TP2: MCA with R

Exercice 1. The pre-processing in MCA

1. Load the dataset **dogs.rda** of the $n = 27$ dogs described on $p = 6$ categorical variables.

```
load("../data/dogs.rda")
print(data[1:5,])
```

2. Check the class of the object **data**. Check the class of the first column of **data**. Use the function **levels** to get the levels of the variable *Size*.

```
class(data)
data$Size #first columns
class(data$Size)
levels(data$Size)
```

3. Use the functions **lapply** to find the number m_j of levels of each variable j and the total number of levels $\ell = \sum_{j=1}^{p} \ell_j$.

```
lj <- unlist(lapply(data,function(x){length(levels(x))}))
l <- sum(lj)
```

4. Build the matrix K of the disjonctive table using the function **tab.disjonctif** of the R package **FactoMineR**.

```
library(FactoMineR)
K <- tab.disjonctif(data)
print(K[1:4,])
```

5. Compute the frequencies n_s and the relative frequencies $\frac{n_s}{n}$ of the levels ?

```
ns <- apply(K,2,sum)
print(ns)
```

```
n <- nrow(K)
fs <- ns/n
print(fs)
```

6. Build the matrix Z of the centered disjonctive table.
7. Perform the variance of the columns of the disjonctive table with the function **var** and then from the formula $\frac{n_s}{n}(1 - \frac{n_s}{n})$.
8. Perform the distance between the two breeds *Pekingese* and *Doberman* descibed in disjonctive table using the metric $M = diag(\frac{n}{n_s})$. Check that the result is the same when using the centered disjonctive table.
9. Perform $I(K)$, the total inertia of the 27 dogs descrived in the disjonctive table.

Exercice 2. The GSVD of Z.

The GSVD of a real matrix Z of dimension $n \times p$ with metrics N on $\mathbb{R}^n$ and M on $\mathbb{R}^p$ gives the following decomposition:

$$Z = U\Lambda V^t,$$

where

- $\Lambda = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ is the $r \times r$ diagonal matrix of the singular values of ZMZ^tN and Z^tNZM, and r denotes the rank of Z;
- U is the $n \times r$ matrix of the first r eigenvectors of ZMZ^tN such that $U^tNU = \mathbb{I}_r$, with $\mathbb{I}_r$ the identity matrix of size r;
- V is the $p \times r$ matrix of the first r eigenvectors of Z^tNZM such that $V^tMV = \mathbb{I}_r$.

The idea is to perform the GSVD of the centered disjonctive table Z of the dogs data with metrics $N = \frac{1}{n}\mathbb{I}_n$ and $M = diag(\frac{n}{n_s}, s = 1, \ldots, \ell)$ used in MCA.

1. Build with the two metrics M and N using the function **diag**.
2. The GSVD of Z can be obtained by performing the standard SVD of the matrix $\tilde{Z} = N^{1/2}ZM^{1/2}$, that is a GSVD with metrics $\mathbb{I}_n$ on $\mathbb{R}^n$ and $\mathbb{I}_p$ on $\mathbb{R}^p$. It gives:

$$\tilde{Z} = \tilde{U}\tilde{\Lambda}\tilde{V}^t$$

   and transformation back to the original scale gives:

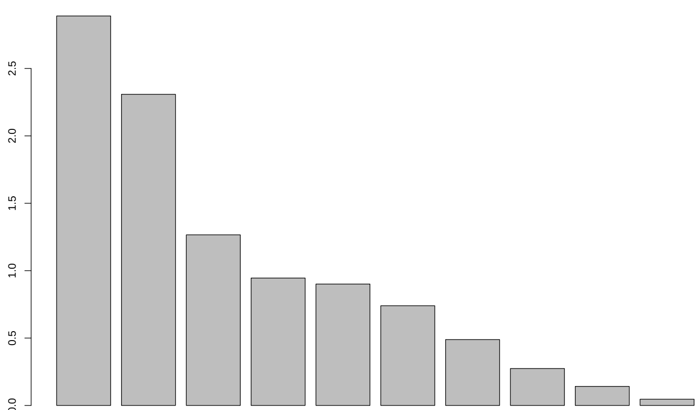
$$\Lambda = \tilde{\Lambda} \quad , \quad U = N^{-1/2}\tilde{U} \quad , \quad V = M^{-1/2}\tilde{V} \quad .$$

