# Multivariate Analysis

## Introduction & PCA

Stéphane Dray

2021-11-03

# Introduction

# Multivariate data

We consider cases where several variables are measured for a number of individuals


$n$ statistical units x $p$ variables

Examples in ecology

- individuals x traits
- species x traits
- sites x species
- sites x environment

```
library(ade4)
data(doubs)
names(doubs$env)
```

```
##  [1] "dfs" "alt" "slo" "flo" "pH"  "har" "pho" "nit" "amm" "oxy" "bdo"
```

# Univariate analysis

```
summary(doubs$env$slo)
```

```
##    Min. 1st Qu.  Median    Mean 3rd
##   1.099   1.831   2.565   2.758   3
```

```
mean(doubs$env$slo)
```

```
## [1] 2.757733
```

```
var(doubs$env$slo)
```

```
## [1] 1.16724
```

```
hist(doubs$env$slo)
```

# Bivariate analysis

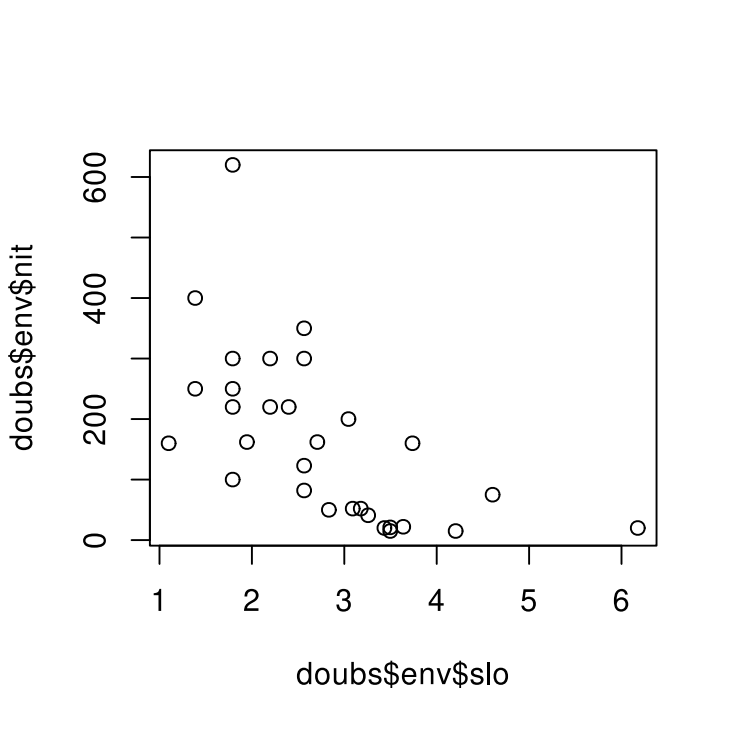
```
cov(doubs$env$slo, doubs$env$nit)
```

```
## [1] -93.28123
```

```
cor(doubs$env$slo, doubs$env$nit)
```

```
## [1] -0.6108798
```

```
plot(doubs$env$slo, doubs$env$nit
```



# Pairwise analysis

```
pairs(doubs$env)
```

# Pairwise analysis

```
round(cor(doubs$env), 2)
```

```
##       dfs   alt   slo   flo    pH   har   pho   nit   amm   oxy   bdo
## dfs  1.00 -0.94 -0.76  0.95  0.00  0.70  0.48  0.75  0.41 -0.51  0.40
## alt -0.94  1.00  0.76 -0.87 -0.04 -0.74 -0.44 -0.76 -0.38  0.36 -0.34
## slo -0.76  0.76  1.00 -0.72 -0.27 -0.65 -0.40 -0.61 -0.35  0.46 -0.32
## flo  0.95 -0.87 -0.72  1.00  0.02  0.70  0.39  0.61  0.29 -0.36  0.25
## pH   0.00 -0.04 -0.27  0.02  1.00  0.09 -0.08 -0.05 -0.12  0.18 -0.15
## har  0.70 -0.74 -0.65  0.70  0.09  1.00  0.36  0.51  0.29 -0.38  0.34
## pho  0.48 -0.44 -0.40  0.39 -0.08  0.36  1.00  0.80  0.97 -0.72  0.89
## nit  0.75 -0.76 -0.61  0.61 -0.05  0.51  0.80  1.00  0.80 -0.63  0.64
## amm  0.41 -0.38 -0.35  0.29 -0.12  0.29  0.97  0.80  1.00 -0.72  0.89
## oxy -0.51  0.36  0.46 -0.36  0.18 -0.38 -0.72 -0.63 -0.72  1.00 -0.84
## bdo  0.40 -0.34 -0.32  0.25 -0.15  0.34  0.89  0.64  0.89 -0.84  1.00
```

# Multivariate analysis

Avoid pairwise analysis to provide a global summary of the full data set. The objective is to answer simultaneously both questions:

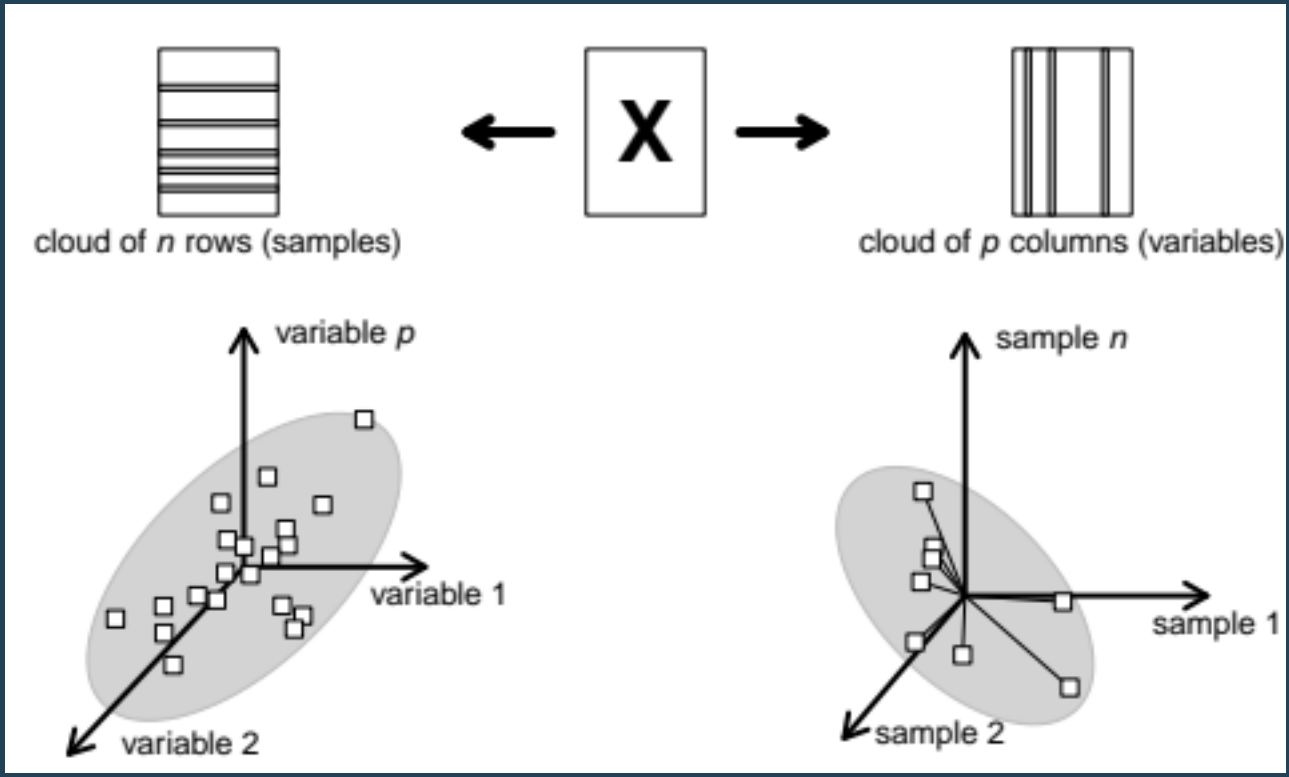
- what are the main similarities and differences between the individuals ?
- what are the main relationships between the variables ?

Multivariate analysis allows to:

- summarize linear relationships between variables
- identify structures among individuals
- reduce the number of variables before new analyses
- replace original variables by new synthetic ones

# Geometric perspective

# Two geometric views

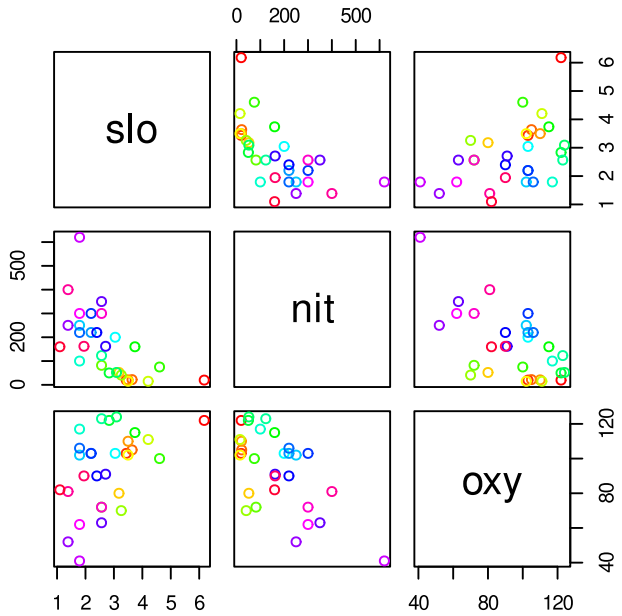



In $\mathbb{R}^p$, what are the main similarities and differences between the individuals ?

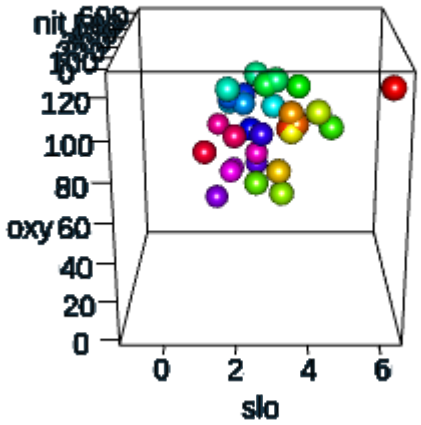
In $\mathbb{R}^n$, what are the main relationships between the variables ?

# A 3D example (space of individuals)

```
tab <- doubs$env[, c(3, 8, 10)]
color <- rainbow(30)
pairs(tab, col = color)
```

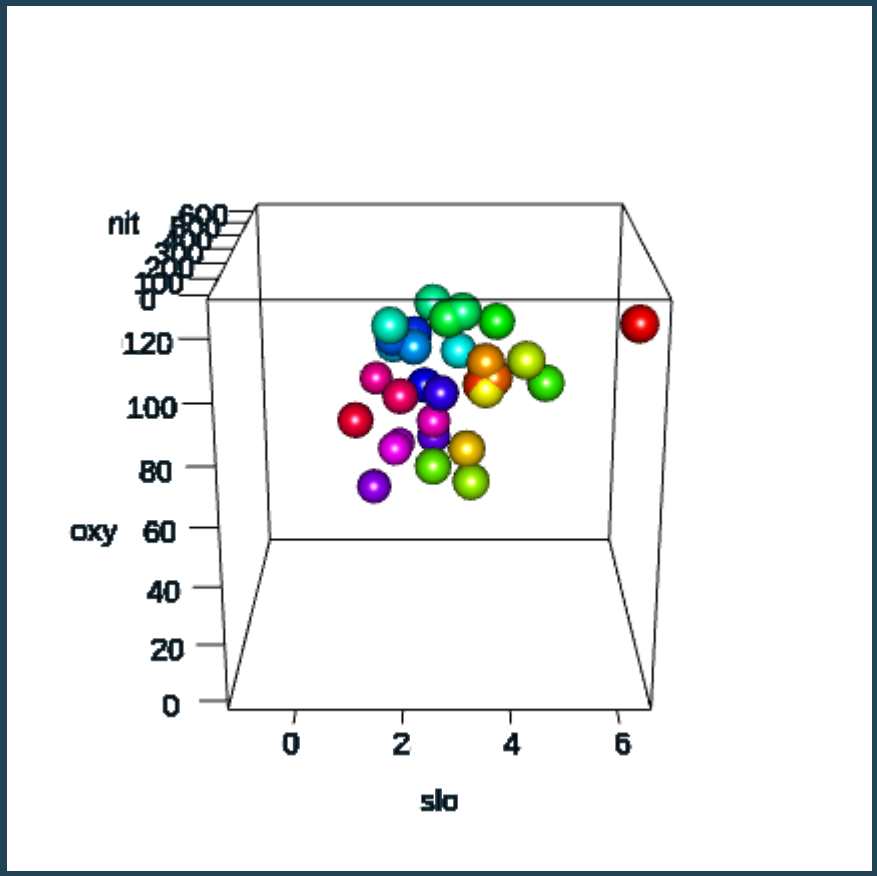



```
source("../../R/3D-utils.R")
plot3d(tab, type = "s", col = col
```



