Exercise 1 : unsupervised multiblock analysis

Nutrimouse dataset

```
library(CCA)
data("nutrimouse")
```

The data sets come from a nutrigenomic study in the mouse (Martin et al., 2007) in which the effects of five regimens with contrasted fatty acid compositions on liver lipids and hepatic gene expression in mice were considered.

Two sets of variables were acquired on forty mice:

- **genes**: expressions of 120 genes measured in liver cells, selected (among about 30,000) as potentially relevant in the context of the nutrition study. These expressions come from a nylon macroarray with radioactive labelling
- **lipids**: concentrations (in percentages) of 21 hepatic fatty acids measured by gas chromatography

Biological units (mice) were cross-classified according to two factors experimental design (4 replicates):

- **genotype**: 2-levels factor, wild-type (WT) and PPARα -/- (PPAR)
- **diet**: 5-levels factor. Oils used for experimental diets preparation were corn and colza oils (50/50) for a reference diet (REF), hydrogenated coconut oil for a saturated fatty acid diet (COC), sunflower oil for an Omega6 fatty acid-rich diet (SUN), linseed oil for an Omega3-rich diet (LIN) and corn/colza/enriched fish oils for the FISH diet (43/43/14)

ComDim analysis of all samples

Question 1: based on lipids and genes data, do we observe clusters of samples ?

Prepare dataset

- concatenate genes and lipids dataframes
- define the number of variables of both block

Run ComDim analysis

- use ComDim() from MBAnalysis package

```
library(MBAnalysis)
ComDim_res <- ComDim(X = ComDim_data,
                     block = n_group,
                     name.block = c("genes", "lipids"),
                     scale = T,
                     scale.block = T)
```

Question 2: how do both blocks contribute to each dimension?

- plot saliences

```
saliences <- ComDim_res$saliences
```

Question 3: observe the samples distributions in the space of the common dimensions, what are the main sources of variation?

- plot scores on Dim.1 vs Dim.2 and Dim.3 vs Dim.4 with percentages of explained variance on axes

```
scores <- data.frame(metadata, ComDim_res$Scor.g)
ComDim_res$cumexplained[1,"%explX"]
```

Question 4: which genes and lipids are responsible of the samples differences?

- plot scores on Dim.1 vs Dim.2 and Dim.3 vs Dim.4 with percentages of explained variance on axes

```
loadings <- data.frame(ComDim_res$Load.g)
```

Discriminant analysis wt vs ppar - mixOmics

Question 1: based on lipids and genes data, can we discriminate wt vs ppar samples ?

Prepare dataset

- as a list of dataframes
- the outcome as a factor

Run block.plsda analysis

- use *block.plsda* () from mixomics package

```
blockPLS_res <- block.plsda(X = blockPLS_data,
                            Y = genotype, design = "full",
                            all.outputs = T, ncomp = 10)
```

Question 2: Choose optimal number of latent variables?

- run perf() function from mixomics package
- plot the results with plot()
- run the analysis with optimal number of latent variables

```
blockPLS_perf <- perf(blockPLS_res, validation =
'Mfold', folds = 7, nrepeat = 10, auc = TRUE, cpus=2)
```

Question 3: Is the model statistically significant?

- run a permutation test with DIABLO.test() from RVAideMemoire package

```
blockPLS_permtest <- DIABLO.test(blockPLS_res)
```

Question 4: what is the variance explained for each block by each latent variable and globally?

```
blockPLS_expl <- do.call("rbind",blockPLS_res$AVE$AVE_X[1:2])
blockPLS_expl <- rbind(blockPLS_expl, blockPLS_res$AVE[["AVE_outer"]])
```

Question 5: observe the samples distributions in the space of the latent variables.

- plot scores with plotIndiv()

```
plotIndiv(blockPLS_res, block = "weighted.average")
```

Question 6: which genes and lipids are discriminant for genotype?

- plot loadings with plotVar()

```
plotVar(blockPLS_res)
```

Discriminant analysis wt vs ppar - consensusOPLS

Question 1: based on lipids and genes data, can we discriminate wt vs ppar samples ?

Prepare dataset

- as a list of dataframes
- the outcome as a matrix of dummy variables

Run ConsensusOPLS analysis

- use *ConsensusOPLS*() from ConsensusOPLS package

```
consensusOPLS_res <- ConsensusOPLS( data = COPLS_data,
Y = dummy_genotype, maxPcomp = 1, maxOcomp = 1, nfold
= 40, cvType = "nfold", nperm = 100, modelType = "da",
mc.cores = 1, verbose = FALSE )
```

Question 2: Is the model statistically significant?

- The results of permutations
- plot histograms for Q^2, dQ^2 and R^{2}Y values obtained from permutations and visualise the results for the optimal model on them

```
COPLS_res$permStats
COPLS_res$optimal$modelCV
COPLS_res$optimal$modelCV$cv
```

Question 3: What is the contribution of each data block?

- plot the blocks contributions of the optimal model as a bar plot

```
COPLS_res$optimal$modelCV$Model$blockContribution
```

Question 4: Show the distribution of samples in the space of the predictive and orthogonal latent variables?

- plot scores of the optimal model

```
COPLS_res$optimal$modelCV$Model$scores
```

Question 5: observe the samples distributions in the space of the latent variables.

- plot loadings of the optimal model

```
COPLS_res$optimal$modelCV$Model$loadings
```

Question 6: which genes and lipids are discriminant for genotype?

- plot Variables Importance in Projection of the optimal model

```
COPLS_res$optimal$VIP
```