Multiple Correspondence Analysis

In practice

Stéphane Dray

2021-11-03

Data structure

- One table with p variables measured on n individuals
- All variables are **qualitative** (categorical)
- For instance
  - sites $\times$ environmental variables (e.g., soil types)
  - species $\times$ traits (e.g., functional groups)

Objectives

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

Data

We consider the `meaudret` data set

```
library(ade4)
library(adegraphics)
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"
```

Categorical variables

The data set contains an environmental table with 20 measurements of 9 environmental variables. For this example, quantitative variables are transformed into categorical variables:

```r
env.categ <- apply(meaudret$env, 2, cut, breaks = 3,
    labels = c("low", "med", "hi"))
env.categ <- as.data.frame(env.categ, stringsAsFactors = TRUE)
head(env.categ, 3)
```

```
##   Temp Flow pH Cond Bdo5 Oxyd Ammo Nitr Phos
## 1  med  low hi  low  low  low  low  low  low
## 2  med  low hi  low  low  med  low  low  low
## 3  med  med hi  low  low  low  low  low  low
```

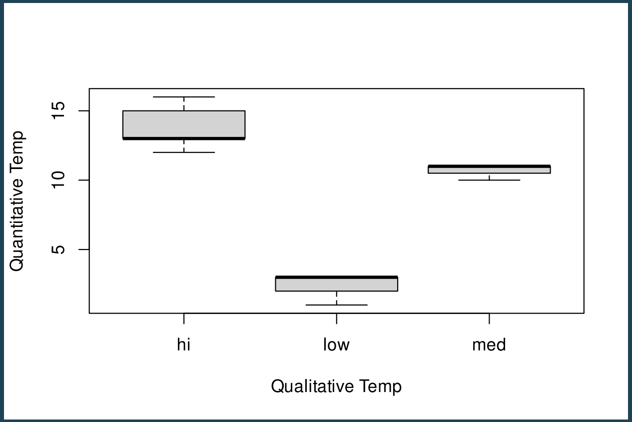
We want to know

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

Quantitative and categorical variables

- Some variables are recorded as categorical
- Quantitative variables can always be recoded as categorical ones
  - 😥 This introduces a loss of information (different values are regrouped into a single category)
  - 😄 It allows to detect non-linear relationships (order of categories is lost)

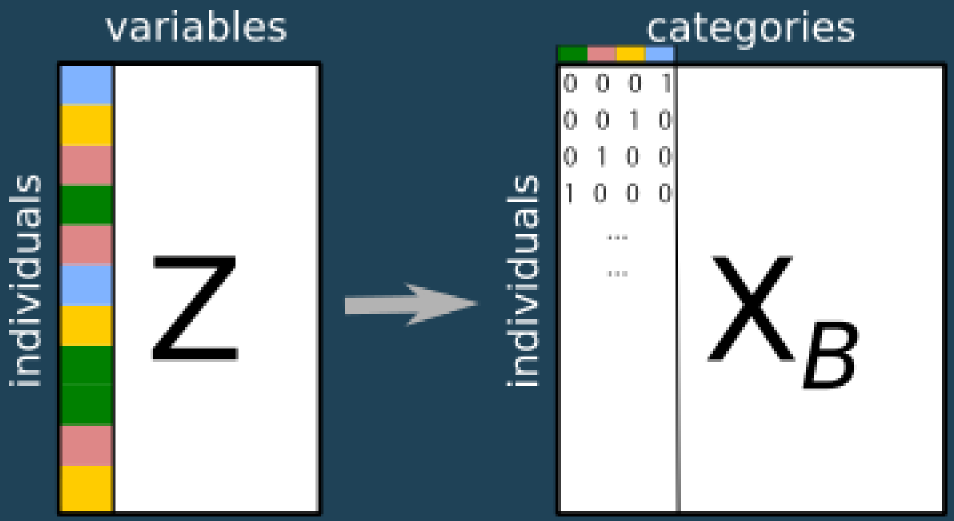
```
plot(meaudret$env[, 1] ~ env.categ[, 1], ylab = "Quantitative Temp",
    xlab = "Qualitative Temp")
```



