Lecture 2: Multivariate visualization

Bénédicte Colnet

2023-02-19

Reminder of session 1

Univariate and bivariate plots

Plots and graphics usually are the starting point for statistical analysis. On the first session we have seen how to plot univariate covariate (e.g. histogram, boxplot, bar plot) and also bivariate analysis (such as scatter plot, or boxplots as a function of a categorical covariate). *What if you have more covariates?* It would still be possible to observe data with a 3D plots[1], but one can hardly go beyond this analysis. Before going on, let's check that everyone is able to reproduce the plot below.

[1] Note that R does this well too. You can try the gg3D library

```r
# Load library for plot
library(ggplot2)

# Load data set (be careful with the path)
immo <- read.csv("./2022.csv")

# Clean data
filter <- !is.na(immo$valeur_fonciere) &
  !is.na(immo$lot1_surface_carrez) &
  immo$valeur_fonciere < 1000000 &
  immo$lot1_surface_carrez < 80


immo <- immo[filter,]


# Plot data
ggplot(immo, aes(x = lot1_surface_carrez,
                 y = valeur_fonciere,
                 color = valeur_fonciere)) +
  geom_point() +
  geom_smooth(method = "lm", color = "purple") +
  xlab("Surface du premier lot en m2") +
  ylab("Valeur foncière") +
  theme_classic() +
  theme(legend.position = "none")
```

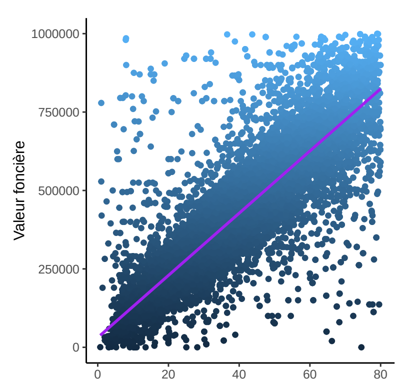


Figure 1: Open data for Paris housing price

Pipe operator in `R`

In the first lab, you were also asked to plot summary of data, such as the average price per year. Here, we will plot the average price of a flat depending on its number of living room. To do this, a first step is to compute this average price per group. This will allow us to present the pipe operator in `R`. The pipe operator is denoted `%>%` and corresponds to "chaining" several functions. It means that you invoke multiple method calls. As each method returns an object, you can actually allow the calls to be chained together in a single statement, without needing variables to store the intermediate results.

```
library(lubridate)

## Loading required package: timechange

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

immo$annee <- year(immo$date_mutation)

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

summary.immo <- immo[immo$nombre_pieces_principales < 6,] %>%
  group_by(nombre_pieces_principales) %>%
  summarise(prix.moyen = mean(valeur_fonciere))

library(dplyr)
summarized.immo <- immo[immo$nombre_pieces_principales < 5,] %>%
  group_by(nombre_pieces_principales) %>%
  summarise(prix.moyen = mean(valeur_fonciere))
```

Then, we observe what has been produced.

```
head(summarized.immo[1:6,])
```

```
## # A tibble: 6 x 2
##   nombre_pieces_principales prix.moyen
##                       <int>      <dbl>
## 1                         0    422752.
## 2                         1    250833.
## 3                         2    410749.
## 4                         3    596760.
## 5                         4    694314.
## 6                        NA        NA
```

```
library(ggplot2)
# Plot data
ggplot(summary.immo, aes(x = nombre_pieces_principales, y = prix.moyen)) +
  geom_point() +
  geom_line() +
  theme(legend.position = 'bottom') +
  xlab("Nombre de pièces principales") +
  ylab("Valeur foncière") +
  theme_bw()
```

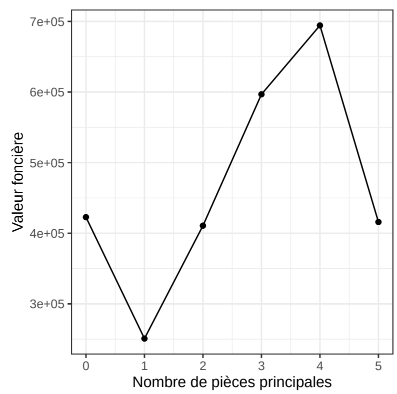


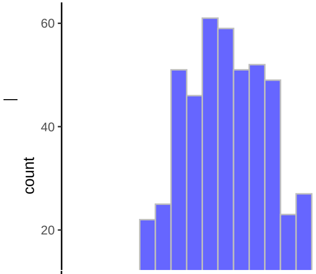
Figure 2: Aggregated data

Why is it interesting to visualize covariates jointly?

Let's look at a funny example. Imagine that we generate two variables X_1 and X_2 from normal distributions. We want these variables to be linked (correlated) and such that $X_j \sim \mathcal{N}(0, 1)$. The following chunk performs the simulation. You can take the output data frame and explore the data first with univariate analysis. And then with a bivariate plot. An outlier is in the dataset. Can we recover it?

```
# simulation -- don't need to understand what is going on here
library(MASS) # for simulations
Sigma <- matrix(c(1,0.8,1,0.8),2,2)
simulated_data <- mvrnorm(n = 500, mu = c(0,0), Sigma)
output <- data.frame(simulated_data)
names(output) <- c("X1", "X2")
output[501,] <- c("X1" = 2, "X2" = -2) # outlier step
```

```
ggplot(output, aes(x = X1)) +
  geom_histogram(bins = 20,
                 fill = "blue",
                 alpha = 0.6,
                 color = "grey") +
  theme_classic()
```




```r
ggplot(output, aes(x = X2)) +
  geom_histogram(bins = 20,
                 fill = "magenta",
                 alpha = 0.6,
                 color = "grey") +
  theme_classic()
```

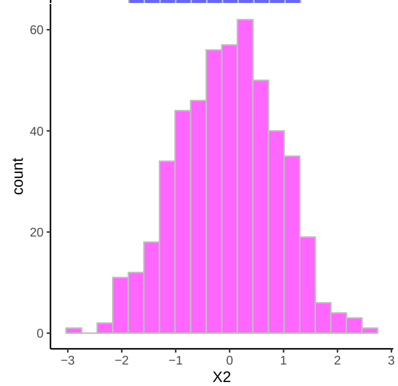


Figure 4: X2

One can rather plot the two covariates at once.

```r
ggplot(output, aes(x = X1, y = X2)) +
  geom_point() +
  theme_classic()
```

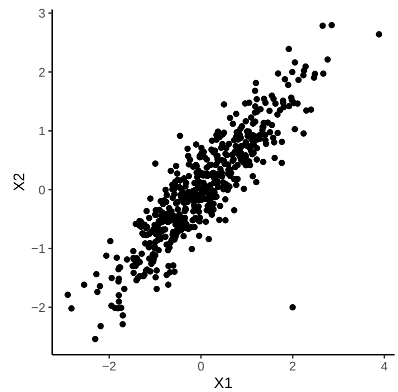


Figure 5: X1 and X2 on a scatter plot

The outlier is clearly identifiable on this scatter plot, but not using only the boxplot or any univariate tool. This is to highlight that multivariate analysis will allow us to see high dimensional outliers.