   This procedure has been implemented in a function **gsvd** avalaible in the file **gsvd.R**. Open this file and read the description of the function and its R code.
3. Perform the GSVD of the centered disjonctive table Z with the metrics M and N.
4. Check that the rank of the centered disjonctive table is is r=min$(n-1, \ell-p)$. Check using %*% (matrix product in R) that the matrix U is N-orthonormal and that the matrix V is M-orthonormal.

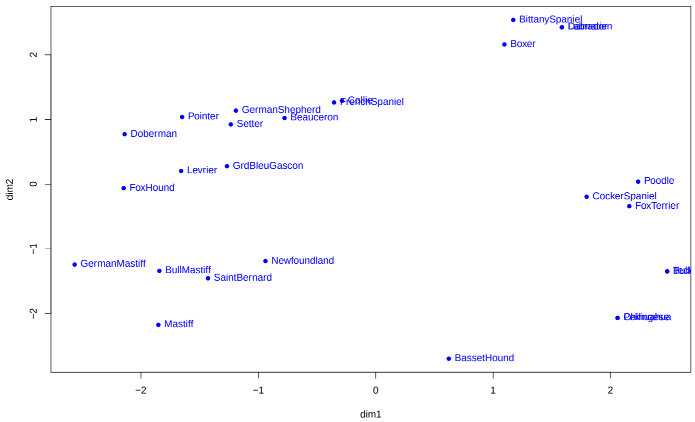
Exercice 3. GSVD and MCA.

We want to perform MCA using the GSVD of the disjonctive table performed in the previous exercice.

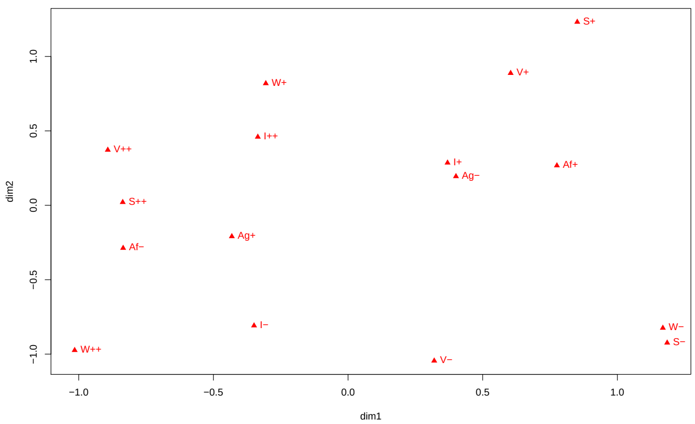
1. Build the matrix F of dimension $n \times r$ of the factor coordinates of the dogs.
2. Build the matrix A of dimension $m \times r$ of the factor coordinates of the levels.
3. Perform the variance of the columns of F and check that you get the eigenvalues of the GSVD.
4. Check that the sum of all the eigenvalues is equal to the total inertia.
5. Plot the eigenvalues with the function **barplot**. How many dimension $q \leq r$ would you keep here ?



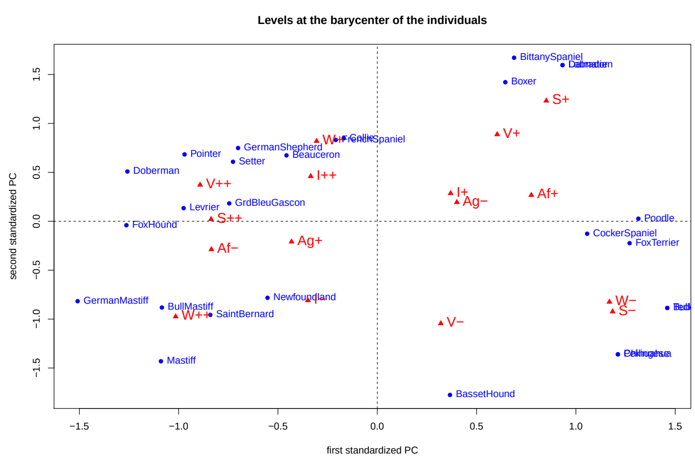
5. Perform the proportion of inertia explained by the q first principal components.
6. Plot of the dogs according to their factor coordinates on dim1-2.



