Multi-tables data analysis

TASK 1 - Multiple tables epidemiological data: Load data called `diet.Rdata` by executing

`load(''data_exercises/diet.Rdata'')`

There are three databases called X1, X2 and X3. X1 contains outcome and confounding variables. X2 contains nutriens and X3 contains food consumption. We aim to identify which table (nutrients or food consumption) can be used to explain the maximum varability of our data.

1. Create a list X to be passed through `mcia` function. As there are missing data, and we are interested in analyzing complete cases, we should execute

```
sel <- complete.cases(X2) & complete.cases(X3)
X2.comp <- X2[sel,]
X3.comp <- X3[sel,]
X <- list(nut=t(X2), food=t(X3))
```

2. Use multiple coinertia analysis (`mcia`) to analyze this data an represent the top-2 variables associated to each axis

3. Perform PCA as we did in the previous tasks (that is, considering an unique table) by executing:

```
XX <- cbind(X2, X3)[sel,]
mm2 <- ord(t(XX))
plotgenes(mm2, nlab = 2)
```

4. What is the IMPORTANT TAKE HOME MESSAGE we can draw from these analyses?

TASK 2 - Multiple tables omic data: 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 by executing `load(''/data_exercises/nci60.Rdata'')`.

2. Perform multi coinertia analysis and provide the top-5 features (they can be proteins, miRNA or mRNA) associated with each tumor. Which are the tables that are explaining the vast majority of the two first axes? (NOTE: look at the pseudoeigen plot)

3. Perform penalized canonical correlation analysis and provide the top-5 features associated with each tumor
4. Compare both results

TASK 3 - Supervised methods: Using the data described in task 2:

1. Perform SGCCA to determine those miRNA that are associated with each tumor. NOTE: the fourth table (e.g. cancer status) can be created by typing:

```
Y <- model.matrix( ~ cancer)[,-1]
```

NOTE: Do not forget to load the R functions that may help you to visualize the results

```
source("Day3-integration_multiple_tables/R/plotInd.R")
source("Day3-integration_multiple_tables/R/selectVars.R")
source("Day3-integration_multiple_tables/R/topVars.R")
```