Multivariate analysis will enable us to summarize highly dimensional data into a simpler 2D plot. This will rely on factorial analysis, where the aim is to summarize a large dataframe. The exact method chosen depends on the nature of the covariates. For example if all covariates are continuous, then Principal Component Analysis (PCA) can be used, but if the covariates are qualitative, then the method is rather correspondance analysis.

Principal Component Analysis (Work from home)

We recommend to watch the videos[2] from François Husson about PCA. Below we recall the main principles.

[2] Here is the link.

- **Context** Principal Component Analysis (usually the shortname is PCA but you can also find ACP in French) focuses on typical data you can find in several domains: *observations* (or individus) in rows, and *variables* in column. Note that the PCA focuses on *quantitative* variables (for example age, or price, but not color or sex). For example we can study the average temperature depending on cities. In that case cities are rows, and in column the average temperature per month.

- **Typical question an ACP answers** A typical question you may ask on your data is: how much the different observations are close to one another considering the variables? (remember that everything you will conclude depends on these variables that you added in your initial model) You can also see PCA as a way to find a low-dimensional representation that captures the "essence" of high-dimensional data

- **What can you interpret from data?** The PCA will group similar individuals together. Information are also learned on variables, with the correlated variables (meaning that you have a linear link between two variables), and also which variables synthetize the most the observations, or which variables bring different information.

Correspondance Analysis (CA)

Typical situation is when you have two qualitative covariates, and in particular a data counting how many times the occurences co-occur in the data. CA proposes you to visualize how the two covariates are associated.

A few historical information: - First applications in the 60's ny Jean-Paul Bensécri (a whole French community on these kind of analysis) - One of the first application is on the characters from the play *Phèdre*.

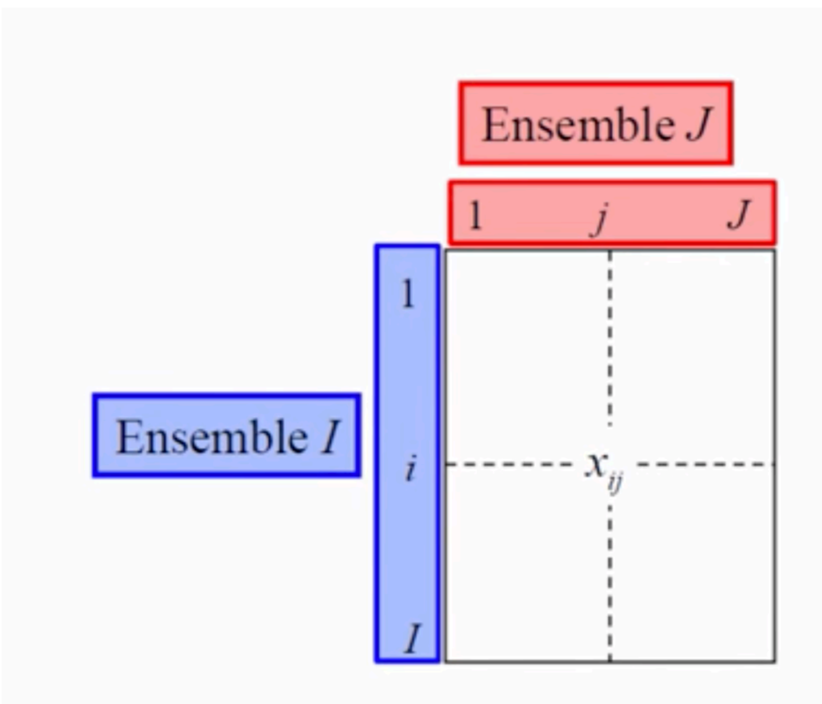


Figure 6: Typical data used: contingency table

Principle

For a concrete example, let's count the number of nobel prize in each domain for the height countries of G8. Is there a specialty depending on the countries? Below we show the example.

- $x_{i,j}$ corresponding to the number of individuals with both characteristics i from column I and j from column J (see Figure).
- Here n, was all the nobel prizes, and the count data are summarizing all these.
- Note that, $\sum_{i=1}^{8}\sum_{j=1}^{7} x_{i,j} = n$.

	Chimie	Econ.	Lettres	Médecine	Paix	Physique	Math
Allemagne	24	1	8	18	5	24	1
Canada	4	3	2	4	1	4	1
France	8	3	11	12	10	9	11
GB	23	6	7	26	11	20	4
Italie	1	1	6	5	1	5	1
Japon	6	0	2	3	1	11	3

From this contingency table, it is possible to go to the probability table, noting that

$$f_{i,j} = \frac{x_{i,j}}{n}.$$

The correspondance analysis will work on this table. For those who have been doing a lecture in statistics or probabilities, one has the joint probability to observe both i and j simulatenously,

$$\mathbb{P}\left[X_i = i, X_j = j\right] = f_{i,j}.$$

We will also look at marginal probabilities, that are,

$$f_i = \sum_{j=1}^{J} f_{i,j}$$

and

$$f_j = \sum_{i=1}^{I} f_{i,j}.$$

Now, the key idea is to observe how much each column (or row) is different than the marginal probability. Two events are said to be independent if

$$\mathbb{P}\left[A, B\right] = \mathbb{P}\left[A\right] \cdot \mathbb{P}\left[B\right].$$

> "The joint probability is equal to the product of the marginal probabilities."

