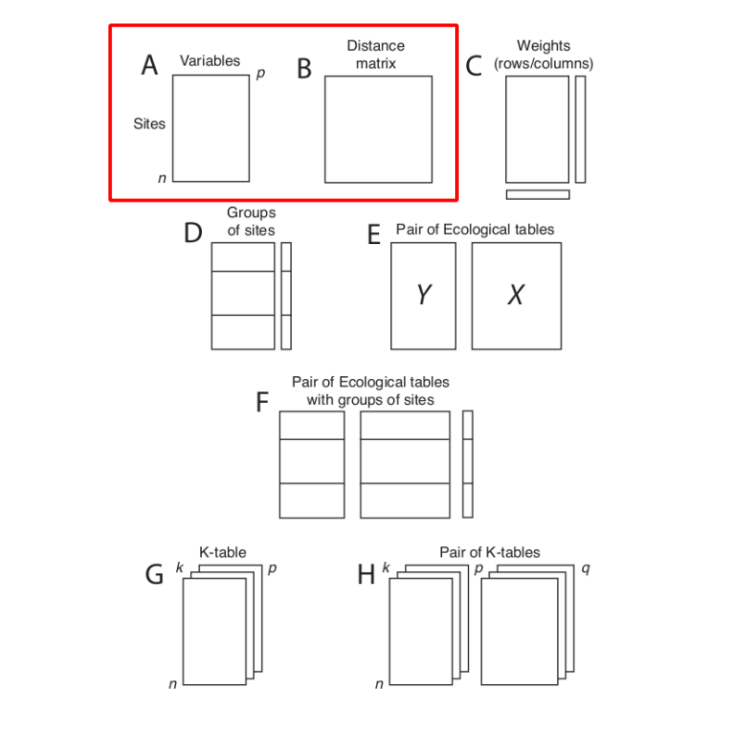


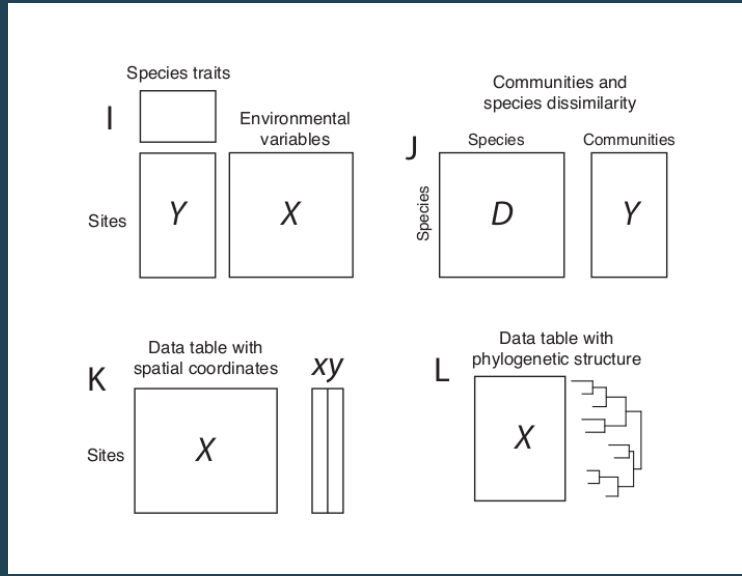
Data structure



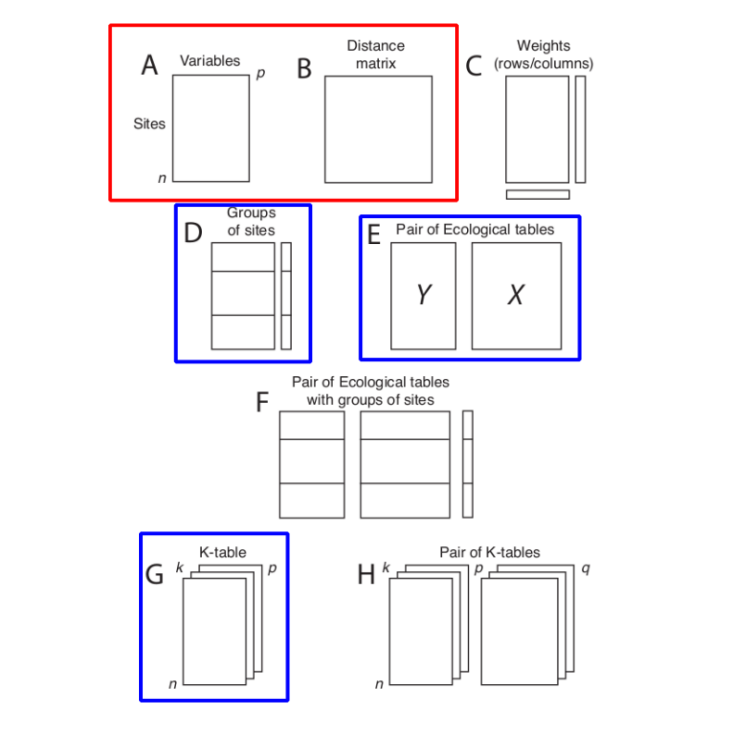


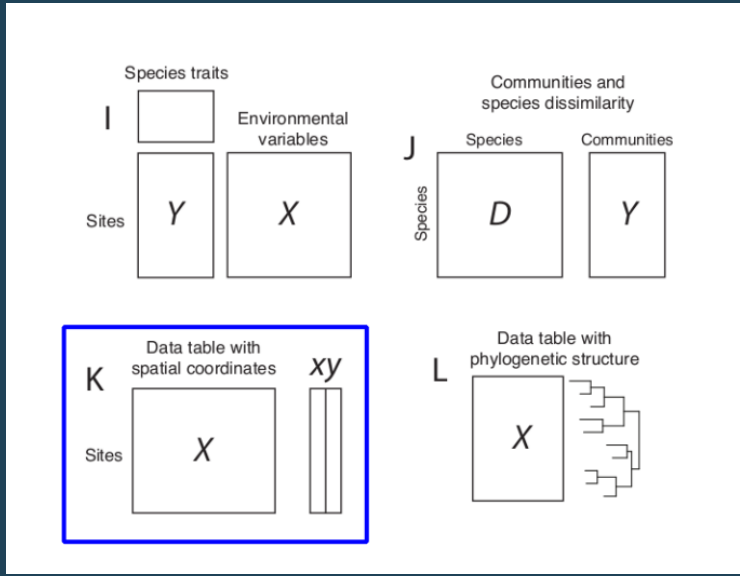
Data structure





Data structure





Module 1

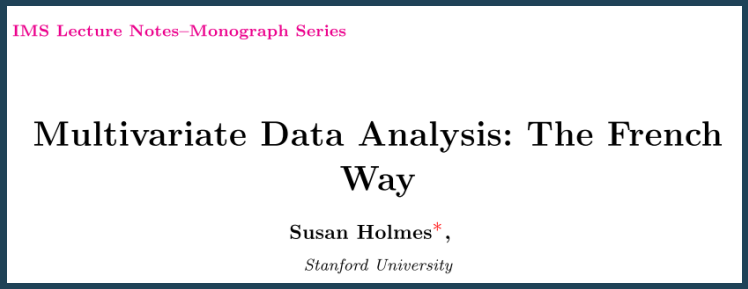
- Environmental variables
  - Quantitative variables → Principal Component Analysis (`dudi.pca`)
  - Categorical variables → Multiple Correspondence Analysis (`dudi.acm`)
  - Mix of both → Hill-Smith Analysis (`dudi.hillsmith`)
- Species table
  - Contingency table → Correspondence Analysis (`dudi.coa`)
  - Distance matrix → Principal Coordinates Analysis (`dudi.pco`)

Module 2: course outline

We will explore the geometric properties, outputs and interpretation of multivariate analysis focusing on one-table methods. Last afternoon for case studies.

