elixir ITALY

Metabolomics and Integrative omics:
from data production to analysis

28-29 Aprile 2022 | BARI

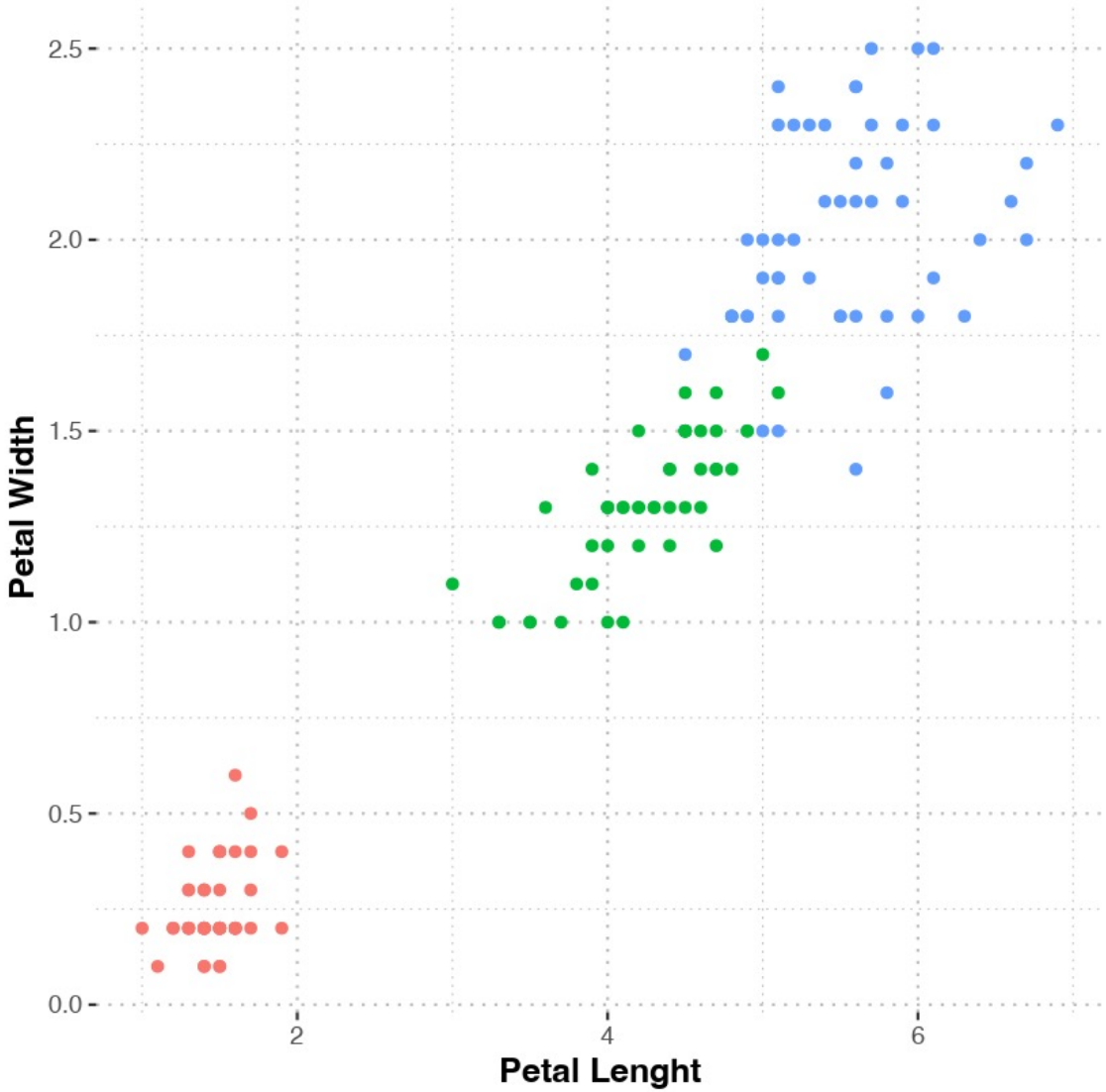
IBIOM CNR.BiOmics ThermoFisher SCIENTIFIC SIB

Few hints about discussed statics

Dr. Bruno Fosso

IBIOM
Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies
Bari, Italy

Principal Component Analysis



An easy way to inspect the relationship among this compound could be to plot a scatter-plot of pairwise compounds comparisons.

Principal Component Analysis

Actually, our experiment allowed to identify about 5,800 compounds.

Considering the number of generated features, how many scatter plots do we need to generate?

So suppose our dataframe (the object containing all our measurements) contains *p* x *n* data, where p are the compounds and n the observation (in our case 9).

$$Plots = \frac{p(p-1)}{2}$$

$$Plots = \frac{5755(5755-1)}{2} = \mathbf{16,557,135}$$

Probably there are to much comparisons to observe!!!

What we need is an approach allowing to summarize the complexity of our data by reducing the number of objects to observe.