Now, the idea is to say that data are not independent if the observed joint probabilities $f_{i,j}$ (observed) are different than the product of the marginal probabilities $f_i \cdot f_j$ (i.e. independence model). Maybe this reminds you the χ^2 test to compare the observed values with the theoretical values.

$$\chi^2_{\text{obs}} = \sum_{i=1}^{1}\sum_{j=1}^{J} \frac{(\text{ obs. num. } - \text{ theor. num. })^2}{\text{theor. num}} = \sum_{i=1}^{1}\sum_{j=1}^{J} \frac{(nf_{ij} - nf_i, f_j)^2}{nf_{i.}f_{.j}} = n\Phi^2.$$

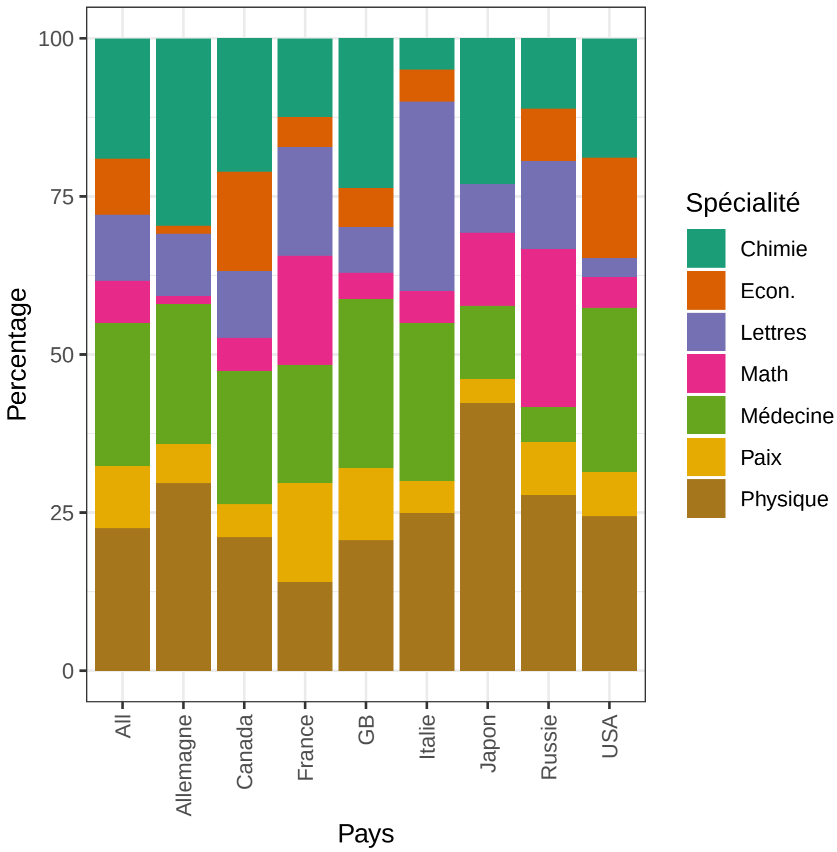
The higher Φ^2, the higher the deviation from independence. In other words Φ^2 is the strength of the relation shipt (it does not depend on n). In this class, we do not focus on whether the link is stastistically different from 0. But we use it to plot the data and understand the link.

In other words, we are comparing each column profile, with its marginal one. This may seem a bit abstract, so let's look at our running example. First, this is the frequency table.

	Chimie	Econ.	Lettres	Médecine	Paix	Physique	Math
Allemagne	0.1983471	0.0082645	0.0661157	0.1487603	0.0413223	0.1983471	0.0082645
Canada	0.0330579	0.0247934	0.0165289	0.0330579	0.0082645	0.0330579	0.0082645
France	0.0661157	0.0247934	0.0909091	0.0991736	0.0826446	0.0743802	0.0909091
GB	0.1900826	0.0495868	0.0578512	0.2148760	0.0909091	0.1652893	0.0330579
Italie	0.0082645	0.0082645	0.0495868	0.0413223	0.0082645	0.0413223	0.0082645
Japon	0.0495868	0.0000000	0.0165289	0.0247934	0.0082645	0.0909091	0.0247934

Now, if we are willing to observe marginal versus conditional distribution.

```
library(ggplot2)
library(tidyr)
nobel.freq %>%
  pivot_longer(cols = c("Chimie", "Econ.", "Lettres", "Médecine", "Paix", "Physique", "Math"), names_to
  ggplot(aes(x = Pays, y = Percentage, fill = Spécialité, group = Spécialité)) +
  geom_bar(stat = "identity") +
  theme_bw() +
  scale_fill_brewer(palette="Dark2") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



See how Italy has a relative more important number in letters. This is the contrary for the USA. The goal is to compare the distribution of Nobel Prize winners.

Each country has a profile. Denoting i the country, for each country we have

$$(\frac{f_{i,1}}{f_i}, \frac{f_{i,2}}{f_i}, \dots, \frac{f_{i,J}}{f_i}).$$

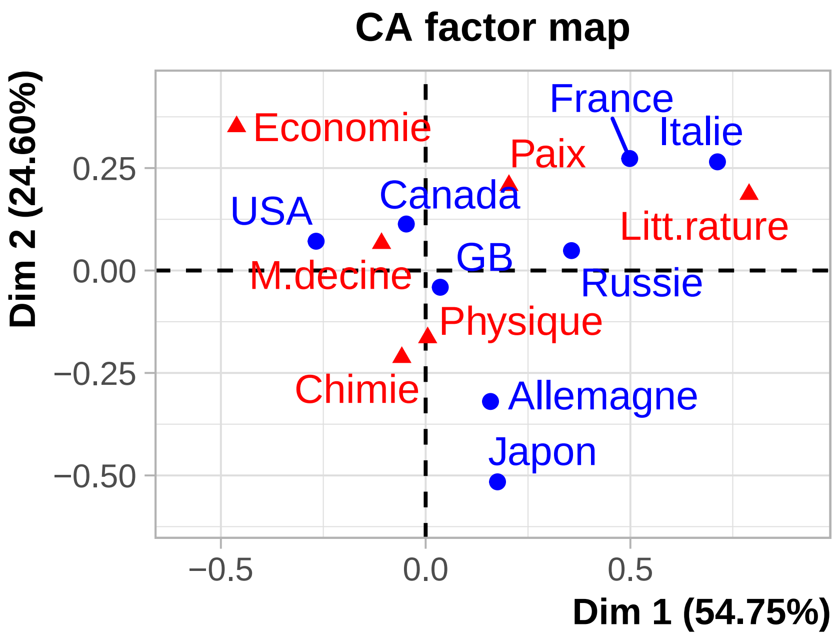
Then, in this space we compute the distance between each vector to the mean vector. In other words, each country is represented by a vector in J dimensional space (i.e. the number or majors).

$$d^2_{\chi^2}(i, i') = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'}} \right)^2.$$

The same can be done in the other direction.

If there is independence, all the vectors are more or less confounded with the mean vector. You can imagine a cloud of points and that all the points are very close to the center of gravity. It can be shown that the inertia of the cloud (i.e. how much it spreads) is linked with Φ^2. Also, it can be shown that rows and columns have the same role.

Then, the process is the same than the PCA, finding plan on which the cloud is the most dispersed.



- UK is super close to the center of gravity (look back to the previous plot);
- Italy and France seems really different;
- Red points are spread out and we could put the assumption that the first axis contrasts science and other categories, while the second axis contrasts natural science with economic science.

Application on Majors and studies

The example of this part are data from the French universities, and in particular how many students are in which specialty, degree (L3, M2, PhD), and their gender. Typical questions we will answer are "Are they major in which students are similar or different?" "Is there an association between the major and the gender?" and so on.

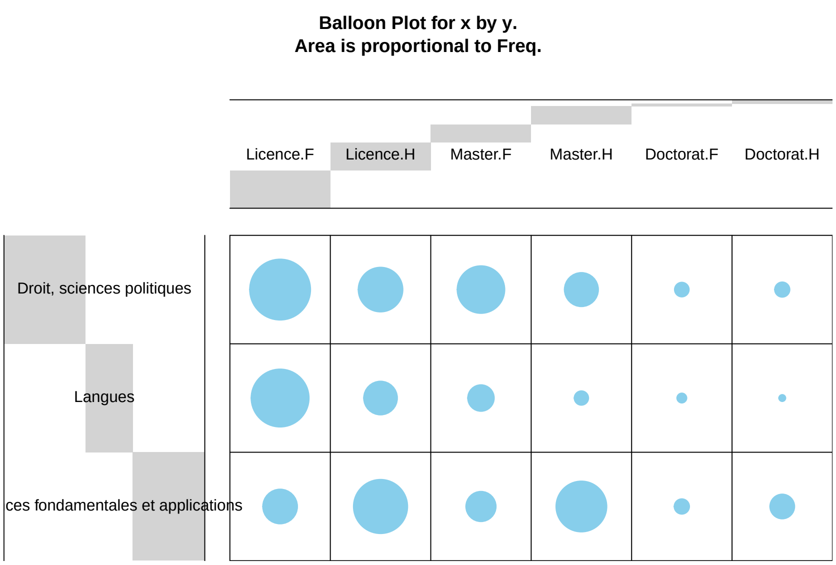
```
univ <- read.csv("./universite.csv", header = TRUE, row.names = 1, skip = 0, sep = ";")
head(univ)
```