Disjunctive table

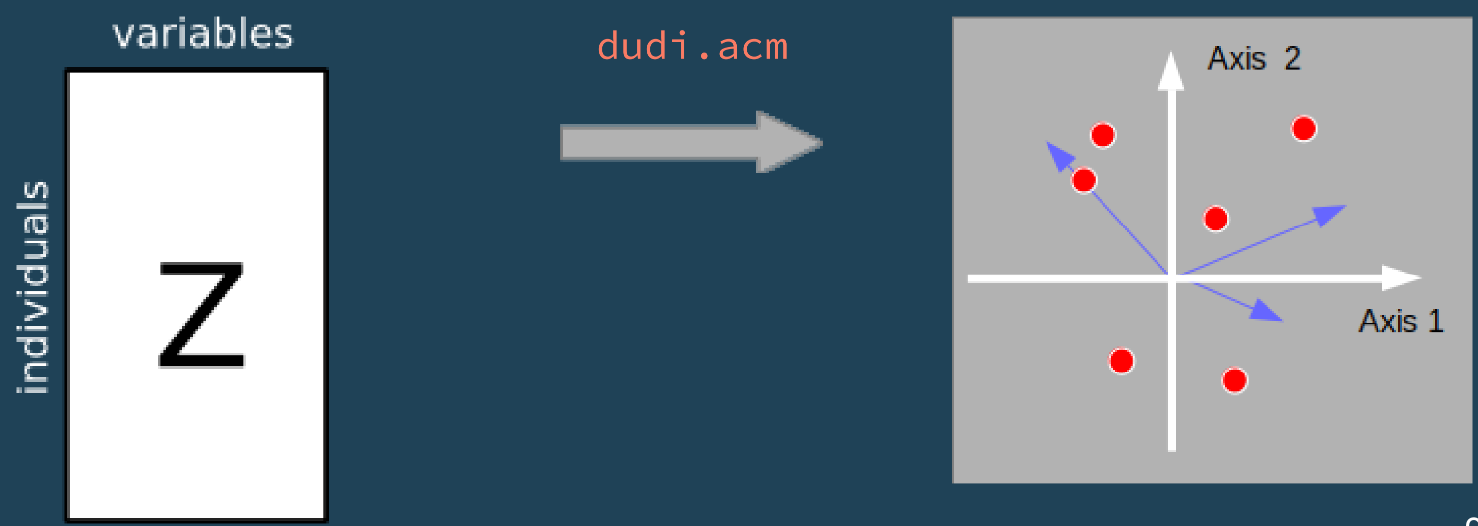
The original data table $\mathbf{Z}$ contains categorical information (words).

The first step is to build a disjunctive table with numbers. Information is stored as a binary table with n rows and m columns (total number of categories).



Multiple correspondence analysis

- $\mathbf{X} = \mathbf{X}_B {\mathbf{D}_m}^{-1} - \mathbf{1}_n {\mathbf{1}_m}^\top$ is the the transformed and centred disjunctive table
- $\mathbf{Q} = \frac{1}{p}\mathbf{D}_m$ where $\mathbf{D}_m = diag({\mathbf{X}_B}^\top \mathbf{D} \mathbf{1}_n)$ contains the category frequencies
- $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$ is the diagonal matrix with $\frac{1}{n}$



Maximized criteria

- For individuals

$$Q(\mathbf{a}) = \|\mathbf{X}\mathbf{Q}\mathbf{a}\|^2_{\mathbf{D}} = \|\mathbf{X}\frac{1}{p}\mathbf{D}_m\mathbf{a}\|^2_{\frac{1}{n}\mathbf{I}_n} = \|\frac{1}{p}\mathbf{X}_B\mathbf{a}\|^2_{\frac{1}{n}\mathbf{I}_n} = var\left(\frac{1}{p}\mathbf{X}_B\mathbf{a}\right) = \lambda$$