Dimension reduction technique

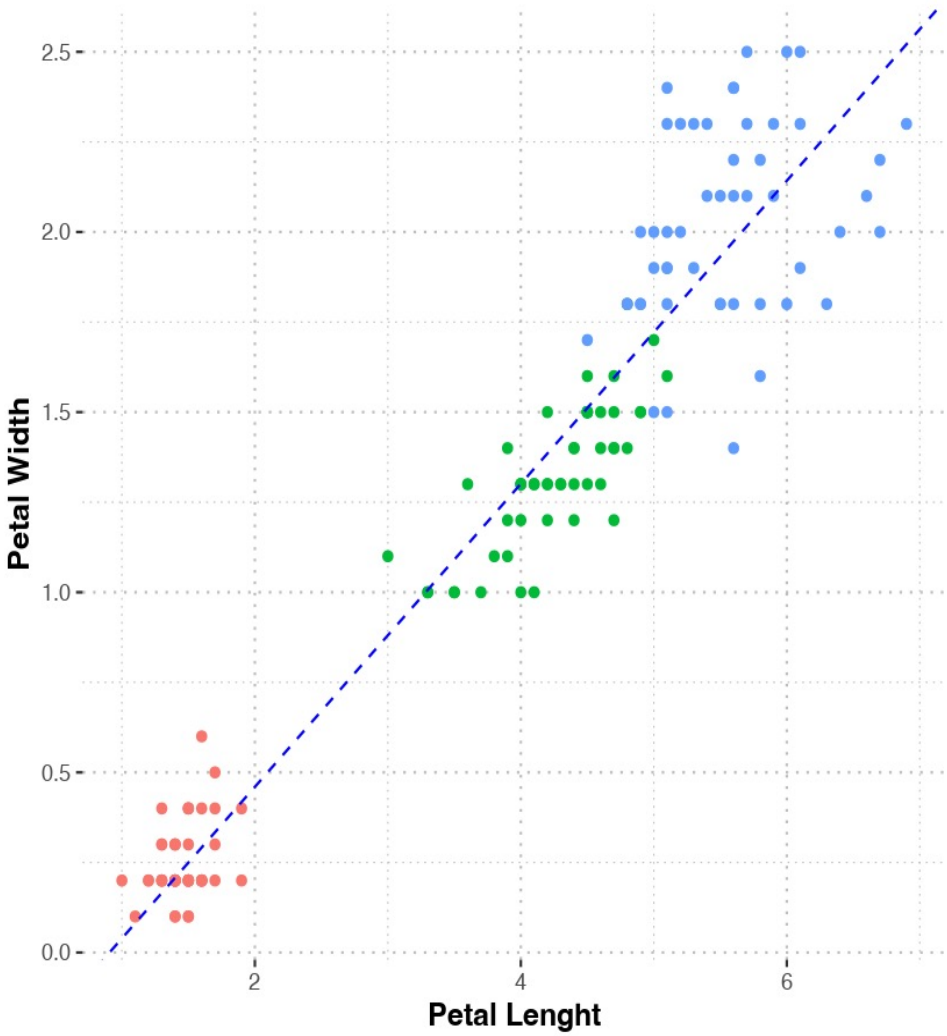
Principal Component Analysis

PCA finds a low-dimensional representation of a data set that contains as much as possible of the variation.

PCA finds a low-dimensional representation of a data set that contains as much as possible of the variation.

The underlying idea is quite simple: each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting. So PCA search for a small number of dimensions that are as interesting as possible.

What does it means interesting in our case? Simply the most variance in data the dimension explains the most important it is



Principal Component Analysis

Principal components (PCs) are obtained by maximizing their variance. Of course the first component explain the highest variance. Following a mathematical representation of how components are computed:

$$\max_{\phi_{11}, \dots, \phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\}$$

Once PCs are computed, the dimension reduction is simply obtained by projecting the data on the PCs. It means that a score is assigned to each sample. This score is a linear combination of the original variable and a weight. This weight is called ***Loading*** Φ. For each PC a loading vector is computed and stored.

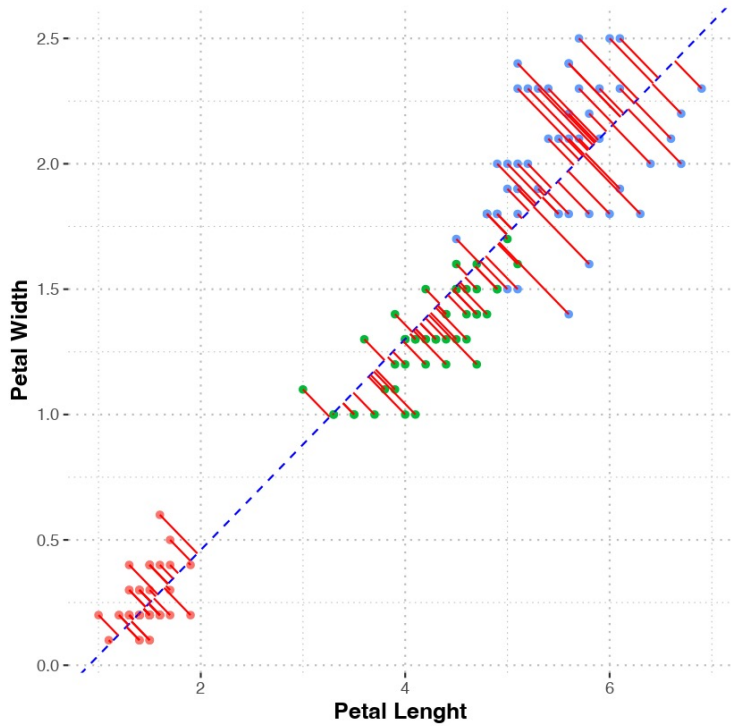
The loadings can be understood as the weights for each original variable when calculating the principal component.

