PCA and Clustering

1 ACP

1. **The data** Save the data olympic which are associated to the performance of people.
2. **Preprocessing** Let denote by `n` the number of individuals.
   (a) Look at each variable independently.
   (b) Compute the empirical mean of each variable and determine the centered data and save them in a matrix named `Oly_centre`.
   (c) Compute the standard deviation of each variable of the table `Oly_centre`. Determine the matrix of the normalized performance and save them in the matrix `Oly_renorm`.
3. **PAC : Representation of the individuals**
   (a) look at the help page of `princomp`.
   Write `acp_olympic=princomp(Oly_renorm,scores=TRUE)`.
   (b) Compare `(1/n)*(t(Oly_renorm)%*%Oly_renorm)` and `(n-1)/n*cov(Oly_renorm)`. What is done?
   (c) What is the output of `summary(acp_olympic)`? See this by computations.
   (d) What are the outputs of `acp_olympic$loadings` ? See this by computations.
   (e) What are the outputs of `acp_olympic$scores` ? See this by computations.
   (f) Plot the individuals in the firts factorial plan.

4. **PCA : Representation of the variables**

   (a) Write `cor(Oly_renorm[,1],acp_olympic$scores)`. What is the norm of this vector ? What is it ?

   (b) Compare `cor(Oly_renorm,acp_olympic$scores[,1])` and `acp_olympic$sdev[1]*acp_olympic$loadings[,1]`. What is done?

   (c) Deduce the correlation circle.

5. **Generally**

   (a) Compare with `biplot` apply to `acp_olympic`.

   (b) What produces `plot(acp_olympic)` ? What can be the use ?

   (c) What happens if the data are not normalized?

2 clustering

1. Write `data(iris)`. Describe the data.

2. Create a matrix `A` where does not appear any more the variable species.

3. Write

   ```
   K=kmeans(A,3,iter.max=1,nstart=1)
   ```

   What is the associated method ?

4. What are the outputs and confirm this by computation.

5. What are the parameters `iter.max` and `nstart`.

6. Is the application of the function correct?

7. Compare the outputs with the reality.

8. Change the number of groups and compare the variances.

9. Write

   ```
   D=dist(A)
   Db=dist(A,method="maximum")
   ```

   What happens?

10. What produces the function `hclust`?

    (a) Write `hc1=hclust(D^2,method='ward.D')`. What does `plot(hc1)` ? What are the informations in `merge` and `height` of `hc` ?

    (b) Write `hc2=hclust(D,method='ward.D2')`. Same things than before.

    (c) Change the distance.

11. How many groups to perform. Combine both methods.

12. **Comparison with the PCA.** Perform a PCA, plot the individuals on the first factorial plan with color points with respect to the value of the species.

    (a) Use the function `princomp` on `A`. Explain the results.

    (b) Do the plot with three colors.