

# Training in ade4 in R - Module I: Basic methods

## Principal component analysis

Stéphane Dray

2023-12-03



# Data structure



- One table with  $p$  variables measured on  $n$  individuals
- All variables are **quantitative**
- For instance
  - sites  $\times$  environmental variables
  - species  $\times$  traits
  - individuals  $\times$  alleles
  - populations  $\times$  alleles

# Objectives

- Identify what is the main information contained in the table
  - Identify which variables are the most linked
  - Identify the principal differences/similarities between individuals

# Data

We consider the **meaudret** data set

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

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

```
dim(meaudret$env)
```

```
## [1] 20  9
```

```
names(meaudret$env)
```

```
## [1] "Temp" "Flow" "pH"    "Cond" "Bdo5" "Oxyd" "Ammo" "Nitr" "Phos"
```

The data set contains an environmental table with 20 measurements of 9 environmental variables. The measurements have been made in 6 sites at each season along a small French stream (see [?meaudret](#))

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

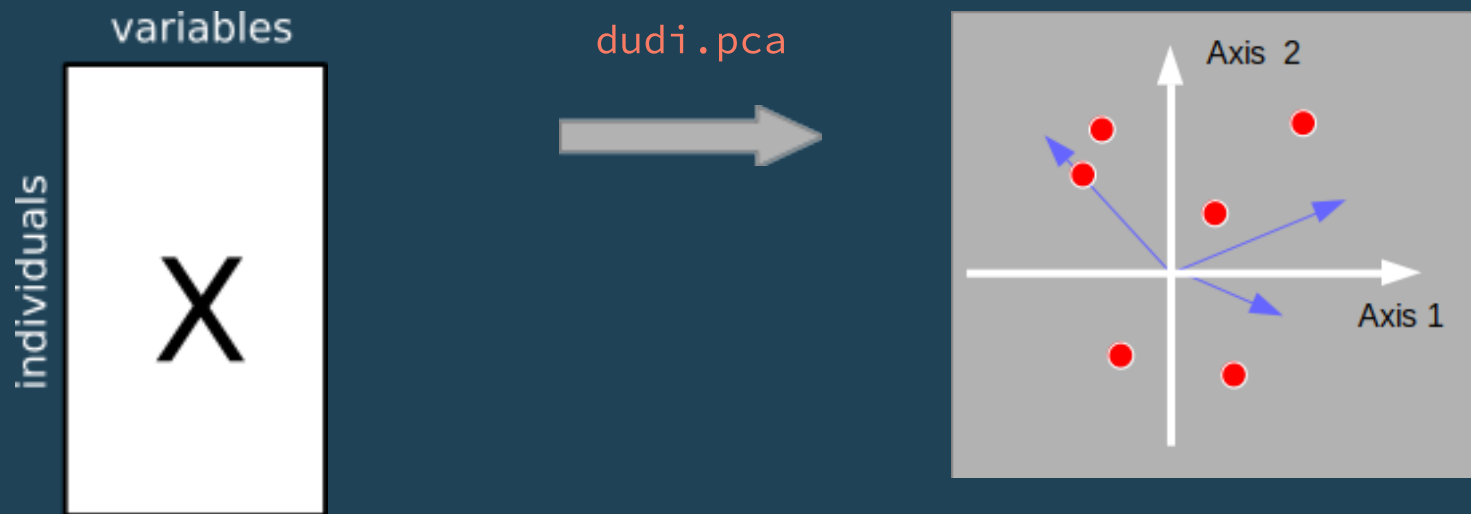
```
##      season site
## sp_1 spring  S1
## sp_2 spring  S2
## sp_3 spring  S3
## sp_4 spring  S4
## sp_5 spring  S5
## su_1 summer  S1
```

We want to know

- what are the main environmental gradients, i.e., which variables co-vary (if any)
- which samples have similar/different environmental conditions

# Principal component analysis

- $\mathbf{X}$  contains centred or scaled variables
- $\mathbf{Q} = \mathbf{I}_p$  is the identity matrix (diagonal matrix with 1s)
- $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$  is the diagonal matrix with  $\frac{1}{n}$