- One table + one categorical variable
- Two tables
- K tables
- One table + spatial information

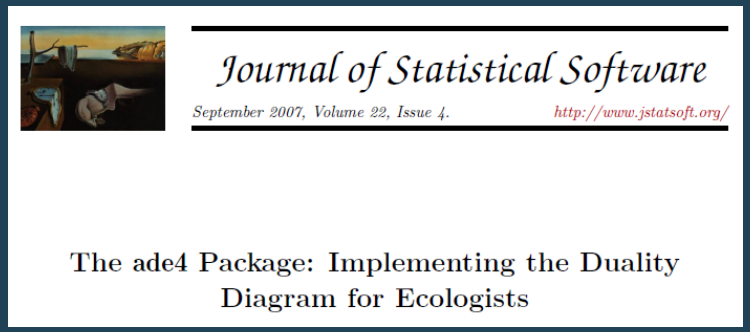
ade4: the French way


IMS Lecture Notes–Monograph Series

Multivariate Data Analysis: The French Way

Susan Holmes*,

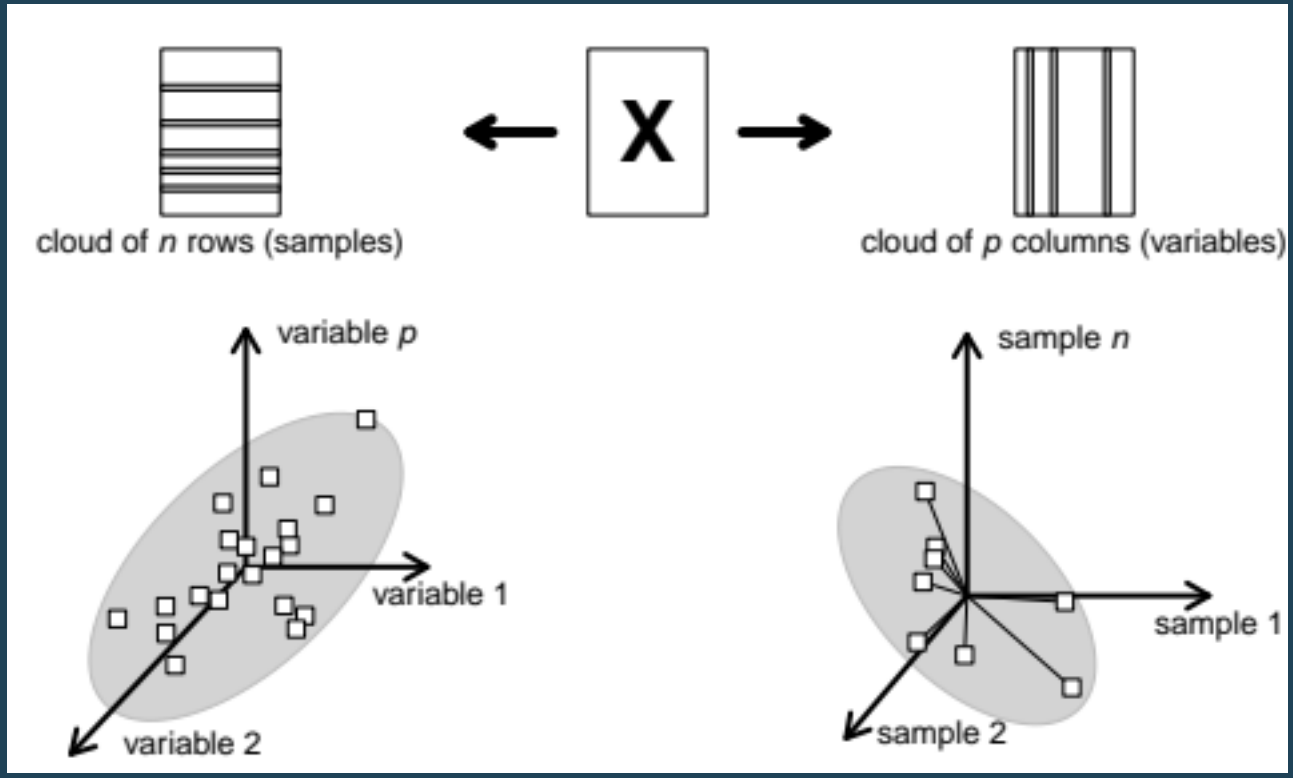
Stanford University



Implementation of functions in ade4 follows the duality diagram theory

More details are provided in the paper published in Journal of Statistical Software available here