- For variables

$$\|\mathbf{X}^\top\mathbf{D}\mathbf{b}\|^2_{\mathbf{Q}} = \|\mathbf{X}^\top\frac{1}{n}\mathbf{I}_n\mathbf{b}\|^2_{\frac{1}{p}\mathbf{D}_m} = \|\frac{1}{n}\mathbf{D}_m^{-1}\mathbf{X}_B{}^\top\mathbf{b}\|^2_{\frac{1}{v}\mathbf{D}_m}$$

The vector $\frac{1}{n}\mathbf{D}_m^{-1}\mathbf{X}_B{}^\top\mathbf{b}$ contains means of $\mathbf{b}$ per category so that:

$$\|\mathbf{X}^\top\frac{1}{n}\mathbf{I}_n\mathbf{b}\|^2_{\frac{1}{p}\mathbf{D}_m} = \frac{1}{p}\sum_{j=1}^{p}\eta^2(\mathbf{z}_j, \mathbf{b})$$

This quantity is the mean of correlation ratios computed for all variables.

The `dudi.acm` function

Arguments

```
args(dudi.acm)
```

```
## function (df, row.w = rep(1, nrow(df)), scannf = TRUE, nf = 2)
## NULL
```

- `df` is a `data.frame` with the categorical data (`factors` in R)
- `row.w` is an optional vector of weights
- `scannf` and `nf` allow to set the number of dimensions to interpret

```
mca.meau <- dudi.acm(env.categ, scannf = FALSE)
```

Returned values

```r
names(mca.meau)
```

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

It returns an object of class `dudi` containing:

- `$eig`: eigenvalues ($\mathbf{\Lambda}$)
- `$cw`: column (i.e., category) weights ($\frac{1}{v}\mathbf{D}_m$)
- `$lw`: row weights ($\mathbf{D} = \frac{1}{n}\mathbf{I}_n$)
- `$tab`: transformed and centred disjunctive data table ($\mathbf{X}$)
- `$c1`: category loadings ($\mathbf{A}$)
- `$li`: row scores ($\mathbf{L} = \frac{1}{p}\mathbf{X}_B\mathbf{A}$)
- `$l1`: principal components ($\mathbf{B}$)
- `$co`: column scores ($\mathbf{C} = \frac{1}{n}\mathbf{D}_m^{-1}\mathbf{X}_B{}^\top\mathbf{B}$)
- `$cr`: correlation ratios between qualitative variables and axes

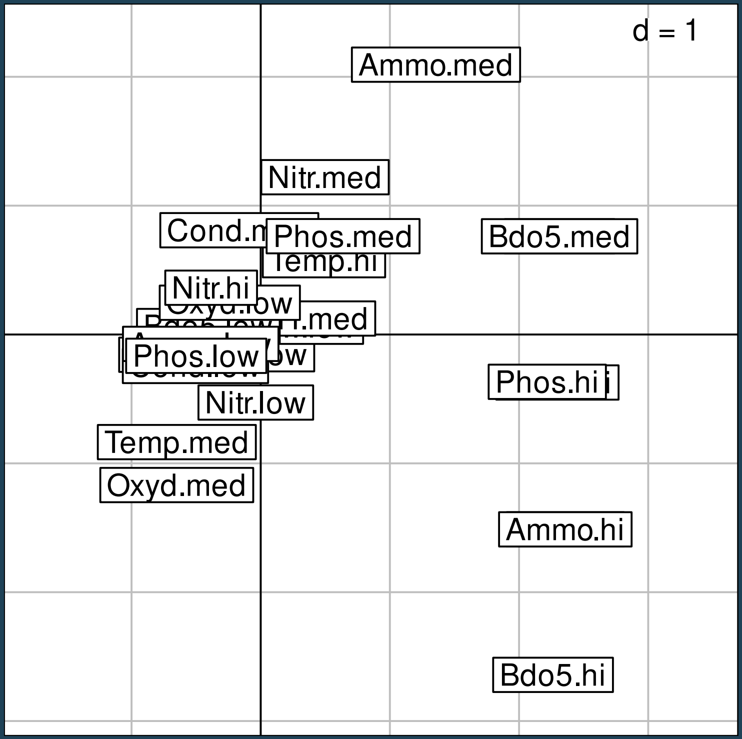
Graphical representation and interpretation

In the first viewpoint, MCA positions categories by a normed score (`$c1`}. A score for individuals (`$li`) is derived from this categories score: an individual is located at the mean of the score of the categories that it carries. This second score provides an ordination of individuals with the highest possible dispersion (maximum variance).