```
##                                        Licence.F Licence.H Master.F Master.H
## Droit, sciences politiques                 69373     37317    42371    21693
## Sciences economiques, gestion              38387     37157    29466    26929
## Administration economique et sociale       18574     12388     4183     2884
## Lettres, sciences du langage, arts         48691     17850    17672     5853
## Langues                                    62736     21291    13186     3874
## Sciences humaines et sociales              94346     41050    43016    20447
##                                        Doctorat.F Doctorat.H Total.F Total.H
```

```
## Droit, sciences politiques                 4029      4342  115773   63352
## Sciences economiques, gestion              1983      2552   69836   66638
## Administration economique et sociale          0         0   22757   15272
## Lettres, sciences du langage, arts         4531      2401   70894   26104
## Langues                                    1839       907   77761   26072
## Sciences humaines et sociales              7787      6972  145149   68469
##                                      Licence Master Doctorat  Total
## Droit, sciences politiques            106690  64064     8371 179125
## Sciences economiques, gestion          75544  56395     4535 136474
## Administration economique et sociale   30962   7067        0  38029
## Lettres, sciences du langage, arts     66541  23525     6932  96998
## Langues                                84027  17060     2746 103833
## Sciences humaines et sociales         135396  63463    14759 213618
```

Be careful as some columns are summing other columns. Here we focus on the variable with gender and level. Before launching the CA, one can still have another view of the data.

```
library("gplots")
# 1. convert the data as a table
dt <- as.table(as.matrix(univ[c(1,5,8),1:6]))
# 2. Graph
balloonplot(t(dt), xlab ="", ylab="",
            label = FALSE, show.margins = FALSE)
```




```
library(FactoMineR)
analysis.ca <- CA(univ, col.sup = 7:12, graph = FALSE)
```

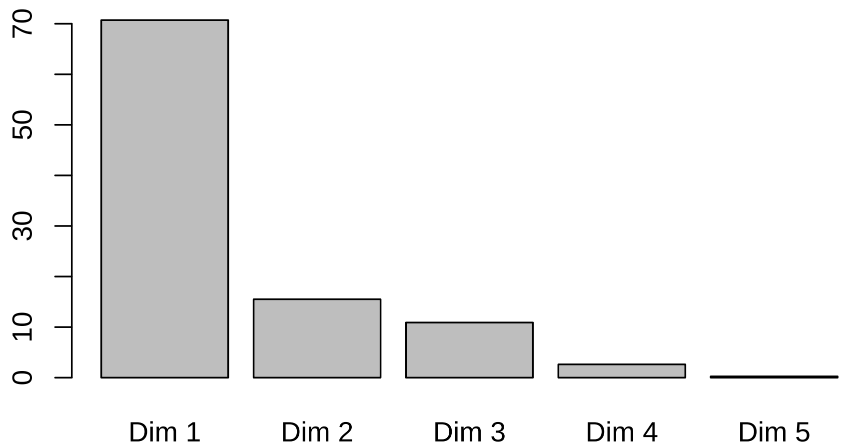
The object `analysis.ca` contains all the results, and automatically it output one plot if `graph = TRUE`.