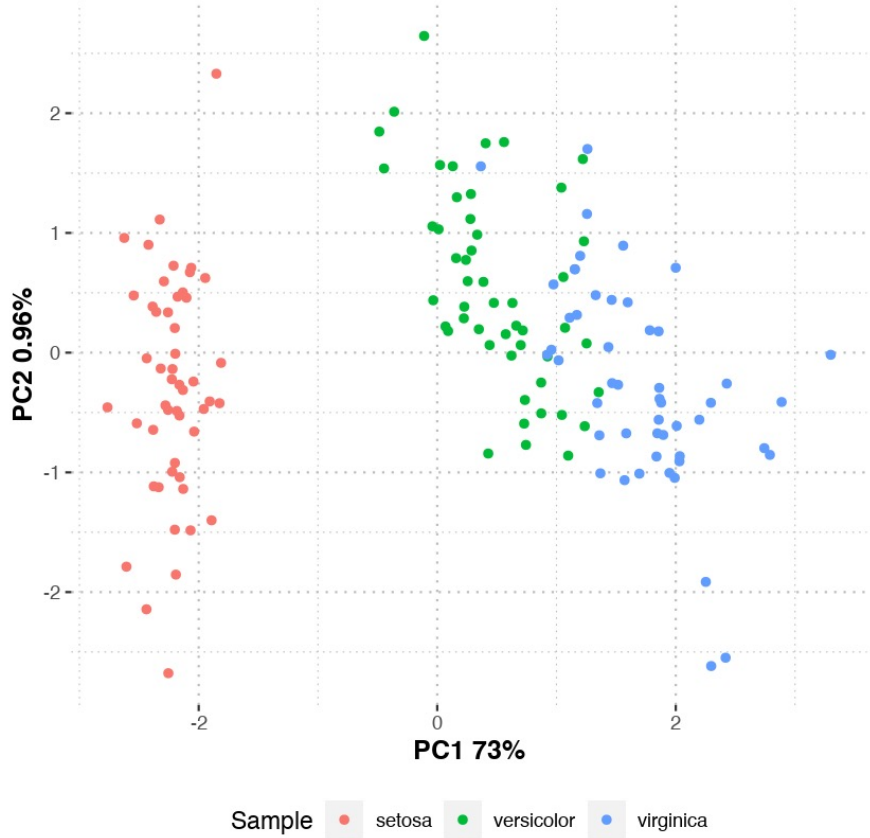
The process relying on the PCs decomposition is based on correlation/covariance matrix so it assumes your data are normally distributed.

Principal Component Analysis

The loadings can be understood as the weights for each original variable when calculating the principal component.

The process relying on the PCs decomposition is based on correlation/covariance matrix so it assumes your data are normally distributed.

