# β Diversity

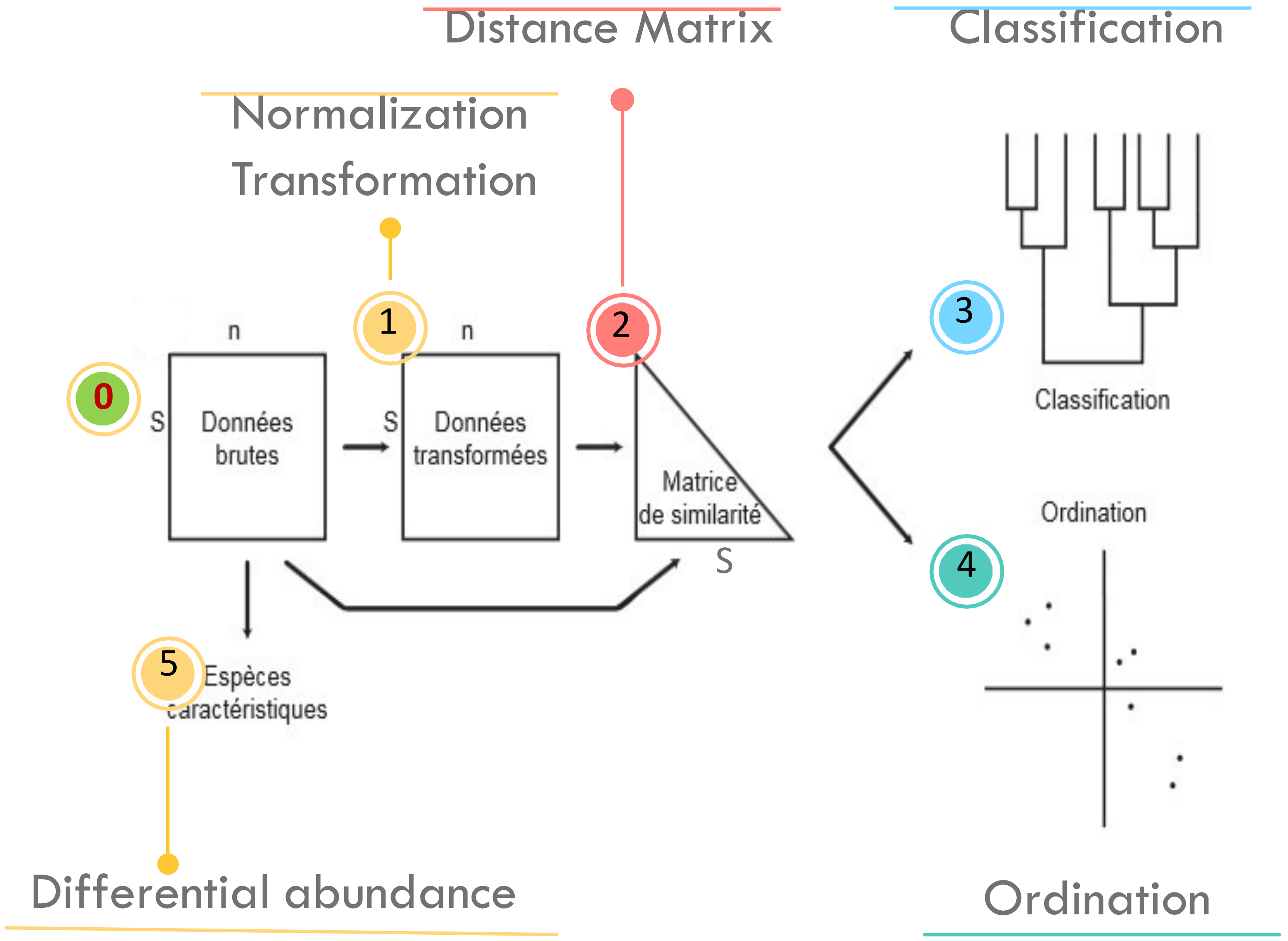
## Inter-sample comparison of the community composition

- **Measure of the similarities/dissimilarities between the samples** according to specific criteria of the MEASURE under consideration (e.g. Unifrac, Bray-curtis)

- **Highlight structure** by **Ordination** Plot (e.g. PCoA, PCA, Db-RDA) or Hierarchical **clustering**

- **Test the structure** differences & **identify main variables/Taxa** e.g. Permanova, Differential Abundance Analysis

# Overview of the Beta-analysis approach



# Some important features of Metabarcoding Data

- The ASV count matrix is :

**SPARSE**, **means** 80- 95% of the counts are **ZEROS**

$$\begin{array}{c|ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 1 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 3 & 0 & 0 & 7 & 0 \\ 4 & 2 & 0 & 0 & 9 & 0 & 0 & 0 \\ 5 & 0 & 0 & 8 & 0 & 0 & 0 & 0 \end{array}$$

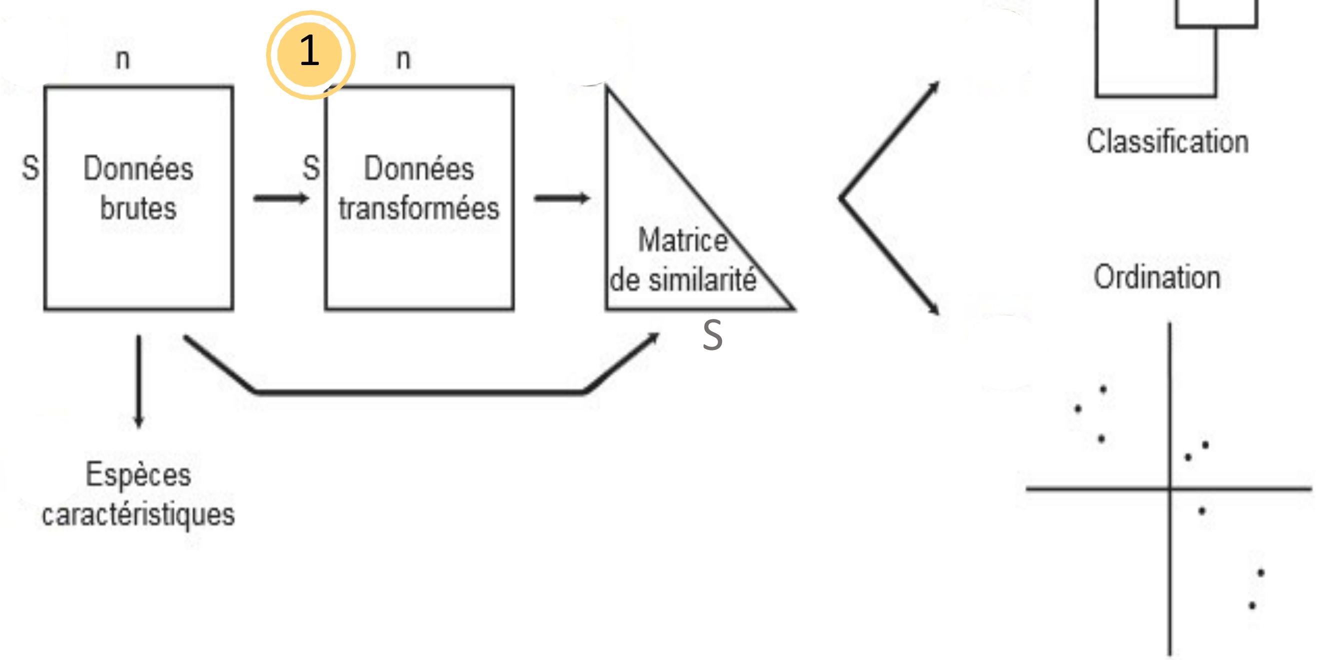
- **Distorded** by experimental bias (i.e. sampling, PCR, sequencing depth limitation), **Overdispersed**

- **Compositional** (ie. a closed system, not independent) = **CoDA**

→ Until recently, these features were **NOT** considered in the analysis of such data!!!!

# Overview of the Beta-analysis approach

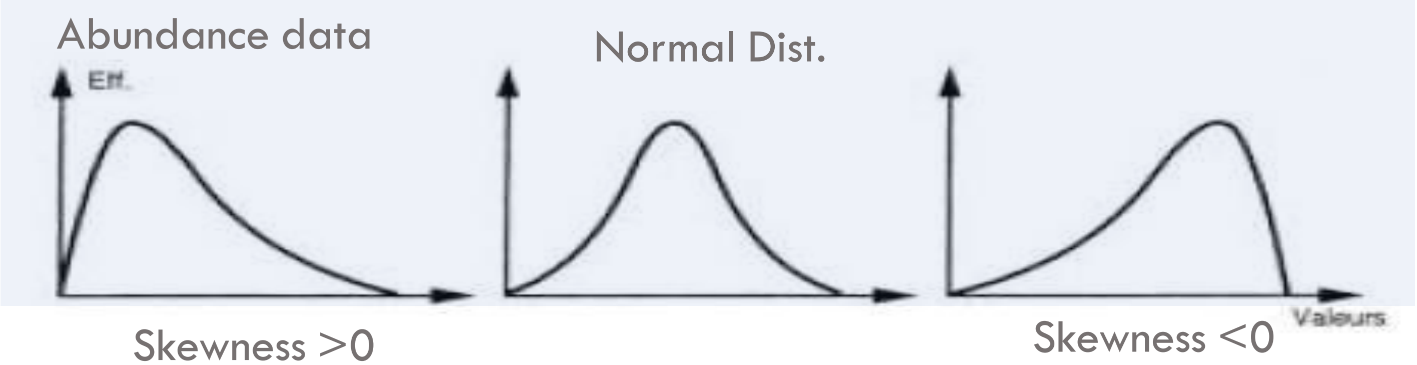
Normalization
Transformation



# First approach for Normalization & Transformations

## Correcting sparsity and overdispertion

Sparse Data
=
contain many Zeros



**Where is this kind of distribution???**

## Why transformation ?

- To **reduce the variation range** (e.g. give low weight to extreme values)
- Transformation motivated by the **type of ordination** (PCA/CA)
- Aid of **comparability** (data are in different units: env parameters) : **Z-score**

# What kind of Transformations for species abundance data

- **Common transformations (avoid)**

  - Log x+1 → *log1p(data)*
  - Square root → *sqrt(data)*
  - double square → root (*sqrt(sqrt(data)*)

decostand() from Vegan

(Thorsen et. al 2016)

**Scale of the reduction** of variation range: Log > double sqrt > sqrt
→ Be careful of the deformation of data

- **Ecologically motivated transformations**

**Hellinger**
→ Gives **low weights** to variables with low counts and many zeros (allow tb-PCA)

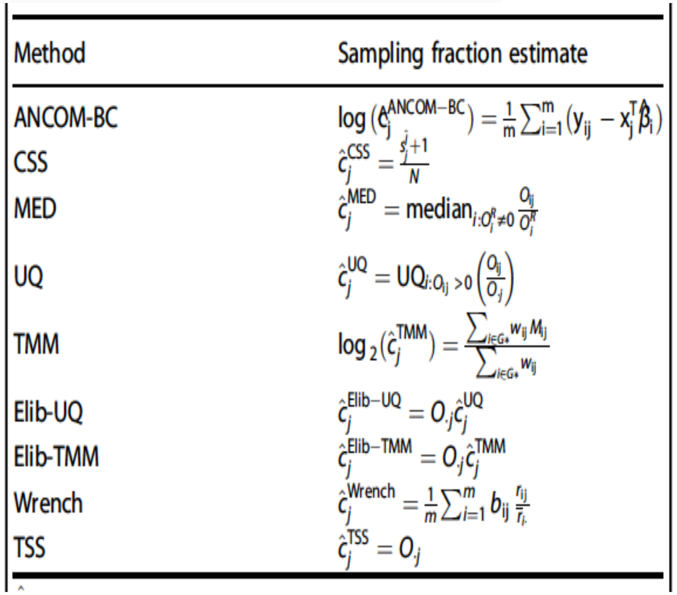
**Chord**
→ similar to Hellinger

# History of normalization : Correcting library size (i.e. sequencing depth)

- **Rarefying** : **Sub-sampling normalization** (alpha diversity)

→ Use rarefaction curves for the minimal libary size, remove samples etc

- **Scaling** : Divide each abundance by a **scaling factor** to eliminate bias from unequal library size
  - → **CSS** : Cumulative Sum Scaling (MetagenomeSeq R)
  - → **TMM**: Trimmed Mean of M-values (Edge R)
  - → **TSS** : Total Sum Scaling = relative abundance



| Method | Sampling fraction estimate |
|---|---|
| ANCOM-BC | $\log(\hat{c}_j^{\mathrm{ANCOM-BC}}) = \frac{1}{m}\sum_{i=1}^{m}(y_{ij} - x_j^T\hat{\beta}_i)$ |
| CSS | $\hat{c}_j^{\mathrm{CSS}} = \frac{s_j^l+1}{N}$ |
| MED | $\hat{c}_j^{\mathrm{MED}} = \mathrm{median}_{i:O_i^R \neq 0} \frac{O_{ij}}{O_i^R}$ |
| UQ | $\hat{c}_j^{\mathrm{UQ}} = \mathrm{UQ}_{i:O_{ij}>0}\left(\frac{O_{ij}}{O_{\cdot j}}\right)$ |
| TMM | $\log_2(\hat{c}_j^{\mathrm{TMM}}) = \frac{\sum_{i \in G*} w_{ij} M_{ij}}{\sum_{i \in G*} w_{ij}}$ |
| Elib-UQ | $\hat{c}_j^{\mathrm{Elib-UQ}} = O_{\cdot j}\hat{c}_j^{\mathrm{UQ}}$ |
| Elib-TMM | $\hat{c}_j^{\mathrm{Elib-TMM}} = O_{\cdot j}\hat{c}_j^{\mathrm{TMM}}$ |
| Wrench | $\hat{c}_j^{\mathrm{Wrench}} = \frac{1}{m}\sum_{i=1}^{m} b_{ij} \frac{r_{ij}}{\bar{r}_{i\cdot}}$ |
| TSS | $\hat{c}_j^{\mathrm{TSS}} = O_{\cdot j}$ |

**BUT**

# Back of an old concept : Compositional Data (CoDA)

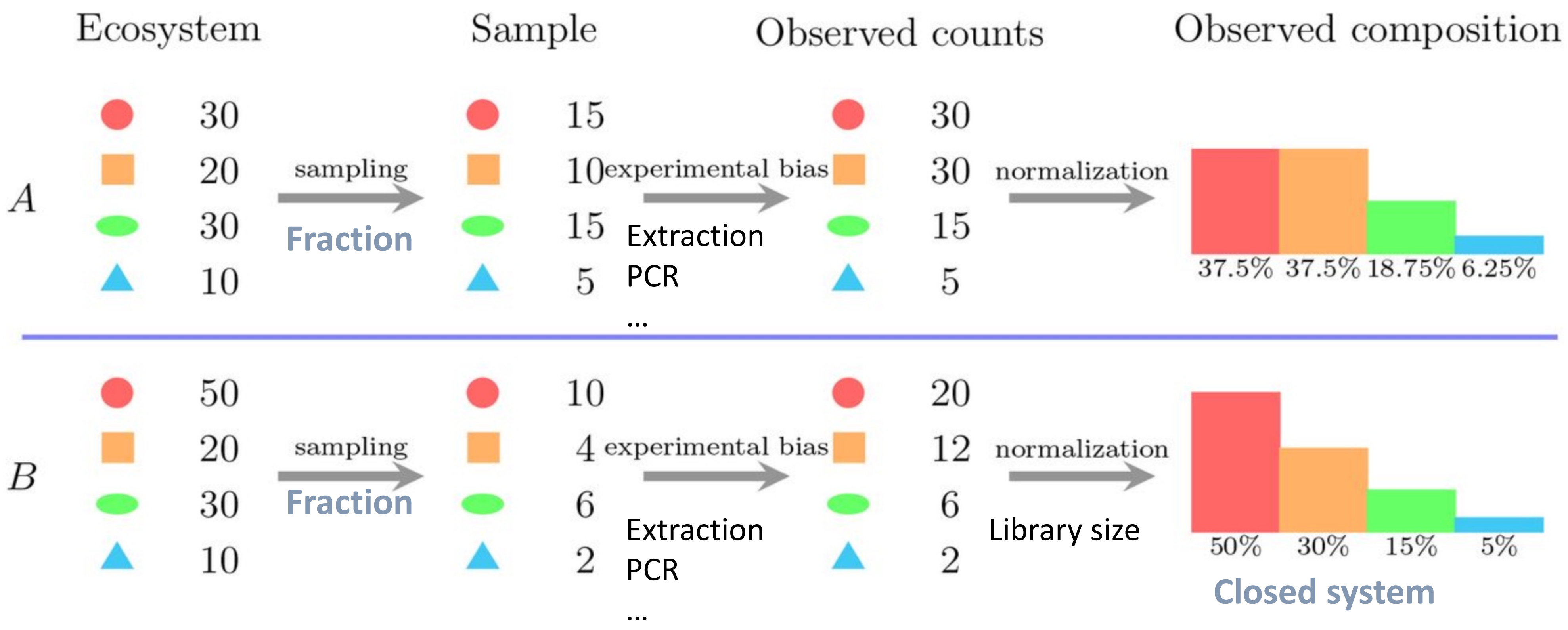
Describe a data set in which the parts in each sample have an arbitrary/constant sum (relative Abundance, pourcentage, probalities…) = A closed System

known as problematic, multivariate data analysis approaches such as ordination, clustering & differential abundance analysis are theoretically invalid!

→ CoDA is still in its infancy, requieres strong mathematical background!!
→ Most of studies/publications do not apply CoDA…

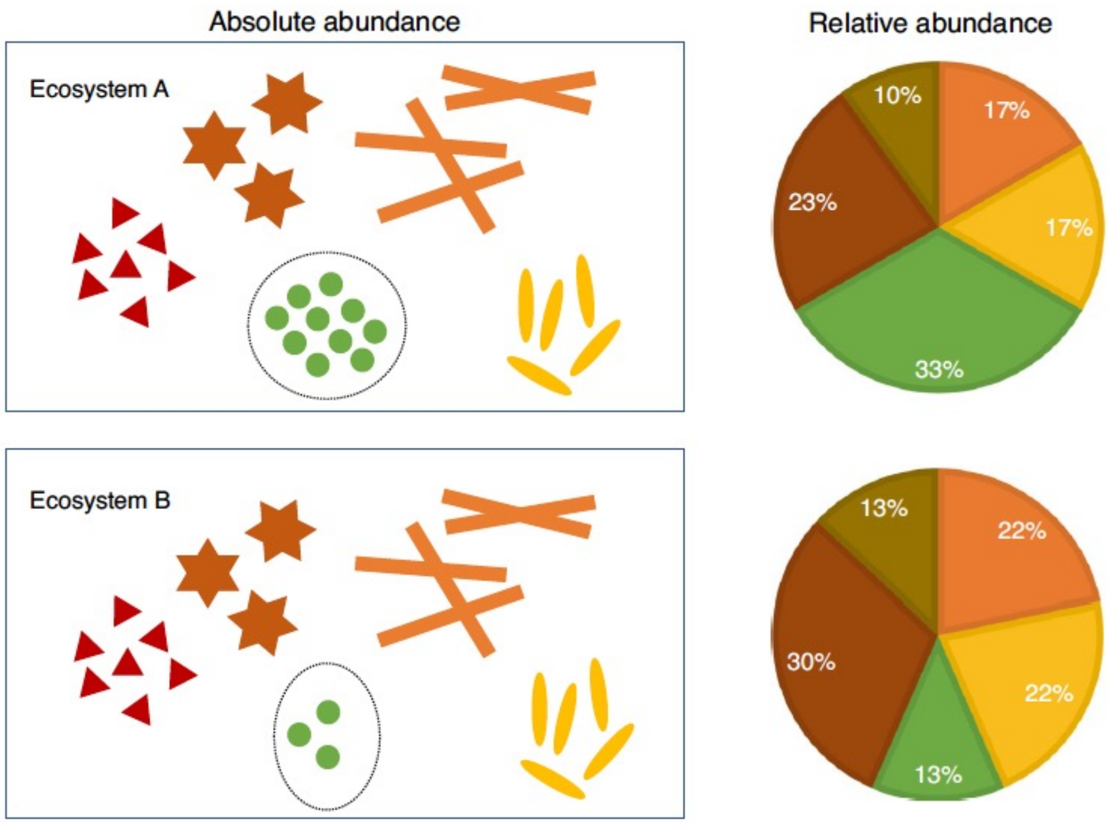
Pionneer addressing CoDa: Pearson & Aitchison

# The limitations inherent in the observed microbial compositional data set



*Wang S. 2021*

# Absolute abundance vs. Relative Abundance (proportions)



Lin H. et al. Nature 2021

Absolute Ab. → Only Green species is different

Relative Ab. → All species are differents

Changing one taxon modifies all the others!!!!!