```
# summary only for the first 2 dimensions
summary(analysis.ca, ncp = 2, dim = 2, nb.dec = 1, nbelements = 2)

##
## Call:
## CA(X = univ, col.sup = 7:12, graph = FALSE)
##
## The chi square of independence between the two variables is equal to 170789.2 (p-value =  0 ).
##
## Eigenvalues
##                       Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
## Variance                0.1   0.0   0.0   0.0   0.0
## % of var.              70.7  15.5  10.9   2.6   0.2
## Cumulative % of var.   70.7  86.2  97.1  99.8 100.0
##
## Rows (the 2 first)
##                                                                 Iner*1000   Dim.1
## Droit, sciences politiques                                     |       5.7 |  -0.1
## Sciences economiques, gestion                                  |       9.8 |   0.2
##                                                                 ctr cos2   Dim.2
## Droit, sciences politiques                                      1.4  0.3 |   0.1
## Sciences economiques, gestion                                   3.9  0.5 |   0.0
##                                                                 ctr cos2
## Droit, sciences politiques                                      2.9  0.1 |
## Sciences economiques, gestion                                   0.1  0.0 |
##
## Columns (the 2 first)
##                                                                 Iner*1000   Dim.1
## Licence.F                                                      |      48.3 |  -0.4
## Licence.H                                                      |      24.3 |   0.2
##                                                                 ctr cos2   Dim.2
## Licence.F                                                      39.7  1.0 |   0.0
## Licence.H                                                      11.5  0.6 |  -0.2
##                                                                 ctr cos2
## Licence.F                                                       2.3  0.0 |
## Licence.H                                                      37.5  0.4 |
##
## Supplementary columns (the 2 first)
##                                                                 Dim.1 cos2
## Total.F                                                        |  -0.3  1.0 |
## Total.H                                                        |   0.4  1.0 |
##                                                                Dim.2 cos2
## Total.F                                                          0.0  0.0 |
## Total.H                                                         -0.1  0.0 |
```

As you can see, the independence test is rejected:

`The chi square of independence between the two variables is equal to 170789.2 (p-value =  0 ).` So one can conclude on the existence of associations between some majors and the level-gender covariates.

```
barplot(analysis.ca$eig[,2],names=paste("Dim",1:nrow(analysis.ca$eig)))
```



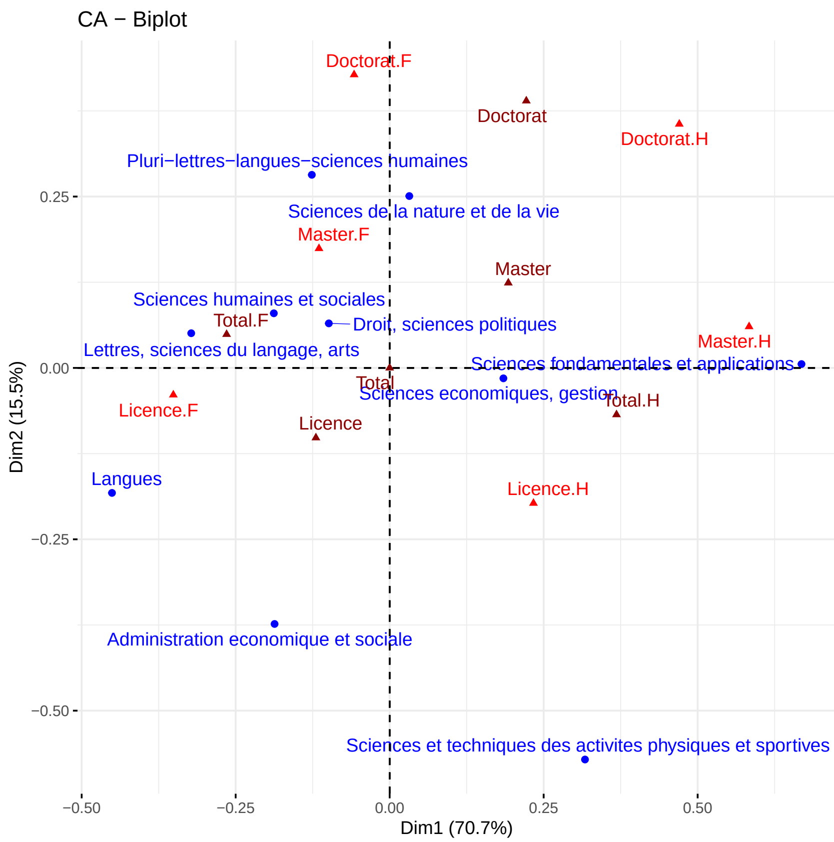
Looking at the percentage of inertia, one can say that the three first dimensions summarizes 97% of the total inertie (almost equal to variability), so only analyzing those three dimensions is enough.

We use another library (but this is not mandatory of course) to plot the results.

```
library("factoextra")
```

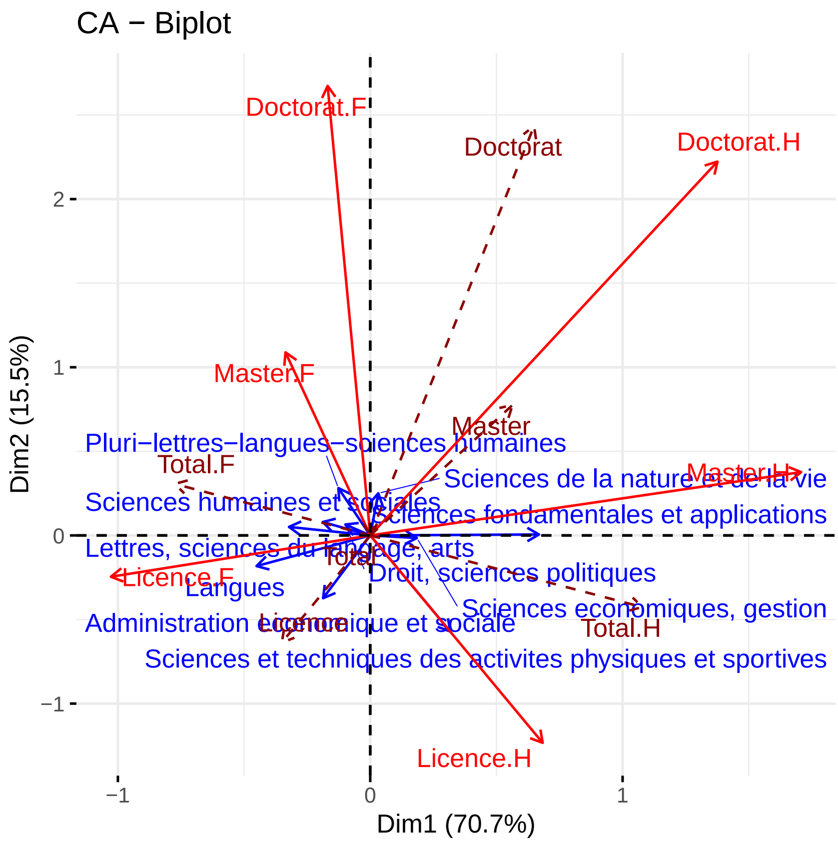
By the way, this package can also allow you to vizualise the data before the analysis.

```
# repel= TRUE to avoid text overlapping (slow if many point)
fviz_ca_biplot(analysis.ca, repel = TRUE)
```



You can also draw what is called an *asymetric biplot*.