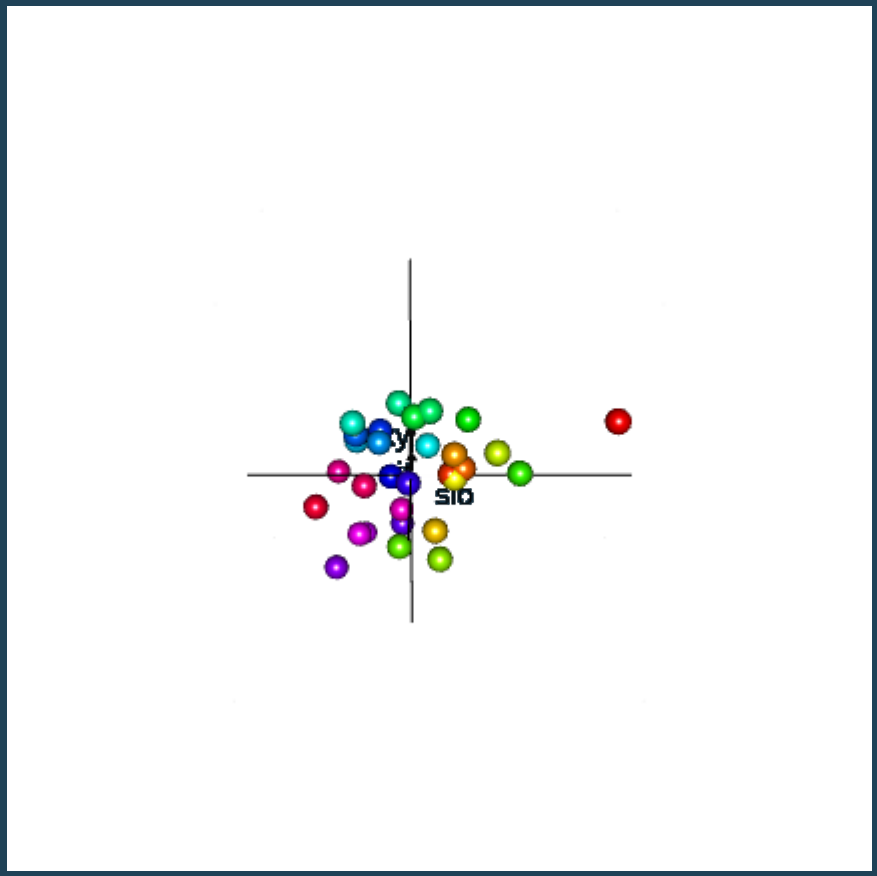
# Change the viewpoint

# Change the viewpoint



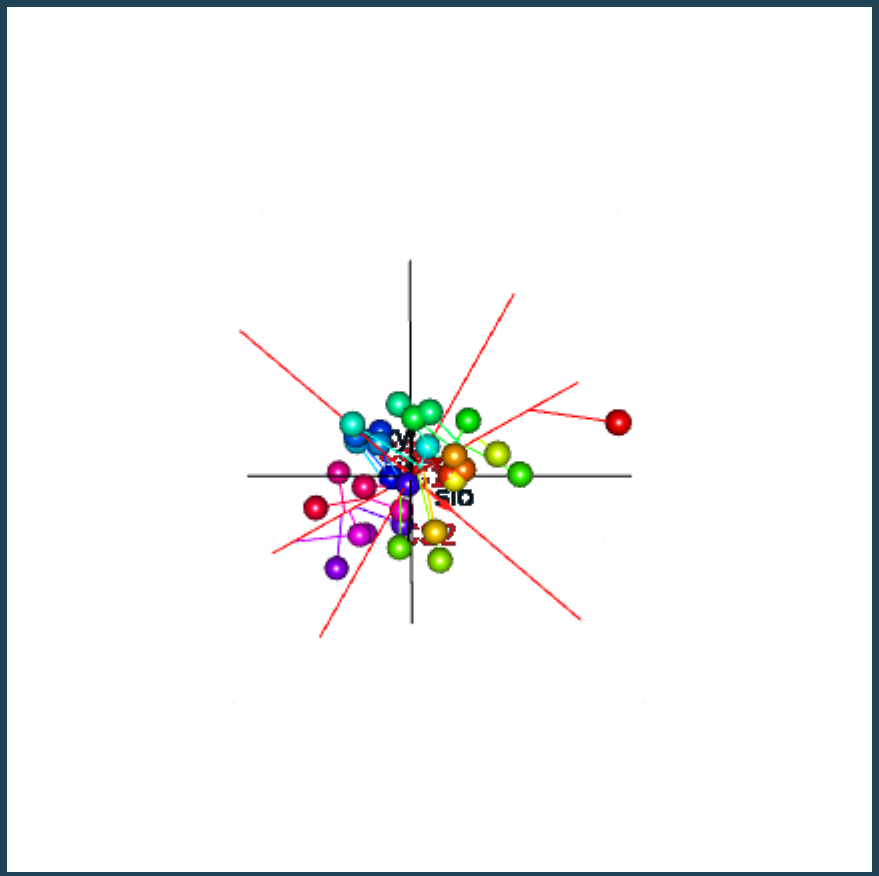

What is the best viewpoint?