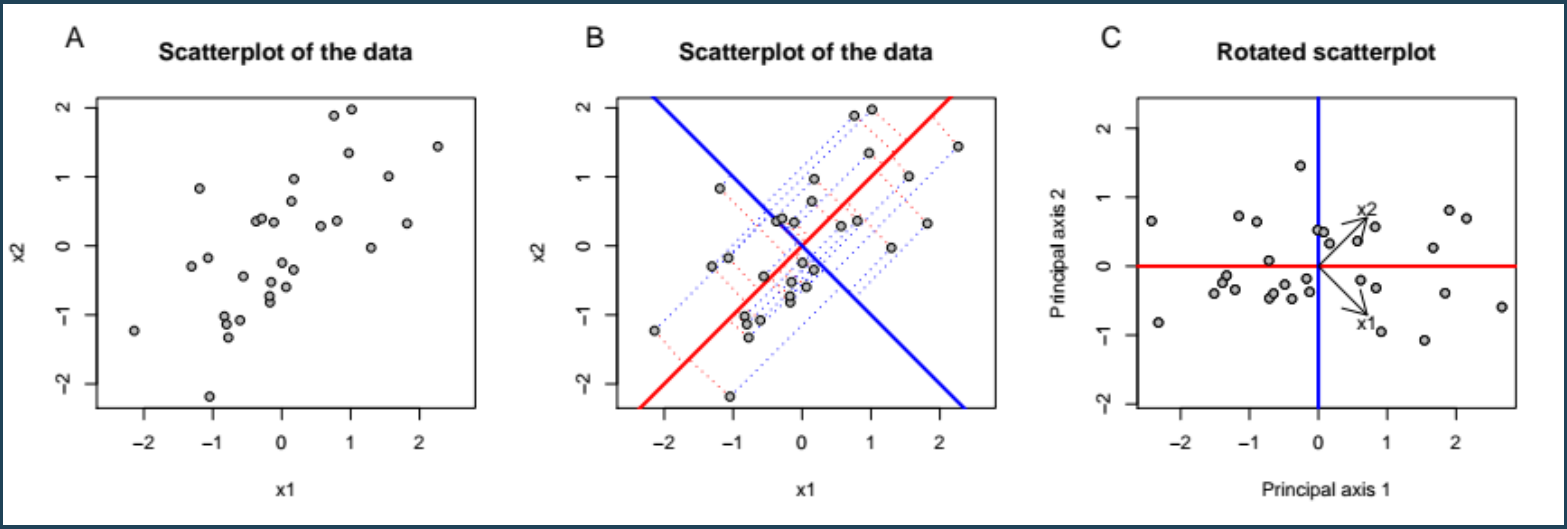
Two geometric views



what are the main similarities and differences between the individuals ?

what are the main relationships between the variables ?

Geometric view for individuals



- Multivariate methods only perform geometric operations (rotations) to obtain the best viewpoint on the data
- When many variables are considered, dimension reduction is also applied to simplify the interpretation

Statistical triplet

Multivariate methods aim to answer these two questions and seek for small dimension hyperspaces (few axes) where the representations of individuals and variables are as close as possible to the original ones.

To answer the two previous questions, we define

- $\mathbf{Q}$, a $p \times p$ positive symmetric matrix, used as an inner product in $\mathbb{R}^p$ and thus allows to measure distances between the n individuals
- $\mathbf{D}$, a $n \times n$ positive symmetric matrix, used as an inner product in $\mathbb{R}^n$ and thus allows to measure relationships between the p variables.

$$(\mathbf{X}, \mathbf{Q}, \mathbf{D})$$

Duality diagram theory

$$\mathbf{XQX}^\top\mathbf{DB} = \mathbf{B}\boldsymbol{\Lambda}$$

$$\mathbf{X}^\top\mathbf{DXQA} = \mathbf{A}\boldsymbol{\Lambda}$$

- $\mathbf{B}$ contains the principal components ($\mathbf{B}^\top\mathbf{DB} = \mathbf{I}_r$).
- $\mathbf{A}$ contains the principal axis ($\mathbf{A}^\top\mathbf{QA} = \mathbf{I}_r$).
- $\mathbf{L} = \mathbf{XQA}$ contains the row scores (projection of the rows of $\mathbf{X}$ onto the principal axes)
- $\mathbf{C} = \mathbf{X}^\top\mathbf{DB}$ contains the column scores (projection of the columns of $\mathbf{X}$ onto the principal components)