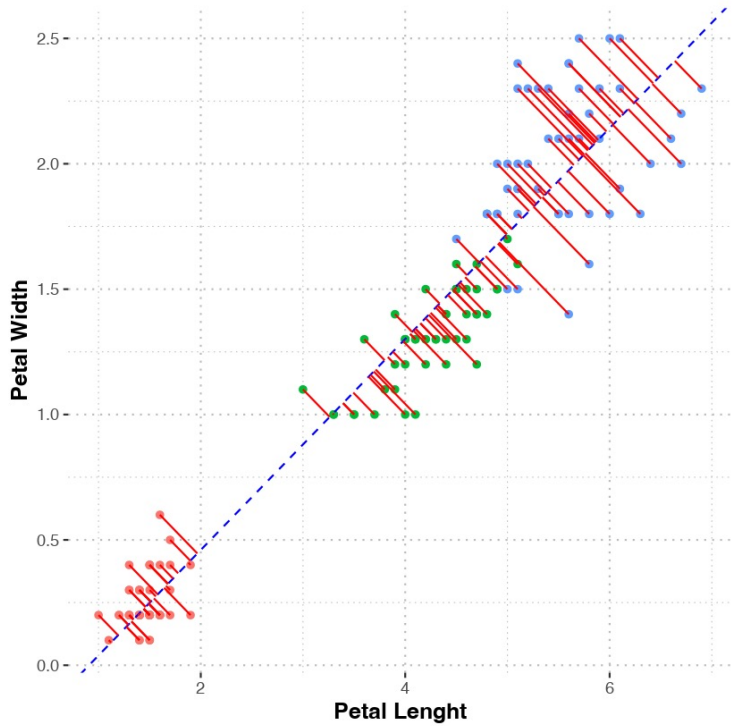


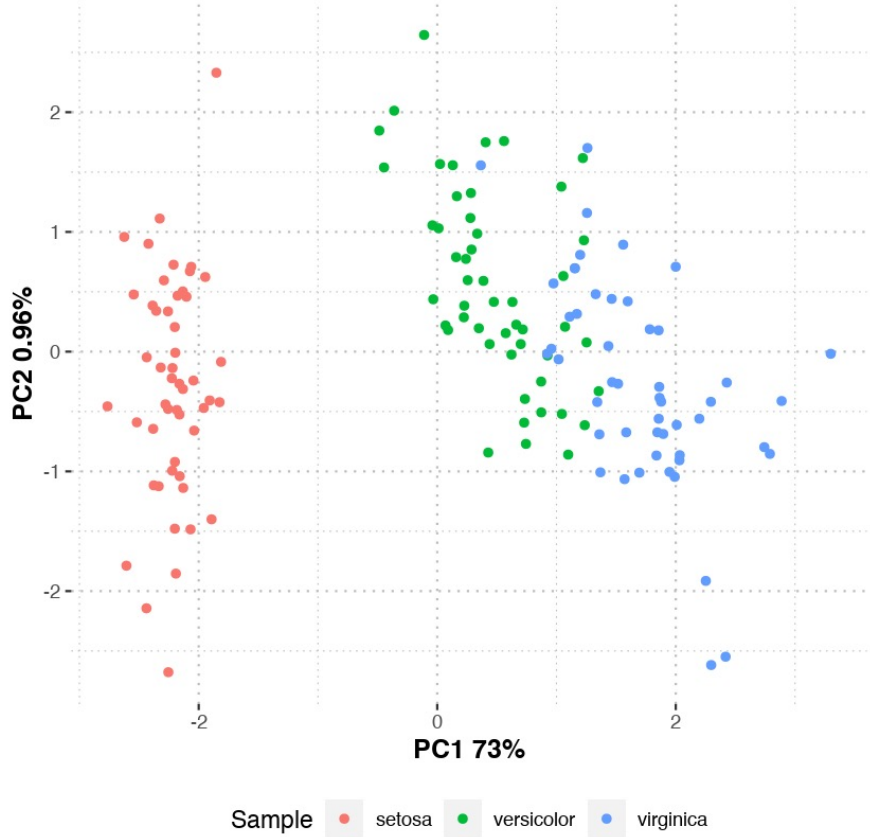


Principal Component Analysis

Once the PC1 is computed we can find PC2 as the linear combination of original values and loadings that has maximal variance out of all linear combinations that are uncorrelated with PC1. This means were forcing the PC2 Loadings vectors direction to be orthogonal to PC1 loadings. → This allows us to represents PCA in a Cartesian plane.

PCA is completely **UNSUPERVISED**: this means it does not take care about sample labels but only process your data.





PLS-DA

Partial Least Squared – Discriminant Analysis (PLS-DA) is a **SUPERVISED** alternative to PCR.

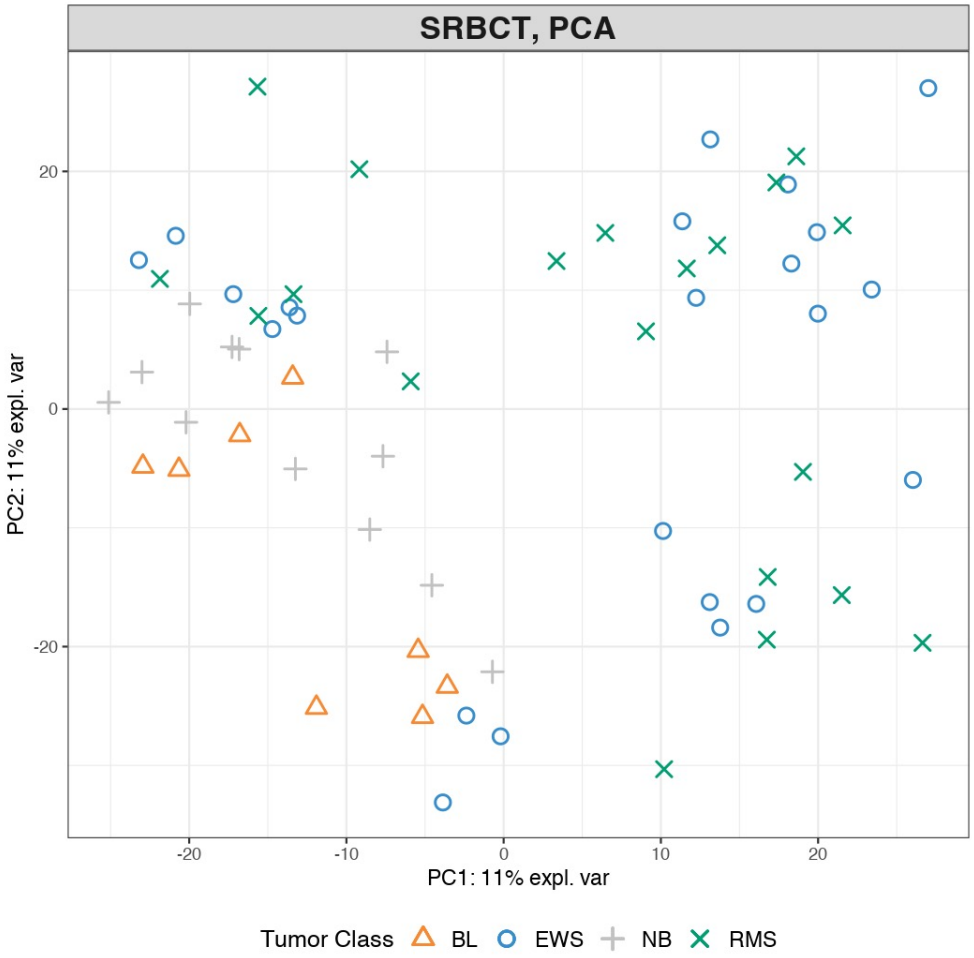
In this specific case were are processing both our n x p matrix (i.e. compounds x samples) and samples groups (e.g. ρ0 vs ρ+). It reduces data dimension by identifying latent variables (such as Components), that corresponds to linear combination of the original variables, and maximize the discrimination according to response variable (sample groups).

