Data mining and multivariate statistical analysis 2017

2nd practical

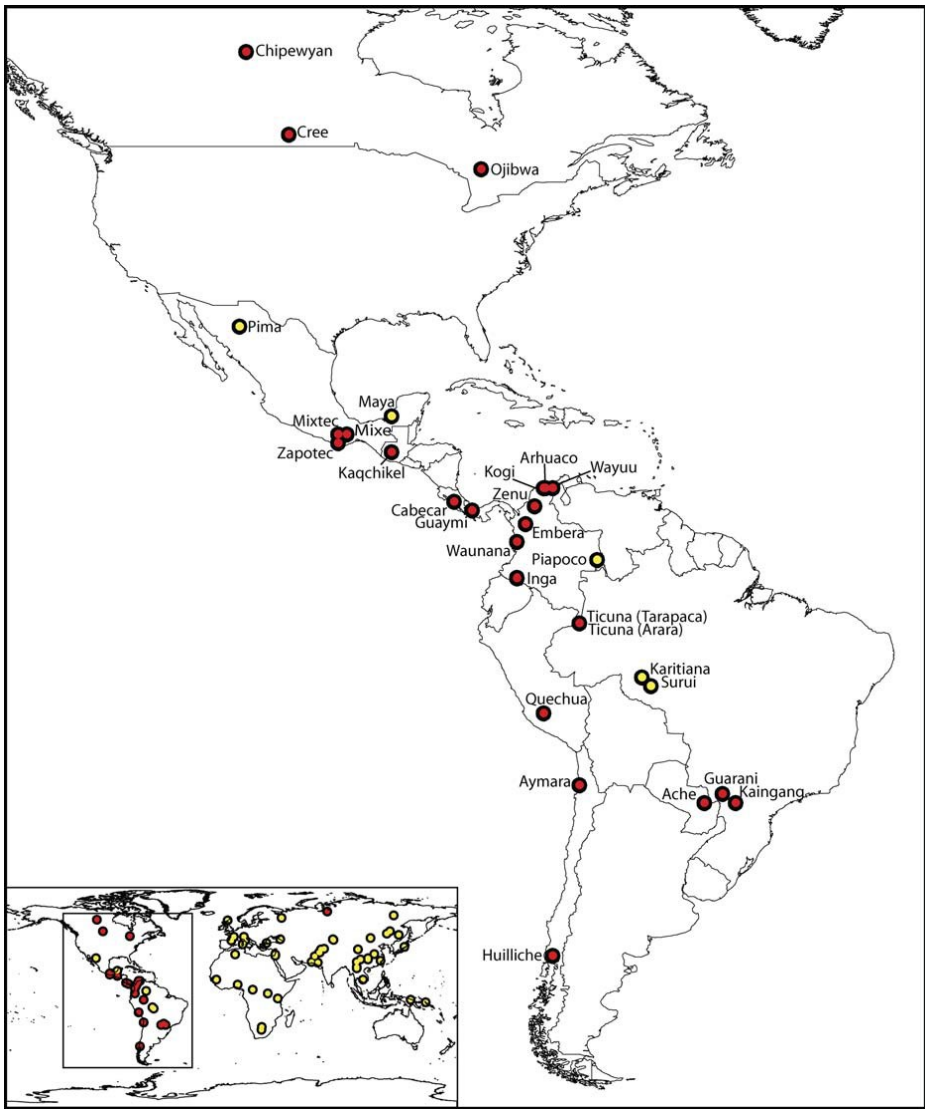
A report must be uploaded on TEIDE before the deadline.
The report should contain graphical representations, which are very important in statistics. For each graph, axis names should be provided as well as a legend when it is appropriate. Figures should be explained by a few sentences in the text. Answer to the questions in order and refer to the question number in your report.

Computations and graphics should be performed with the software R.
The report should be written using the Rmarkdown format, as for the first Lab work.

2nd Practical: PCA-regression in genetics

The goal of this practical is to use genetic markers to predict the geographical origin of an individual. Individuals are Indians from America. We propose to build predictive linear models to predict latitude and longitude of an individual from its genetic markers. Because the number of markers (p=5709) is larger than the number of samples (n=494), the explicative variables will be the outputs of PCA performed on genetic markers. A genetic marker is encoded 1 if the individual has the mutation, 0 elsewhere.

Fig. 1



1. **Data :**

Take the dataset NAm2.txt from chamilo
chamilo2.grenet.fr/inp/courses/ENSIMAG4MMFDAS6/ and load it using
`NAM2 = read.table("NAm2.txt", header=TRUE).`
Each row is an individual. Check up that columns have explicit names.
The third column contains the population of origin for that individual. Columns 7 and 8 contain the latitude and the longitude. Each columnn from 9th is a genetic marker.

