Ensimag – Grenoble INP – UGA
Statistical Analysis and Document Mining
Pedro L. C. Rodrigues
Jacopo Iollo

Year 2022-2023

pedro.rodrigues@inria.fr
jacopo.iollo@inria.fr

⚠ General guidelines for TPs

Each team shall upload its report on Teide before the deadline indicated at the course website. Please **include the name of all members of the team** on top of your report.

The report should contain graphical representations. 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 have to be performed with `R`.

The report should be written using the `Rmarkdown` format. This is a file format that allows users to format documents containing text and `R` instructions. You should include all of the `R` instructions that you have used in the `rmd` document so that it may be possible to replicate your results. From your `rmd` file, you are asked to generate an `html` file for the final report. In Teide, you are asked to submit both the `rmd` and the `html` files. In the `html` file, you should limit the displayed `R` code to the most important instructions.

TP 2: Principal components regression in genetics

The goal of this practical session is to use genetic markers to predict the geographical origin of a set of indians from South, Central, and North America. We propose to build two regression linear models to predict the latitude and longitude of an individual based on its genetic markers. Because the number of markers ($p = 5709$) is larger than the number of samples ($N = 494$), the predictors of the regression model will be the outputs of a principal component analysis (PCA) performed on the genetic markers. A genetic marker is encoded 1 if the individual has a mutation, 0 elsewhere.

▶ Exercise 1: Data

Download the dataset `NAm2.txt` from Chamilo. Each row corresponds to an individual and the columns have explicit names. The third column contains the names of the tribes to which each individual pertains. Columns 7 and 8 contain the latitude and the longitude and from Column 9 onwards are genetic markers.

Describe what the code below does and how it works (you can take a look at `help(unique)`). You should get the same figure as the one shown below.

```
NAm2 = read.table("NAm2.txt", header=TRUE)
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","darkgreen"),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, thus the map is not deformed
legend("bottomleft",legend=names,col=colPalette,lty=-1,
       pch=pch,cex=0.75,ncol=2,lwd=2)
library(maps); map("world",add=T)
```