Maximization of:

$Q(\mathbf{a}) = \mathbf{a}^\top\mathbf{Q}^\top\mathbf{X}^\top\mathbf{DXQa} = \lambda$ and $S(\mathbf{b}) = \mathbf{b}^\top\mathbf{D}^\top\mathbf{XQX}^\top\mathbf{Db} = \lambda$

$$\langle\mathbf{XQa}|\mathbf{k}\rangle_{\mathbf{D}} = \langle\mathbf{X}^t\mathbf{Db}|\mathbf{a}\rangle_{\mathbf{Q}} = \sqrt{\lambda}$$

Inertia

- The total amount of information (variation) contained in the data is called the inertia

$$I_{(\mathbf{X},\mathbf{Q},\mathbf{D})} = Trace(\mathbf{X}^\top \mathbf{D}\mathbf{X}\mathbf{Q})$$

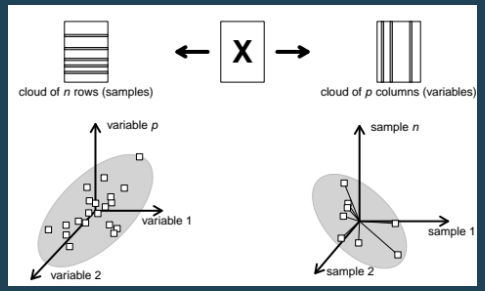
- Multivariate analysis aims to find new axes maximizing the projected inertia (i.e., the inertia of the projections).
- Total inertia is equal to the sum of eigenvalues
- In the case of PCA, total inertia is a sum of variances and an eigenvalue is equal to the variance of the projections on the associated axis

Implementation in ade4

Computations are performed by the function `as.dudi`. This functions takes 3 arguments defining the statistical triplet and returns an object of class `dudi` that contains:

ade4	theory	Definition
`tab`	$\mathbf{X}$	(transformed) data table
`cw`	$\mathbf{Q}$	inner product for rows
`lw`	$\mathbf{D}$	inner product for columns
---	---	---
`eig`	$\mathbf{\Lambda}$	eigenvalues
`l1`	$\mathbf{B}$	principal components
`c1`	$\mathbf{A}$	principal axes
`li`	$\mathbf{L}$	row scores
`co`	$\mathbf{C}$	column scores

From the theory



- The principal axes

$$\mathbf{X}^\top\mathbf{DXQA} = \mathbf{A\Lambda}$$

- The row scores

$$\mathbf{L} = \mathbf{XQA}$$

- Maximization of

$$Q(\mathbf{a}) = \mathbf{a}^\top\mathbf{Q}^\top\mathbf{X}^\top\mathbf{DXQa} = \lambda$$

$$Q(\mathbf{a}) = \|\mathbf{XQa}\|^2_{\mathbf{D}} = \lambda$$

- The principal components

$$\mathbf{XQX}^\top\mathbf{DK} = \mathbf{B\Lambda}$$

- The column scores

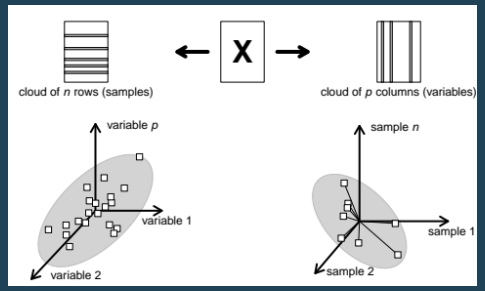
$$\mathbf{C} = \mathbf{X}^\top\mathbf{DB}$$