**NOT INDEPENDENT**

# Consequences …

- Relative Ab. of one taxon impact all the others : not independent

→ Compositional data have a **negative correlation bias**

→Increase **spurious correlations‼**

→ impact in Differential abundance analysis

## Solutions

- Normalized the data: **Sampling fraction** and not only library size
- Log- ratio : Independence of variables (taxon)

- Deal with sampling fraction and Compositional data → CoDA

e.g. ANCOM-BC (Nature, 2022)

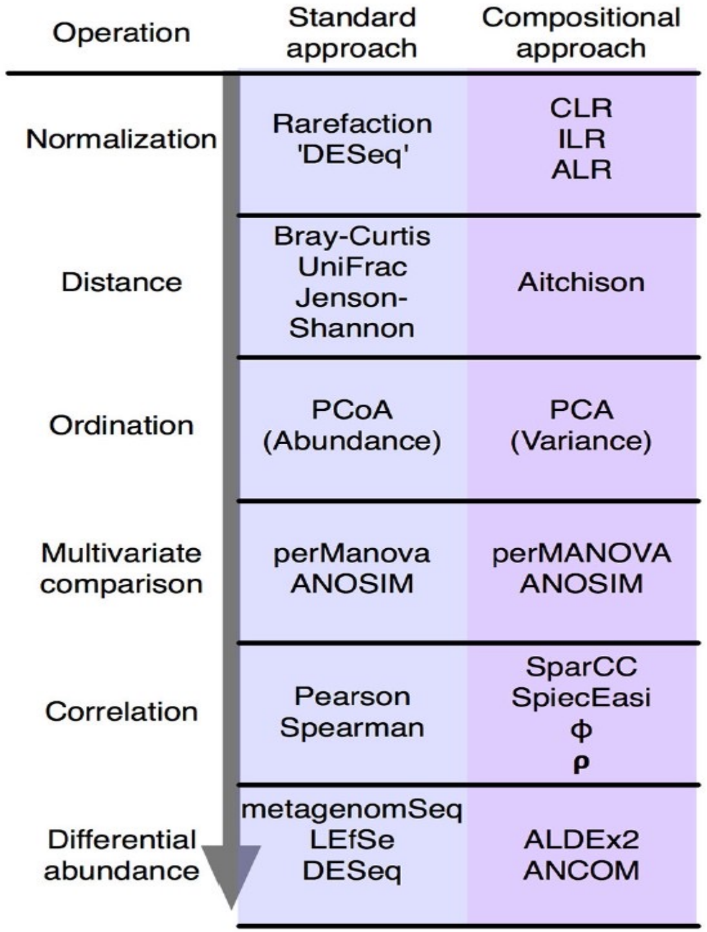
# CoDA : Log ratio transformation

CoDA: Aitchison's Log-ratio based-methods

- Isometric log-ratio (ILR)
- Centered log-ratio (CLR)
- Additive log-ratio (ALR)
- Phylogenetic Isometric Log-Ratio (phILR)

zcompositions, easycoda R packages

Differential Abundance CoDA : Aldex2, ANCOM-BC



| Operation | Standard approach | Compositional approach |
|---|---|---|
| Normalization | Rarefaction<br>'DESeq' | CLR<br>ILR<br>ALR |
| Distance | Bray-Curtis<br>UniFrac<br>Jenson-Shannon | Aitchison |
| Ordination | PCoA<br>(Abundance) | PCA<br>(Variance) |
| Multivariate comparison | perManova<br>ANOSIM | perMANOVA<br>ANOSIM |
| Correlation | Pearson<br>Spearman | SparCC<br>SpiecEasi<br>ϕ<br>ρ |
| Differential abundance | metagenomSeq<br>LEfSe<br>DESeq | ALDEx2<br>ANCOM |

Gloor B., Frontiers 2017

If you can, Log-ratio methods should be favored!

Especially for Differential Abundance Analysis

# Centered log-ratio use geometric mean

- An **“average”** is supposed to describe the **“central tendency”** of data

- **Medians** ignore the values of everything, except from the middle element!

- **Arithmetic mean** is sensitive to extreme/outlier values

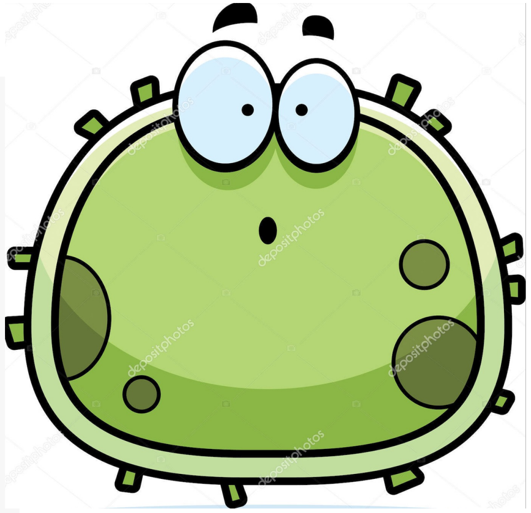
- **Geometric mean** is known to give a **more precise value of the central tendency** of data (this is common in data analysis!!)

# Centered Log-Ratio = CLR (Aitchison, 1986)

For a sample X : CLR is the log ratio of each abundance (x1,x2,...) divided by the geometric mean(Gx)

$$\mathbf{x}_{clr} = [\overset{\text{Log}}{\overset{\downarrow}{log}}\overbrace{(x_1/G(\mathbf{x}))}^{\text{ratio}}, log(x_2/G(\mathbf{x}))\dots log(x_D/G(\mathbf{x}))],$$

$$\underset{\text{Geometric mean}}{\underset{\uparrow}{G(\mathbf{x})}} = \sqrt[D]{x_1 \,\mathrm{X}\, x_2 \,\mathrm{X} \dots \mathrm{X}\, x_D}$$



You & me

- Ratios are the same whether the data are counts or proportions
- Standard statistical methods can be done (mathematical propriety, PCA)
- Data become symetric

# Handling zeros

- CoDA methods depend on logarithms that do not compute for **zeros**!

Removal : Components with zeros get **excluded**

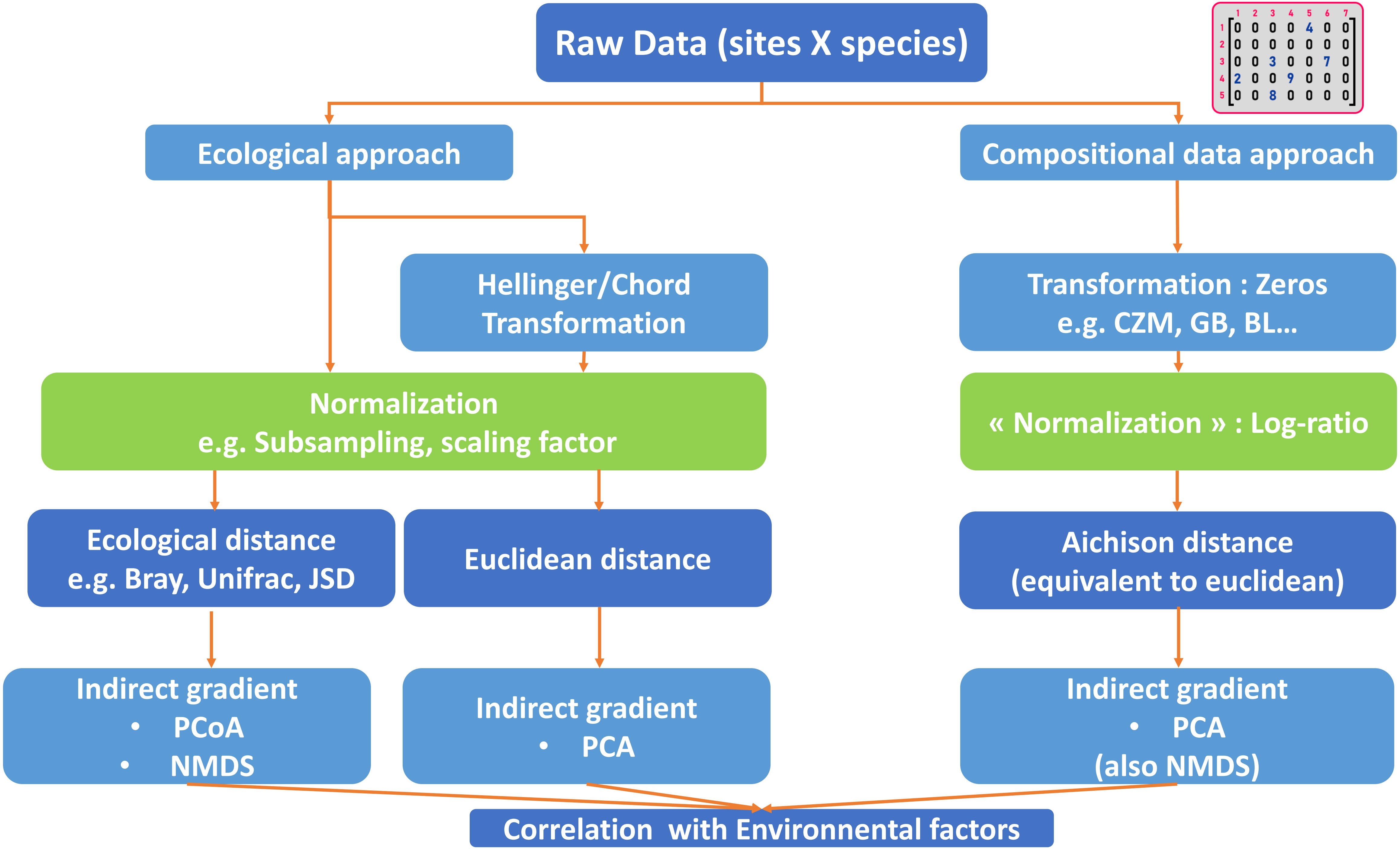
→ sub-composition analyzed by CoDA method (bof)

Modification : Zeros get **replaced** with a non-zero value, **with or without** modification to non-zeros

modification of the non-zero will preserves the ratios between the non-zero components (best)

- **Bayesian**-multiplicative replacement (preserves the ratios, GBM, SB and BL)
- Multiplicative simple replacement (i.e. **CMZ**, do not perserves ratio)

→ ***See cmultRepl R function***



Raw Data (sites X species)

$$\begin{array}{c|ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 1 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 3 & 0 & 0 & 7 & 0 \\ 4 & 2 & 0 & 0 & 9 & 0 & 0 & 0 \\ 5 & 0 & 0 & 8 & 0 & 0 & 0 & 0 \end{array}$$

Ecological approach

Hellinger/Chord Transformation

Normalization
e.g. Subsampling, scaling factor

Ecological distance
e.g. Bray, Unifrac, JSD

Indirect gradient
- PCoA
- NMDS

Euclidean distance

Indirect gradient
- PCA

Compositional data approach

Transformation : Zeros
e.g. CZM, GB, BL…

« Normalization » : Log-ratio

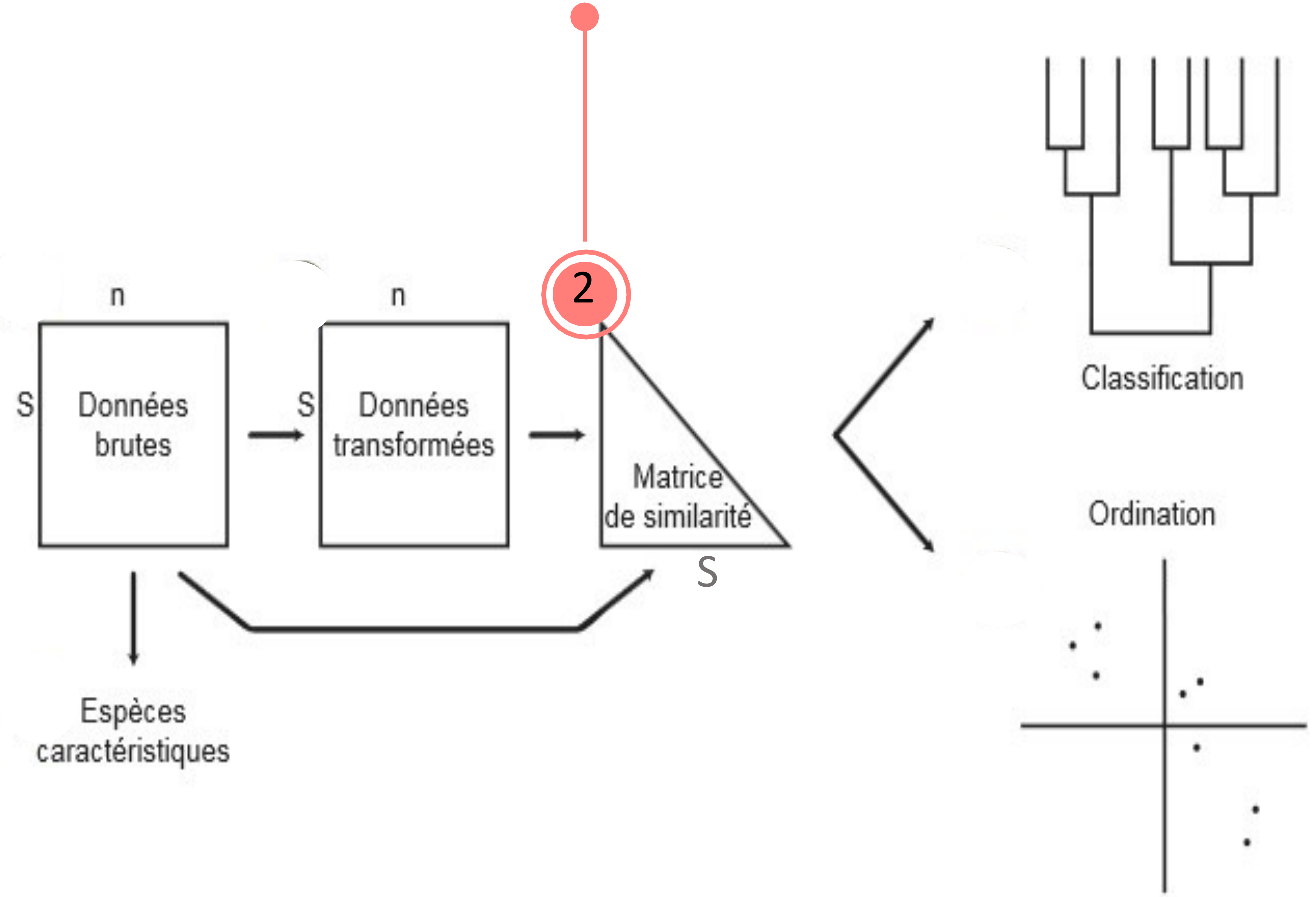
Aichison distance
(equivalent to euclidean)

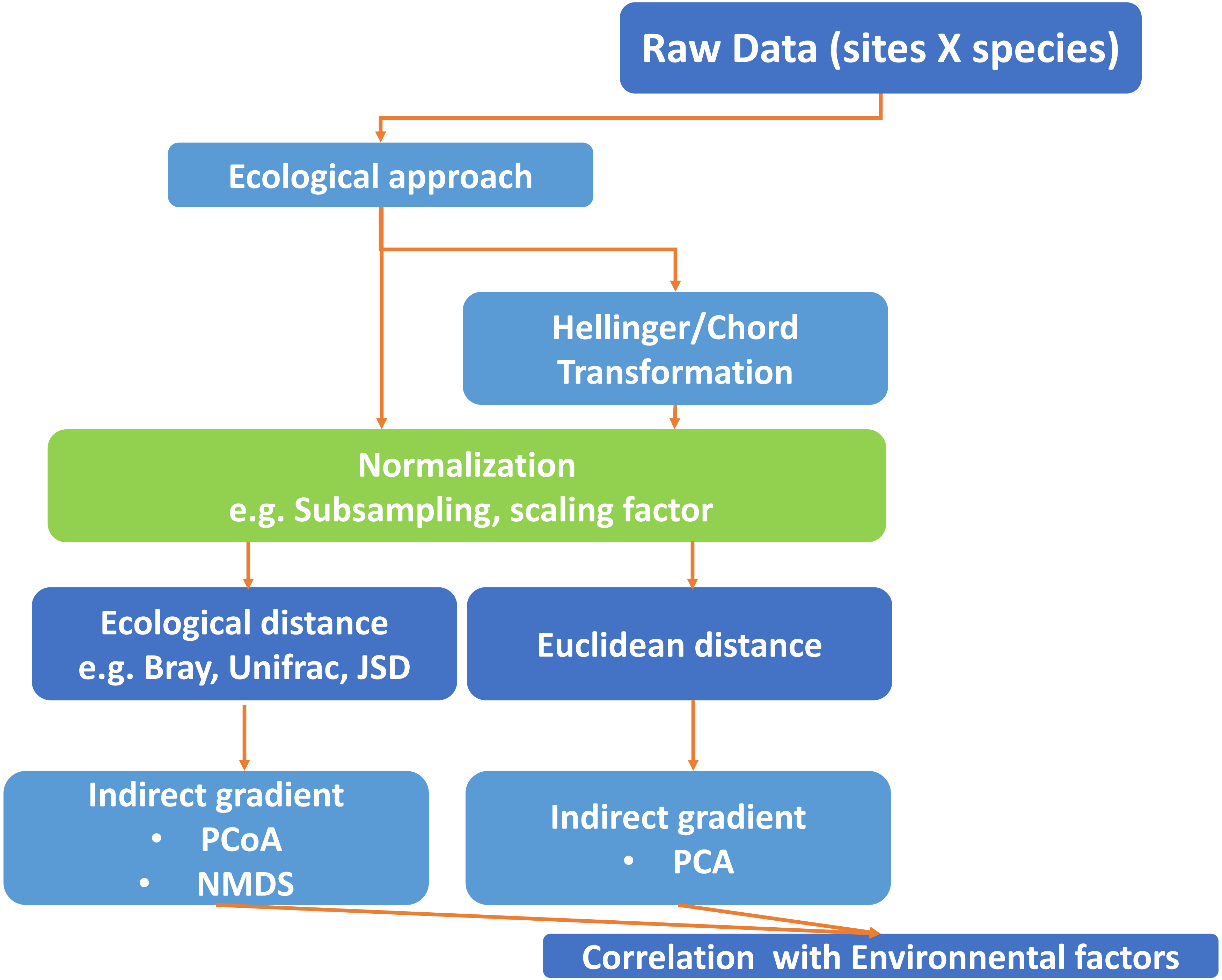
Indirect gradient
- PCA

(also NMDS)

Correlation  with Environmental factors

# Overview of the Beta-analysis approach





Raw Data (sites X species)

Ecological approach

Hellinger/Chord Transformation

Normalization
e.g. Subsampling, scaling factor

Ecological distance
e.g. Bray, Unifrac, JSD

Euclidean distance

Indirect gradient

- PCoA
- NMDS

Indirect gradient

- PCA

Correlation with Environnental factors

# Distance matrix

## Similarity & Distance: Evaluate the ecological resemblance

**Find** metrics (i.e. indices) that describe how similar samples/sites/species might be is the first step for multivariate analysis!!

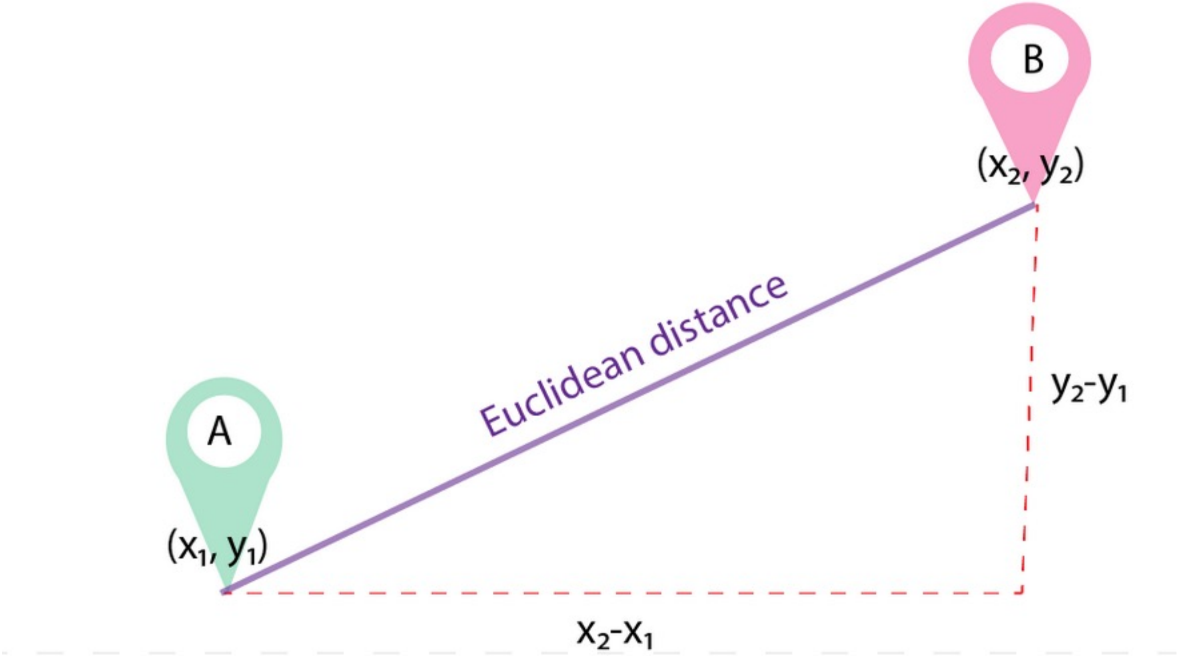
includes similarities (S) and dissimilarities (or distances)

- **Similarity decreases with the differences in species composition**
- **Multivariate analysis operates with distances (e.g. D= 1-S)**

# Distance  matrix versus Dissimilarity matrix

- What is a distance (e.g. Euclidean)?
  - D1: d(i,j) >= 0
  - D2: d(i,i) = 0
  - D3: d(i,j) = d(j,i)
  - D4: d(i,j) <= d(i,h) + d(h,j) (triangle inequality)

Not respected by
dissimilarity index (Bray)



# Distance matrix

## The global process: ASV Abundance to Distance to Ordination/Clustering of samples

Variables

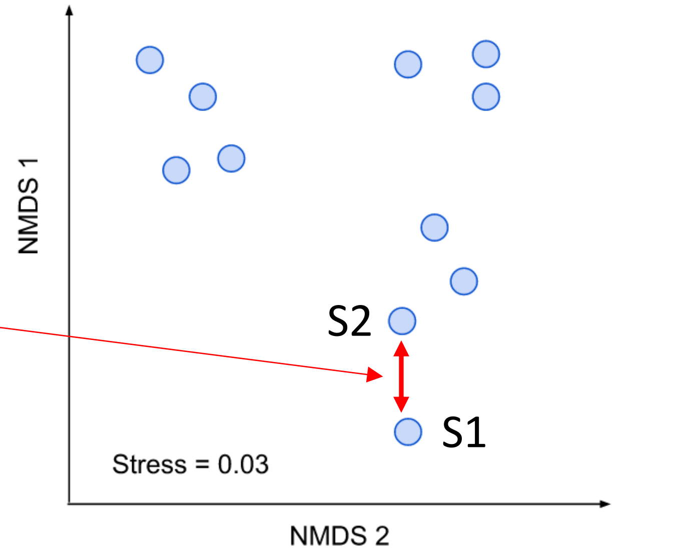
| Samples | X1 | X2 | X3 | X4 |
|---|---|---|---|---|
| S1 | 14 | 2 | 14 | 14 |
| S2 | 10 | 14 | 0 | 8 |
| S3 | 0 | 5 | 0 | 2 |
| S4 | 0 | 0 | 1 | 0 |