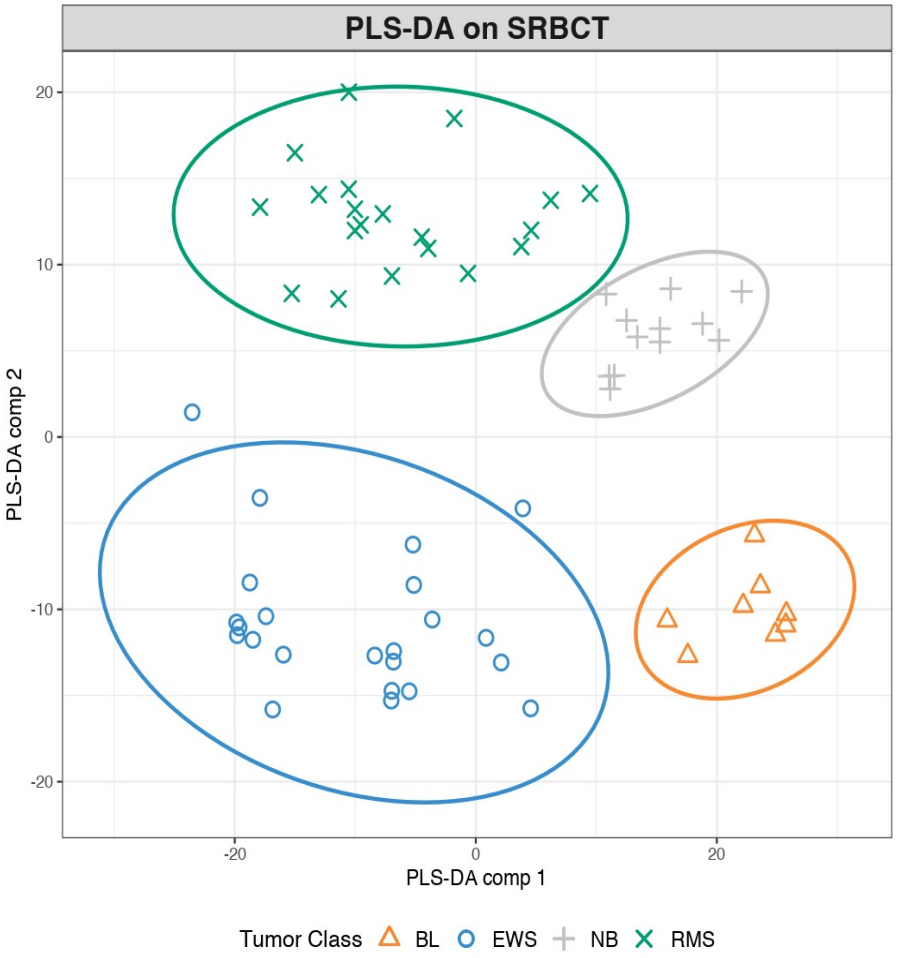
Practically speaking, the first components gives an higher weight to the variables that are more related to sample groups.

We obtain the following data:

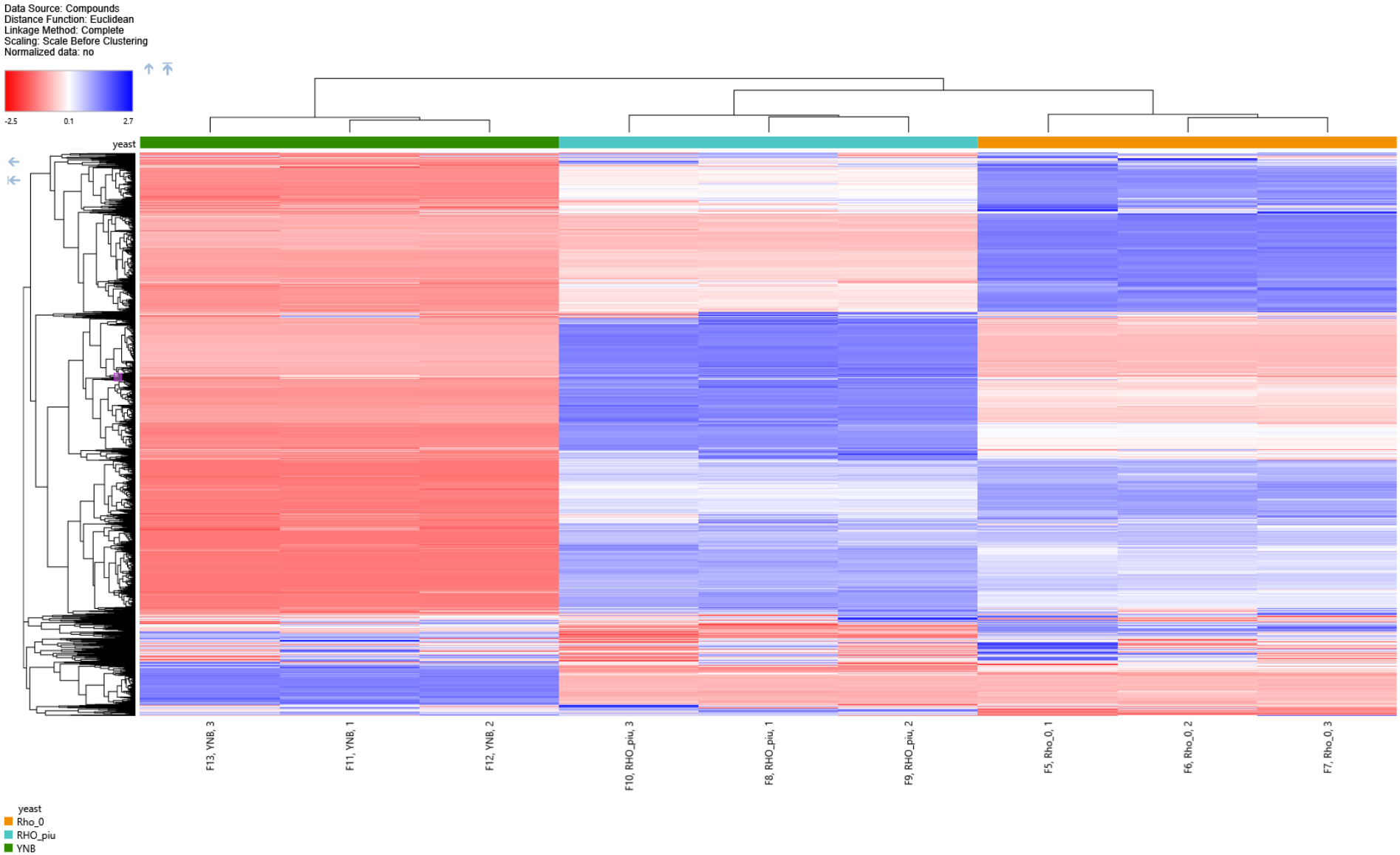
- Components;
- loading vectors: they indicate the importance of each variable in PLS-DA. Loading vectors are obtained by maximizing the covariance among the linear combination of the variables from our matrix and the factor of interest Y (sample groups)
- A list of selected variables from X and associated to each component.