# Maximized criteria

- For individuals

$$Q(\mathbf{a}) = \|\mathbf{XQa}\|_{\mathbf{D}}^2 = \|\mathbf{Xa}\|_{\frac{1}{n}\mathbf{I}_n}^2 = \text{var}(\mathbf{Xa}) = \lambda$$

- For variables

- Centred data (  $x_{ij} - \bar{x}_j$  )

$$S(\mathbf{b}) = \|\mathbf{X}^\top \mathbf{D} \mathbf{b}\|_{\mathbf{Q}}^2 = \left\| \frac{1}{n} \mathbf{X}^\top \mathbf{b} \right\|_{\mathbf{I}_p}^2 = \sum_{j=1}^p \text{cov}^2(\mathbf{x}_j, \mathbf{b}) = \lambda$$

- Scaled data (  $(x_{ij} - \bar{x}_j)/s_j$  )

$$S(\mathbf{b}) = \|\mathbf{X}^\top \mathbf{D} \mathbf{b}\|_{\mathbf{Q}}^2 = \left\| \frac{1}{n} \mathbf{X}^\top \mathbf{b} \right\|_{\mathbf{I}_p}^2 = \sum_{j=1}^p \text{cor}^2(\mathbf{x}_j, \mathbf{b}) = \lambda$$



# The `dudi.pca` function

## Arguments

```
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.meau <- dudi.pca(meaudret$env, scannf = FALSE)
```

# Returned values

```
names(pca.meau)
```

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

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

- **\$eig**: eigenvalues (  $\mathbf{\Lambda}$  )
- **\$cw**: column weights (  $\mathbf{Q} = \mathbf{I}_p$  )
- **\$lw**: row weights (  $\mathbf{D} = \frac{1}{n} \mathbf{I}_n$  )
- **\$tab**: transformed data table (  $\mathbf{X}$  )
- **\$c1**: principal axes or variable loadings (  $\mathbf{A}$  )
- **\$li**: row scores (  $\mathbf{L} = \mathbf{X}\mathbf{A}$  )
- **\$l1**: principal components (  $\mathbf{B}$  )
- **\$co**: column scores (  $\mathbf{C} = \frac{1}{n} \mathbf{X}^\top \mathbf{B}$  )

# Graphical representation and interpretation

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

- **distance biplot** where  $\mathbf{A}$  and  $\mathbf{L} = \mathbf{X}\mathbf{A}$  ( $\$c1$ ,  $\$li$ ) are superimposed.
- **correlation biplot** where  $\mathbf{B}$  and  $\mathbf{C} = \frac{1}{n}\mathbf{X}^\top\mathbf{B}$  ( $\$l1$ ,  $\$co$ ) are superimposed.

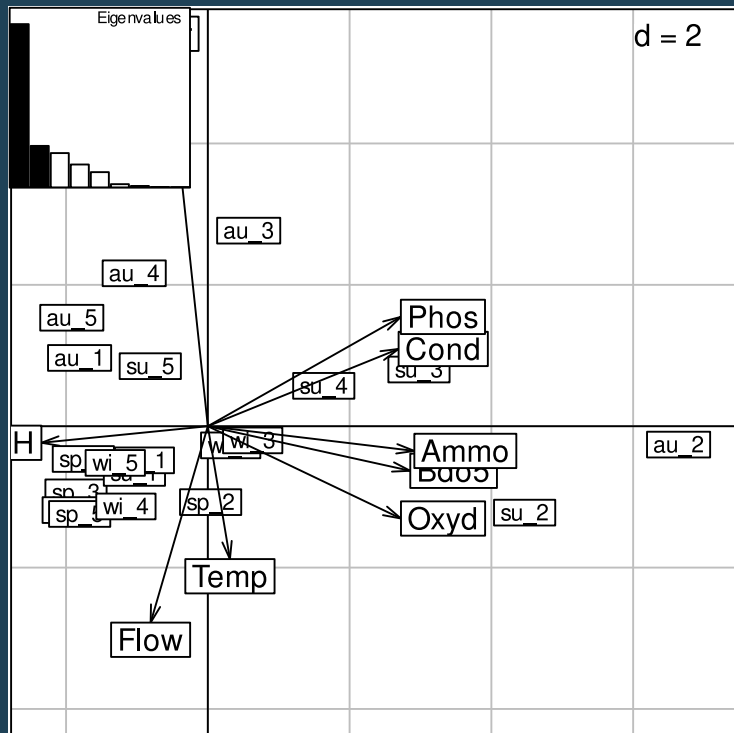
In the first interpretation, PCA finds coefficients for variables ( $\$c1$ ) to compute a linear combination ( $\$li$ ) that provides an ordination of individuals with the greatest dispersion (maximum variance).

