Multidimensional Statistical Analysis -
Optional FDASM TP - 2017

Course supervisors: Michael.Blum@imag.fr
Jean-Baptiste.Durand@imag.fr

TP supervisors: Cecile.Amblard@imag.fr
li.liu@gipsa-lab@grenoble-inp.fr

Calculations, graphics and programming will be done with software R.

Create a TP3 directory where you will load the data and open the program R.

ASM-TP Classification supervised by linear discriminant analysis

The objective of this TP is to use the TP2 genetic data to show how one can predict a continent's membership (North America, South America or Central America) from genetic data. This classification will be done with 2 different supervised classification techniques.

Retrieve the TP2 data available on the Kiosk intranet.ensimag.fr/KIOSK/Matieres/4MMFDASM/ and load the object with read.table ("NAm2.txt", header = T), call the NAm2. Each line corresponds to an individual. Verify that the columns have explicit names. Column 3 contains the original population of the individual. Each column from the 9th corresponds to a genetic marker.

To create a vector containing the continent of origin of each individual, run the following code

```
cont<-function(x)
  { if (x %in% c("Canada"))
    cont<-"NorthAmerica"
    else if (x %in% c("Guatemala","Mexico","Panama","CostaRica"))
    cont<-"CentralAmerica"
    else
    cont<-"SouthAmerica"
    return (factor(cont))
  }
contID<-sapply(as.character(NAm2[,4]),FUN=cont)
```

Discriminant linear analysis. LDA
To do this TP, the MASS and class packages must be loaded.

```
library("class")
library("MASS")
```

a) Perform a discriminant linear analysis using half the individuals as labels (labels <-rep (1: 2, each = 494/2); set = sample (labels, 494)).
Use the lda function with the subset argument (Boolean list). To learn the classification model, we can create the following table which contains the continent of origin in the first column followed by the genetic markers.

```
NAcont<-cbind(contID=contID,NAm2[,-(1:8)])
NAcont[,1]<-factor(NAcont[,1])
```

Find out that lda does not work. To better see the error message you can restrict yourself to the first 1000 variables.
Explain why lda does not work (consider reversing the intra_group covariance matrix).

b) To run lda, all genetic markers with a variance of 0 in at least one of the continents will be removed.

```
withinvar<-apply(NAcont[(set==1),-
1],FUN=function(x){tapply(x,NAcont[(set==1),1],FUN=var)},MARGIN=2)
bool<-as.logical(apply(withinvar,FUN=function(x){prod(x!=0)},MARGIN=2))
NAcont2<-cbind(contID=contID,(NAm2[,-(1:8)])[,bool])
NAcont2[,1]<-factor(NAcont[,1])
```

You can ignore the warnings.

Count the number of markers remaining and comment. Learn the classification model.

c) Now predict the original populations of individuals who have not been trained.
(`predict` function, `class` attribute)

d) Using the `table` function, you can display the confusion matrix for these individuals (cf poly).
`table(real pop, pop predicted)`. Represent it as a barplot and comment.

e) From the confusion matrix, calculate the rate of misclassified individuals (called classification error or prediction error).

f) Propose and implement an improvement to the method implemented in the previous question to validate this classification model.