```
fviz_ca_biplot(analysis.ca,
               map ="rowprincipal", arrow = c(TRUE, TRUE),
               repel = TRUE)
```

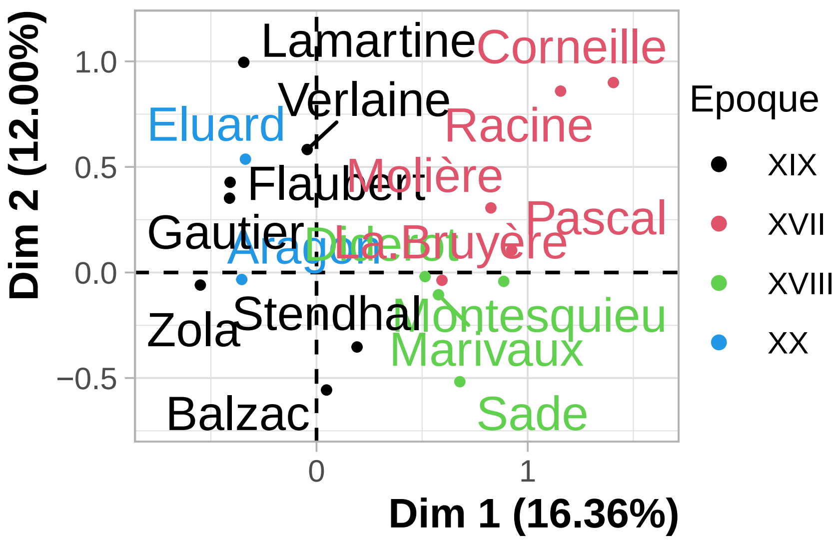


- Recall that two majors are close if they attract similar profiles (here gender and studies length)
- Langues, Lettres, Science du Language: attract women in licence.
- Women and men seems to be separated along the first axis, men on the right, women on the left. Second dimension is more related to studies length: from licence at the bottom and PhD in the upper part.
- Major on the left are mostly occuppied by women, and on the right by men.
- It is not always easy to interpret the axis (here it seems possible): in general, you can focus on which entities are close or not.

Correspondance analysis and text data analysis

```
books = read.table("./litterature.csv", header=TRUE, row.names=1, sep=";", check.names=FALSE, quote="\"

res.ca = CA(books, quanti.sup=1, quali.sup=2:3, graph = F)

plot(res.ca,
     invis = c("col","quali"),
     hab=2, cex=1.2,
     title="",cex.axis=1.2,
```

```
cex.lab=1.2,
palette=palette(c("black","green3","blue","darkred","orange")),shadow=TRUE)
```



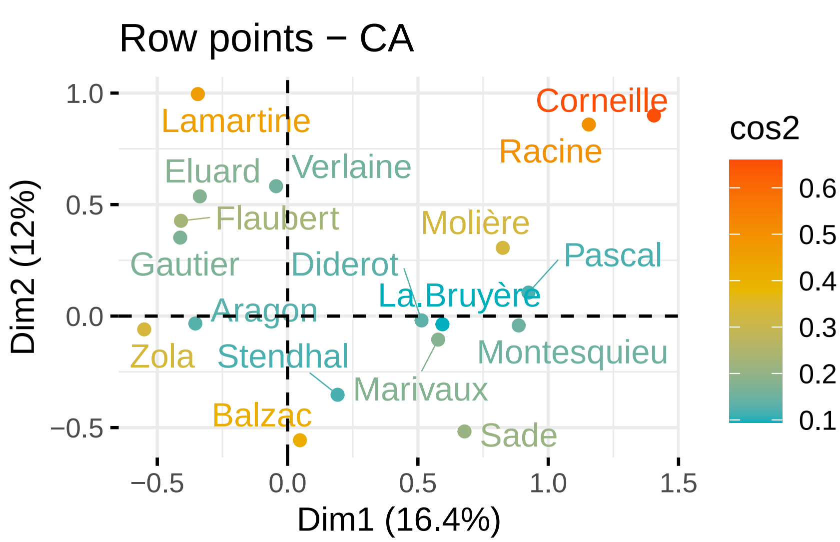

```
summary(res.ca)
```

```
##
## Call:
## CA(X = books, quanti.sup = 1, quali.sup = 2:3, graph = F)
##
## The chi square of independence between the two variables is equal to 2768153 (p-value =  0 ).
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               0.285   0.209   0.190   0.174   0.163   0.145   0.116
## % of var.             16.359  11.997  10.912   9.961   9.369   8.334   6.685
## Cumulative % of var.  16.359  28.356  39.268  49.230  58.598  66.932  73.617
##                        Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
## Variance               0.108   0.082   0.063   0.052   0.042   0.037   0.027
## % of var.              6.187   4.708   3.612   3.011   2.392   2.136   1.545
## Cumulative % of var.  79.804  84.512  88.124  91.135  93.528  95.664  97.209
##                       Dim.15  Dim.16  Dim.17
## Variance               0.023   0.019   0.007
## % of var.              1.302   1.068   0.421
## Cumulative % of var.  98.511  99.579 100.000
##
## Rows (the 10 first)
##                       Iner*1000     Dim.1     ctr    cos2     Dim.2     ctr
## Aragon              |   148.513 |  -0.354   6.620   0.127 |  -0.033   0.079
## Balzac              |   138.225 |   0.047   0.144   0.003 |  -0.557  27.062
## Corneille           |   160.564 |   1.406  26.364   0.468 |   0.900  14.724
```