# Scale the data



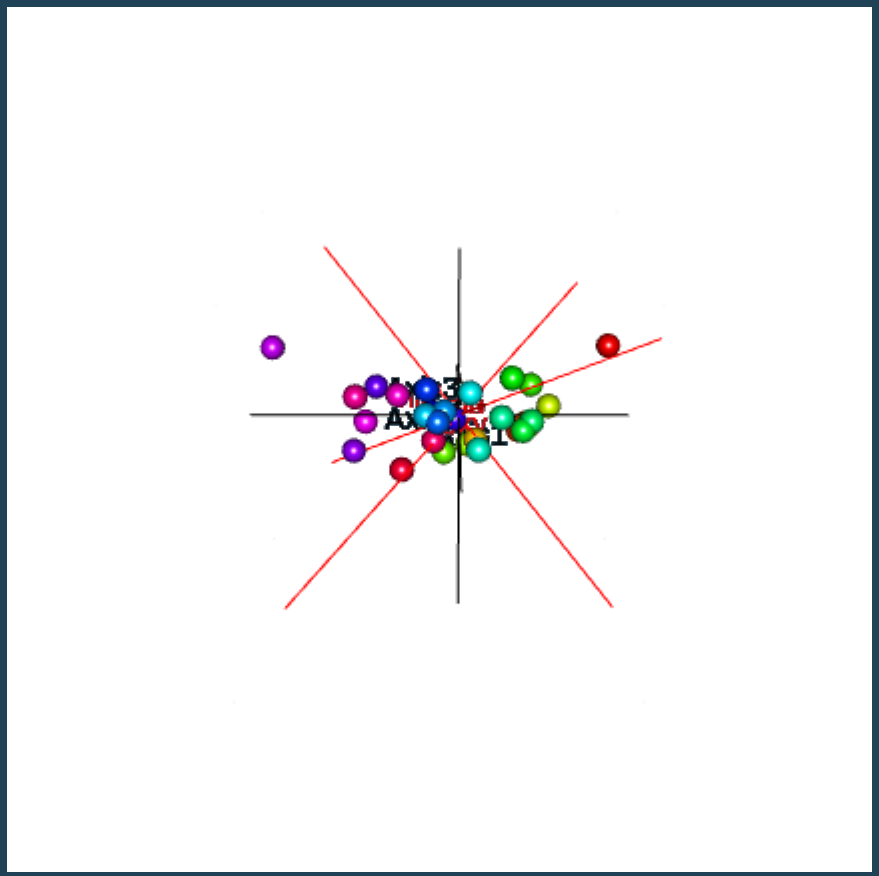
# Axes of main variation

Axes of main variations describe directions where the projections of individuals are the most scattered



# New system of axes

Lastly, we can use the principal axes as a new system of coordinates and represent the data in this new system



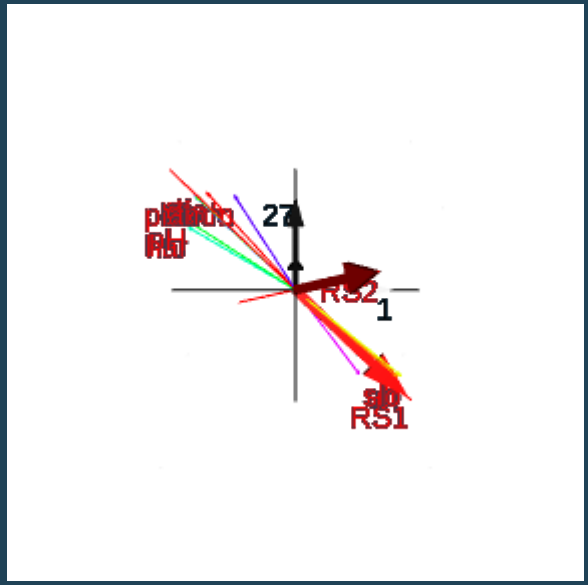
# Dimension reduction

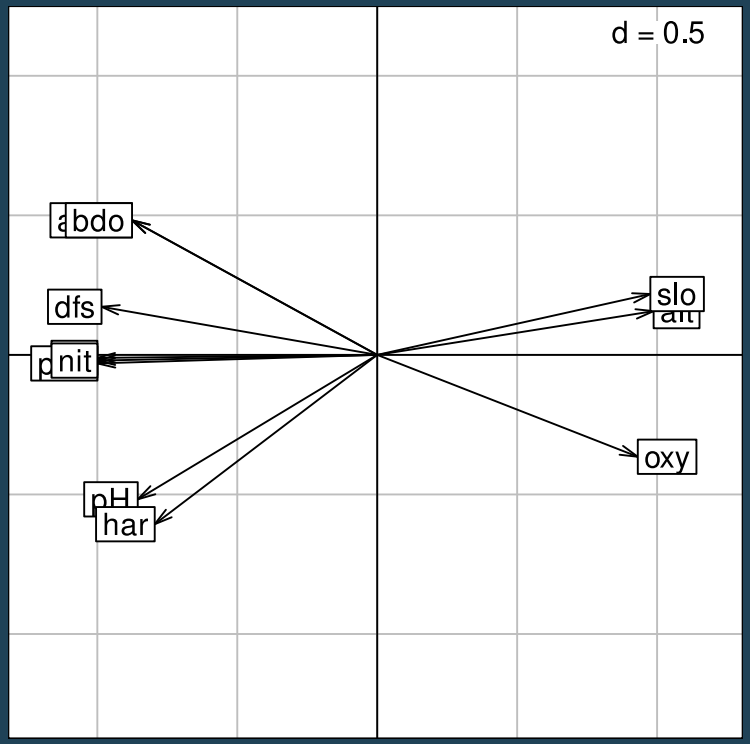
Only the first few axes (axes of main variation) can be used to describe the main structures of the data. Very useful when a big number of variables/individuals are considered.

```
s.label(pca$li)
```

# Space of variables

The same approach can be used to search for the best representation of variables.



# Multivariate analysis in short

- Data transformation (e.g., centering, scaling)
- Best viewpoint (rotation, projection)
- Summarize (dimension reduction)

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

To achieve these goals, we need to define some ways to compute distances:

- $\mathbf{Q}$, 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.