PLS-DA


SRBCT, PCA
PC2: 11% expl. var
PC1: 11% expl. var
20
0
-20
-20
-10
0
10
20
Tumor Class
BL
EWS
NB
RMS


PLS-DA on SRBCT
PLS-DA comp 2
PLS-DA comp 1
20
10
0
-10
-20
-20
0
20
Tumor Class
BL
EWS
NB
RMS

Hierarchical Clustering Analysis



Hierarchical Clustering Analysis

Hierarchical Clustering refers to a grouping approach relying on samples aggregation according to pairwise similarities represented into a dendrogram.

More precisely, it is usually achieved by using a *bottom-up (or agglomerative)* approach, in which samples are grouped according to decreasing similarities. Practically, clustering is performed by aggregating closest objects from higher similarities (dendrogram leaf) to lower (top of the tree).

It is crucial to define two concepts:

- Sample comparisons: refers to metrics used to compute the similarities among single observations. *Distance metrics.*
- Groups comparisons: refers to metrics used to compute the similarities among group of observation. *Linkage metrics.*

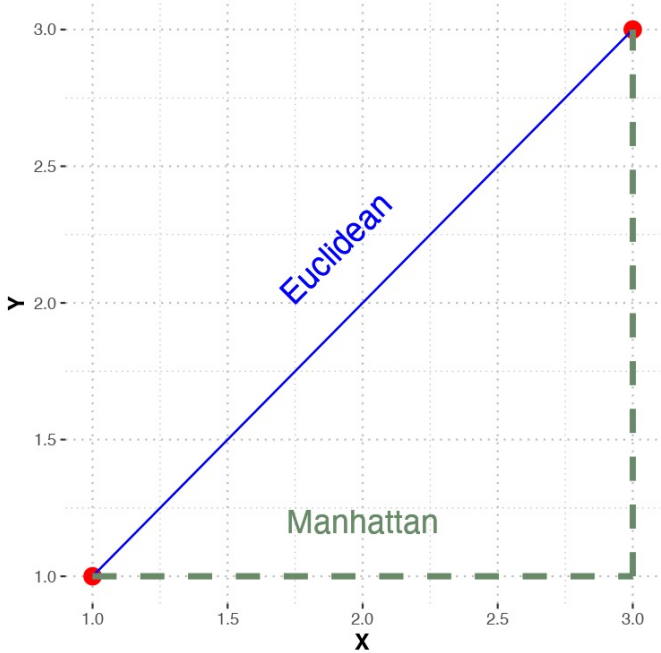
Hierarchical Clustering Analysis

Distance metrics

- Euclidean distance: conventional distance among vectors
- Squared Euclidean distance: Euclidean distance without the square root. Emphasize higher differences;
- Manhattan: Absolute distance between the two vectors;
- Binary: values vector are converted into a binary representation.

A	0.39	-0.39	0.76	0.18	1.85	-1.09	0.13	1.33	1.25	0.11
B	0.61	1.20	-1.00	-1.18	0.99	0.74	0.98	-0.93	-0.10	-0.36

Distance Metrics	Observed distance
Euclidean	4.420277
Squared euclidean	19.53885
Manhattan	12.56158
Binary	0