In the second interpretation, PCA provides a linear combination ( $\$l1$ ) that maximise the correlations ( $\$co$ ) with all variables (or covariances for centred PCA). Hence, it is the best summary of the variables.

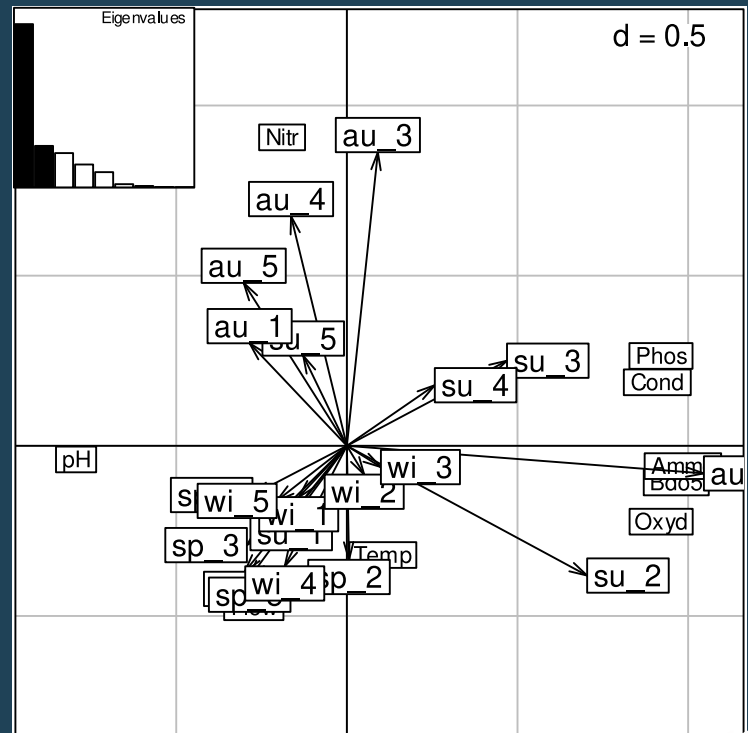
# The biplot function

```
library(adegraphics)
```

```
biplot(pca.meau)
```

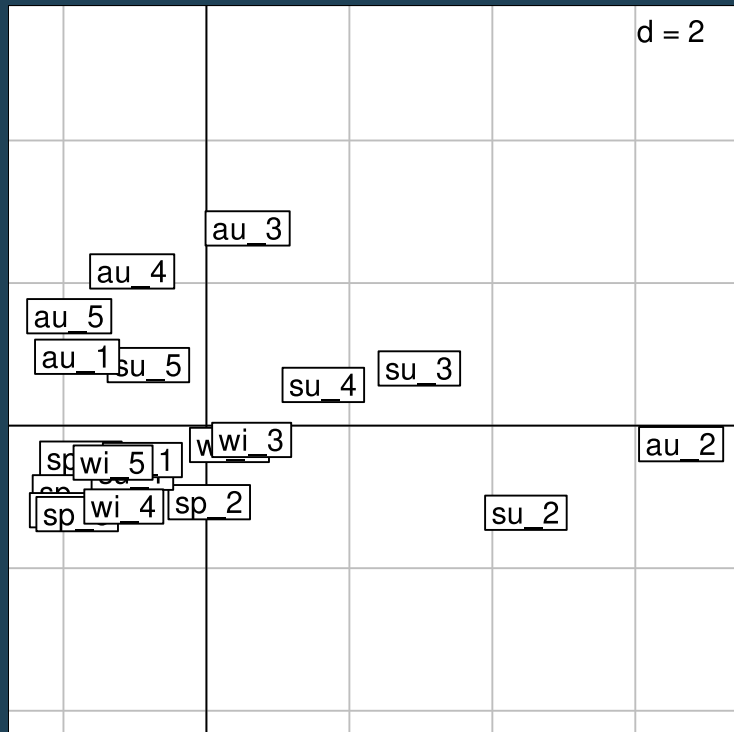


```
biplot(pca.meau, permute = TRUE)
```