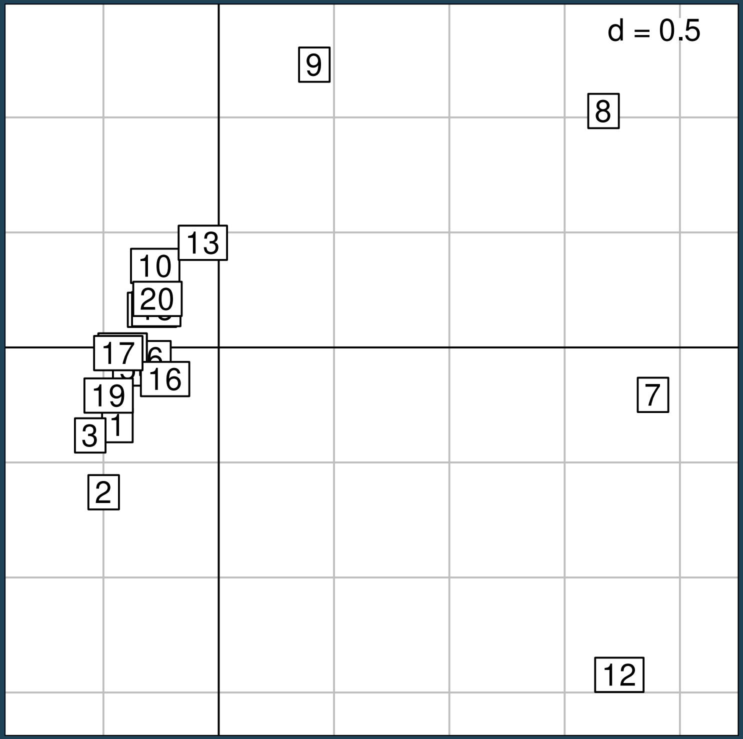
In the second type of interpretation, MCA finds normed coordinates for individuals (`$l1`) and positions categories at the mean of the individual scores that belong to them (`$co`). This maximises the mean of the variance of the categories for all variables. In other words, it maximises the mean of the correlation ratios.