```
## Diderot          |    52.930 |   0.514   2.458   0.132 |  -0.019   0.005
## Eluard           |    46.859 |  -0.337   0.822   0.050 |   0.537   2.853
## Flaubert         |    68.150 |  -0.409   2.632   0.110 |   0.427   3.914
## Gautier          |   115.002 |  -0.412   3.945   0.098 |   0.352   3.929
## La.Bruyère       |    26.605 |   0.594   0.879   0.094 |  -0.037   0.005
## Lamartine        |   153.156 |  -0.344   2.634   0.049 |   0.996  30.028
## Marivaux         |    84.755 |   0.578   5.150   0.173 |  -0.106   0.234
##                     cos2     Dim.3     ctr    cos2
## Aragon             0.001 |   0.119   1.122   0.014 |
## Balzac             0.409 |  -0.330  10.465   0.144 |
## Corneille          0.192 |  -0.661   8.731   0.103 |
## Diderot            0.000 |   0.116   0.187   0.007 |
## Eluard             0.127 |   0.214   0.500   0.020 |
## Flaubert           0.120 |   0.162   0.621   0.017 |
## Gautier            0.071 |   0.236   1.941   0.032 |
## La.Bruyère         0.000 |   0.121   0.055   0.004 |
## Lamartine          0.410 |   0.312   3.250   0.040 |
## Marivaux           0.006 |  -0.225   1.172   0.026 |
##
## Columns (the 10 first)
##                    Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## accord           |     0.913 |  0.571  0.039  0.123 |  0.349  0.020  0.046 |
## affaire          |     1.566 |  0.089  0.011  0.021 | -0.461  0.412  0.550 |
## âge              |     0.287 |  0.049  0.002  0.018 | -0.068  0.005  0.033 |
## ah               |     0.777 | -0.663  0.021  0.078 | -0.073  0.000  0.001 |
## air              |     1.387 | -0.324  0.290  0.596 | -0.078  0.023  0.035 |
## allemagne        |     1.221 | -0.434  0.017  0.039 | -0.074  0.001  0.001 |
## allemand         |     1.689 | -0.663  0.046  0.078 | -0.073  0.001  0.001 |
## amant            |     1.876 |  0.637  0.297  0.452 |  0.036  0.001  0.001 |
## âme              |     3.739 |  0.417  0.372  0.284 |  0.350  0.359  0.201 |
## ami              |     1.136 |  0.164  0.057  0.144 | -0.260  0.197  0.362 |
##                   Dim.3    ctr   cos2
## accord           -0.223  0.009  0.019 |
## affaire          -0.254  0.137  0.166 |
## âge               0.168  0.031  0.202 |
## ah                0.273  0.005  0.013 |
## air              -0.080  0.026  0.036 |
## allemagne         0.314  0.013  0.021 |
## allemand          0.273  0.012  0.013 |
## amant            -0.359  0.142  0.144 |
## âme              -0.032  0.003  0.002 |
## ami               0.046  0.007  0.011 |
##
## Supplementary continuous variable
```

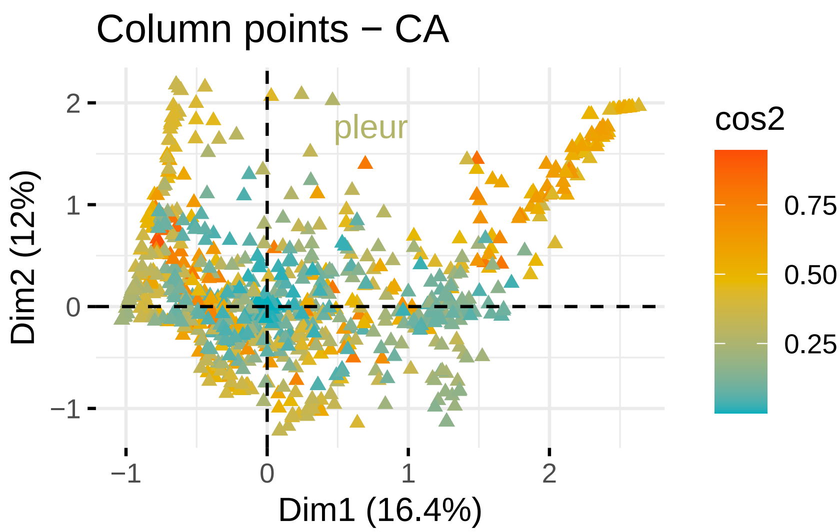
```
##                            Dim.1    cos2    Dim.2    cos2    Dim.3    cos2
## Décès                    | -0.863  0.744 | -0.125  0.016 |  0.143  0.020 |
##
## Supplementary categorical variables
##                              Dim.1      cos2   v.test      Dim.2      cos2   v.test
## Epoque.XIX             |    -0.234    0.406 -352.769 |   -0.043    0.014  -65.062 |
## Epoque.XVII            |     1.110    0.633  441.534 |    0.584    0.175  232.322 |
## Epoque.XVIII           |     0.655    0.442  345.097 |   -0.226    0.053 -119.064 |
## Epoque.XX              |    -0.352    0.148 -201.647 |    0.036    0.002   20.444 |
## Courant.Classicisme |     1.110    0.633  441.534 |    0.584    0.175  232.322 |
## Courant.Lumières       |     0.655    0.442  345.097 |   -0.226    0.053 -119.064 |
## Courant.Naturalisme |    -0.550    0.328 -308.118 |   -0.060    0.004  -33.409 |
## Courant.Réalisme       |     0.008    0.000    6.522 |   -0.361    0.364 -290.476 |
## Courant.Romantisme  |    -0.379    0.132 -184.274 |    0.666    0.408  324.154 |
## Courant.Surréalisme |    -0.352    0.148 -201.647 |    0.036    0.002   20.444 |
##                            Dim.3      cos2   v.test
## Epoque.XIX               -0.112    0.093 -169.013 |
## Epoque.XVII              -0.400    0.082 -159.092 |
## Epoque.XVIII              0.537    0.296  282.736 |
## Epoque.XX                 0.131    0.020   74.800 |
## Courant.Classicisme      -0.400    0.082 -159.092 |
## Courant.Lumières          0.537    0.296  282.736 |
## Courant.Naturalisme      -0.216    0.050 -120.744 |
## Courant.Réalisme         -0.227    0.144 -182.761 |
## Courant.Romantisme        0.273    0.069  132.923 |
## Courant.Surréalisme       0.131    0.020   74.800 |
```

```r
fviz_ca_row(res.ca, col.row = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```