All methods consider these different steps but differ in the transformation of the data ( $\mathbf{X}$ ), metrics ( $\mathbf{Q}$ and $\mathbf{D}$ ) and thus on the mathematical criteria that are maximized

# Principal component analysis

# Data

- PCA can be applied when the data table contains only quantitative variables
- Data are usually centered and can be scaled
- Two metrics (scalar products) are defined to compute distances in the spaces of variables and individuals
  - $\mathbf{Q} = \mathbf{I}_p$ allows to measure distances between the $n$ individuals in $\mathbb{R}^p$.
  - $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$ allows to measure relationships between the $p$ variables in $\mathbb{R}^n$.

# From geometry to statistics

Several geometric operations translate into statistical concepts:

- the mean is computed by a scalar product and corresponds to a Euclidean projection

$$\mathrm{m}(\mathbf{x}) = \langle \mathbf{x} | \mathbf{1}_n \rangle_{\frac{1}{n}\mathbf{I}_n}$$

- the variance is equal to the squared norm of the centred vector $\mathbf{x}^*$

$$\mathrm{var}(\mathbf{x}) = \|\mathbf{x} - \mathrm{m}(\mathbf{x})\mathbf{1}_n\|^2_{\frac{1}{n}\mathbf{I}_n} = \|\mathbf{x}^*\|^2_{\frac{1}{n}\mathbf{I}_n}$$

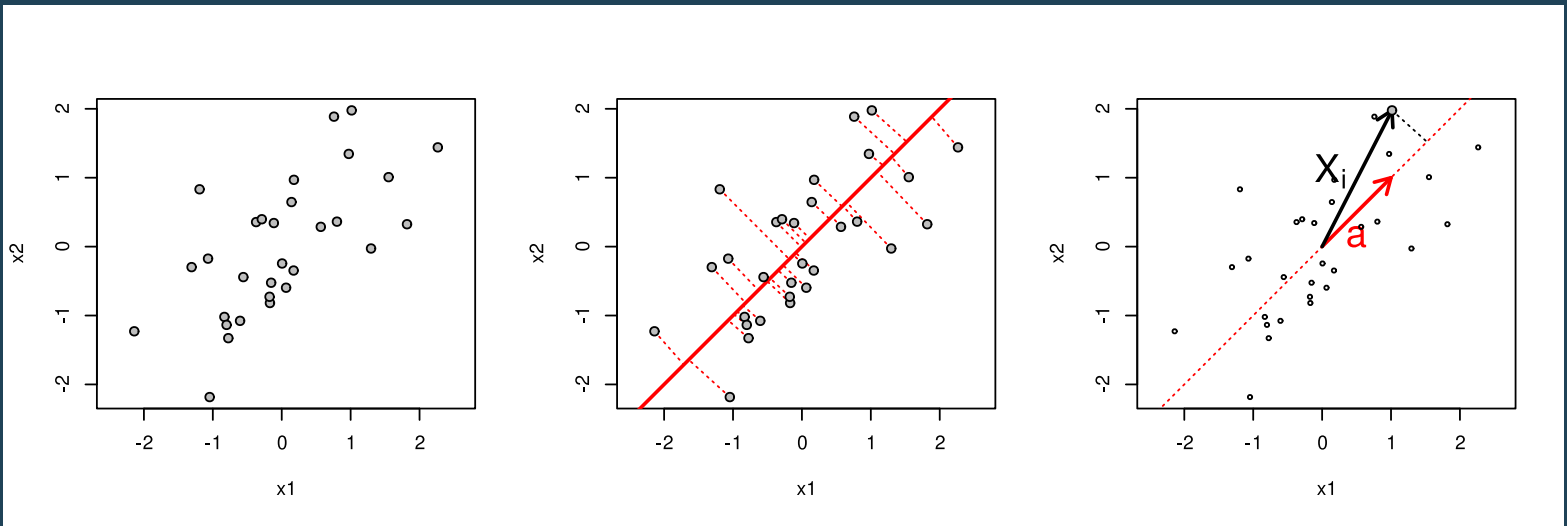
- the covariance is a scalar product between two centred vectors

$$\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}^* | \mathbf{y}^* \rangle_{\frac{1}{n}\mathbf{I}_n} = \|\mathbf{x}\|_{\frac{1}{n}\mathbf{I}_n} \|\mathbf{y}\|_{\frac{1}{n}\mathbf{I}_n} \cos(\theta_{\mathbf{xy}})$$

- the correlation is a cosine of the angle formed by the two vectors

$$\mathrm{cor}(\mathbf{x}, \mathbf{y}) = \cos(\theta_{\mathbf{xy}})$$

# Geometry for individuals




We aim to find a vector $\mathbf{a}$ of $\mathbb{R}^p$ maximizing the sum of the norms of the projections of the rows of $\mathbf{X}$:

$$\sum_{i=1}^{n} d_i \|\mathbf{P_a} \mathbf{X}_i\|_{\mathbf{Q}}^2$$

where $d_i$ is the $i$-th diagonal element of $\mathbf{D}$ and $\mathbf{P_a}$ is the projection operator onto the vector $\mathbf{a}$

# Projection on a (normed) vector

The projection step is obtained by

$$\mathbf{P}_{\mathbf{a}}\mathbf{X}_i = \frac{\langle \mathbf{a}|\mathbf{X}_i\rangle_{\mathbf{Q}}}{\langle \mathbf{a}|\mathbf{a}\rangle_{\mathbf{Q}}}\mathbf{a}$$

If $\mathbf{a}$ is $\mathbf{Q}$-normed ( $\mathbf{a}^\top\mathbf{Q}\mathbf{a} = 1$ ), it simplifies to

$$\mathbf{P}_{\mathbf{a}}\mathbf{X}_i = \langle \mathbf{a}|\mathbf{X}_i\rangle_{\mathbf{Q}}\mathbf{a}$$

So that the maximized quantity can be rewritten as:

$$\sum_{i=1}^{n} d_i \|\mathbf{P}_{\mathbf{a}}\mathbf{X}_i\|^2_{\mathbf{Q}} = \sum_{i=1}^{n} d_i \langle \mathbf{a}|\mathbf{X}_i\rangle^2_{\mathbf{Q}} = \|\mathbf{X}\mathbf{Q}\mathbf{a}\|^2_{\mathbf{D}} = \mathrm{var}(\mathbf{X}\mathbf{Q}\mathbf{a})$$

# Diagonalization

We denote $\lambda$ as the maximum possible value:

$$\lambda = \|\mathbf{XQa}\|_{\mathbf{D}}^2 = \mathbf{a}^\top \mathbf{Q}\mathbf{X}^\top \mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{a}$$

As $\mathbf{a}$ is $\mathbf{Q}$-normed, we can write:

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

and it follows that:

$$\lambda \mathbf{a} = \mathbf{X}^\top \mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{a}$$

This corresponds to matrix diagonalization.