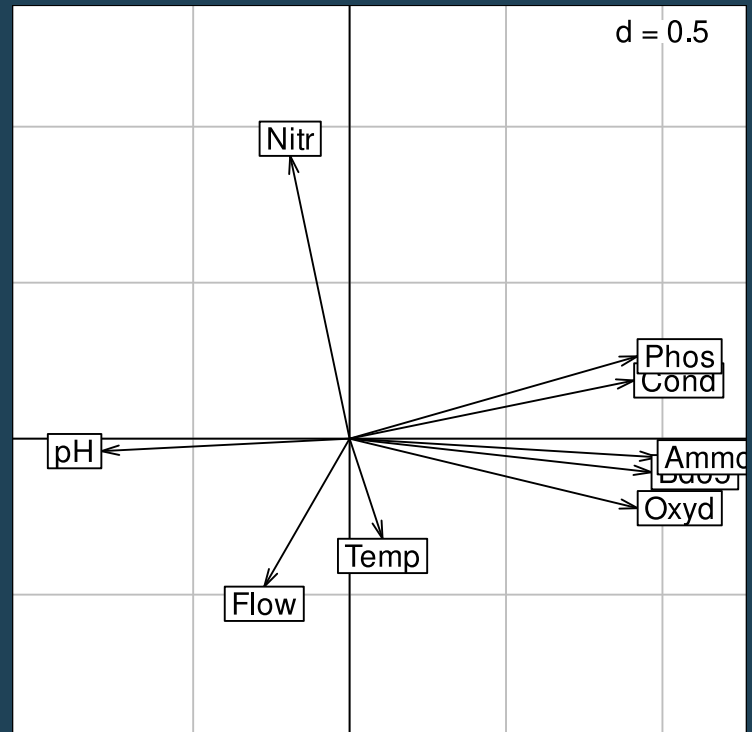


# Separate representations

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

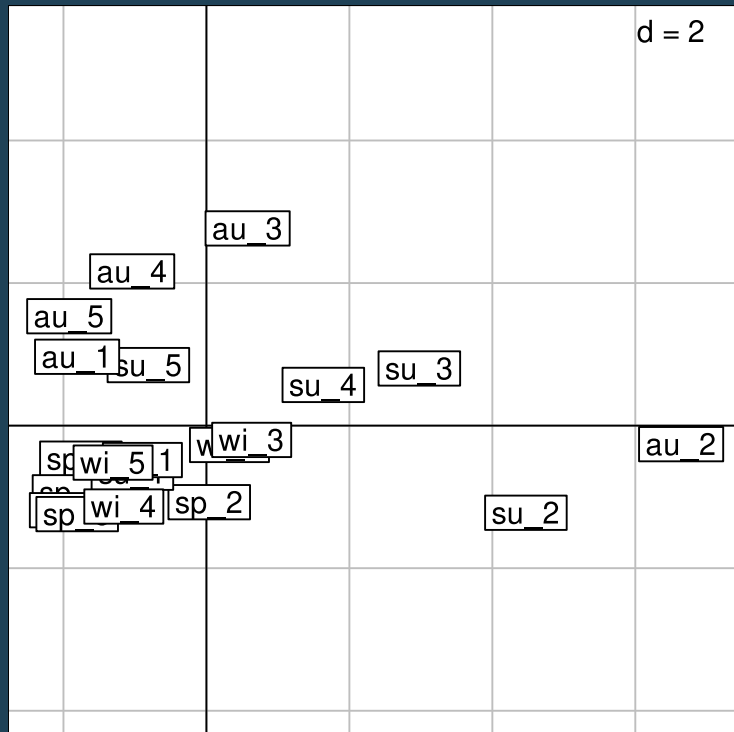


```
s.arrow(pca.meau$co)
```

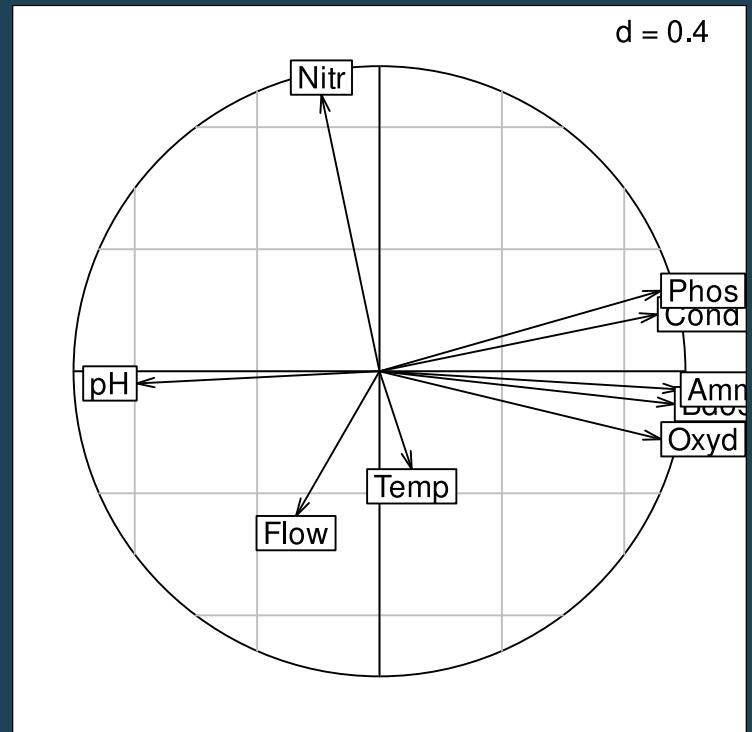


# Separate representations

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

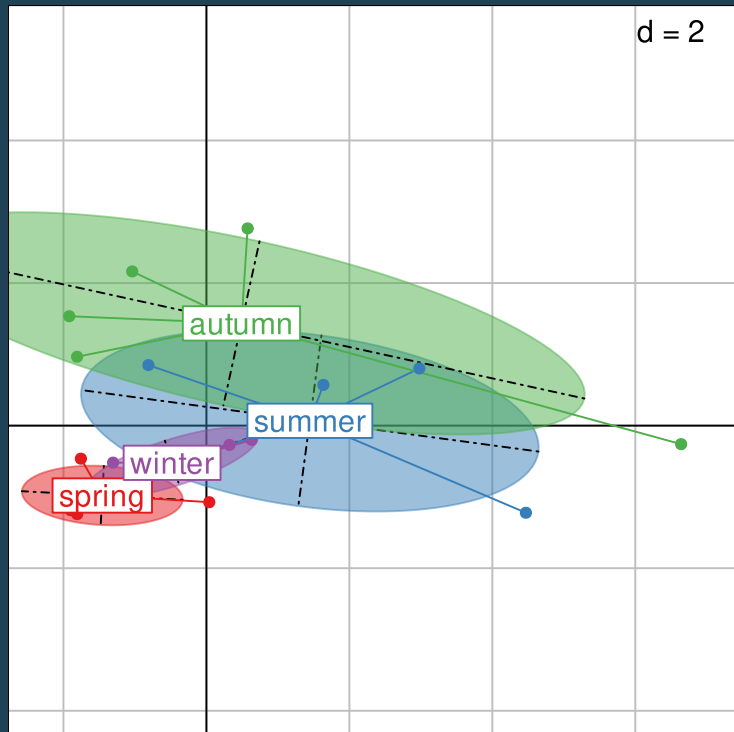


```
s.corcircle(pca.meau$co)
```

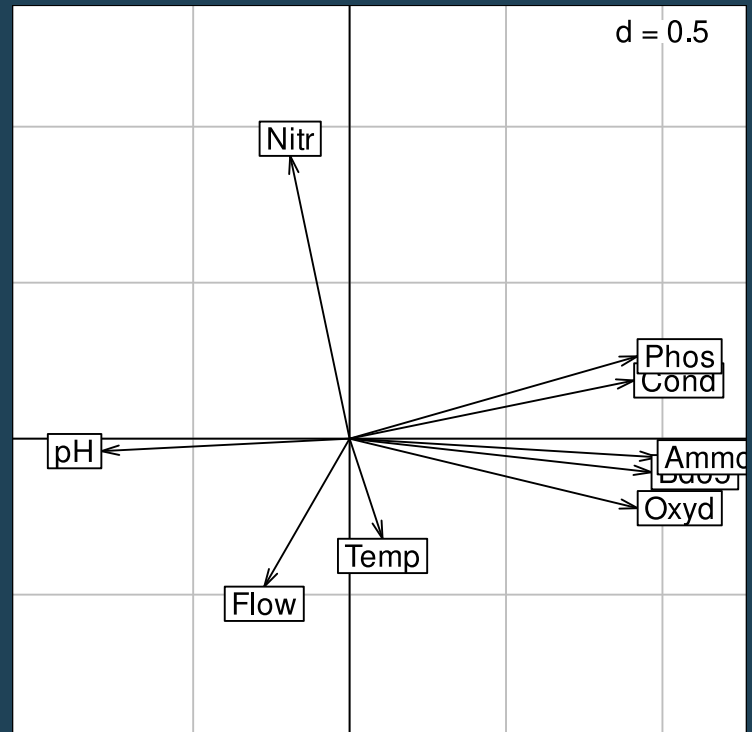


# Separate representations

```
