Dimensionality reduction

Exercices

PCA

1. Load the `nutrimouse` data from the `mixOmics` R package and investigate its structure.

A data object provided by an R package can be loaded with `data`. Its structure can be obainted with `str`, `length`, `dim`, etc.

2. Take the gene expression dataset in *samples* x *variables* matrix format. Investigate their distribution.

3. Perform PCA and investigate variances, sample distribution and variable relationship with plots.

A number of methods in different R packages can perform PCA, e.g. `stats::prcomp`, `stats::princomp`, `mixOmics::pca`, `multiblock::pca`, `psych::principal`, `FactoMineR::PCA`, etc.

Variances = eigenvalues of the covariance matrix = (standard deviation)^2

Scree plot: plot of variances.

Scree plot on variance percentage.

Scores: sample coordinates in the new reference (rotated axes or principal components).

Score plot: plot of sample distribution.

Loadings: contributions of variables to principal components (eigenvectors of covariance matrix).

Loading plot: plot of variables' contribution, revealing their relationship.

Both score and loading plot can be plot altogether with the `biplot` function.

4. Visually investigate the sample distribution with coloring by metadata or expression of certain genes.

The samples can be colored with some metadata, e.g *genotype* or *diet*,

or by some gene expression.

PLS

1. Perform PLS (`mixOmics::pls`) and investigate the output, sample distribution and variable relationship with plots.

The sample distribution plot can be performed with **variates**, sample coordinates in the new reference (rotated axes) for each of the two blocks.

which is also produced with `plotIndiv`.

Loading plot: plot of variables' contribution in each data block to each variate, after deflating more *important* variates.

which is the same as with `plotLoadings`.

The plot of variable relationship could be obtained from **loadings.star**.

Both sample distribution and variable relationship plot could be done with `biplot` function.

2. Observe the difference between the two modes *regression* and *canonical* of PLS.

CCA

1. Perform CCA (`mixOmics::rcc`) between 20 genes and all lipids. Investigate correlations, sample distribution and variable relationship with plots.

The gene expression data is reduced to 20 genes so that the number of variables is less than the number of samples, to perform an unregularized CCA.

The sample distribution plot can be performed with **variates**, sample coordinates in the new reference (rotated axes) for each of the two blocks.

Variable relationship is obtained from **loadings** or with `plotVar`.

2. Perform CCA with scaled datasets and observe the difference

3. Perform regularized CCA with all genes and lipids.