Hence, the best axis can be identified as the eigenvector ( $\mathbf{a}$ ) associated to the highest eigenvalue ( $\lambda$ ) of $\mathbf{X}^\top \mathbf{D}\mathbf{X}\mathbf{Q}$

Other axes can be obtained and correspond to the next eigenvectors/eigenvalues. They maximize the same quantity but should be orthogonal to the previous ones.

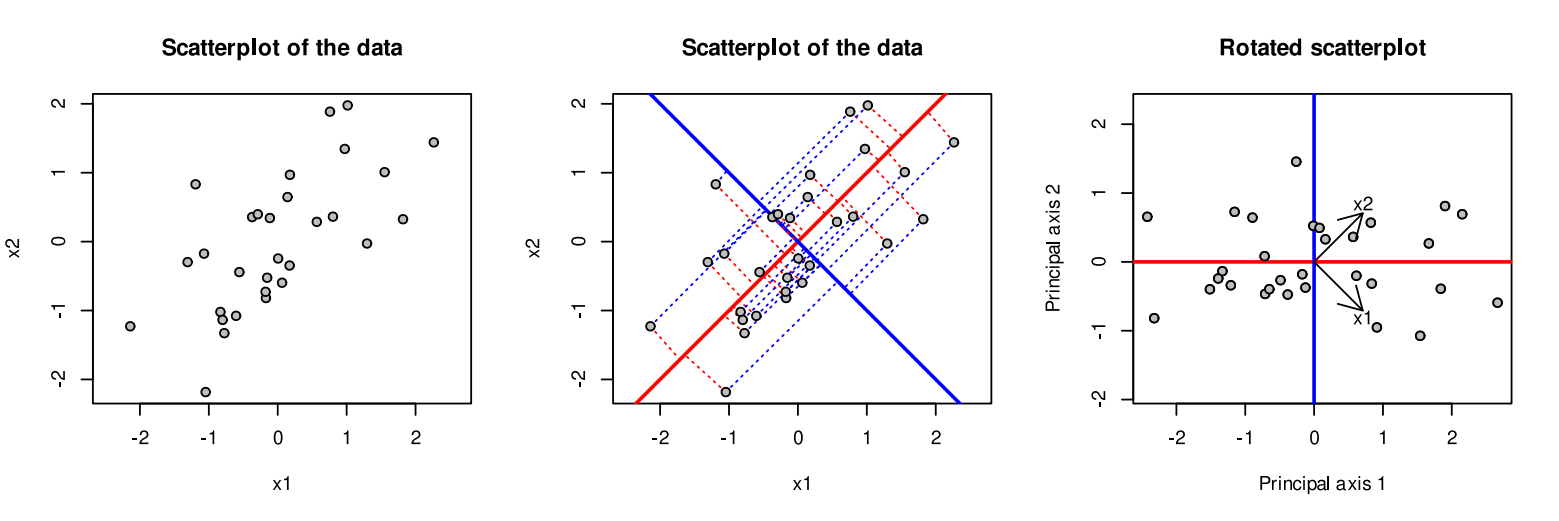
# PCA in the space of individuals

As $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$ and $\mathbf{Q} = \mathbf{I}_p$, we have:

$$\lambda \mathbf{a} = \frac{1}{n}\mathbf{X}^\top \mathbf{X}\mathbf{a}$$

- If data have been centered, this corresponds to the diagonalization of the covariance matrix.
- If data have been centered and scaled, this corresponds to the diagonalization of the correlation matrix.
- In both cases, we have the following interpretations:
  - geometric: PCA seeks for an axis ( $\mathbf{a}$ ), called the **principal axis**, on which individuals are projected ( $\mathbf{XQa}$ ) so that the points are the most scattered ( $\|\mathbf{XQa}\|^2_{\mathbf{D}}$ ).
  - statistical: PCA seeks for coefficients for variables ( $\mathbf{a}$ ) to compute a score for individuals ( $\mathbf{XQa}$ ) with maximal variance ( $\mathrm{var}(\mathbf{XQa})$ ).