s.class(pca.meau$li, meaudret$des
```



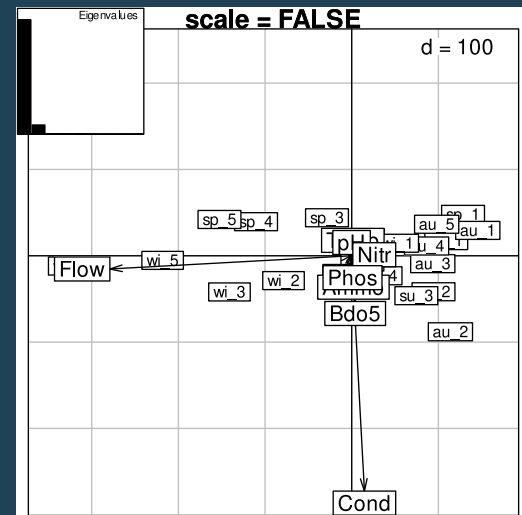
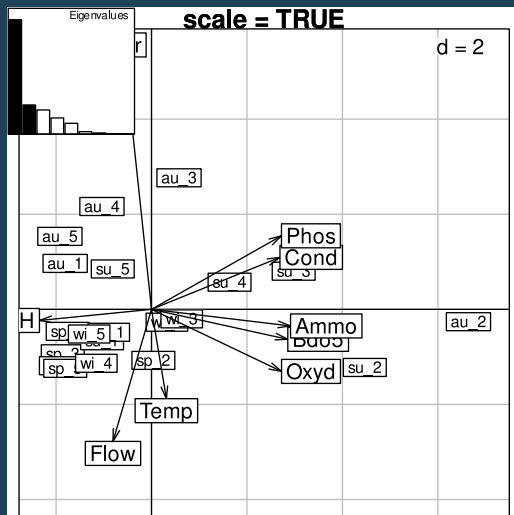
```
s.arrow(pca.meau$co)
```



# To scale or not to scale

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

```
pca.meau.c <- dudi.pca(meaudret$env, scannf = FALSE,  
  scale = FALSE)
```



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



# Inertia statistics

```
summary(pca.meau)
```

```
## Class: pca dudi
## Call: dudi.pca(df = meaudret$env, scannf = FALSE)
##
## Total inertia: 9
##
## Eigenvalues:
##      Ax1      Ax2      Ax3      Ax4      Ax5
##  5.1747  1.3204  1.0934  0.7321  0.4902
##
## Projected inertia (%):
##      Ax1      Ax2      Ax3      Ax4      Ax5
##  57.497  14.671  12.149   8.135   5.447
##
## Cumulative projected inertia (%):
##      Ax1  Ax1:2  Ax1:3  Ax1:4  Ax1:5
##   57.50   72.17   84.32   92.45   97.90
##
## (Only 5 dimensions (out of 9) are shown)
```

# PCA in practice

[Go to practical 2](#)