Non-supervised methods

TASK 1 - Multidimensional reduction: File `nhanes.Rdata` contains two tables (`nhanes.nut` and `nhanes.air`) including variables about nutrients and air pollution obtained from NHANES project. The file also contains two objects describing the column names of those tables (`nut.desc` and `air.desc`, respectively). These tables can be loaded into R by executing

```
load(‘‘data_exercises/nhanes.Rdata’’)
```

1. Perform a principal component analysis of nutrient variables (columns 1:29 of table `nhanes.nut`) using variable as the grouping variable and determine those variables that are associated with each category (`normal` and `hypercol`). Use the default method of the `ord` function and do not forget to scale the data.

TASK 2 - Multidimensional reduction: File `nci60.Rdata` contains miRNA, mRNA and protein data of melanoma, leukemia and CNS disease. Data are encapsulated in a list where each components stands for a given omic data (NOTE: features are in rows and samples in columns). Data corresponds to cells lines from the NCI-60 panel available at TCGA project. 21 cell lines are providing information about 537 miRNAs, 12,895 gene expression and 7,016 proteins. We are interested in obtaining omic profiles to characterize those diseases. NOTE: The vector `cancer` is a factor variable indicating the type of cancer of each sample.

1. Load data into R and select miRNA table by executing

```
load(’’data_exercises/nci60.Rdata’’)
miRNA <- nci60$miRNA
```

2. Perform a PCA using of miRNA dataset and give the top-5 features associated with each tumor.
3. How much variability is explained by the first two axes?
4. Determine how many axes are necessary to be selected to properly reduce the dimensionallity of this data.

TASK 3 - Hierarchical analysis: Load data called `diet.Rdata` by executing

```
load(‘‘data_exercises/diet.Rdata’’)
```

There are three databases calles X1, X2 and X3. X1 contains outcome and confounding variables. X2 contains nutriens and X3 contains food consumption. We aim to identify cluster of individuals given nutrients and food consumption variables. Once clusters are identify we will assess whether these clusters are associated with any colorectal cancer. After that, we will perform an analysis to provide an interpretation of clusters with regard to our variables. Let us do these steps

1. Create a table having nutrients and food variables by executing

   ```
   X <- cbind(X2, X3)
   ```

2. Perform a hierarchical clustering of X and select the optimal number of clusters. How many sample are in each cluster?

3. Assess association between this variable cluster and the variable `cascoc` of table X1. NOTE: use `glm` function.

4. Perform a PCA and use the cluster variable as a grouping variable to illustrate whether samples are separated with regard to it.