Abundance Matrix
Contingency table
OUT/ASV table

Samples

| Samples | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | 0 | ... | ... | ... |
| S2 | 0.47 | 0 | ... | ... |
| S3 | 0.84 | 0.64 | 0 | ... |
| S4 | 0.96 | 1 | 1 | 0 |

Dissimilarity/Distance matrix



Ordination plot in a reduced dimensional space

= Coordinates in euclidean space

# Distance matrix

## Similarity : How do deal with Double-zeros? Co-absence

- Species composition data are **sparse matrix**, which means that it contains lot of **zeros, double zeros**
- Double zero" is a situation when **certain species are missing** in both compared community → similarity/distance will be next calculated!

| | Species A | Species B | Species C |
|---|---|---|---|
| Site 1 | 0 | 44 | **0** |
| Site 2 | 11 | 50 | **0** |

**Really absent ? Both ? Only one?**

**Does not say anything about ecological similarity or difference between both samples…**

# Distance matrix (case with no log ratio transformation)

## Similarity : How do deal with Double-zeros? Co-absence

You can not conclude about the relationship because of :

- **Dispersal limitation** (present in the ecosystem but not in sample), **Sampling fraction**
- **Depth sequencing bias** (rare)

- Recommendation is to use **dissimilarity indices or distance-based** method that do **not take into account the double zero as a resemblance!!!**

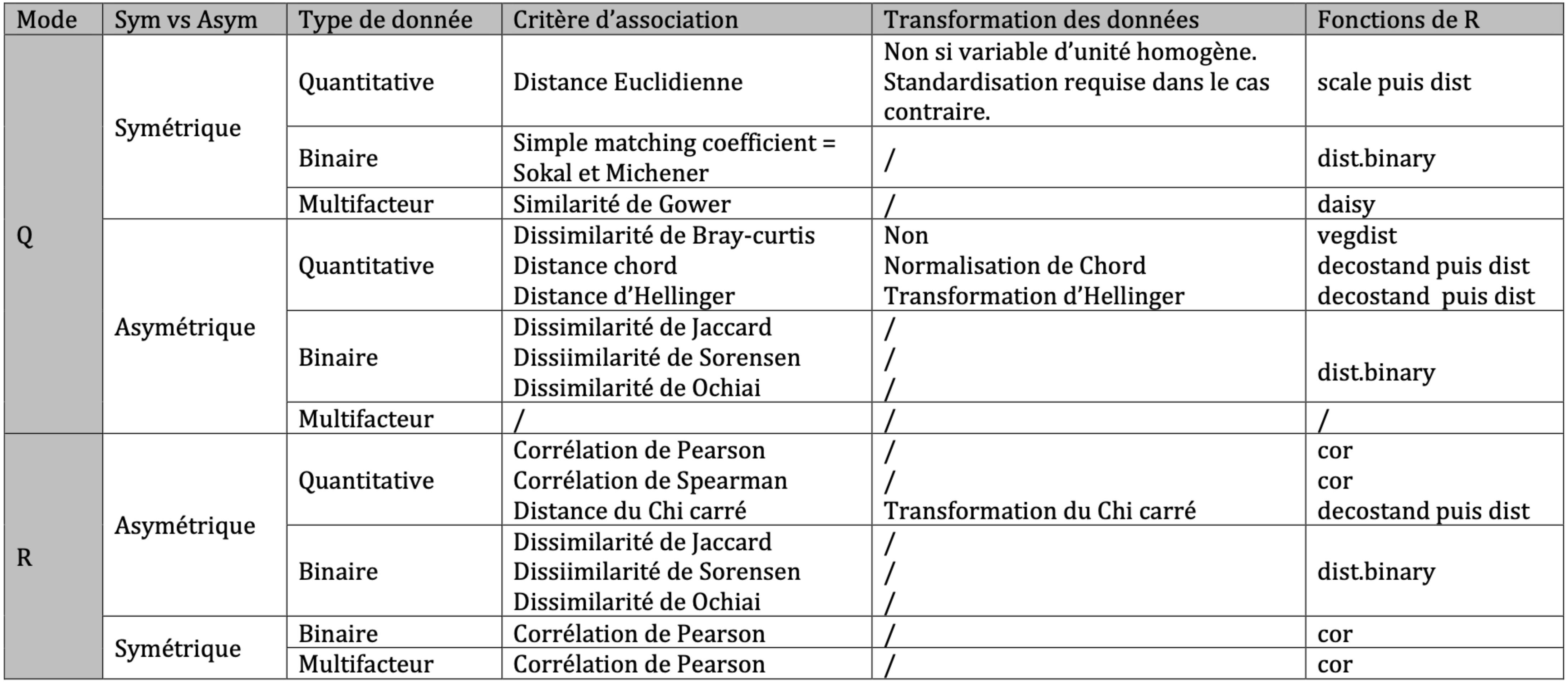
**Symmetrical vs. Asymmetrical indices**

- **Asymmetrical indices** ignore the double-zeros (e.g. bray-Curtis, Weighted Unifrac)
- **Symmetrical indices** consider the double-zeros as important (PCA!)! (e.g. Euclidian without transformation)

# Distance matrix

Three broad categories of **dissimilarity or distance index** :

- For binary data (presence/absence)
- For quantitative data (e.g. metabarcoding)
- For a mix of numerical and categorical data (multifactor)



| Mode | Sym vs Asym | Type de donnée | Critère d'association | Transformation des données | Fonctions de R |
|---|---|---|---|---|---|
| Q | Symétrique | Quantitative | Distance Euclidienne | Non si variable d'unité homogène. Standardisation requise dans le cas contraire. | scale puis dist |
| | | Binaire | Simple matching coefficient = Sokal et Michener | / | dist.binary |
| | | Multifacteur | Similarité de Gower | / | daisy |
| | Asymétrique | Quantitative | Dissimilarité de Bray-curtis<br>Distance chord<br>Distance d'Hellinger | Non<br>Normalisation de Chord<br>Transformation d'Hellinger | vegdist<br>decostand puis dist<br>decostand  puis dist |
| | | Binaire | Dissimilarité de Jaccard<br>Dissiimilarité de Sorensen<br>Dissimilarité de Ochiai | /<br>/<br>/ | dist.binary |
| | | Multifacteur | / | / | / |
| R | Asymétrique | Quantitative | Corrélation de Pearson<br>Corrélation de Spearman<br>Distance du Chi carré | /<br>/<br>Transformation du Chi carré | cor<br>cor<br>decostand puis dist |
| | | Binaire | Dissimilarité de Jaccard<br>Dissiimilarité de Sorensen<br>Dissimilarité de Ochiai | /<br>/<br>/ | dist.binary |
| | Symétrique | Binaire | Corrélation de Pearson | / | cor |
| | | Multifacteur | Corrélation de Pearson | / | cor |

# Distance matrix

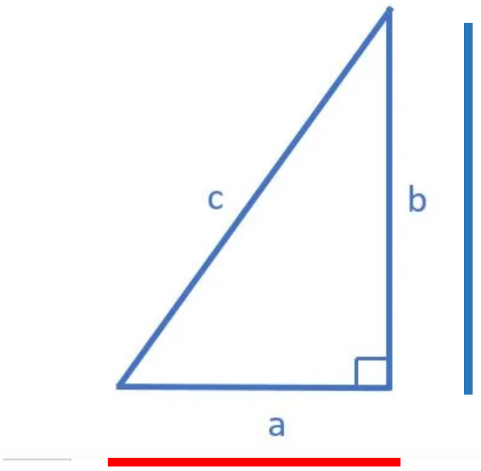
## Most common dissimilarities/distance used for species data

| Dissimilarities<br>Distances | Taxonomic | Phylogenetic |
|---|---|---|
| Compositional<br>(Binary) | Sorensen<br>Jaccard<br>Ochiai | Unweighted Unifrac<br>PhyloSor |
| Structural<br>(Quantitative) | Bray-Curtis<br>Chord<br>Hellinger<br>Aitchison<br>Euclidean | Weighted Unifrac<br>Allen |

# Distance matrix : you know it!

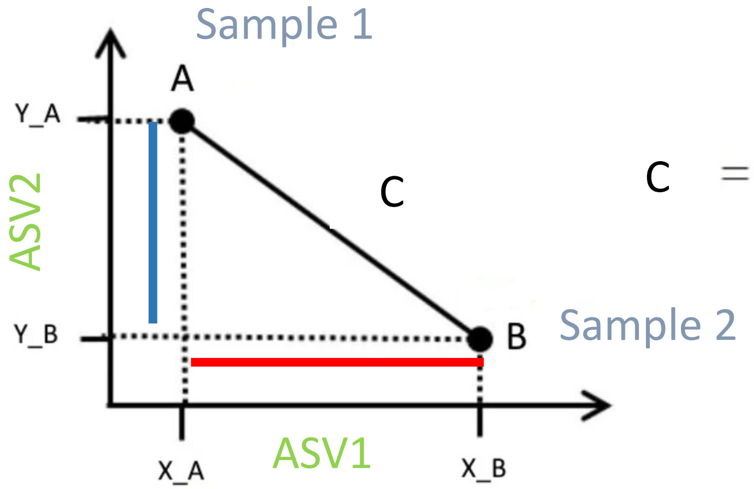
$$d(A, B) = \sqrt{(u_A - u_B)^2 + (v_A - v_B)^2 + \ldots (z_A - z_B)^2}$$

## Euclidean Distance



$c^2 = a^2 + b^2$

Pythagore

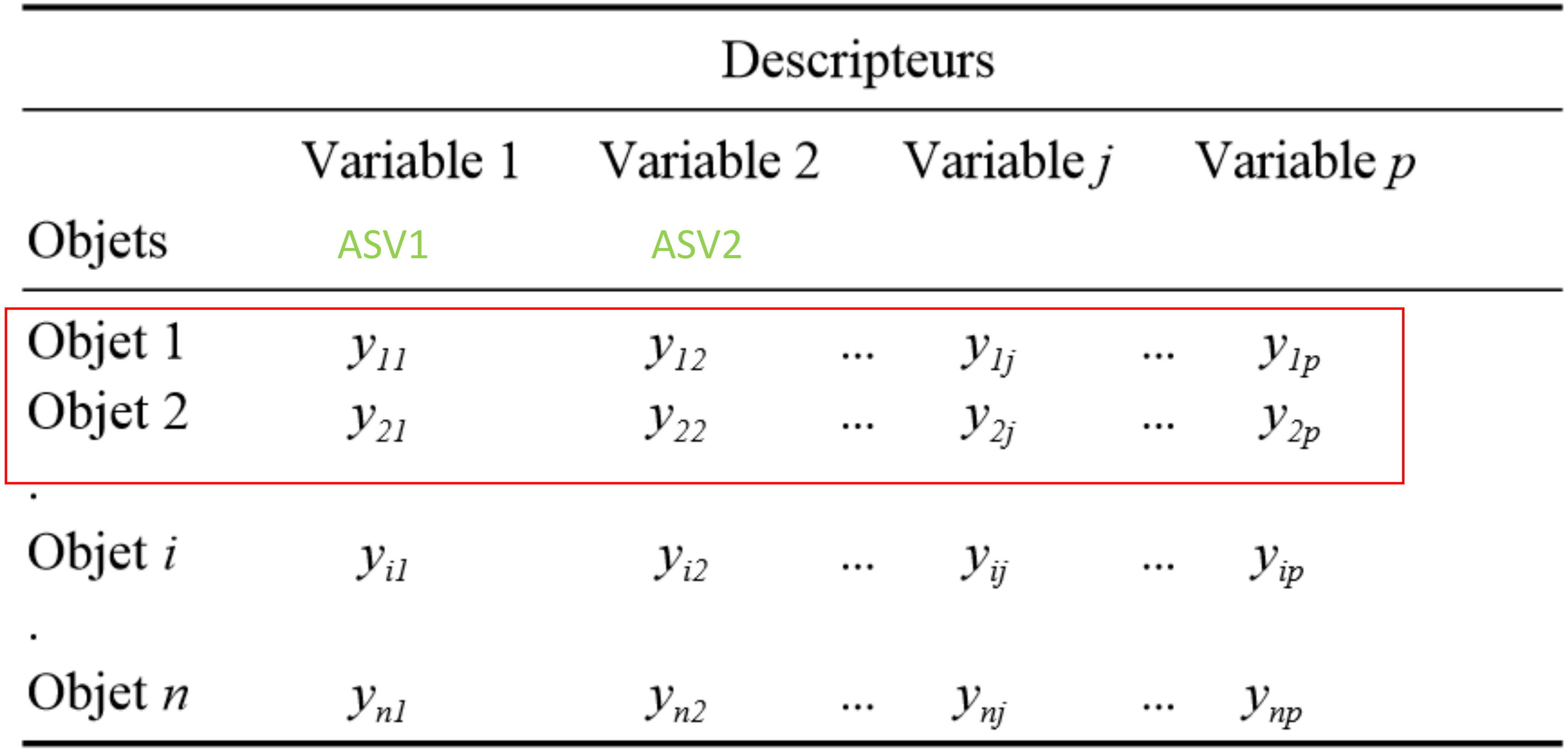


$$C = \sqrt{\underline{(x_B - x_A)^2} + \underline{(y_B - y_A)^2}}$$

**Generalization of Pythagore theorem**
**In space of 2 dimensions**

- Used by PCA

# For n dimensions…



|  |  | Descripteurs |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  | Variable 1 | Variable 2 |  | Variable $j$ |  | Variable $p$ |
|  | Objets | ASV1 | ASV2 |  |  |  |  |
| Site1 | Objet 1 | $y_{11}$ | $y_{12}$ | … | $y_{1j}$ | … | $y_{1p}$ |
| Site2 | Objet 2 | $y_{21}$ | $y_{22}$ | … | $y_{2j}$ | … | $y_{2p}$ |
|  | . |  |  |  |  |  |  |
|  | Objet $i$ | $y_{i1}$ | $y_{i2}$ | … | $y_{ij}$ | … | $y_{ip}$ |
|  | . |  |  |  |  |  |  |
|  | Objet $n$ | $y_{n1}$ | $y_{n2}$ | … | $y_{nj}$ | … | $y_{np}$ |

**Distance Site1-Site2 = ….**

# Distance matrix

## Hellinger distance (to do for PCA)

$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}}$$

Hellinger transformation

Squared root of proportions!

| | Sp1 | Sp2 | Sp3 |
|---|---|---|---|
| Com1 | 3 | 3 | 0 |
| Com2 | 4 | 4 | 2 |

→ Hellinger transformation

| | Sp1 | Sp2 | Sp3 |
|---|---|---|---|
| Com1 | 0.7 | 0.7 | 0 |
| Com2 | 0.6 | 0.6 | 0.4 |

→ Euclidean distance

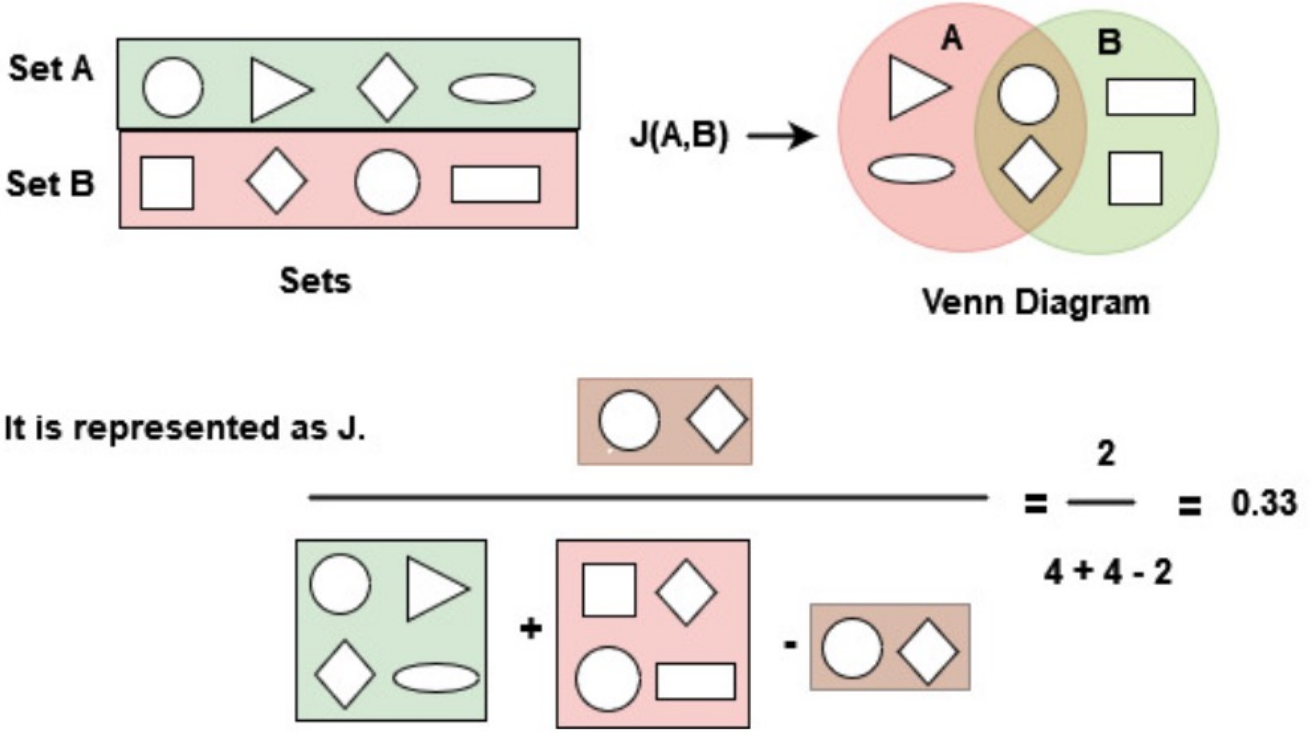
| | Com1 | Com2 |
|---|---|---|
| Com1 | 0 | 0.42 |
| Com2 | 0.42 | 0 |

- Particularly suited to species abundance data, this transformation gives low weights to variables with low counts and many zeros

- Reduce the effects of values that are extremely large

# Distance matrix : Dissimilarities (Binary = does not take into account the relative abundance!)

## Jaccard Similarity = Jaccard Index : measure of similarity!



$$\frac{\text{Intersection (A-B)}}{\text{Specific A + Specific B - Union}}$$

Jaccard Distance? 1-S = **0.67**

# Distance matrix

## Bray-Curtis dissimilarities

**The Bray-Curtis dissimilarity assumes that the two sites are of equal size!!**

I and J are Sites

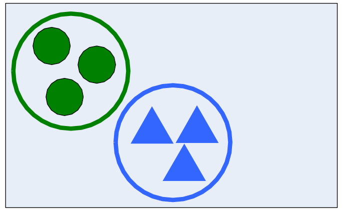
$$BC_{ij} = 1 - (2^*C_{ij}) / (S_i + S_j)$$

D = 1- S

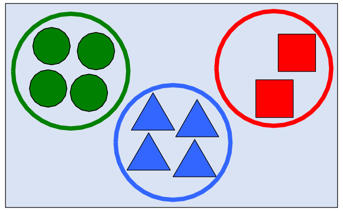
- **$C_{ij}$:** The sum of the lesser values for each species
- **$S_i$:** The total number of specimens counted at site *i*
- **$S_j$:** The total number of specimens counted at site *j*

**BC : is a value range from 0 to 1**
**0 is the maximum similarity**
**Same sampling depth**

## Community 1



## Community 2



**Count of Species**

| | A | B | C | D | E |
|---|---|---|---|---|---|
| **Site 1** | 4 | 0 | 2 | 7 | 8 |
| **Site 2** | 3 | 6 | 0 | 4 | 11 |

The minimum for each species

$C_{ij} = 3 + 0 + 0 + 4 + 8 = \mathbf{15}$

$S_i = 4 + 0 + 2 + 7 + 8 = \mathbf{21}$

$S_j = 3 + 6 + 0 + 4 + 11 = \mathbf{24}$

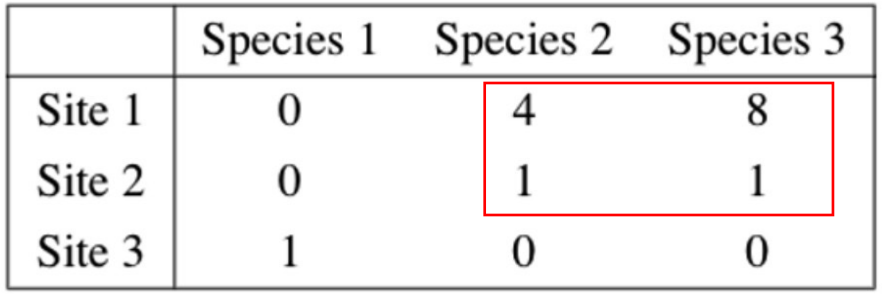
Total Count of species by site

- $BC_{ij} = 1 – (2^*C_{ij}) / (S_i + S_j)$
- $BC_{ij} = 1 – (2^*15) / (21 + 24)$
- $BC_{ij} = 0.33$