Describe what this script does and how it works (`help(unique)`) ?
Check up that you obtain the same map as in Fig. 1.

```
names=unique(NAm2$Pop)
npop=length(names)
coord=unique(NAm2[,c("Pop","long","lat")]) #coordinates for each pop

colPalette=rep(c("black","red","cyan","orange","brown","blue","pink","purple","dar
kgreen"),3)
pch=rep(c(16,15,25),each=9)

plot(coord[,c("long","lat")],pch=pch,col=colPalette,asp=1)
# asp allows to have the correct ratio between  axis  longitude and latitude
# Then the map is not deformed
legend("bottomleft",legend=names,col=colPalette,lty=-
1,pch=pch,cex=.75,ncol=2,lwd=2)
library(maps);map("world",add=T)
```

Remark: The last line works in the Ensimag's rooms because the package `maps` has been installed.
To install this package on your own computer: `install.packages("maps")`

2. **Regression :**

   Using all genetic markers as predictors, predict the longitude using a linear regression model. You may need to create a new data.frame, keeping the relevant variables only: `NAaux = NAm2[,-c(1:7)]`
   What happens? (you can use `sink` to show all error messages sent by `R` or consider a subsample).

3. **PCA :**

   a) Explain quickly the principle of PCA.

   b) Perform PCA on genetic data (**and only on these ones)** with all samples. Keep on the result in the object `pcaNAm2`. Do we need to use the argument `scale` of `prcomp`?

   c) This script plots the populations on the first 2 PCA axes.

```
caxes=c(1,2)
plot(pcaNAm2$x[,caxes],col="white")
for (i in 1:npop) {
       print(names[i])
       lines(pcaNAm2$x[which(NAm2[,3]==names[i]),caxes],
              type="p",col=colPalette[i],pch=pch[i])
}
legend("top",legend=names,col=colPalette,lty=-
1,pch=pch,cex=.75,ncol=3,lwd=2)
```

   Describe and interpret the obtained graphs. Which populations are easly identified using the first 2 axes? Answer to the same question using the 5th and 6th axes.

   d) Which percentage of variance (inertia) is captured by the first 2 principal components? How many principal components would you keep if you would like to represent genetic markers using a minimal number of principal components?

4. **PCR Principal Components Regression :**

   a) Predict the latitude and the longitude using the scores of the first 250 PCA axes. Let denote the results of these regressions by `lmlat` et `lmlong`.

   Plot the graph of predicted spatial coordinates:

```
plot(lmlong$fitted.values,lmlat$fitted.values,col="white")
for (i in 1:npop) {
       print(names[i])
       lines(lmlong$fitted.values[which(NAm2[,3]==names[i])],lmlat$fitted.va
       lues[which(NAm2[,3]==names[i])],type="p",col=colPalette[i],pch=pch[i]
       )
          }
legend("bottomleft",legend=names,col=colPalette,lty=-
1,pch=pch,cex=.75,ncol=3,lwd=2)
map("world",add=T)
```

Compare with the map of question 1. What can you see? Does this map illustrate too optimistically or too pessimistically the ability to find geographical origin of individuals outside the database from its genetic markers?

b) We choose to quantify the error using the mean distance between real and predicted coordinates (of origin populations). Be careful, use the orthodromic distance, « great circle distance»).
Calculate the mean error of the previous model built using 250 axes.
Help : `??rdist.earth` (using option `miles=F)`. This function is in the package `fields`
`library("fields")`

5. PCR and cross-validation:

Our goal is to build the best predictive model to predict individual geographical coordinates. To choose the number of axes we will keep (`naxes`), we will apply the 10-folds cross validation method.

a) Recall quickly the principle of cross validation method. Explain why this method is interesting to build a predictive model.
The dataset has to be divided into ten subsets, which will be used in turns as validation sets.
Create a vector `set` that contains, for each individual, the index of the subset he/she belongs to.

Example for 9 individuals eand a e 3-fold cross validation :
`set = c(1,2,3,1,2,3,1,2,3)` or s`et=c(1,3,1,3,3,2,1,2,2)`
You can randomly build this vector, with the same number of individuals in each validation set :
```
labels=rep(1:3,each=3)
set=sample(labels,9)
```

b) We will study the models using `naxes` from 2 to the maximum value (10 by 10 for example). Firstly, we focus to the case `naxes=4`.

i. Create an empty matrix predictedCoord with 2 columns ("longitude", "latitude") and as many rows as there are individuals.

ii. Using as predictors the scores of the first 4 PCA axes, explain latitude and longitude using the individuals who do not belong to validation set n. 1.
Help : you can introduce the following table
`pcalong=data.frame(cbind(long=NAm2[,c("long")],pcaNAm2$x))`
and use the argument `subset` when you perform the regression.

iii. Using the built model, predict latitude and longitude for individuals belonging to the validation set n. 1. Store the predicted coordinates into `predictCoord` (in rows corresponding to the individual indices, in order to be able to compare real and predicted coordinates). Be careful, the function `predict` needs a `data.frame` of input points to predict.

iv. Repeat for all other validation sets. In the end, the matrix `predictCoord` must be full. Calculate the prediction error (cf. 4.c).

c) Repeat all steps of b), changing `naxes` from 2 to 440. Plot the prediction errors and the error obtained on the training set versus the number of components.
Help : `seq(2,440,by=10)`

d) Which model would you keep? What is the prediction error for this model? Compare it with the training error. (cf 4.c) Plot the predicted coordinates on a map (cf. 4.b).

6. Conclusion :

Propose a conclusion to the study. You could write some paragraph about the quality of predictors, versus the number of factors, possible improvements to the approach , … We expect some thorough presentation of the final model and interpretation, not exclusively R code lines.