Principal Component Analysis

`FactoMineR` Package may be useful for the follwing exercises.

Exercise 1 Coding PCA

1. Code your own PCA. From a data frame X, your function will produce
   - the principal components (for representing the individuals) ,
   - the rotation matrix (definition of the new variables).
2. Apply you function to the following data

```
X<-read.table(text="
               math   scie   fran   lati    d-m
  jean          6.0    6.0    5.0    5.5    8.0
  aline         8.0    8.0    8.0    8.0    9.0
  annie         6.0    7.0   11.0    9.5   11.0
  monique      14.5   14.5   15.5   15.0    8.0
  didier       14.0   14.0   12.0   12.5   10.0
  andré        11.0   10.0    5.5    7.0   13.0
  pierre        5.5    7.0   14.0   11.5   10.0
  brigitte     13.0   12.5    8.5    9.5   12.0
  evelyne       9.0    9.5   12.5   12.0   18.0
")
```

   and compare to output with the `prcomp` R function.

Exercise 2 PCA and size effect

The dataset considered consist of 200 crabs described by eight variables (3 qualitative and 5 quantitative). Load the dataset and select the quantitative variables using the following R code:

```
> library(MASS)
> data(crabs)
> crabsquant<-crabs[,4:8]
```

The purpose of this study is to use PCA to find a representation of crabs that make it possible to visually distinguish different groups, related to species and sex.

1. Test a PCA `crabsquant` without any preliminary transformation. What do you observe.
2. Find a solution to improve the quality of your representation for representing the four different groups.
3. What about the quality of representation of this new PCA? How many axes do you keep ? What for ?
4. How do you interpret the axes selected from the circle of Correlations?
5. what can you deduce concerning the characterization of Male/female, orange/blue crabs?

Exercise 3 Phylogeny of Globins

Let us carry out a factorial analysis of the dissimilarities between Protein sequence of several globins from different Species and compare the results obtained to the phylogenetic tree of Figure 1.

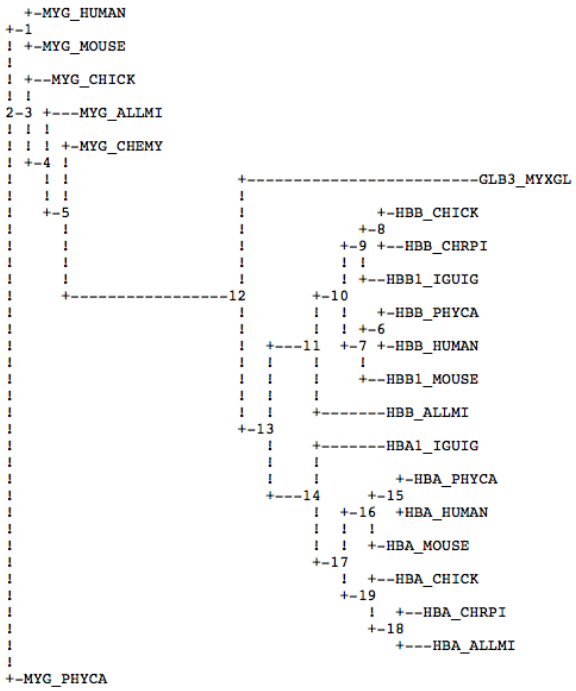

Figure 1: Phylogeny of Globins

1. Download the file `neighbor_globin.txt` and import the data into `R` in a `data.frame` d that contains pairs alignment scores of Various globins in different species as described in the File `Globines_liste.txt`.
2. Check that these scores correspond well to dissimilarities. Name the columns.
3. Compute the matrix Δ of squared dissimilarities.
4. Compute the centering matrix J defined by:
$$J = I - \frac{1}{n} 1_{(n,n)}$$
5. Compute $B = -\frac{1}{2} J \Delta J$. How to interpret B ?
6. Perfom the spectral decomposition of B:
$$B = U \Lambda U^T$$
7. What are the main factors in this decomposition? How much do you keep for the rest of the analysis? What do you also conclude from the observation of eigenvalues?
8. Calculate the main components associated with the main axes selected and represent the corresponding main plans by differentiating the types of globins by type of point (1 = myoglobin, 2 = hemoglobin β, 3 = hemoglobin α, 4 = Globin-3) and species per color (the same color will be set for both turtle species).
9. What do you notice by comparing these different planes to the phylogenetic tree?
10. Perform the same analysis on the subsets of Hemoglobins α, β then myoglobins, always comparing the results to the phylogenetic tree. To save time, you can help with the `cmdscale` function.