# Distance matrix

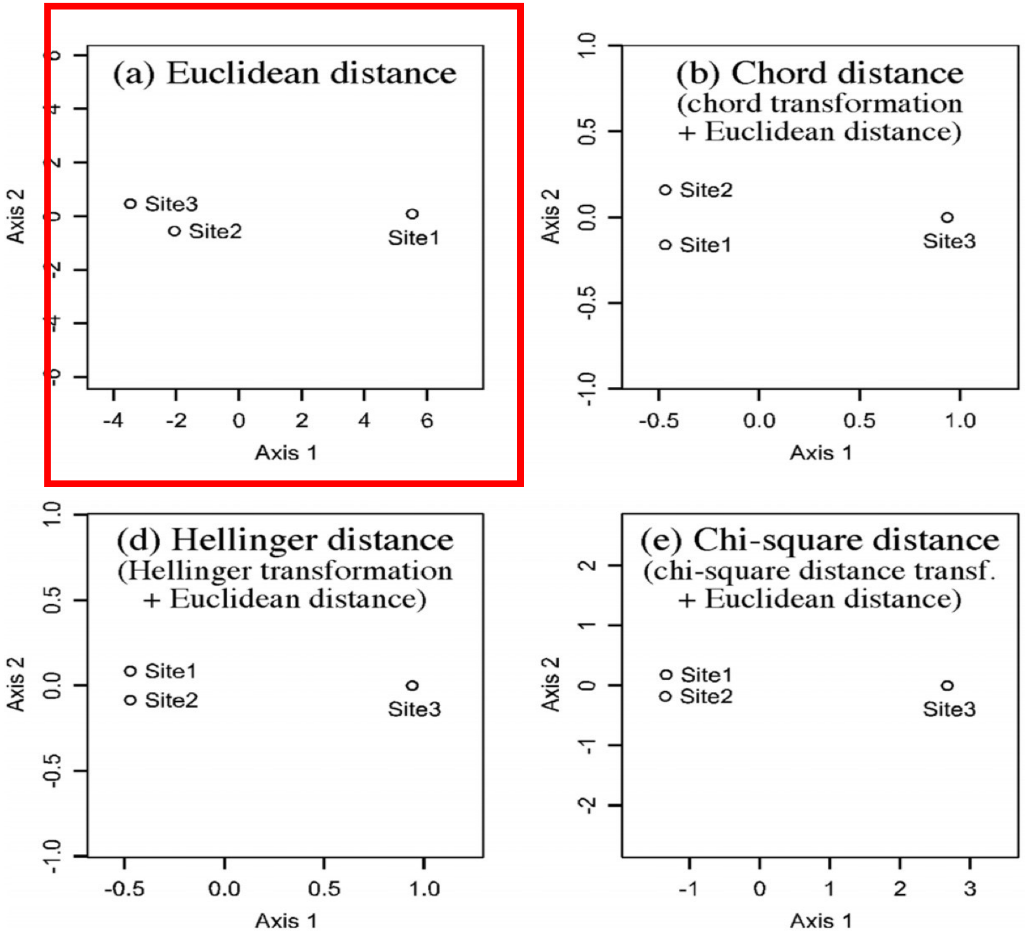
## Choose the right distance/dissimilarity

Species abundance paradox data ⇒
(3 sites, 3 species)



| | Species 1 | Species 2 | Species 3 |
|---|---|---|---|
| Site 1 | 0 | 4 | 8 |
| Site 2 | 0 | 1 | 1 |
| Site 3 | 1 | 0 | 0 |

It’s clear that Site1 and Site2 are more similar … but



Without any transformation of data (i.e. Hellinger/Chord), Euclidean distance not appropriate for ecological data

# Distance matrix

$$u = \frac{\sum_{i=1}^{N} l_i |A_i - B_i|}{\sum_{i=1}^{N} l_i \max(A_i, B_i)}$$

**UNIFRAC:** Comparison of microbial communities using phylogenetic information

Measure the difference between the composition of communities from diverse environments using **phylogenetic distance** by :

- Estimate the proportion of **branch length** unique to an environment
- **Unique vs. Shared**

Two modes :

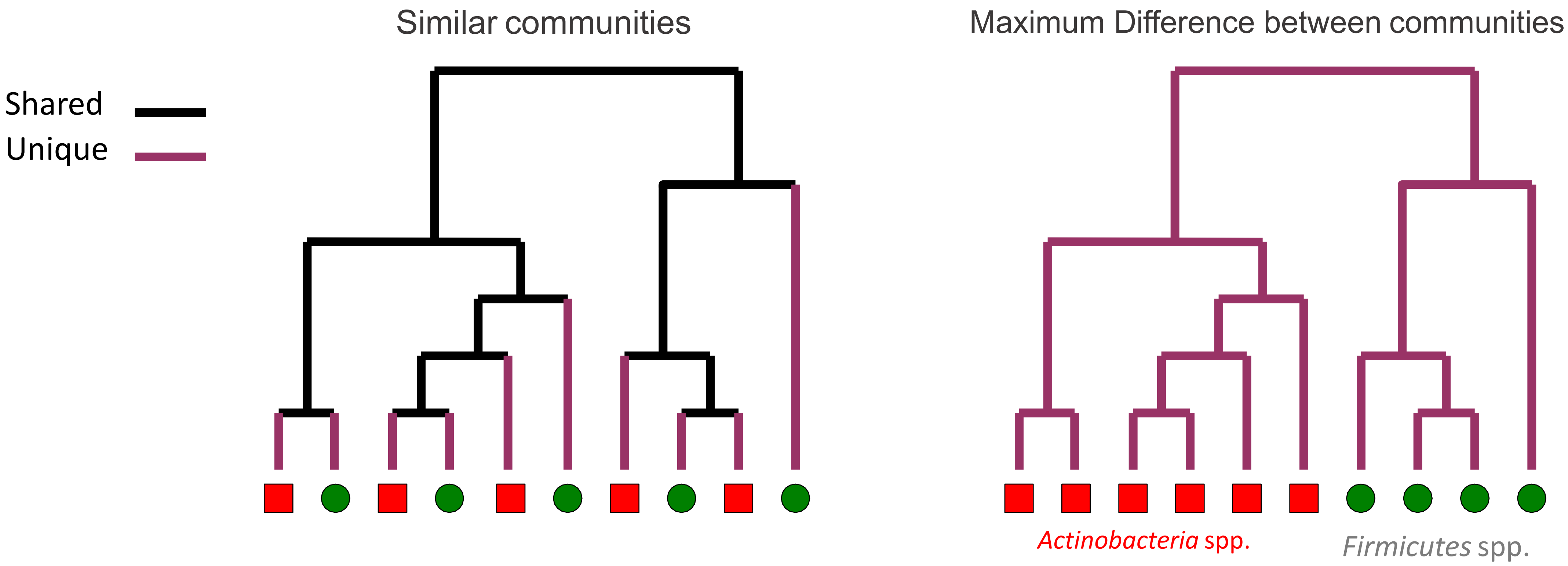
**Unweighted Unifrac**
**Weighted Unifrac** (takes into account the relative abundance of taxa)

# Distance matrix

## Unweighted Unifrac

🟥 ENV A

🟢 ENV B

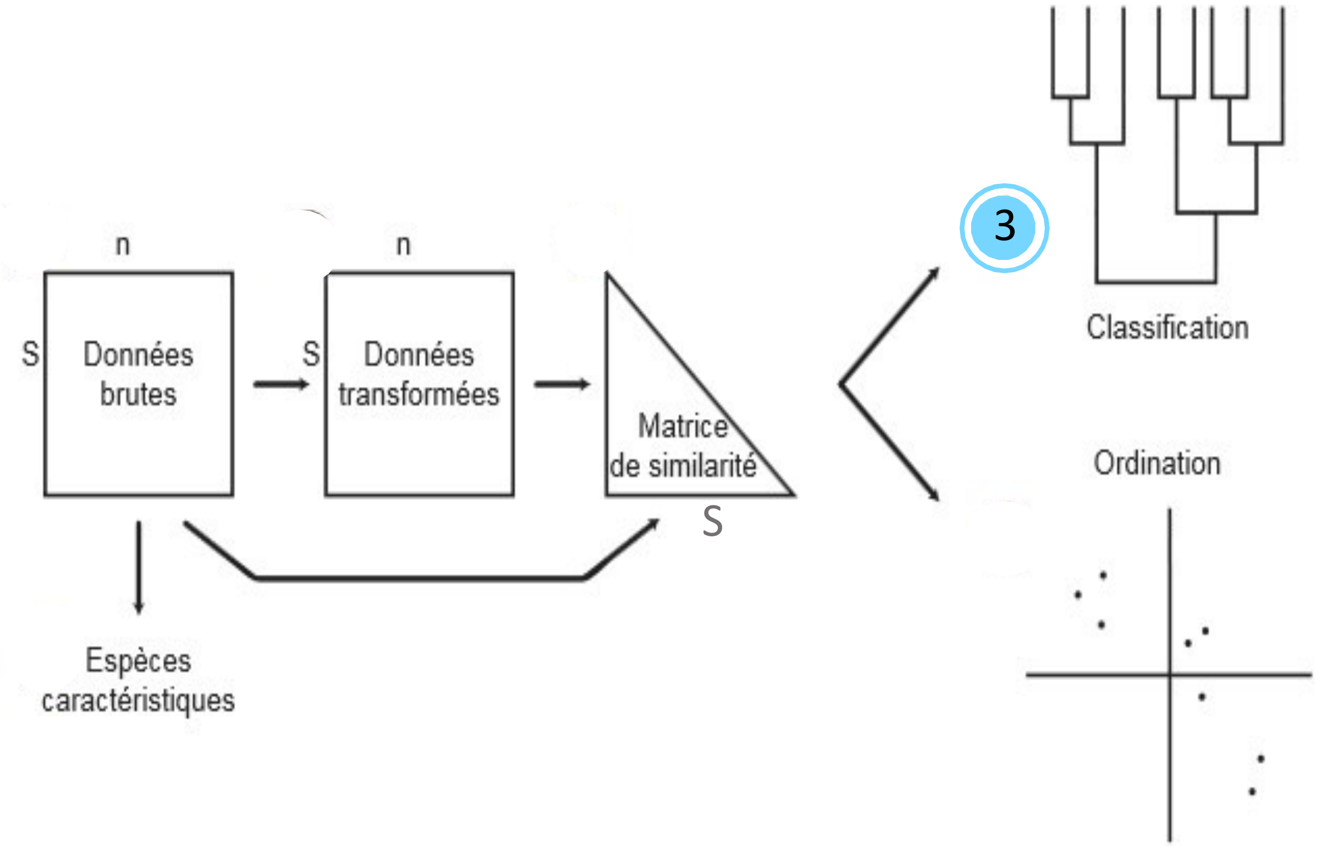


Distance Measure of UniFrac = (-━━ ) / (━━ + ━━)

UniFrac measures the amount of evolutionary divergence between two communities by dividing the length of the **Uniq branches** by the total branch length of the tree.

# Overview of the Beta-analysis approach

Classification

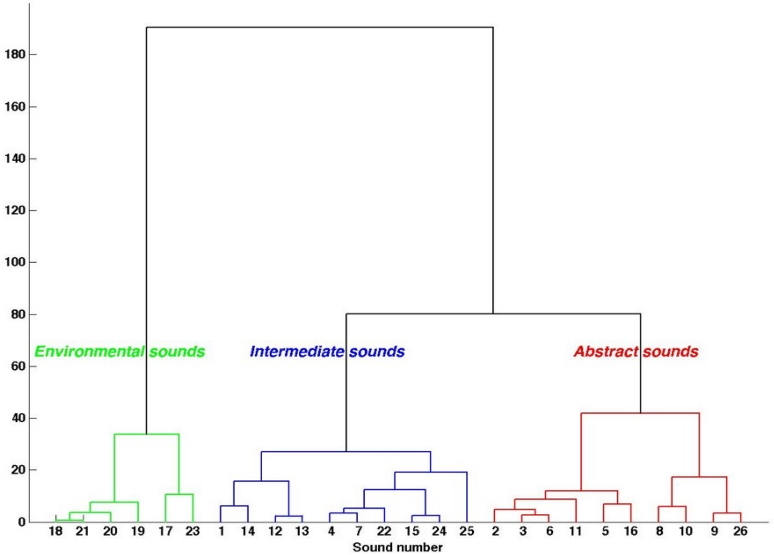


# Classification methods : Clustering Analysis

- Group objects (sites, communities) that are similar
- The final result is a dendrogram that can be very different depending on:

1) the similarity or dissimilarity criterion used to calculate the distance matrix