7. Plot of the levels according to their factor coordinates on dim1-2.



8. Check the barycentric property for the level S-.
9. Plot the levels at the barycenter of the dogs on dim1-2.

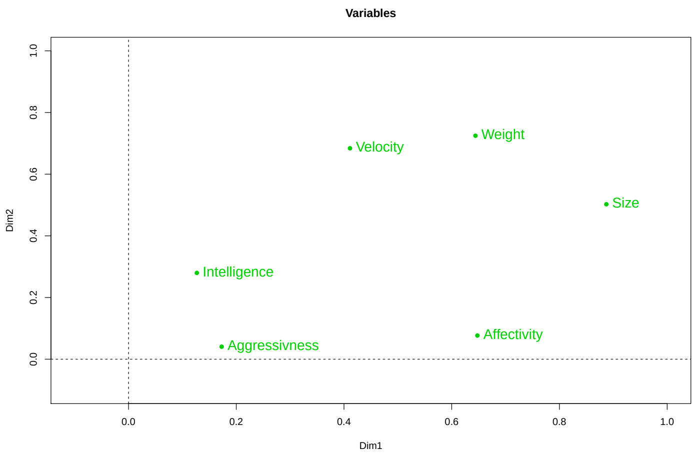


10. Perform the matrix C of dimension $p \times 2$ of the contributions of the categorical variables to the inertia of the two first principal components.

```
eta2 <- function(x, gpe) {
  moyennes <- tapply(x, gpe, mean)
  effectifs <- tapply(x, gpe, length)
  varinter <- (sum(effectifs * (moyennes - mean(x))^2))
  vartot <- (var(x) * (length(x) - 1))
```

```
  res <- varinter/vartot
  return(res)
}
```

11. Plot the variables according to their contributions to the two first principal components.



Exercice 3. MCA with R functions.

We want now to perform MCA using the functions **MCA** of the R package **FactoMineR** and **PCAmix** of the R package **PCAmixdata**.

1. Apply the function **MCA** to the dogs dataset. Explain and comment the three graphical output obtained by default.
2. Put the result in an object **res**. What is the class of this R object ? Two functions (methods) are associated with this class of R objects : **plot.MCA** and **print.MCA**. Check that is is equivalent to execute:

    a. **res** or **print.MCA(res)**
    b. **plot(res)** or **plot.MCA(res)**

3. Find in the object **res**:

    a. the numerical results used to build the previous 3 graphical representations.
    b. the numerical results used to interpret these graphics.

4. With the method **plot** associated with the objects of class **MCA**, plot on the map 1-2 the dogs, then the levels, then the levels and the dogs on the same map, then the variables.
5. Compare the factor coordinates obtained via the GSVD (in the exercice 3) and via the function **MCA** of **FactoMineR**. More precisely:

    a. Check that the factor coordinates of the levels and variables are identical.
    b. Check that the factor coordinates of the individuals are identical up to a multiplicative constant (to be defined).
    c. Check the consequence on the inertia of the principal components and on the total ineria.

6. Apply now the function **PCAmix** of the R package **PCAmixdata**. Answer the same questions as previously with the function **MCA** of the package **FactoMineR**.

Exercice 4. PCA of a mixture of quantitative and qualitative data.

We want now to use a method called **PCAmix** wich performs a Principal Component Analysis of a mixture of numerical and categorical data. This function is implemented in the R package **PCAmixdata**.

1. First check that the function **PCAmix** performs a **PCA** if all the data are numerical and an **MCA** if all the data are categorical. Use the examples provided in the help of the function.
2. Use the vignette of the package to see the main possibilities of the function **PCAmix** (prediction and supplementary variables for instance).
3. Use the vignette to discover the possibilities of the functions **PCArot** and **MFAmix** of the package.