# Summary

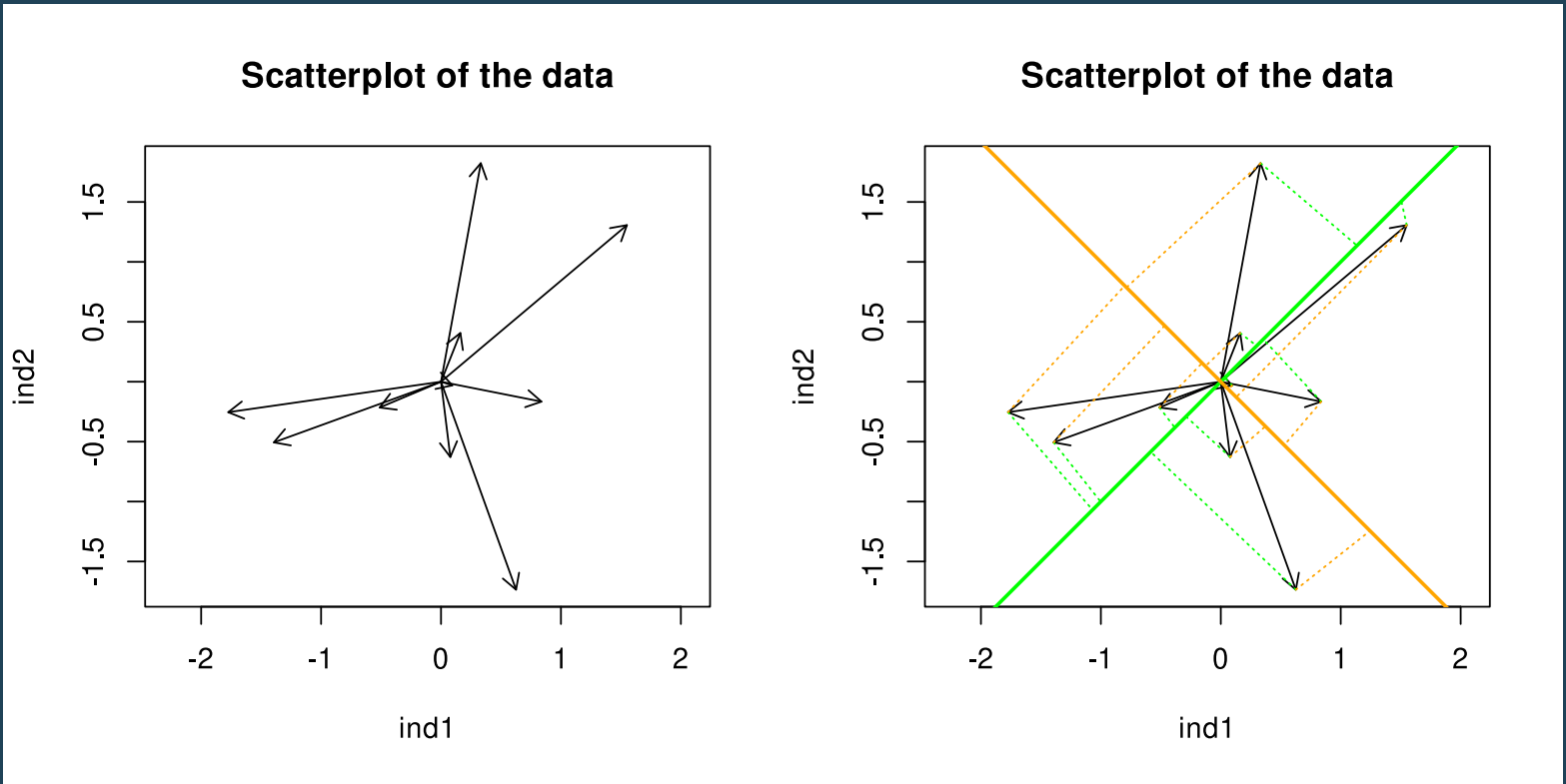



The last plot corresponds to the PCA factorial map and is usually called a **distance biplot** as it best preserves the distances between individuals.

- individuals are represented by $\mathbf{XQa}$
- variables are represented by $\mathbf{a}$
- if more variables (here 2) are considered, only the first few axes can be kept so that dimension reduction is performed. The first axes preserves the most important part of information (maximized projected inertia/variance).

# Space of variables

In $\mathbb{R}^n$, we can follow the same rationale for variables than for individuals in $\mathbb{R}^p$.



# PCA in the space of individuals

$$\lambda \mathbf{b} = \frac{1}{n} \mathbf{X} \mathbf{X}^\top \mathbf{b}$$

- geometric interpretation: PCA seeks for an axis ( $\mathbf{b}$ ), called the **principal component** on which variables are projected ( $\mathbf{X}^\top \mathbf{D} \mathbf{b}$ ) so that they are the most scattered ( $\|\mathbf{X}^\top \mathbf{D} \mathbf{b}\|_{\mathbf{Q}}^2$ ) or collinear.

- statistical interpretation:

  - If data have been centered, PCA seeks for a principal component ( $\mathbf{b}$ ). Variables are represented by their covariances with the component ( $\mathrm{cov}(\mathbf{x}_j, \mathbf{b})$ ) whose sum of squares is maximized ( $\sum_{j=1}^{p} \mathrm{cov}^2(\mathbf{x}_j, \mathbf{b})$ ).
  - If data have been scaled, PCA seeks for a principal component ( $\mathbf{b}$ ). Variables are represented by their correlations with the component ( $\mathrm{cor}(\mathbf{x}_j, \mathbf{b})$ ) whose sum of squares is maximized ( $\sum_{j=1}^{p} \mathrm{cor}^2(\mathbf{x}_j, \mathbf{b})$ ).

# PCA with ade4

```
args(dudi.pca)
```

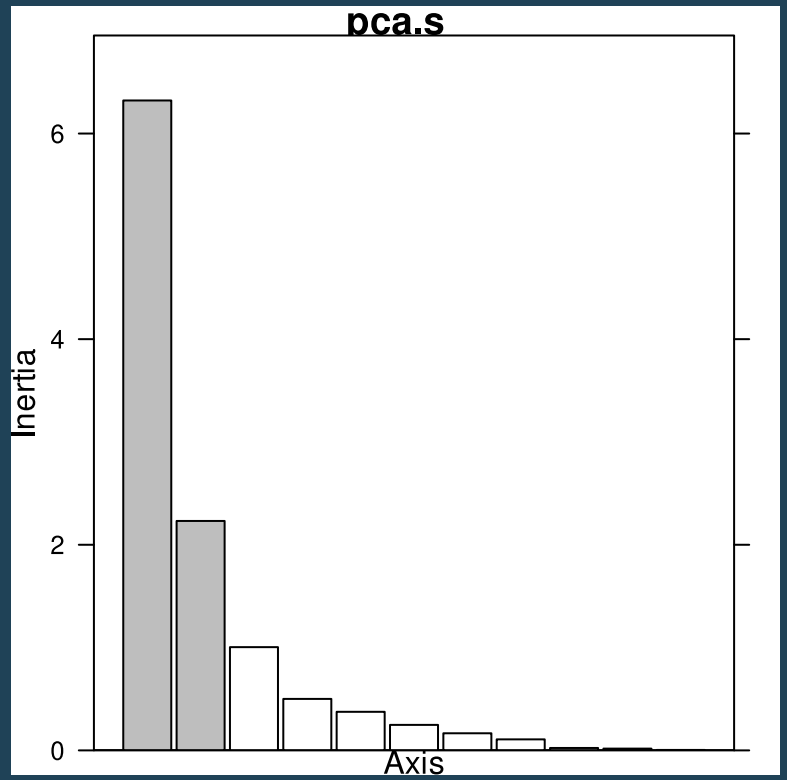
```
## function (df, row.w = rep(1, nrow(df))/nrow(df), col.w = rep(1,
##     ncol(df)), center = TRUE, scale = TRUE, scannf = TRUE, nf = 2)
## NULL
```