2) the aggregation/clustering criterion chosen for the partitions formed

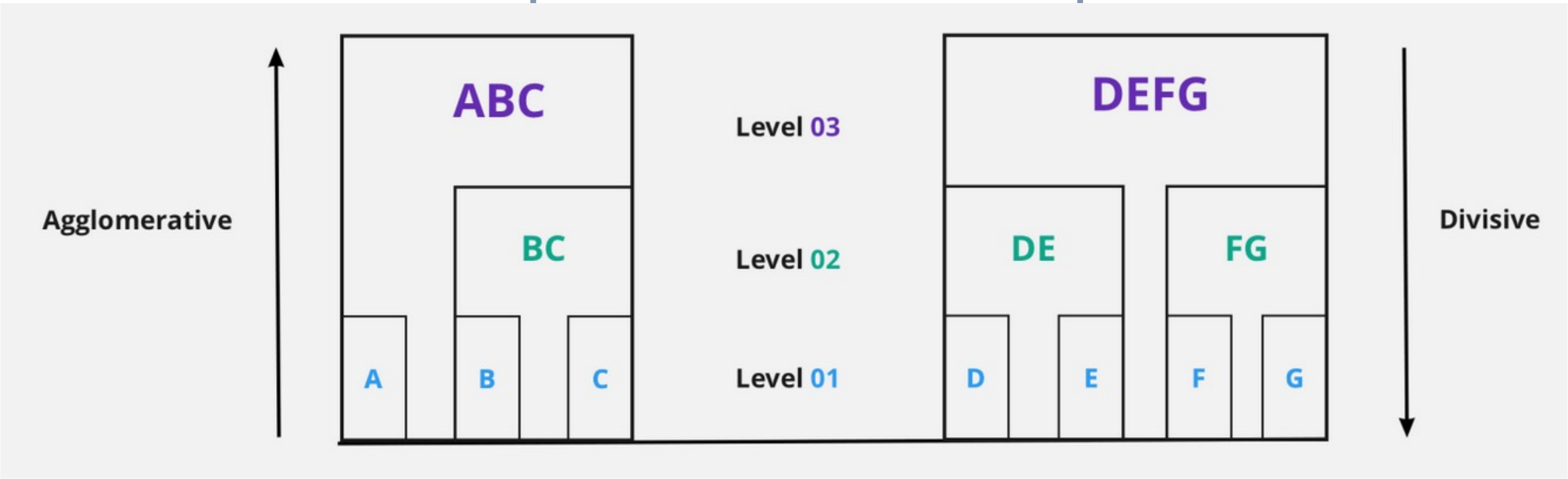


# Hierarchical Clustering Analysis

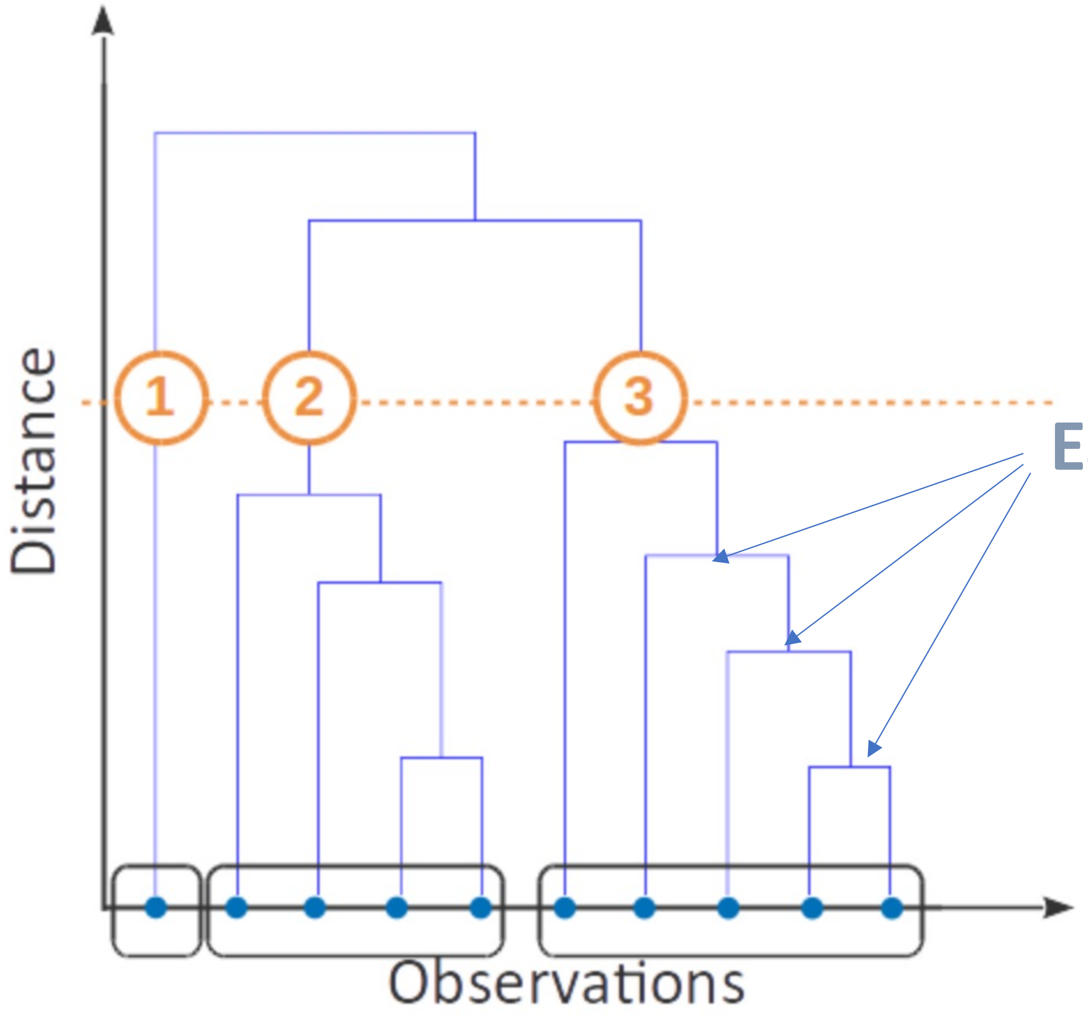
## Unsupervised method

Agglomerative HC
Bottom -> Top

Divisive HC
Top -> bottom

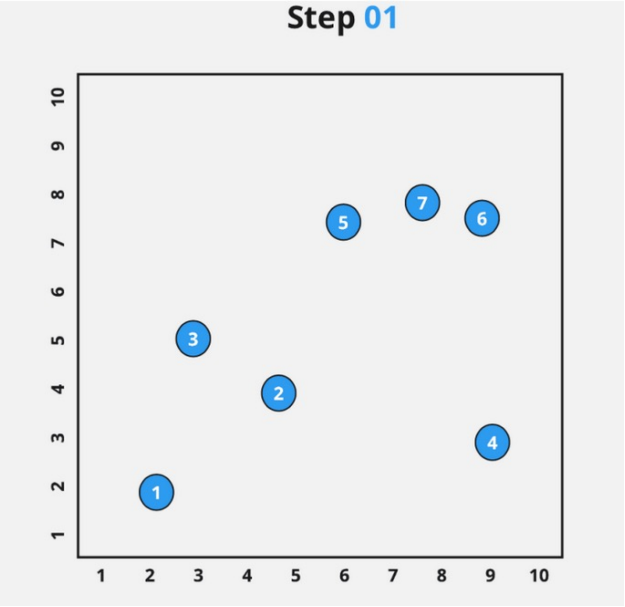


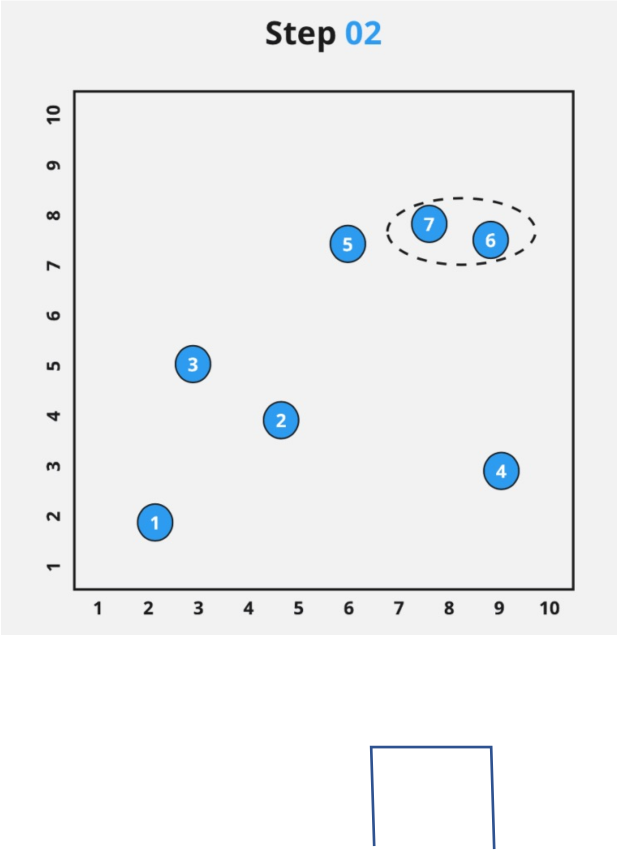
# Dendrogram

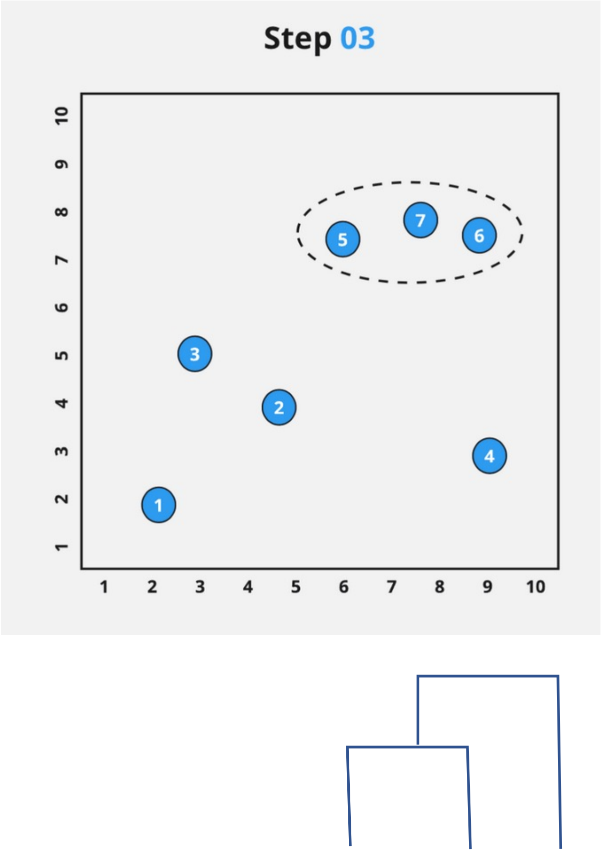


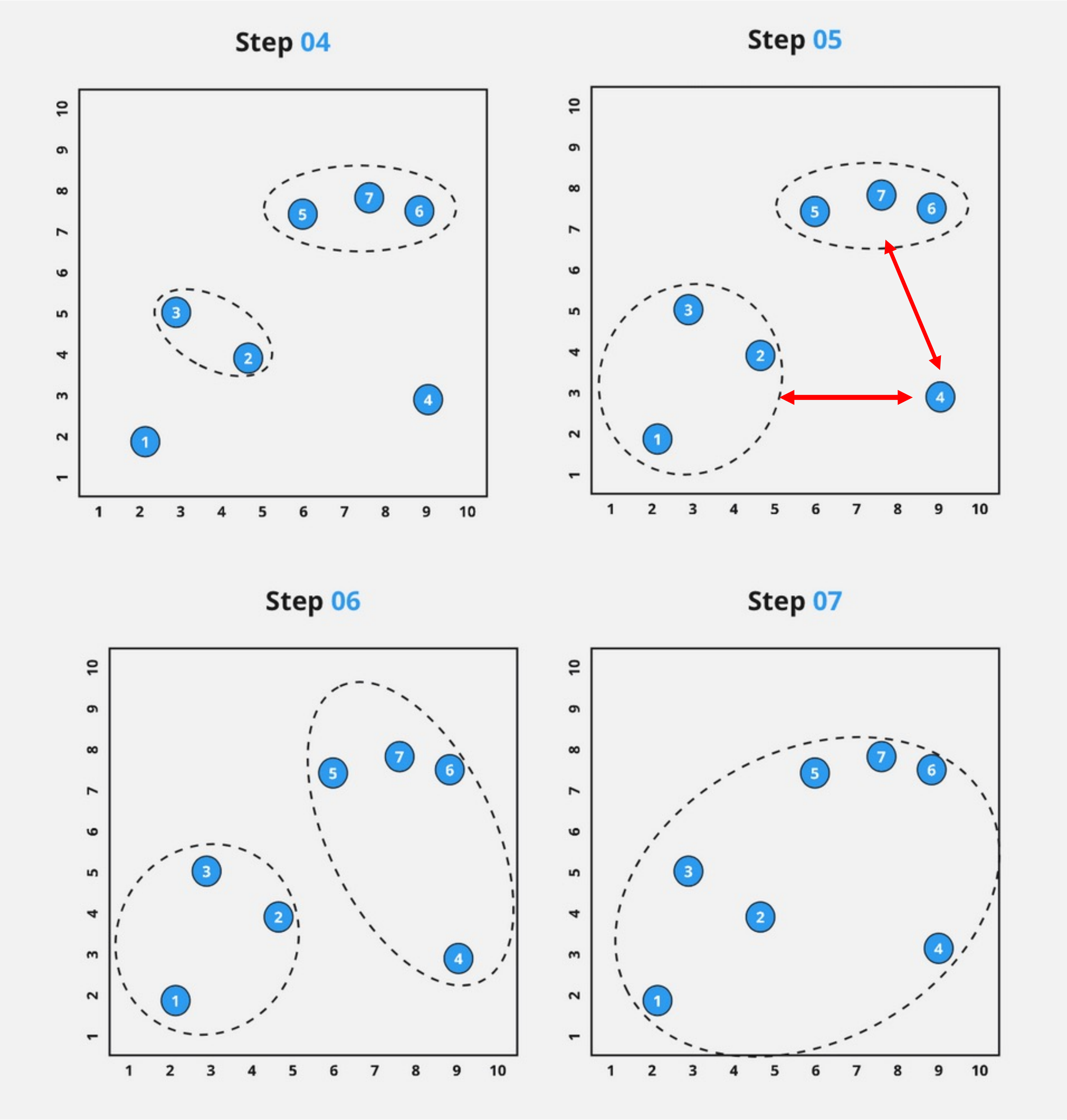
Each node is a cluster

# Iterative : Find the closest objects and clustering them





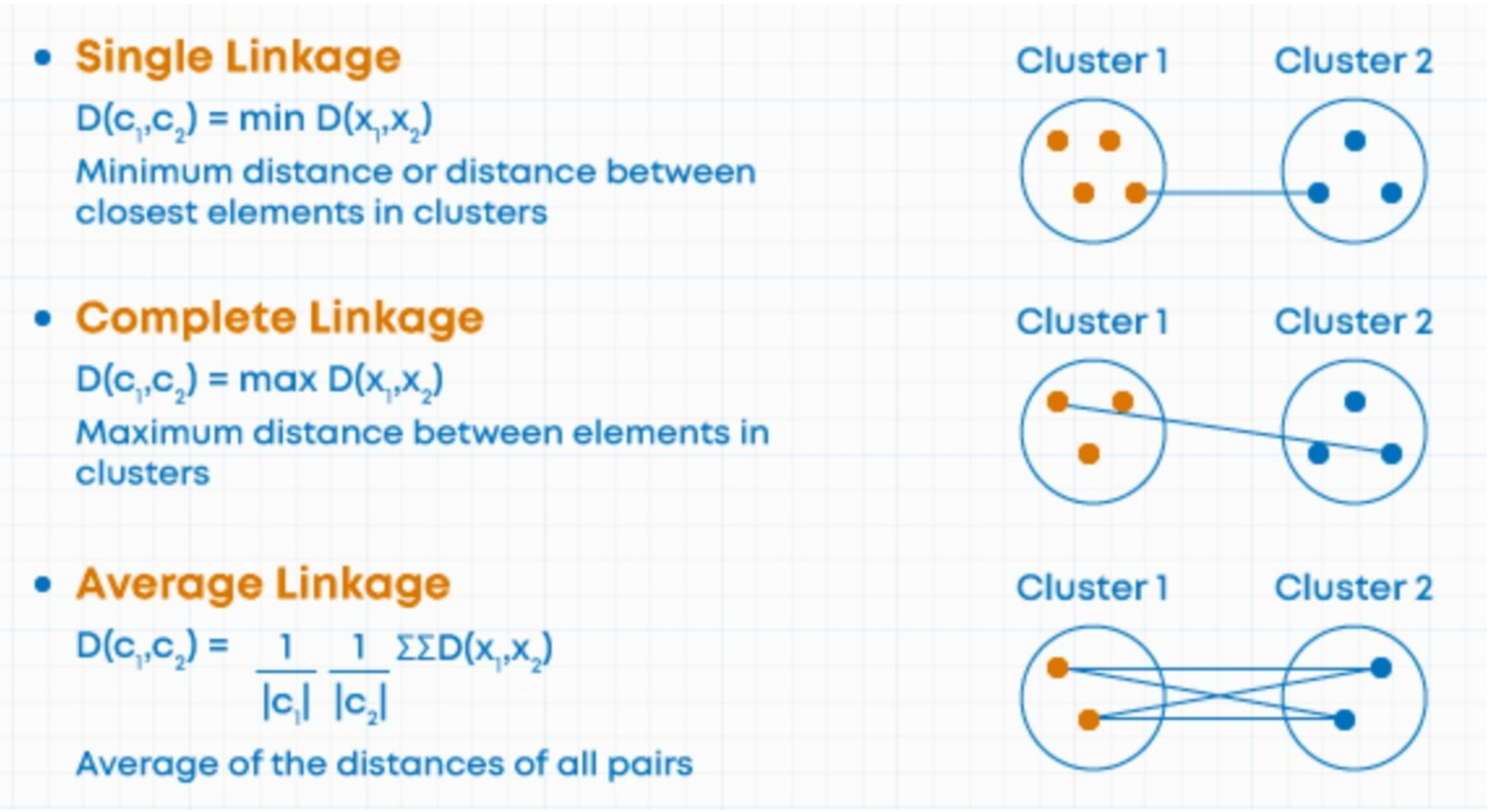




**Belong to which cluster the 4?**
**Rules to agglomerate?**

- The nearest?
- The farest?
- The average?
- …

# Distance between clusters : Rules that define the way for clustering



- **Single Linkage**

  $D(c_1,c_2) = \min D(x_1,x_2)$

  Minimum distance or distance between closest elements in clusters

- **Complete Linkage**

  $D(c_1,c_2) = \max D(x_1,x_2)$

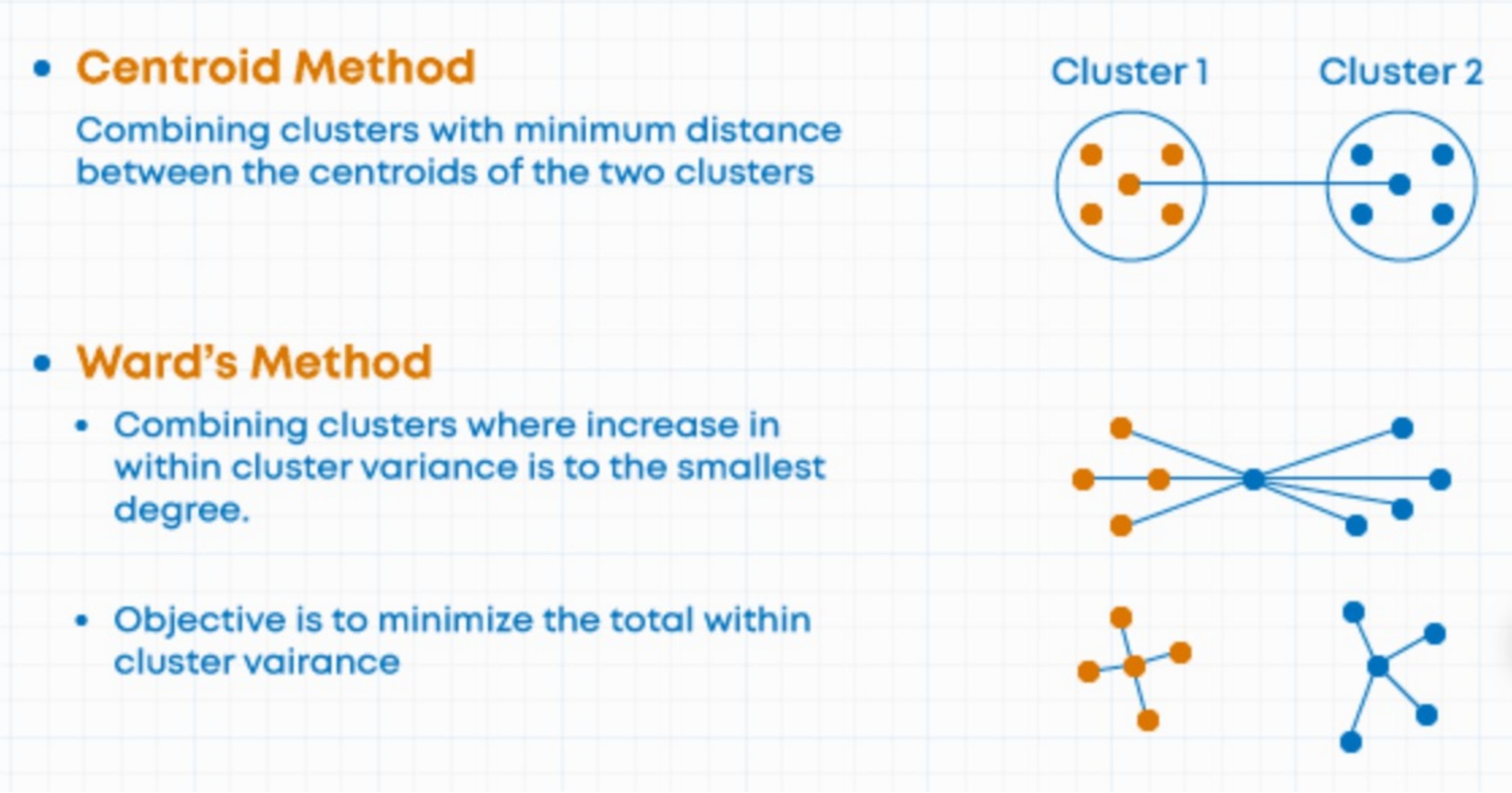
  Maximum distance between elements in clusters

- **Average Linkage**

  $$D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\Sigma\Sigma D(x_1,x_2)$$

  Average of the distances of all pairs

# Distance between clusters :
# Rules that define the way for clustering

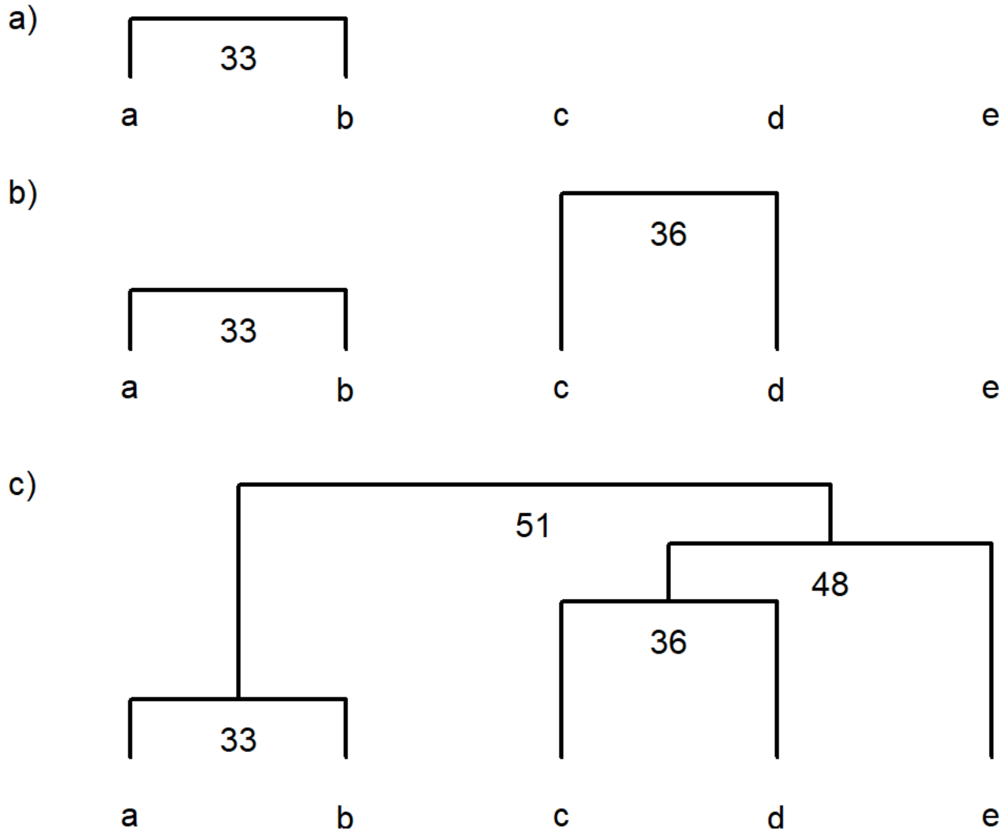


- **Centroid Method**

  Combining clusters with minimum distance between the centroids of the two clusters

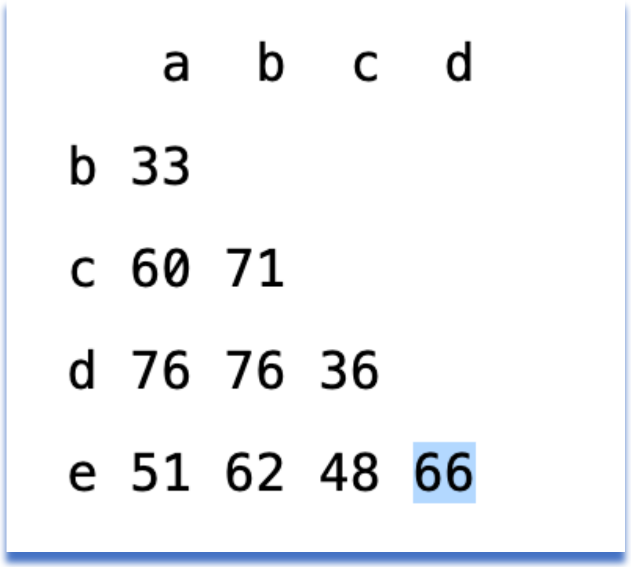
- **Ward's Method**
  - Combining clusters where increase in within cluster variance is to the smallest degree.
  - Objective is to minimize the total within cluster vairance

# To Understand : Single Linkage Example



Clustering (Single Linkage)

## Original Distance Matrix



```
     a  b  c  d
b 33
c 60 71
d 76 76 36
e 51 62 48 66
```

|   | a | b | c | d |
|---|---|---|---|---|
| b | 33 |  |  |  |
| c | 51 |  |  |  |
| d | 51 | 51 | 36 |  |
| e | 51 | 51 | 48 | 48 |

Clustering/cophenetic matrix

= distance between clusters!

# Cophenetic correlation coefficient : How good is the clustering?

## Classification methods modify the original distances

**Cophenetic** distance matrix

|  | Obj1 | Obj2 |
|---|---|---|
| Obj1 |  |  |
| Obj2 |  |  |

vs

**Original** distance matrix

|  | Obj1 | Obj2 |
|---|---|---|
| Obj1 |  |  |
| Obj2 |  |  |

**Pearson Correlation**

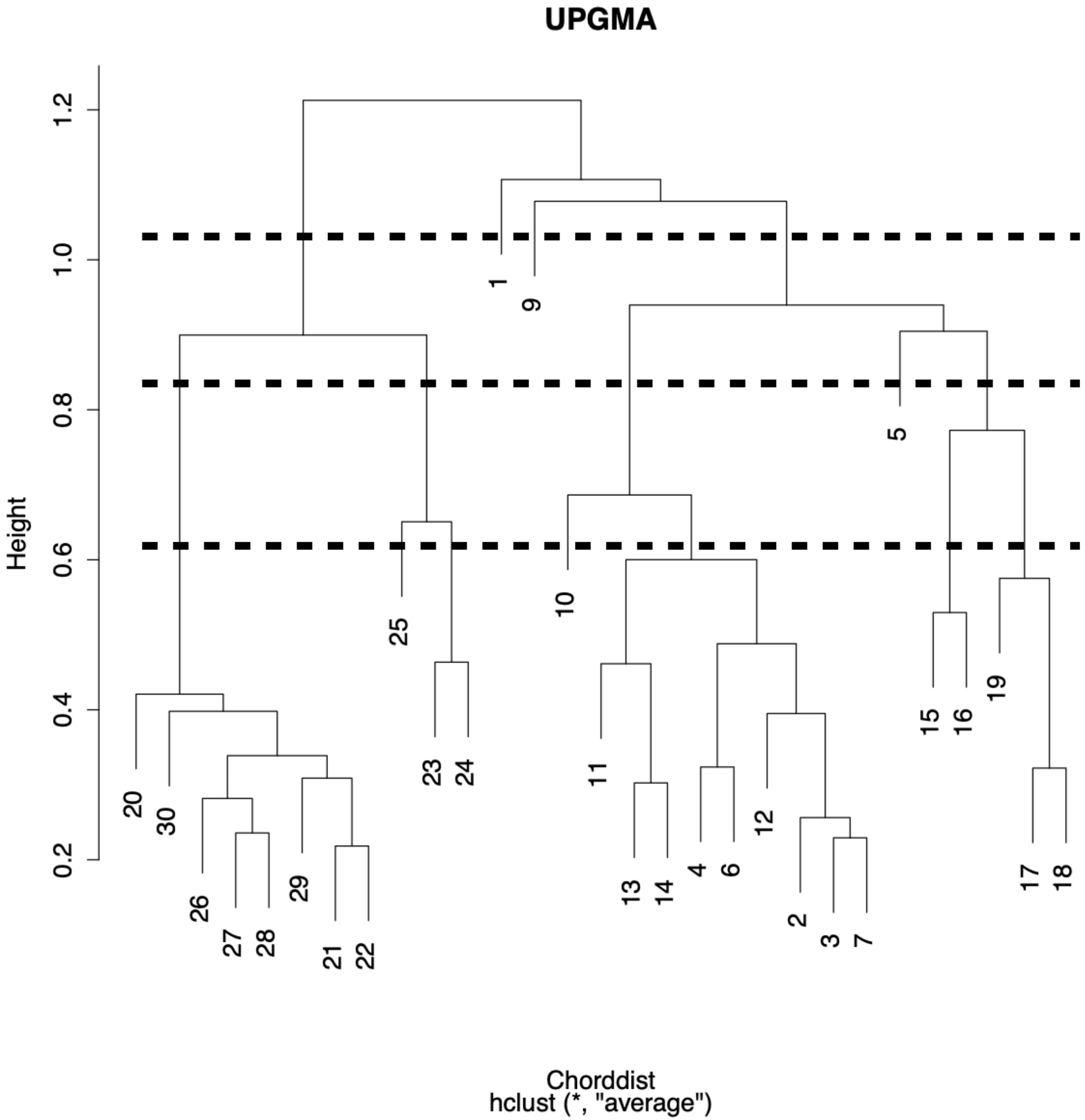
**The more corelated is, the best representation you have!**