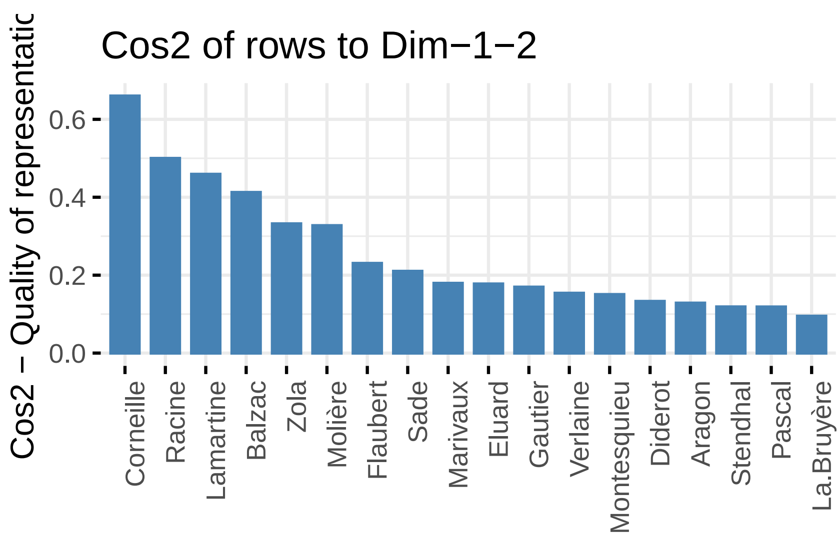


```
fviz_ca_col(res.ca, col.col = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```

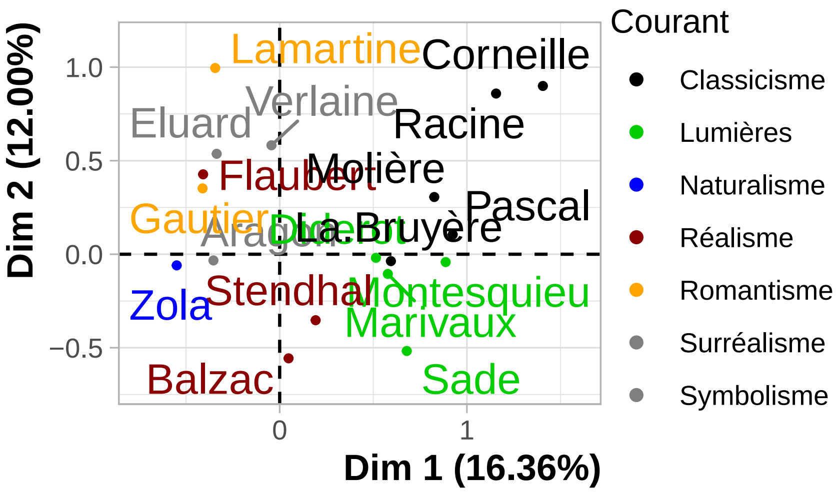
```
## Warning: ggrepel: 974 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



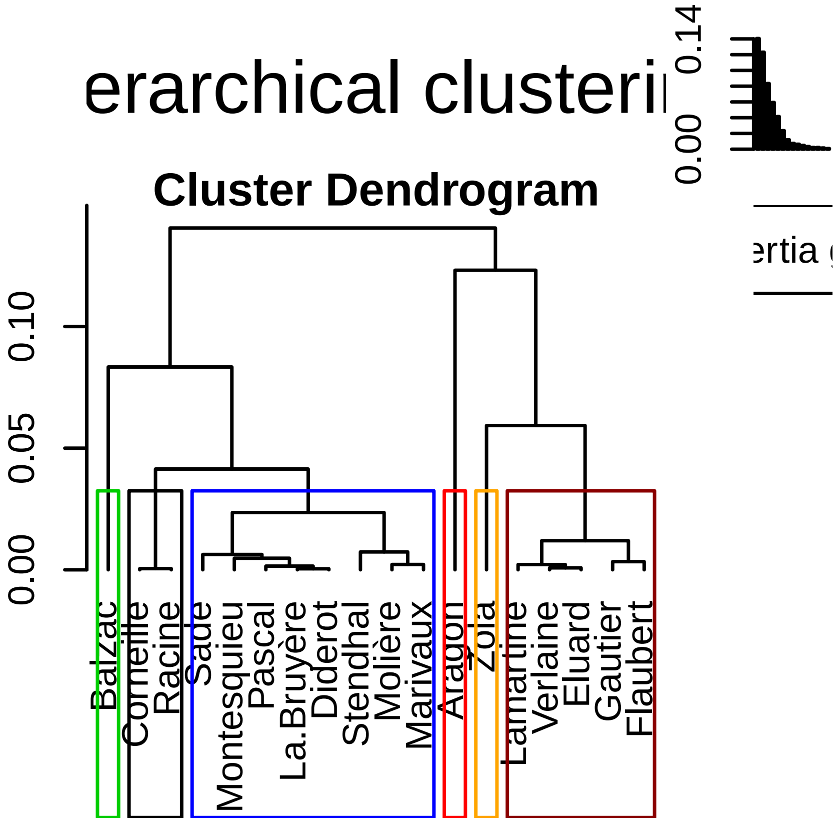

```
fviz_cos2(res.ca, choice = "row", axes = 1:2, xtickslab.rt = 90)
```




```
plot(res.ca,
     invis=c("col","quali"), hab=3,
     cex=1.2,title="",
     cex.axis=1.2, shadow=TRUE, palette=palette(c("black","darkred","orange","lightblue","blue","green3
```




```
res.ca.faster = CA(books[, 1:200], quanti.sup=1, quali.sup=2:3, graph = F)
### classif
res.hcpc = HCPC(res.ca.faster, nb.clust=-1, graph=FALSE, consol=FALSE)

plot(res.hcpc, choice="tree", palette=palette(c("black","green3","blue","darkred","orange","red","grey"

## Warning in graphics:::plotHclust(n1, merge, height, order(x$order), hang, :
## "palette" is not a graphical parameter

## Warning in graphics:::plotHclust(n1, merge, height, order(x$order), hang, :
## "palette" is not a graphical parameter

## Warning in axis(2, at = pretty(range(height)), ...): "palette" is not a
## graphical parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "palette" is not a graphical parameter
```




```
# bb$row$coord[,1]=-bb$row$coord[,1]
# bb$col$coord[,1]=-bb$col$coord[,1]
# bb$quali.sup$coord[,1]=-bb$quali.sup$coord[,1]
# plot(bb,invis=c("col","quali"), hab=4,cex=1.2,title="",cex.axis=1.2,cex.lab=1.2,palette=palette(c("ye
```

```
res.hcpc$desc.ind$para
```

```
## Cluster: 1
##    Racine Corneille
## 0.1234185 0.1234185
## ------------------------------------------------------------
## Cluster: 2
##  Balzac
##       0
## ------------------------------------------------------------
## Cluster: 3
## La.Bruyère    Diderot   Marivaux     Pascal   Stendhal
##  0.1227552  0.2079247  0.3548788  0.3965549  0.4569940
## ------------------------------------------------------------
## Cluster: 4
## Lamartine    Eluard  Flaubert   Gautier  Verlaine
## 0.2383384 0.3165550 0.3313708 0.4123533 0.4149375
```

```
## ----------------------------------------------------------
## Cluster: 5
##  Zola
##     0
## ----------------------------------------------------------
## Cluster: 6
##  Aragon
##       0
```