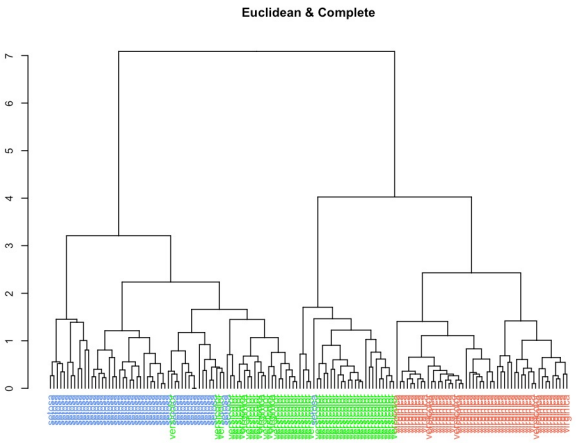


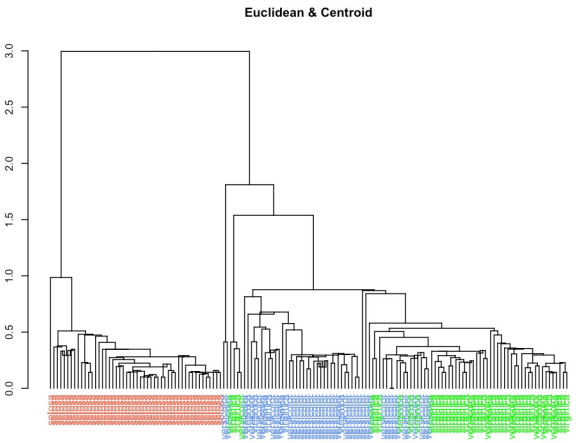
Hierarchical Clustering Analysis

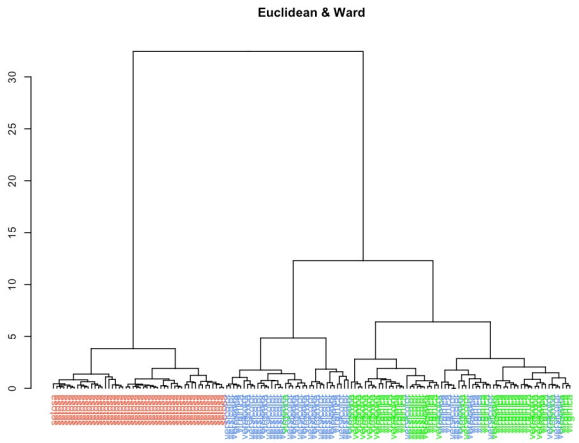
<u>Linkage Metrics</u>

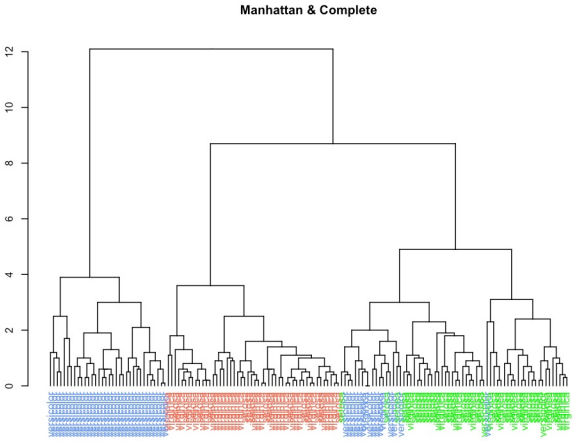
- Complete: Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between observation in cluster A and cluster B, and stores the largest one.
- Single: Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between observation in cluster A and cluster B, and stores the smallest one.
- Average: Mean intercluster dissimilarity. Compute all pairwise dissimilarities between observation in cluster A and cluster B, and stores the average value.
- Centroid: Distance between the cluster A and B centroids. Centroid is a mean vector of length *p*.
- Ward: Minimum cluster variance. It relies on an algorithm which tries to minimize the variance increase when adding a element to a group.