# Looking for Interpretable Clusters

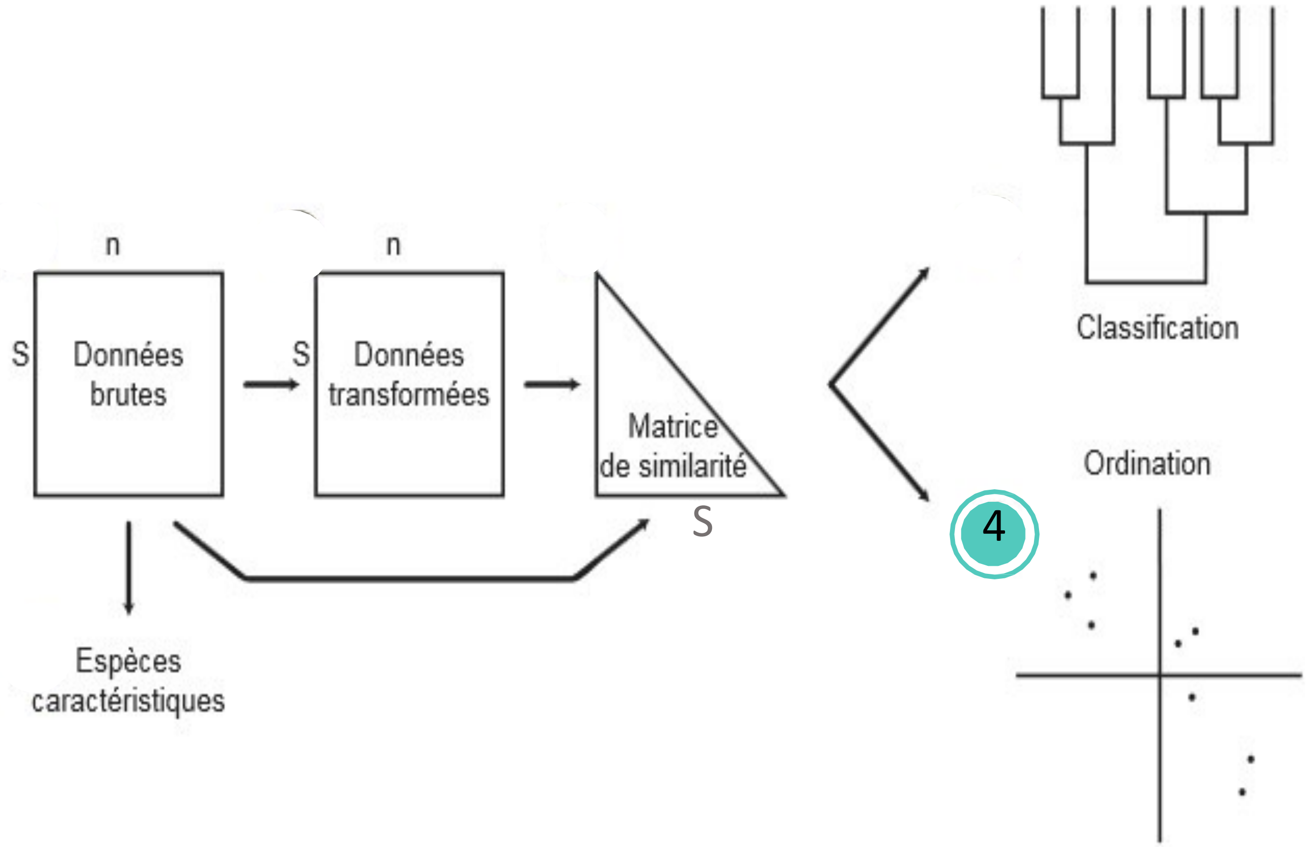
A decision must be made: at what level should the dendrogram be cut?

Many indices (more than 30) has been published in the literature for finding the right number of clusters in a dataset.

→ TP Use NbClust R



# Overview of the Beta-analysis approach

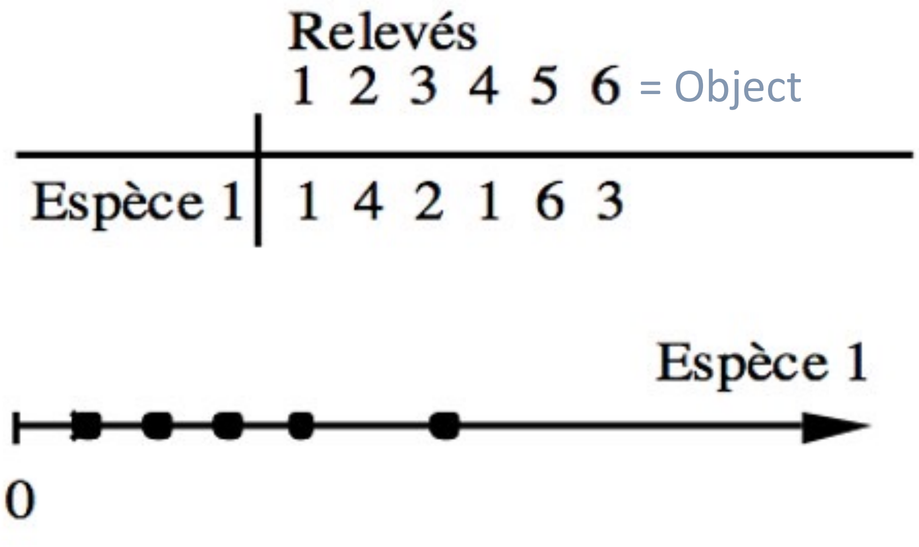


# Ordination : The meaning of this approach?

<u>Objective</u> : Represent relationships between **Objects** and **Variables** in a **reduce space…**

- **Let see why!!!**

## Unidimensional Data



Relevés

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Espèce 1 | 1 | 4 | 2 | 1 | 6 | 3 |

= Object

Data Table

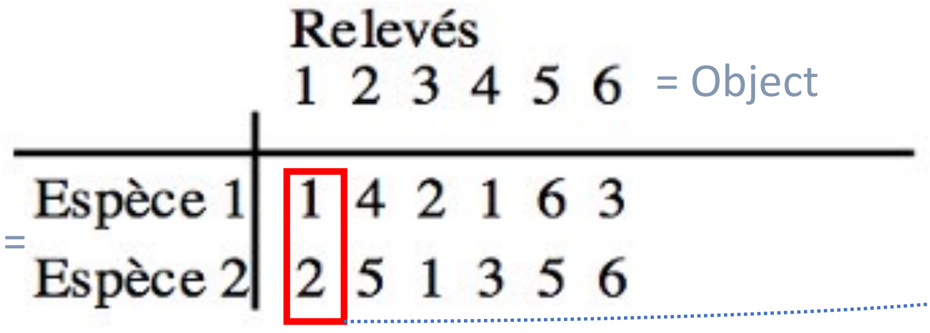
Graph Representation

**I CAN DO IT!**

# Ordination : The meaning of this approach?

## Bidimensional Data

Data Table

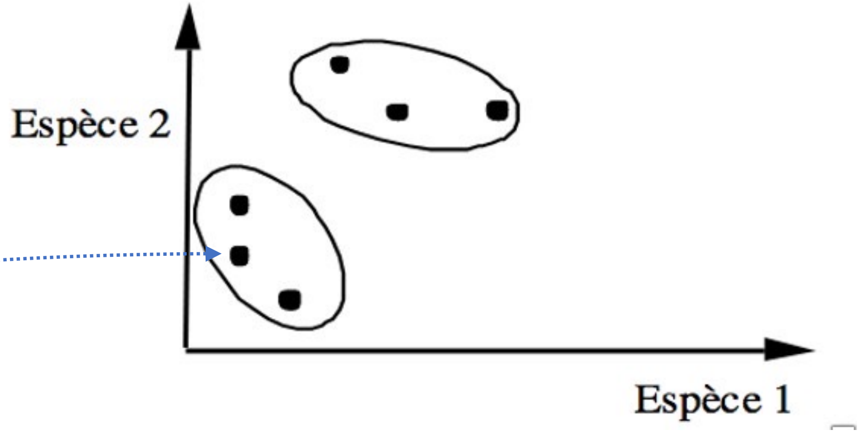


|  | Relevés 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Espèce 1 | 1 | 4 | 2 | 1 | 6 | 3 |
| Espèce 2 | 2 | 5 | 1 | 3 | 5 | 6 |

= Object

Variables =

Graph Representation



- Coordinates of Relevé 1 are (x,y)=(1,2)
- Coordinates of Relevé 2 are (x,y)=(4,5)

…

…

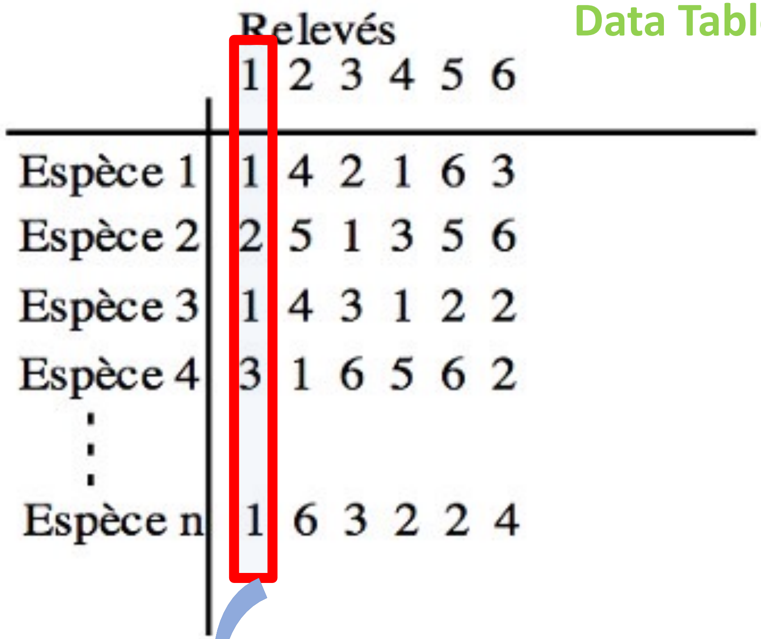
…

I CAN DO IT!

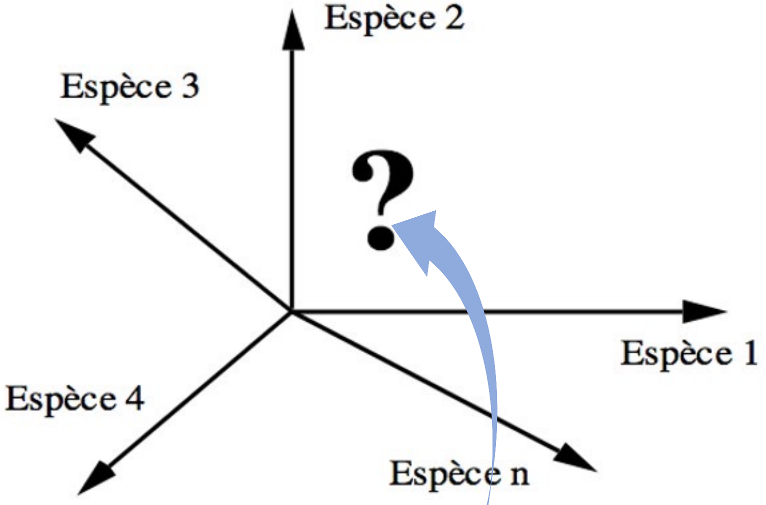
# Ordination : The meaning of this approach?

## Multidimensional Data (e.g. Metabarcoding)

Data Table



|  | Relevés |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  | 1 | 2 | 3 | 4 | 5 | 6 |
| Espèce 1 | 1 | 4 | 2 | 1 | 6 | 3 |
| Espèce 2 | 2 | 5 | 1 | 3 | 5 | 6 |
| Espèce 3 | 1 | 4 | 3 | 1 | 2 | 2 |
| Espèce 4 | 3 | 1 | 6 | 5 | 6 | 2 |
| ⋮ |  |  |  |  |  |  |
| Espèce n | 1 | 6 | 3 | 2 | 2 | 4 |

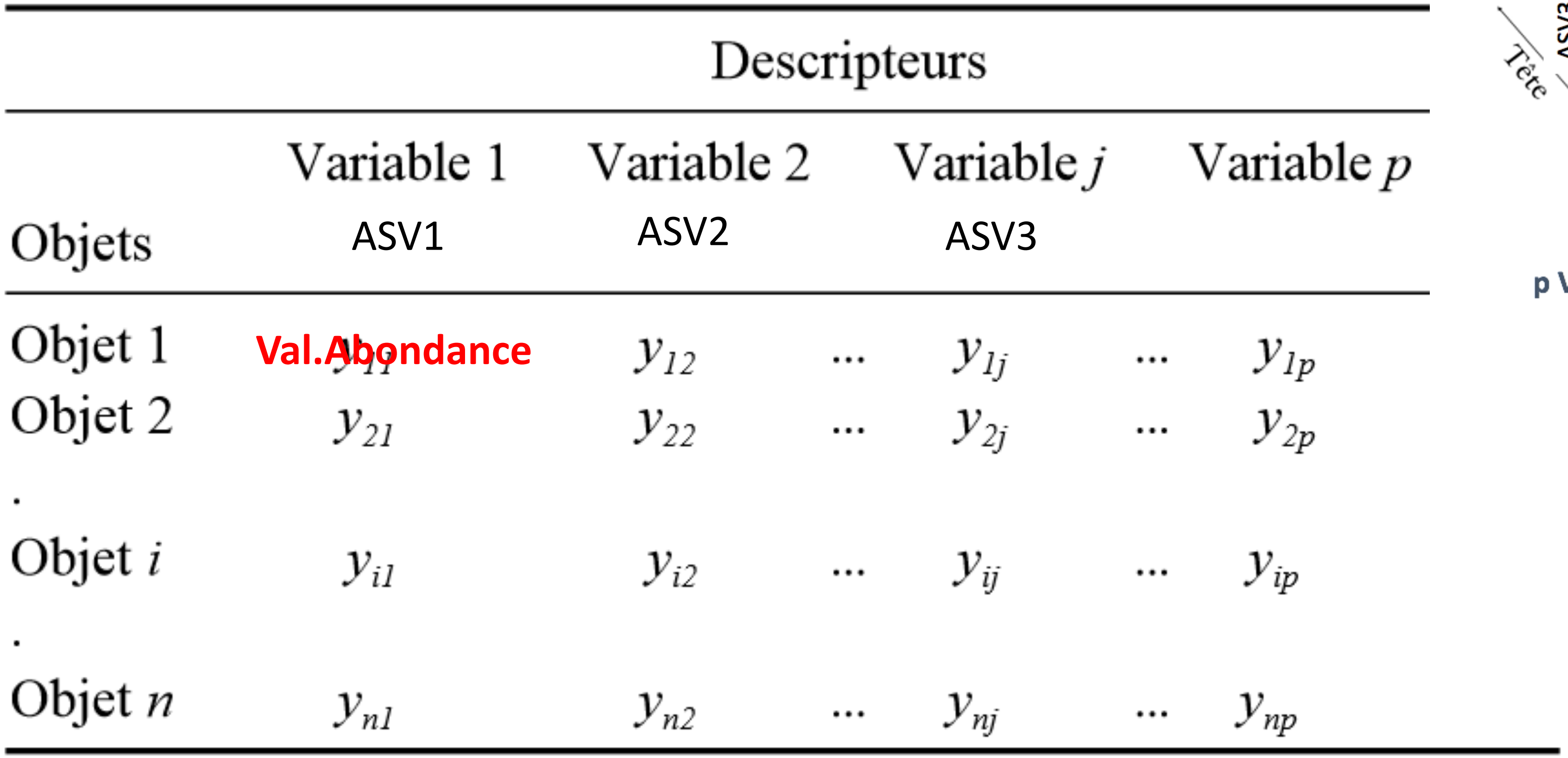


Coordinates Relevé1 are(x,y,z, …)= (1,2,1,3 …1)

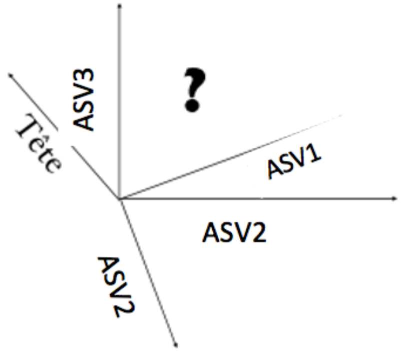
**Impossible** to graphically display all the axes! ☹

# Resume … How to visualize data in more than 3 dimensions ??

## Metabarcoding



| | Descripteurs | | | | | |
|---|---|---|---|---|---|---|
| Objets | Variable 1<br>ASV1 | Variable 2<br>ASV2 | | Variable *j*<br>ASV3 | | Variable *p* |
| Objet 1 | Val.Abondance $y_{11}$ | $y_{12}$ | … | $y_{1j}$ | … | $y_{1p}$ |
| Objet 2 | $y_{21}$ | $y_{22}$ | … | $y_{2j}$ | … | $y_{2p}$ |
| . | | | | | | |
| Objet *i* | $y_{i1}$ | $y_{i2}$ | … | $y_{ij}$ | … | $y_{ip}$ |
| . | | | | | | |
| Objet *n* | $y_{n1}$ | $y_{n2}$ | … | $y_{nj}$ | … | $y_{np}$ |



p Variables

Variables = Descriptors (Taxa/ASV)
Objets = Observations (Site, Stations)

→ Plot? Need a number of axes equal to the number of Descriptors!!!!! ☹

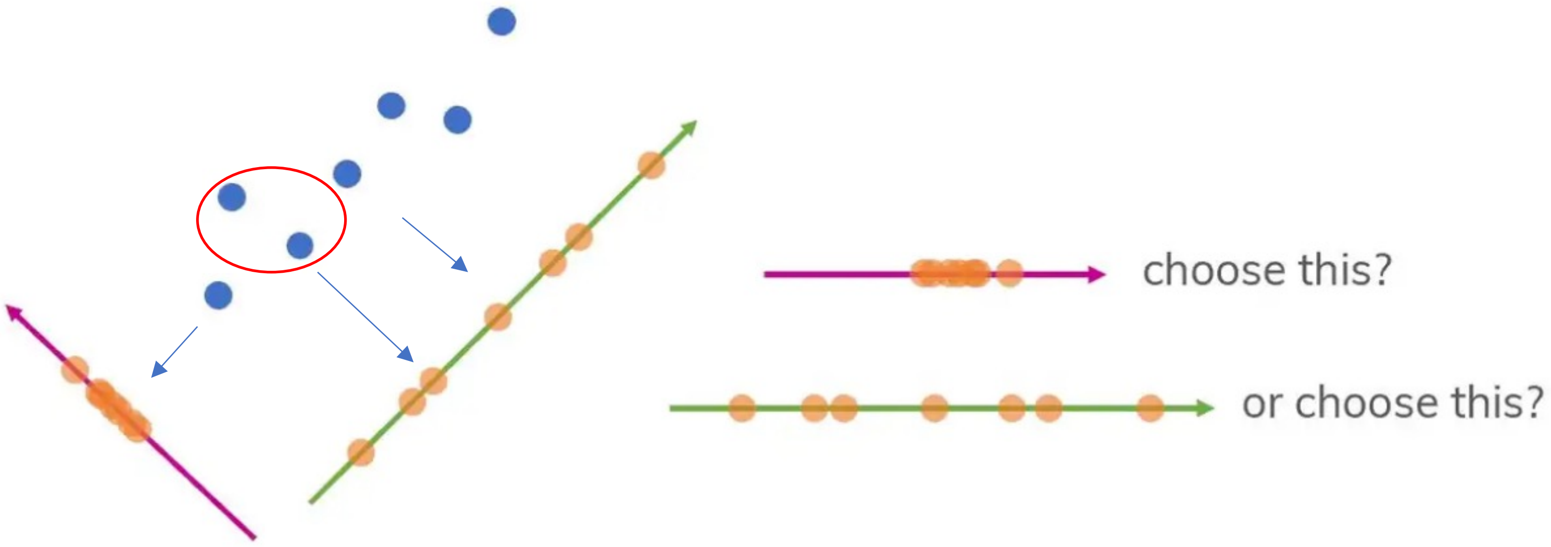
# Objective of ordination methods

Impossible to graphically display all the axes!
The ordination methods respond to this problem by projecting the variability of all these axes over 2 or even 3 axes/dimensions that can be visualized! = DIM reductions!

Obtain plots that provide the best possible summary of the information contained in your large data table
→ Minimize the loss of informationby the DIM reduction ‼ because there will be!

HOW?

# How to minimize the lost of information in data projection?



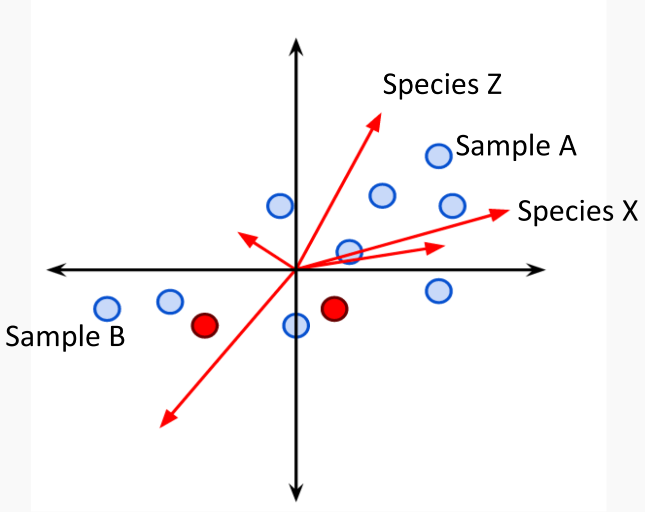
Choose Axis that Maximizes the variance (dispersion), is the more informative

→ Ordinations identifies the axes that MAXIMIZE the VARIANCE of data!