- `df` is a `data.frame` with the data
- `row.w` and `col.w` are optional vectors of weights
- `center` and `scale` define the standardization of the data
- `scannf` and `nf` allow to set the number of dimensions to interpret

```
pca.s <- dudi.pca(doubs$env, scannf = FALSE)
```

# Eigenvalues

```
library(adegraphics)
screeplot(pca.s)
```



# Inertia statistics

```
summary(pca.s)
```

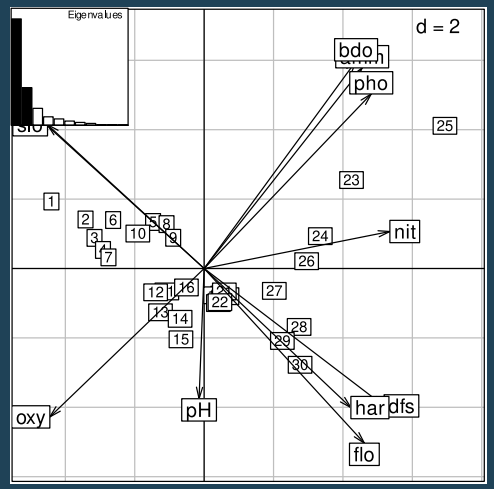
```
## Class: pca dudi
## Call: dudi.pca(df = doubs$env, scannf = FALSE)
##
## Total inertia: 11
##
## Eigenvalues:
##     Ax1     Ax2     Ax3     Ax4     Ax5
##  6.3216  2.2316  1.0042  0.5007  0.3752
##
## Projected inertia (%):
##     Ax1     Ax2     Ax3     Ax4     Ax5
##  57.469  20.287   9.129   4.552   3.411
##
## Cumulative projected inertia (%):
##     Ax1   Ax1:2   Ax1:3   Ax1:4   Ax1:5
##   57.47   77.76   86.89   91.44   94.85
##
## (Only 5 dimensions (out of 11) are shown)
```

# Graphical representations

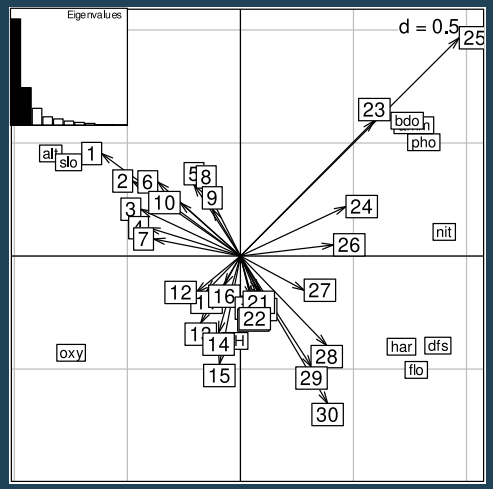
As we have *two* analyses (individuals and variables spaces), two representations can be defined:

- **distance biplot** where $\mathbf{A}$ and $\mathbf{XQA}$ are superimposed.
- **correlation biplot** where $\mathbf{B}$ and $\mathbf{X}^\top\mathbf{DB}$ are superimposed.

```
biplot(pca.s)
```



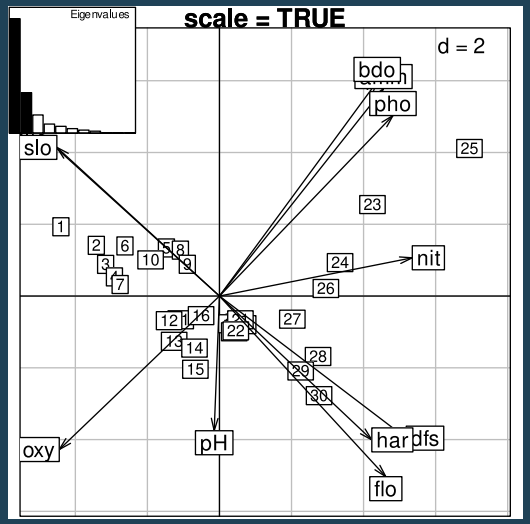

```
biplot(pca.s, permute = TRUE)
```

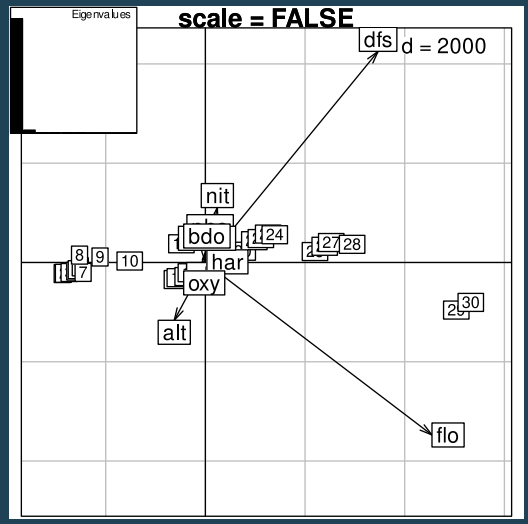


# To scale or not to scale

Scaling should be performed when we do not want that differences in variances affect the results

```
pca.c <- dudi.pca(doubs$env, scannf = FALSE, scale = FALSE)
pca.s <- dudi.pca(doubs$env, scannf = FALSE, scale = TRUE)
```




In this case, we must scale the data as differences in variances are mainly due to differences in units

# Conclusions

# PCA as a particular case in the 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 $S(\mathbf{b}) = \mathbf{b}^\top\mathbf{D}^\top\mathbf{XQX}^\top\mathbf{Db} = \lambda$

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

# Available methods

Different definitions of a statistical triplet correspond to different methods. Several are available in ade4

| 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 |