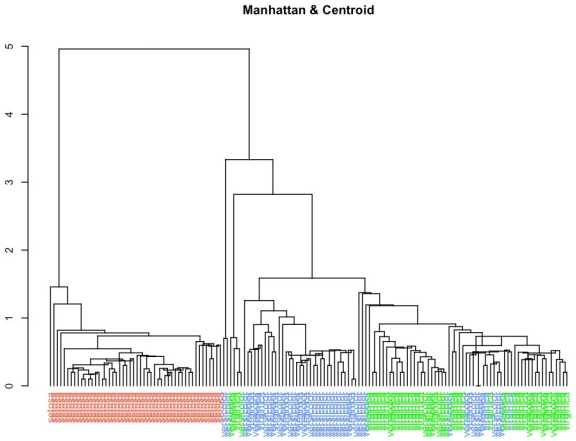
Hierarchical Clustering Analysis

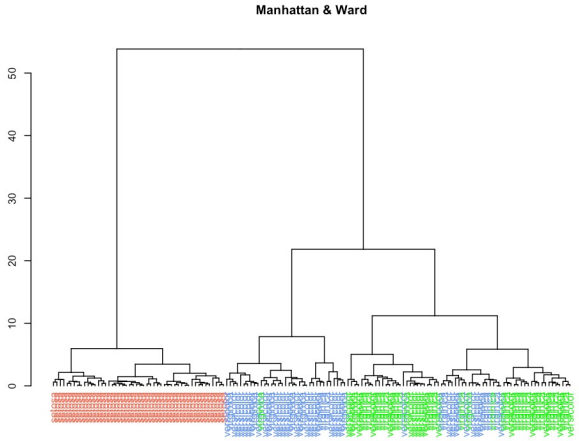




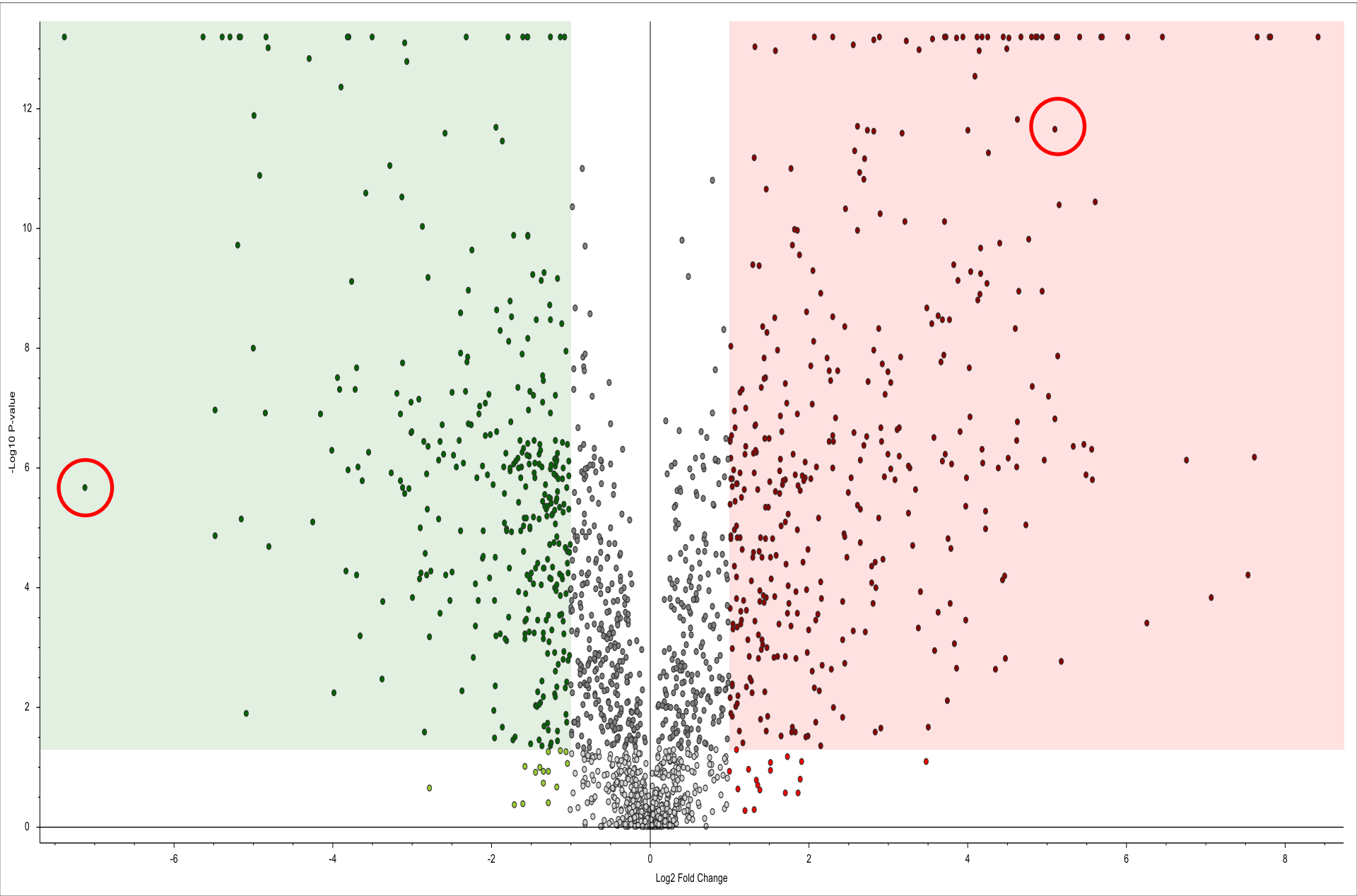




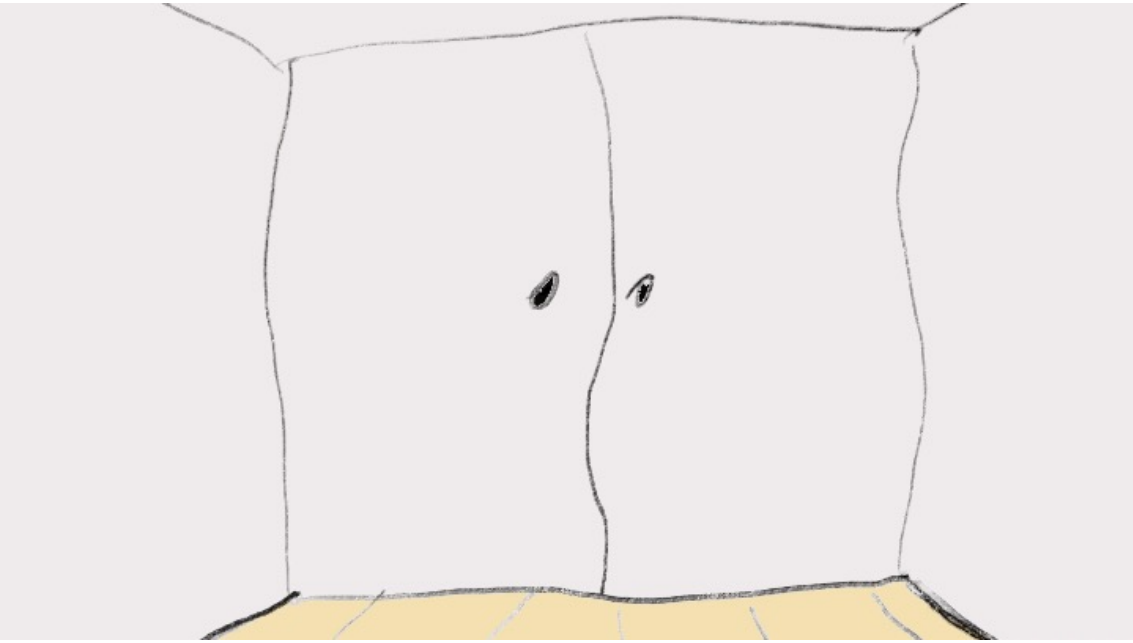




Volcano plot



Thanks



Any Questions?