# Ordination & Ecology

- **Definition**

Ordination **summarizes** community data (abundance data: samples by species) by producing a **low-dimensional** ordination space



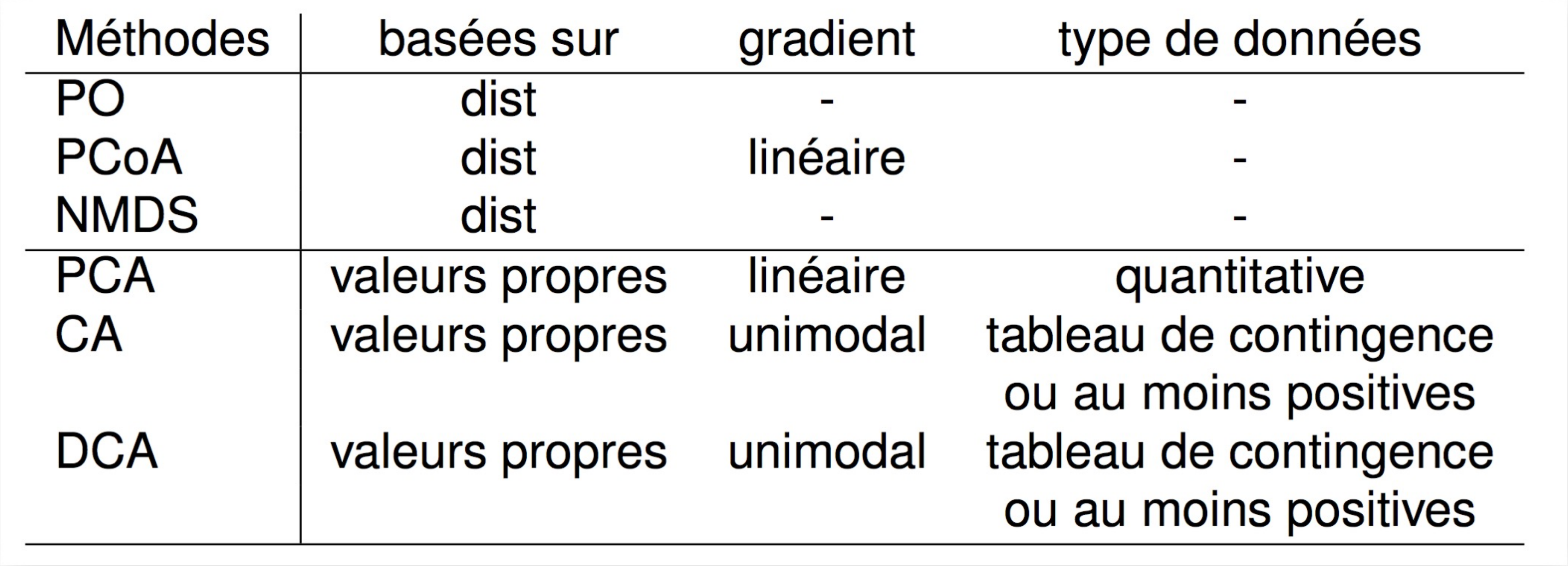
- **Consequences**

- **Similar** species/samples are plotted **close** together
- **Dissimilar** species/samples are placed **far** apart

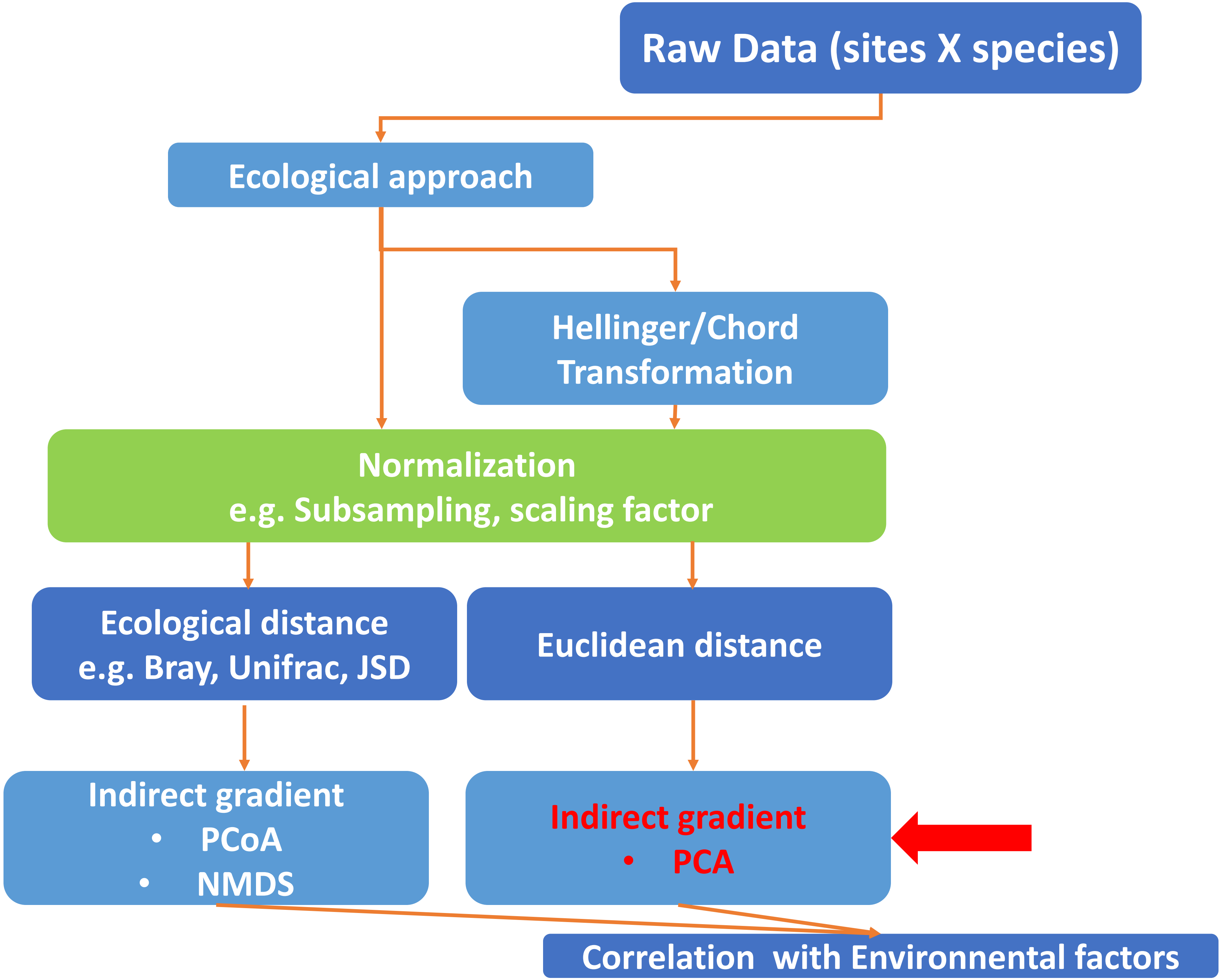
- This low dimensional space should represent important and **interpretable species patterns**
- **The axes defined a GRADIENT (i.e. species composition or environmental)**
- **Major contributors to the axes can be shown**

# Unconstrained Ordination (= Indirect Gradient, Exploratory)

- Ordination **IS NOT influenced** by environmental variables
- Relationships among objects (e.g. sites) and variables (e.g. species) without constraint
- Env variables can be tested **AFTER** the computation of the ordination (e.g. envfit R)



| Méthodes | basées sur | gradient | type de données |
|---|---|---|---|
| PO | dist | - | - |
| PCoA | dist | linéaire | - |
| NMDS | dist | - | - |
| PCA | valeurs propres | linéaire | quantitative |
| CA | valeurs propres | unimodal | tableau de contingence ou au moins positives |
| DCA | valeurs propres | unimodal | tableau de contingence ou au moins positives |



Raw Data (sites X species)

Ecological approach

Hellinger/Chord Transformation

Normalization
e.g. Subsampling, scaling factor

Ecological distance
e.g. Bray, Unifrac, JSD

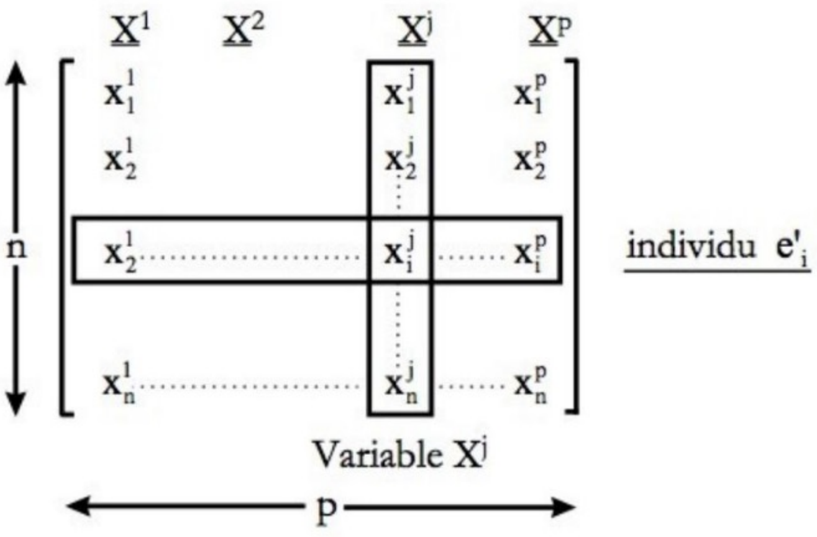
Euclidean distance

Indirect gradient
- PCoA
- NMDS

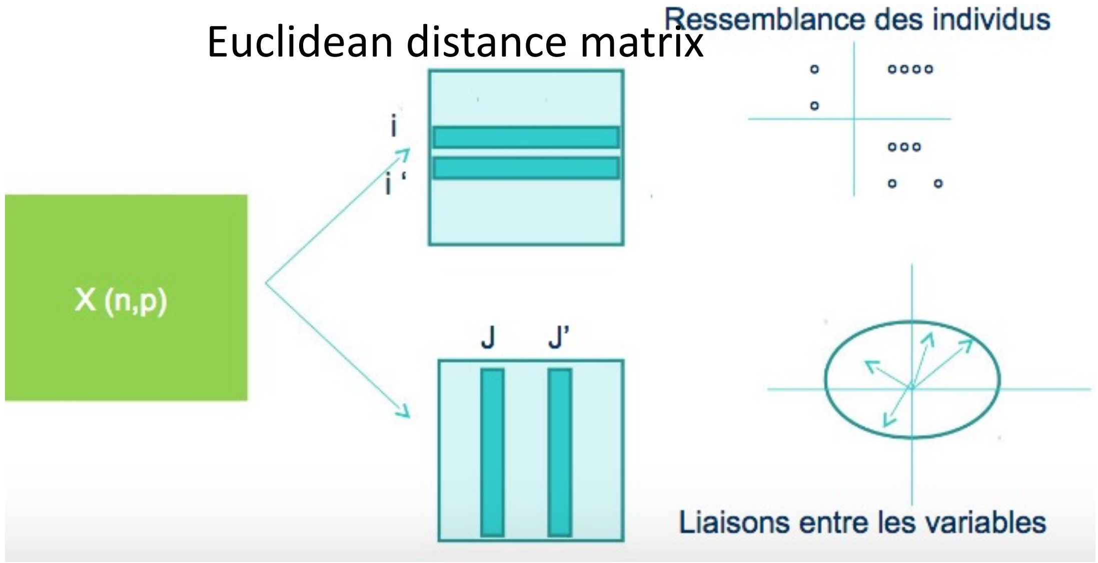
Indirect gradient
- PCA

Correlation with Environnental factors

# ACP. Remember you have to reduce dimensions!



Euclidean distance matrix



- Find a **linear combination of the original variables** (X, taxa) for which the **variance** of the **individuals** (n, site) is **maximal**

  → Is the **first Principal Component** (i.e. PC1)

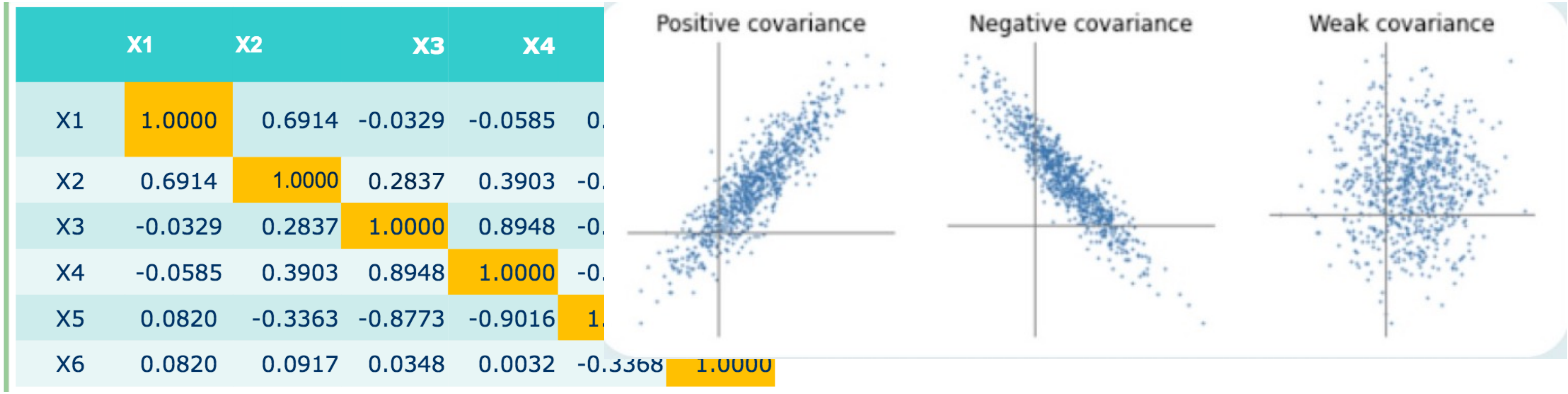
- Find a second PC

Which is not correlated with the PC1 & **Has the Next highest variance**

- Find a third PC … etc

# Dim Reduction? Covariance/correlation matrix of variables (species/ASV)

→ Idée des associations existantes entre les variables



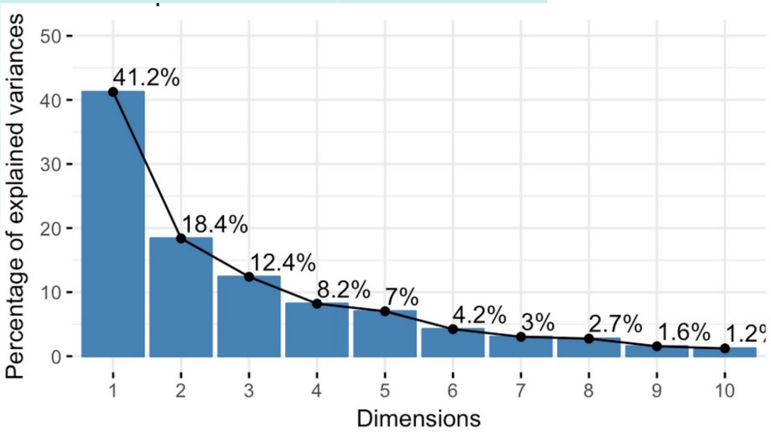
|  | X1 | X2 | X3 | X4 |  |  |
|---|---|---|---|---|---|---|
| X1 | 1.0000 | 0.6914 | -0.0329 | -0.0585 | 0. |  |
| X2 | 0.6914 | 1.0000 | 0.2837 | 0.3903 | -0. |  |
| X3 | -0.0329 | 0.2837 | 1.0000 | 0.8948 | -0. |  |
| X4 | -0.0585 | 0.3903 | 0.8948 | 1.0000 | -0. |  |
| X5 | 0.0820 | -0.3363 | -0.8773 | -0.9016 | 1. |  |
| X6 | 0.0820 | 0.0917 | 0.0348 | 0.0032 | -0.3368 | 1.0000 |

- Association between variables
- Build linear combinaison of variables

# Reduction of dimension & Keep associations between variables/Descriptors

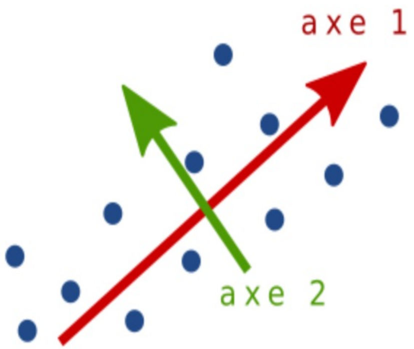
## Transformation in **eigenvalues** and **eigenvector**

| Axe | Valeur propre | Variance | Variance cumulée |
|---|---|---|---|
| 1 | 2.9914 | 49.90% | 49.90% |
| 2 | 1.6125 | 26.90% | 76.80% |
| 3 | 1. | | |
| 4 | 0. | | |
| 5 | 0. | | |
| 6 | 0. | | |



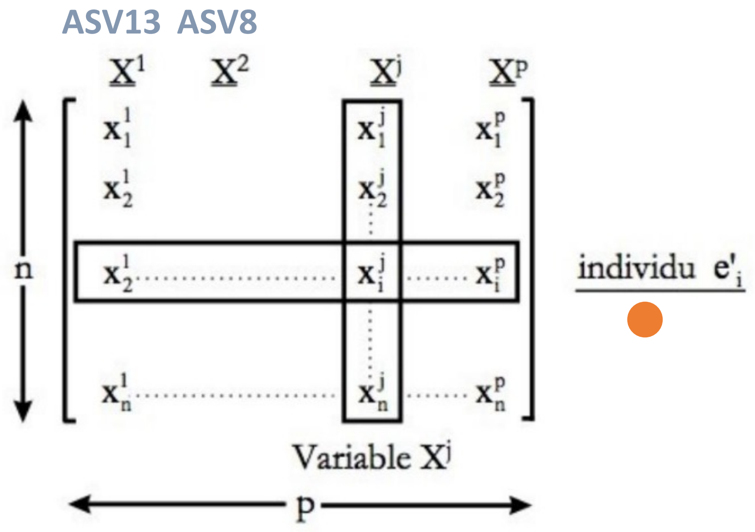
The **eigenvalue** represents the **variance** « explained” by the kth axis

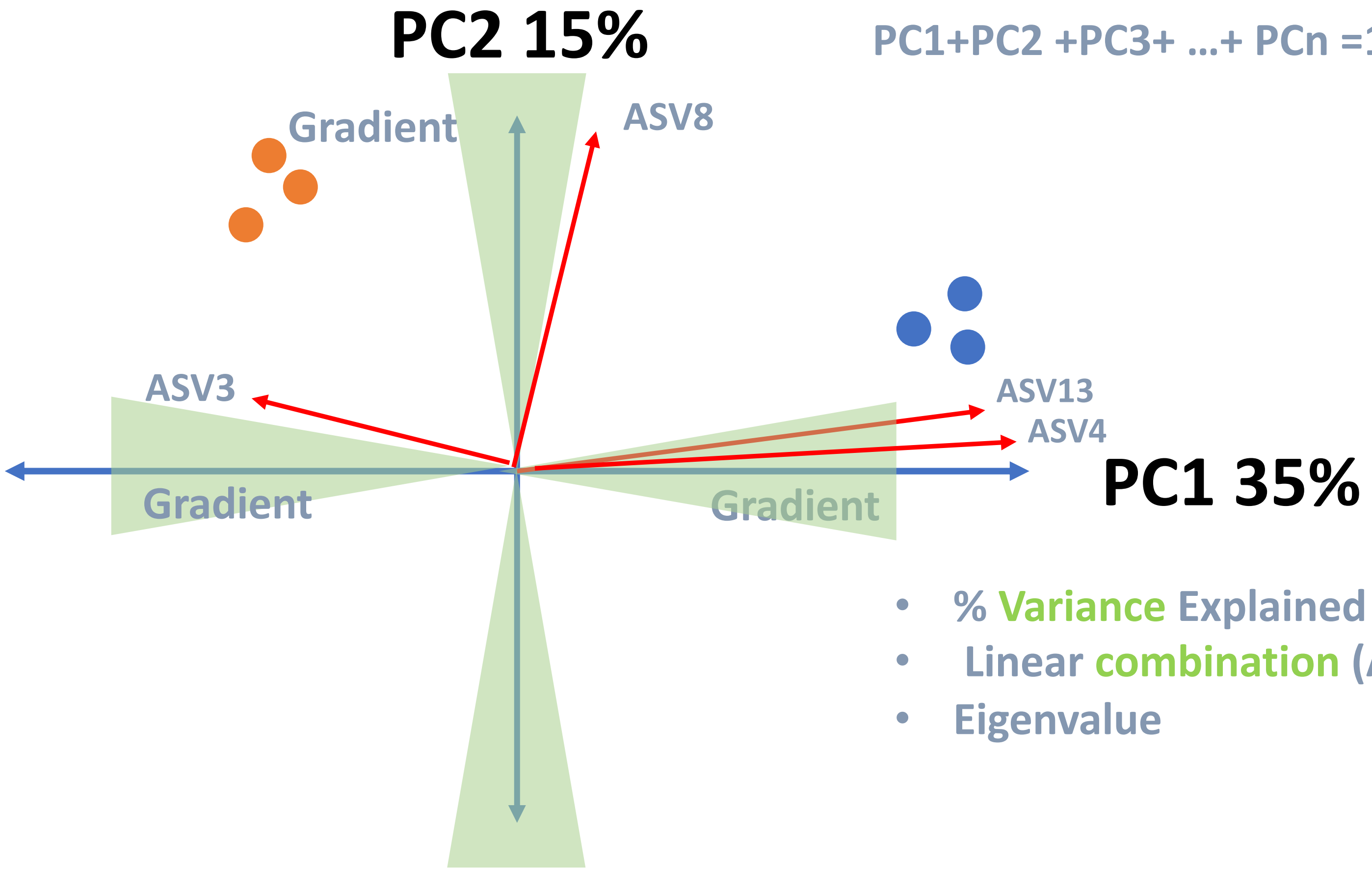
| | vec1 | vec2 | vec3 | vec4 | | |
|---|---|---|---|---|---|---|
| X1 | 0.063 | 0.743 | 0.060 | 0.597 | | |
| X2 | 0.304 | 0.609 | 0.117 | -0.643 | -0.331 | 0.019 |
| X3 | 0.534 | -0.164 | 0.137 | 0.461 | -0.646 | 0.200 |
| X4 | 0.548 | -0.138 | 0.176 | -0.130 | 0.595 | 0.528 |
| X5 | -0.552 | 0.147 | 0.172 | 0.032 | -0.193 | 0.778 |
| X6 | 0.120 | 0.100 | -0.950 | 0.007 | -0.040 | 0.266 |



→**Calcul of Principal components (highest coeff)**

Each **eigenvector** consists of coefficient which represents the **contribution** to PC axis (combination)





PC1+PC2 +PC3+ ...+ PCn =100%

- % Variance Explained
- Linear combination (ASV)
- Eigenvalue

Reduction of dimensions (variables numbers)

PC1 = x ASV13+ z ASV4+ y ASVn...
PC2 = x ASV3+ z ASV13+ y ASVn...
PC= ....