- Maximization of

$$S(\mathbf{b}) = \mathbf{b}^\top\mathbf{D}^\top\mathbf{XQX}^\top\mathbf{Db} = \lambda$$

$$S(\mathbf{b}) = \|\mathbf{X}^\top\mathbf{Db}\|^2_{\mathbf{Q}} = \lambda$$

To the practice in ade4



- The principal axes

  `$c1`

- The row scores

  `$li`

- Maximization of

  `$eig`

- The principal components

  `$l1`

- The column scores

  `$co`

- Maximization of

  `$eig`

Available methods

Different definitions of a statistical triplet correspond to different methods

Function name	Analysis name
dudi.pca	Principal component analysis
dudi.pco	Principal coordinate analysis
dudi.coa	Correspondence analysis
dudi.acm	Multiple correspondence analysis
dudi.dec	Decentered correspondence analysis
dudi.fca	Fuzzy correspondence analysis
dudi.fpca	Fuzzy PCA
dudi.mix	Mixed nalysis
dudi.hillsmith	Hill-Smith analysis
dudi.nsc	Non-symmetric correspondence analysis

Graphical functions

- Outputs of multivariate methods are usually provided as plots
- `ade4` contains several graphical functions
- they have been re-implemented in a much more flexible way in the package `adegraphics`

A comprehensive overview of the package is available in its vignette available online or in R by:

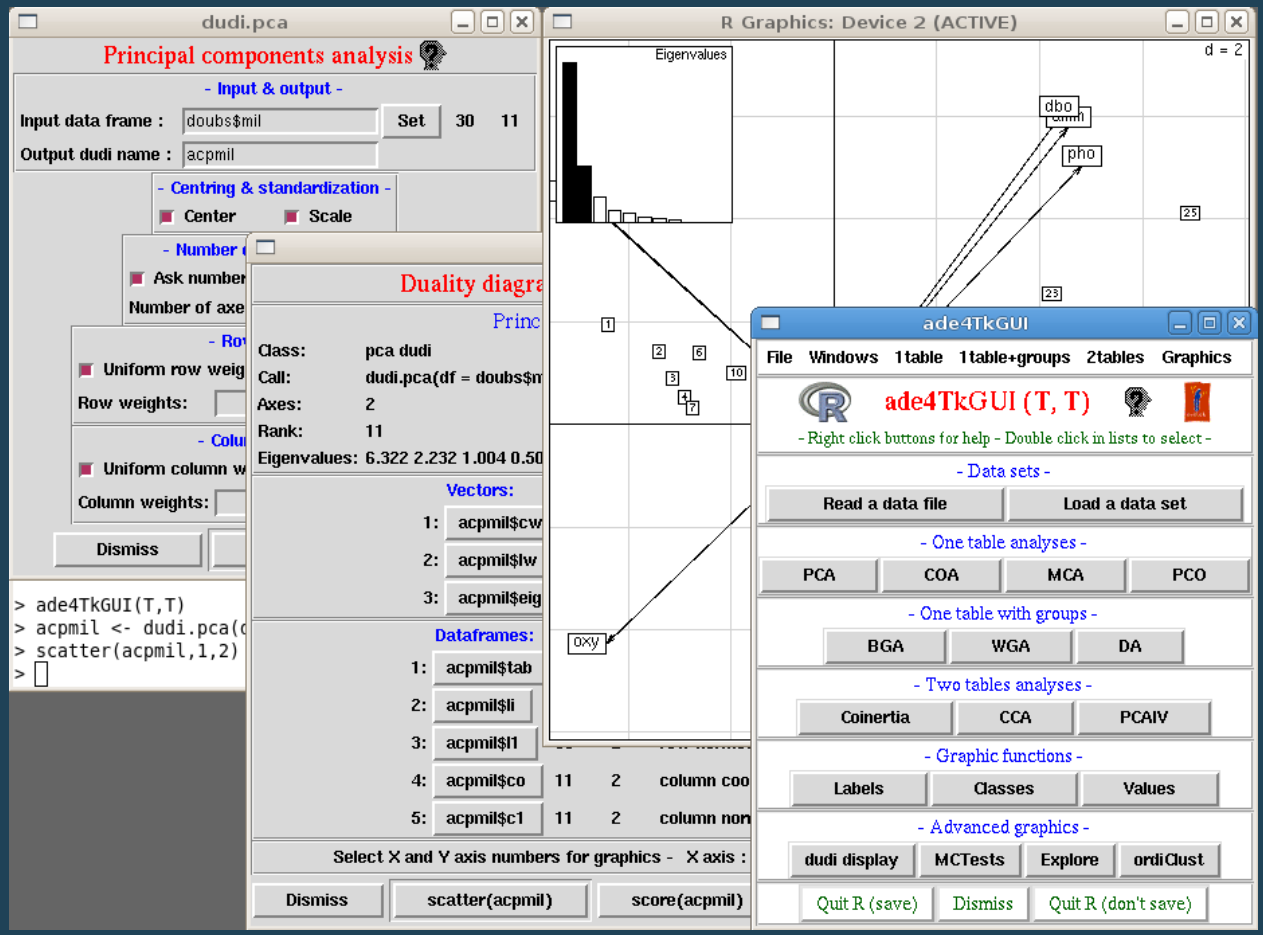
```
vignette("adegraphics")
```

See also the paper published in the R Journal here

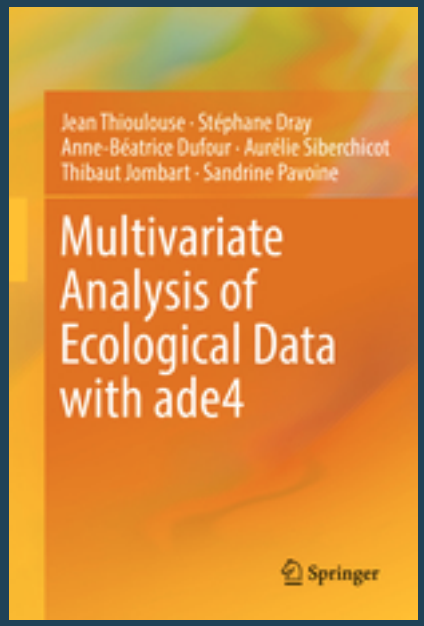
The ade packages

- `adegraphics`: S4-lattice based multivariate graphics
- `adespatial`: spatial multiscale multivariate analysis
- `adiv`: analysis of diversity
- `adehabitat`: analysis of habitat selection by animals
- `adegenet`: classes and methods for the multivariate analysis of genetic markers
- `adephylo`: exploratory analyses for the phylogenetic comparative method
- `ade4TkGUI`: graphical interface

ade4TkGUI


dudi.pca
Principal components analysis
- Input & output -
Input data frame :
doubs$mil
Set
30
11
Output dudi name :
acpmil
- Centring & standardization -
Center
Scale
Ask number
Number of axe
Uniform row weig
Row weights:
Uniform column w
Column weights:
Dismiss
> ade4TkGUI(T,T)
> acpmil <- dudi.pca(
> scatter(acpmil,1,2)
Duality diagra
Princ
Class: pca dudi
Call: dudi.pca(df = doubs$m
Axes: 2
Rank: 11
Eigenvalues: 6.322 2.232 1.004 0.50
Vectors:
1: acpmil$cw
2: acpmil$lw
3: acpmil$eig
Dataframes:
1: acpmil$tab
2: acpmil$li
3: acpmil$l1
4: acpmil$co
11
2
column coo
5: acpmil$c1
11
2
column nor
Select X and Y axis numbers for graphics - X axis :
Dismiss
scatter(acpmil)
score(acpmil)
R Graphics: Device 2 (ACTIVE)
Eigenvalues
d = 2
dbo
pho
oxy
ade4TkGUI
File Windows 1table 1table+groups 2tables Graphics
ade4TkGUI (T, T)
- Right click buttons for help - Double click in lists to select -
- Data sets -
Read a data file
Load a data set
- One table analyses -
PCA
COA
MCA
PCO
- One table with groups -
BGA
WGA
DA
- Two tables analyses -
Coinertia
CCA
PCAIV
- Graphic functions -
Labels
Classes
Values
- Advanced graphics -
dudi display
MCTests
Explore
ordiClust
Quit R (save)
Dismiss
Quit R (don't save)

Resources



https://www.springer.com/fr/book/9781493988488

- Mailing list:

  http://listes.univ-lyon1.fr/wws/info/adelist

- Development:

  https://github.com/sdray/ade4

- Courses (in French):

  http://pbil.univ-lyon1.fr/R/enseignement.html