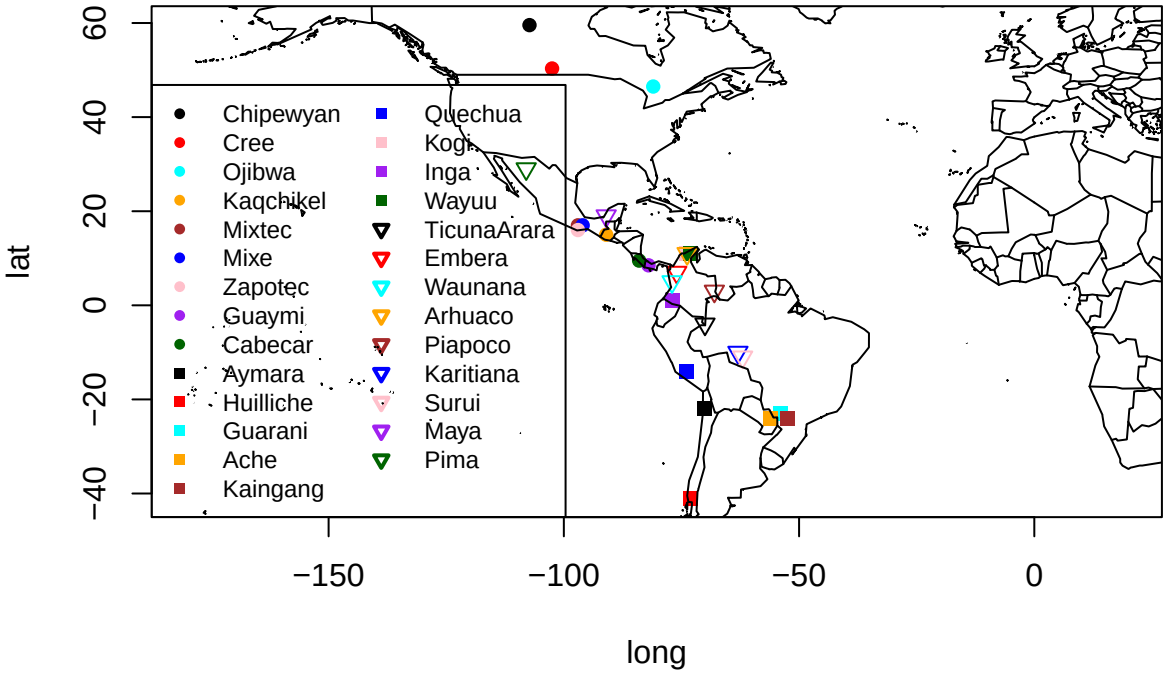

Figure 1: Populations of indians from America

Remark: The last line of the code above works in the ENSIMAG computers because the package `maps` has been installed beforehand. To install it in your own computer you should use `install.packages("maps")`.

▶ Exercise 2: Multiple linear regression

Using all genetic markers as predictors, use a multiple linear regression model to predict the longitude of each individual. You may need to create a new `data.frame` containing the relevant variables, i.e.: `NAaux = NAm2[,-c(1:7)]`. What happens? Why?

Remark: You should relate the results with the fact that $\text{rank}(\mathbf{X}) < p$, where $\mathbf{X} \in \mathbb{R}^{N \times p}$ is the data matrix.

▶ Exercise 3: Principal component analysis

(a) Explain in a few words the main concepts and ideas underlying principal component analysis (PCA).

(b) Use command `prcomp` to apply PCA on the data matrix $\mathbf{X}$. Remember that the features of interest in $\mathbf{X}$ are only the genetic markers for each individual. Store the results in a object called `pcaNAm2`. Should we use the argument `scale` in `prcomp`?

(c) The code below plots the populations on the first two principal axes of PCA. Interpret and compare the results of the PCA using `scale=T` versus `scale=F` and explain why they are different.

```
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=0.75,ncol=3,lwd=2)
}
```

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

of principal components?

▶ Exercise 4: Principal components regression (PCR)

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

Plot the graph of predicted spatial coordinates using the code:

```
plot(lmlong$fitted.values,lmlat$fitted.values,col="white",asp=1)
for (i in 1:npop)
{
    print(names[i])
    lines(lmlong$fitted.values[which(NAm2[,3]==names[i])],
        lmlat$fitted.values[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 your results with the map of Figure 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 of the linear regression model using the mean distance between real and predicted coordinates (of source populations). Be careful, use the orthodromic distance, ("great circle distance' '). Calculate the mean error of the previous model built using (the first) 250 principal axes.

Remark: Look at `?rdist.earth` for a function for that calculates the orthodromic distance. Consider using option (`miles = F`). Note that this function is in package `fields` that you should load using `library("fields")`.

▶ Exercise 5: PCR and cross-validation

Our goal now is to build the best predictive model to predict individual geographical coordinates. We will use 10-fold cross-validation to helps us choose the number (`naxes`) of principal axes that we should keep.

(a) Recall in a few words the principle of cross-validation. Explain why this procedure is useful when building a predictive model. We will divide the dataset 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 to which he/she belongs. You can randomly build this vector, with the same number of individuals in each validation set.

(b) We first assess the quality of the PCR fit for `naxes=4`. For this, you should proceed as follows:

1. Create an empty dataframe `predictedCoord` with 2 columns `("longitude", "latitude")` and as many rows as there are individuals.
2. Using as predictors the scores of the first 4 PCA axes, explain `latitude` and `longitude` using the individuals who do not belong to the validation set number 1.
3. Using the estimated model, predict `latitude` and `longitude` for individuals belonging to the validation set number 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 and they should be different from those used to fit the model.
4. Repeat for all the other validation sets. At the end, the matrix `predictCoord` must be full. Calculate the prediction error as in Exercise 4(b).

(c) Repeat the steps of 5(b) but changing `naxes` between 2 and 440 in steps of 10. Plot the prediction errors and the error obtained on the training set versus the number of principal components. **Remark**: You might need to use `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. Plot the predicted coordinates on a map as in Exercise 4.

▶ Exercise 6: Conclusion

Propose a conclusion to the study. You can write a paragraph about the quality of predictors versus the number of factors, possible improvements to the approach, etc. Note that we expect a thorough presentation of the final predictive model as well as an interpretation of it, not simply a bunch of `R` code lines.