Graphical representations

```
s.label(mca.meau$co)
```

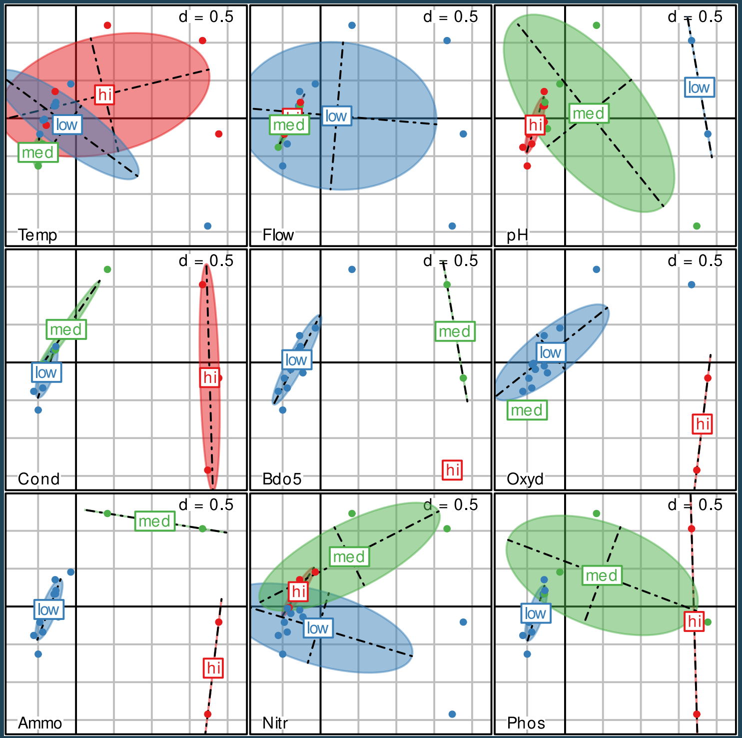



```
s.label(mca.meau$li)
```



Optimal representation

```r
plot(mca.meau, col = TRUE)
```




```r
mca.meau$cr
```

```
##              RS1         RS2
## Temp 0.15821537 0.232782758
## Flow 0.14545664 0.008532622
## pH   0.75017631 0.079641855
## Cond 0.95131327 0.220320278
## Bdo5 0.93106310 0.410689421
## Oxyd 0.62593624 0.348643718
## Ammo 0.91289231 0.671527779
## Nitr 0.07988202 0.491733356
## Phos 0.74796252 0.149118142
```

Inertia statistics

```r
summary(mca.meau)
```

```
## Class: acm dudi
## Call: dudi.acm(df = env.categ, scannf = FALSE)
##
## Total inertia: 2
##
## Eigenvalues:
##     Ax1     Ax2     Ax3     Ax4     Ax5
##  0.5892  0.2903  0.2505  0.1971  0.1561
##
## Projected inertia (%):
##     Ax1     Ax2     Ax3     Ax4     Ax5
##  29.461  14.517  12.525   9.857   7.804
##
## Cumulative projected inertia (%):
##     Ax1   Ax1:2   Ax1:3   Ax1:4   Ax1:5
##   29.46   43.98   56.50   66.36   74.16
##
## (Only 5 dimensions (out of 14) are shown)
```

Mix of variables

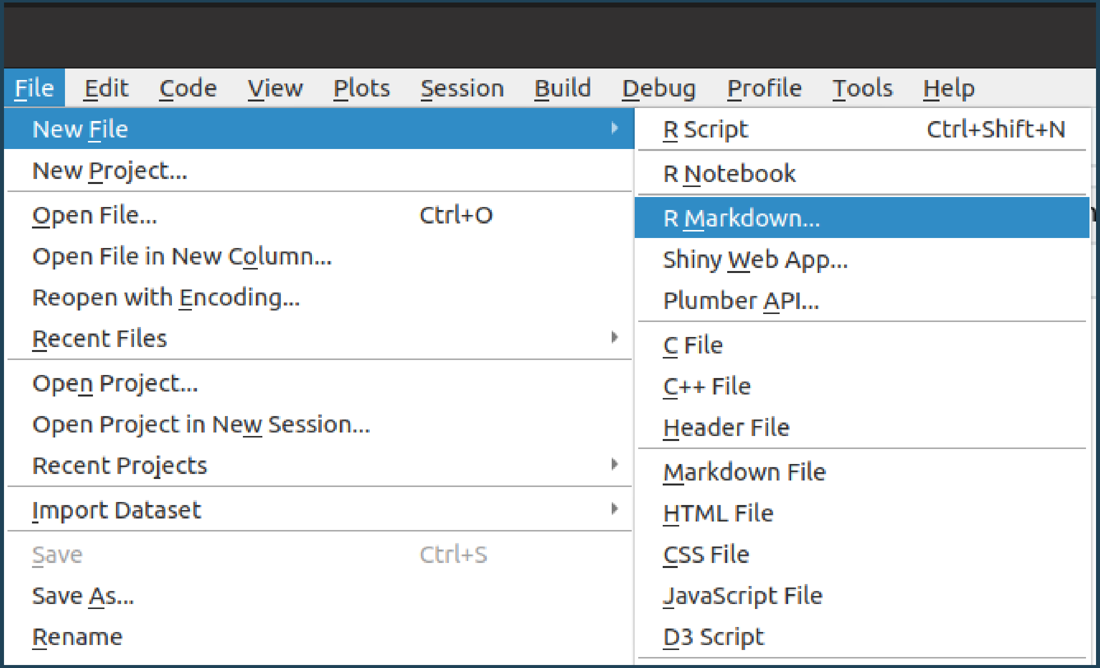
If a table contains both quantitative and categorical variables, Hill and Smith analysis (HSA) can be applied. See `dudi.hillsmith`

This method is a compromise between PCA and MCA.

- If all the variables are quantitative, then the results of HSA are identical to those of PCA.
- If all the variables are qualitative, then the results are identical to those of MCA.
- If there is a mix of variables, then the analysis is an optimal combination of the properties of the two analyses (maximizing the squared correlations for quantitative variables and correlation ratios for categorical ones)

Your turn

Write a report with Rmarkdown



Data

We will analyze the `doubs` data set (see `?doubs`)

```
library(ade4)
library(adegraphics)
data(doubs)
names(doubs)
```

```
## [1] "env"     "fish"    "xy"      "species"
```

```
names(doubs$env)
```

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

Tranformation into categorical variables

```
fenv <- apply(doubs$env, 2, cut, breaks = 4, labels = 1:4)
fenv <- as.data.frame(fenv, stringsAsFactors = TRUE)
```

Multiple Correspondence Analysis

- Perform MCA
- Display the barplot of eigenvalues

Graphical representation of MCA results

- Plot the results using the `plot` function

PCA scores on the geographical map

- Draw maps of PCA scores on the first two axes
- Interpret the maps to describe the environmental structure of the river

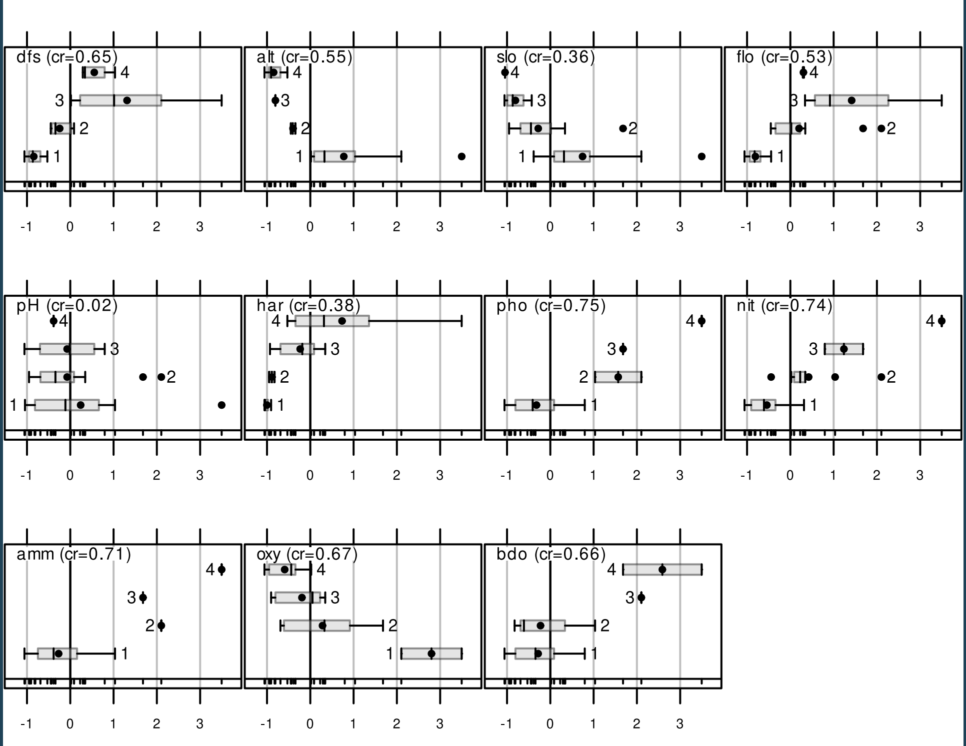
A look to variables

- Which variables are the most discriminated by the first axes

A look to variables

The generic function `score` provides an optimal representation of the maximized criteria

```r
score(acm1, type = "boxplot")
```



Hill-Smith analysis

- Build a table mixing quantitative and categorical variables

```
menv <- cbind(fenv[, 1:6], doubs$env[, 7:11])
```

- Perform Hill-Smith analysis

Graphical representation

```
score(hs1, type = "boxplot")
```