- Drivers of the indirect Gradient = Decompostion of the contributors of the PC

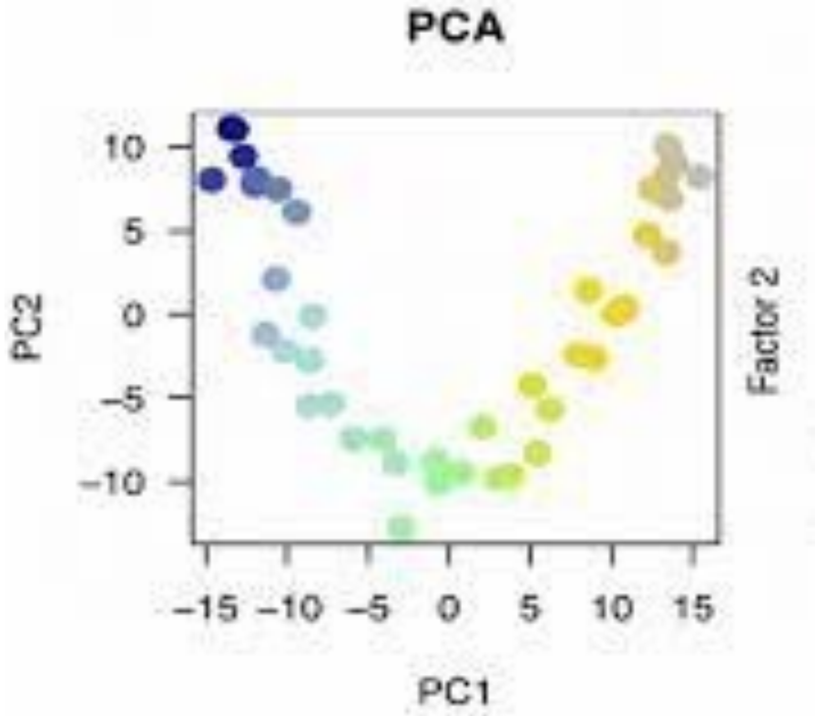
→ Eigen Vector = Major contribution to the PC axis
Exple: ASV3 and ASV4 are major contributors to the PC1 axis (contrib, $\cos^2$)

Exploratory = indirect = No hypothesis about the gradient (unknown)

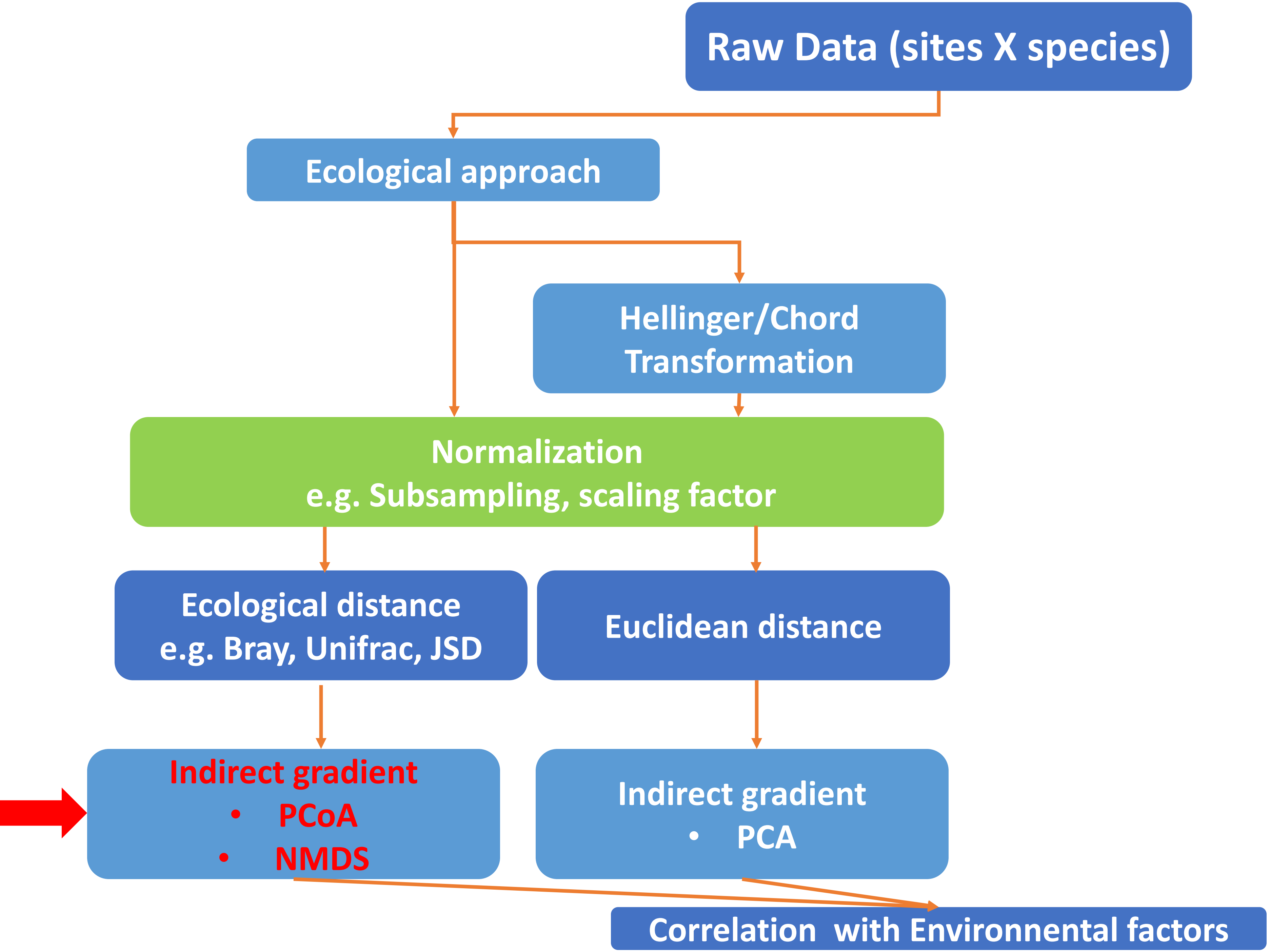
# What are the assumptions of PCA?

- Assumes relationships among variables are **LINEAR**
- Use **Euclidean distance** (Double zero issue → Hellinger transformation)

If the structure in the data is **NONLINEAR**
→ the cloud of points twists and curves its way through p-dimensional space, the principal axes will not be an efficient and informative summary of the data



Arch effect



Raw Data (sites X species)

Ecological approach

Hellinger/Chord Transformation

Normalization
e.g. Subsampling, scaling factor

Ecological distance
e.g. Bray, Unifrac, JSD

Euclidean distance

Indirect gradient

- PCoA
- NMDS

Indirect gradient

- PCA

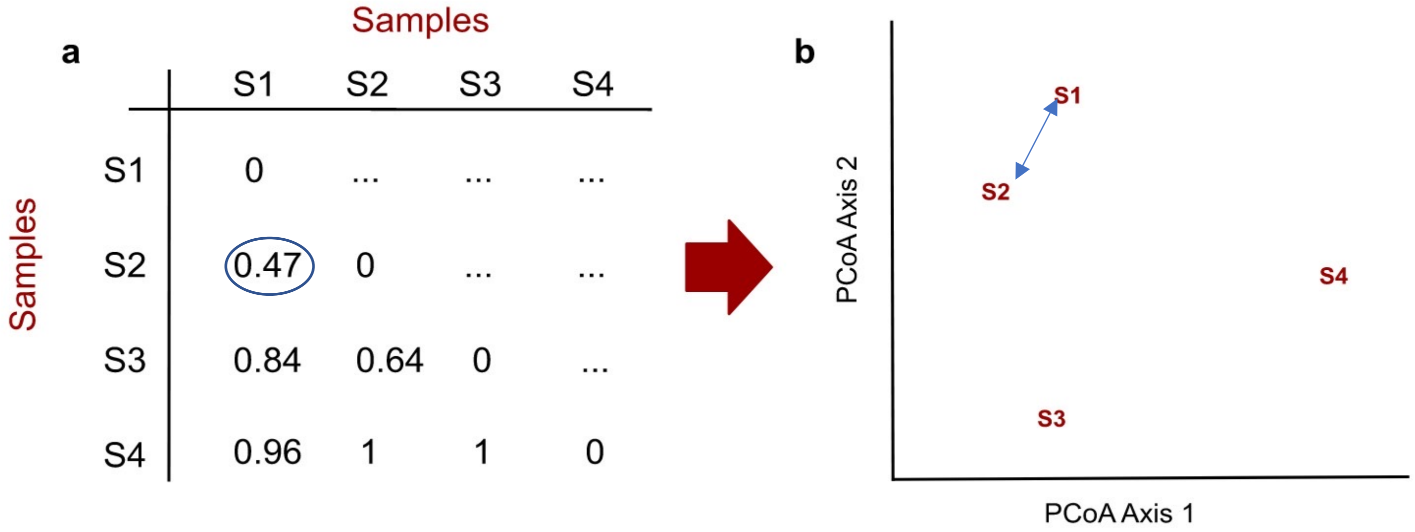
Correlation  with Environnental factors

Unconstrained Ordination based on distances

# Principal Coordinate Analysis (PCoA or MDS)

It Euclidean representation (distance are preserved) of a set of **objects** whose relationships are measured by **any similarity or distance measure (excepted euclidean! Why?)**

**Important : Does not use** original/raw data (e.g PCA)…

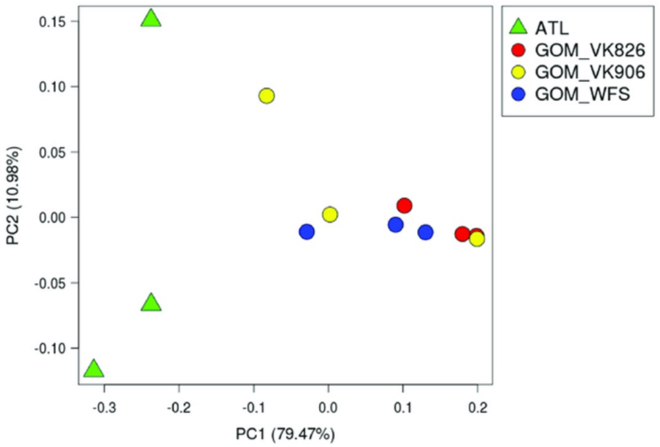


|  | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | 0 | ... | ... | ... |
| S2 | 0.47 | 0 | ... | ... |
| S3 | 0.84 | 0.64 | 0 | ... |
| S4 | 0.96 | 1 | 1 | 0 |

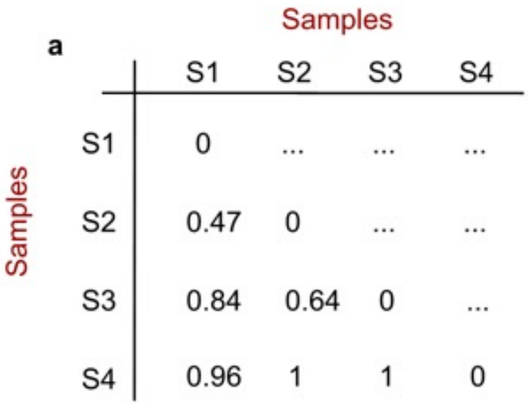
Like PCA, PCoA produces a set of orthogonal axes which **maximize the correlation between the dissimilarity matrix and the distance among samples in ordination space.**

# PCoA : Where are the species??

- Because PCoA is based on a distance matrix, the analysis **never “see”** any taxa data
- Distance matrix has no information about original column variables/Taxa



→ **Because**



a

Samples

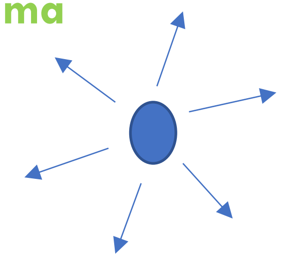
| Samples | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | 0 | ... | ... | ... |
| S2 | 0.47 | 0 | ... | ... |
| S3 | 0.84 | 0.64 | 0 | ... |
| S4 | 0.96 | 1 | 1 | 0 |

- **Solution to have species scores/information**

→ add species by going back and calculating **weighted averages!**
→ **weighted average of species** positions according to **abundance across all samples**

→ **You will obtain a biplot**

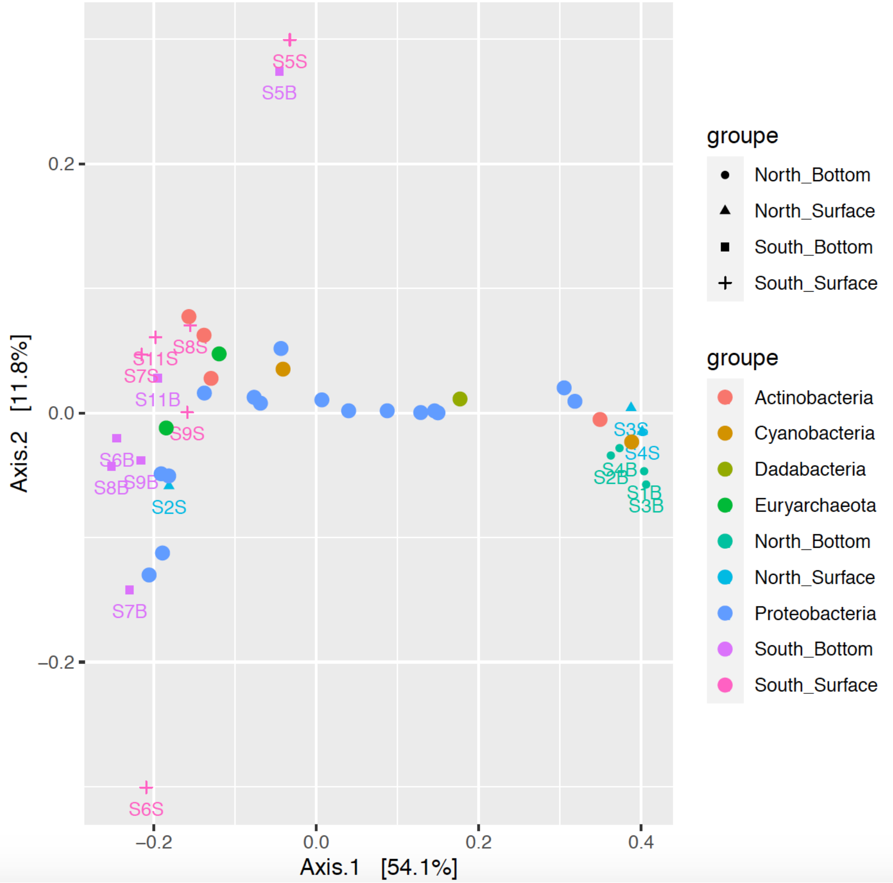
**Optima**



# Biplot PCoA-Species variables

- A biplot simultaneously displays, in two dimensions, rows (samples) of a data matrix as points, and columns (variables) as arrows/points

## PCoA: weighted average of species positions



## Unconstrained Ordination based on distances

# Non Metric Multidimensional Scaling (NMDS)

NMDS represent dissimilarity between objects in a low-dimensional space. **Any dissimilarity coefficient or distance measure** may be used!

**NMDS is an iterative algorithm.** Begin by random placement of data objects in ordination space. Refine this placement by an iterative process, attempting to find an ordination in which **ordinated object distances match at best the order of object dissimilarities in the original distance matrix.**
→ The **stress value** reflects this!

Stress values >0.2 are generally poor and potentially uninterpretable**, whereas values <0.1 are good and <0.05 are excellent**, leaving little danger of misinterpretation.

# Unconstrained Ordination based on distances

## Non Metric Multidimensional Scaling (NMDS)

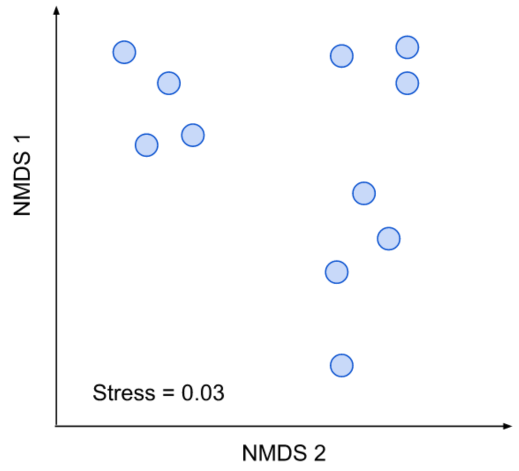
**NMDS is a rank-based approach.** This means that the original distance data is substituted with ranks. While information about the magnitude of distances is lost, rank-based methods are generally **more robust to data which do not have an identifiable distribution**

| | | Samples | | | |
|---|---|---|---|---|---|
| | | S1 | S2 | S3 | S4 |
| Samples | S1 | 0 | ... | ... | ... |
| | S2 | 0.47 | 0 | ... | ... |
| | S3 | 0.84 | 0.64 | 0 | ... |
| | S4 | 0.96 | 1 | 1 | 0 |

Dissimilarity /Distance

| | | Samples | | | |
|---|---|---|---|---|---|
| | | S1 | S2 | S3 | S4 |
| Samples | S1 | 0 | ... | ... | ... |
| | S2 | 1 | 0 | ... | ... |
| | S3 | 3 | 2 | 0 | ... |
| | S4 | 4 | 5.5 | 5.5 | 0 |

Rank calcul

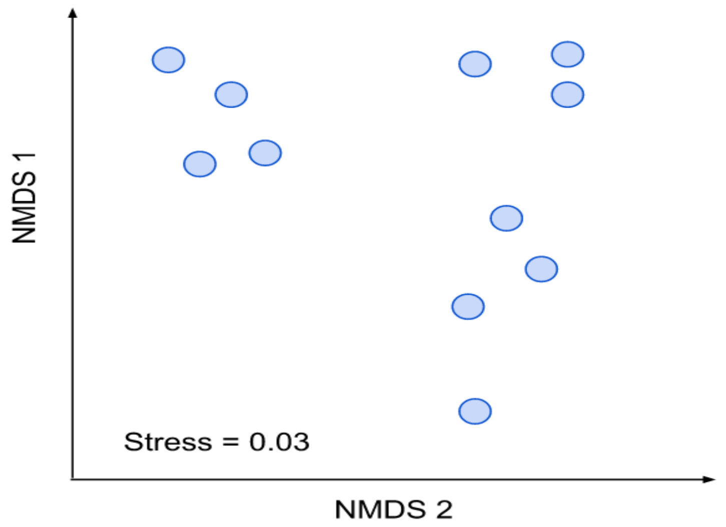


NMDS
Axes are arbitrary
No % of inertia/ variance

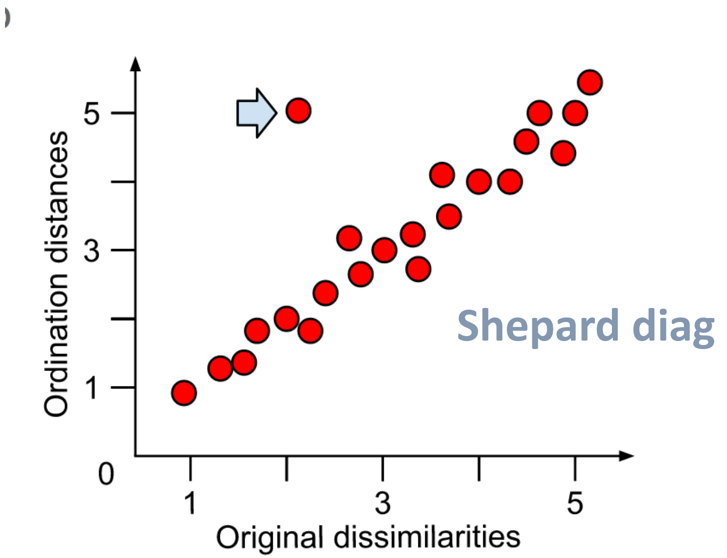
**The axes of an NMDS ordination are entirely arbitrary**

# Stress

|  |  | Samples |  |  |  |
|---|---|---|---|---|---|
|  |  | S1 | S2 | S3 | S4 |
| Samples | S1 | 0 | ... | ... | ... |
|  | S2 | 0.47 | 0 | ... | ... |
|  | S3 | 0.84 | 0.64 | 0 | ... |
|  | S4 | 0.96 | 1 | 1 | 0 |



**Ordinated object distances Vs. object dissimilarities in the original distance matrix…. Must fit at best → Stress**



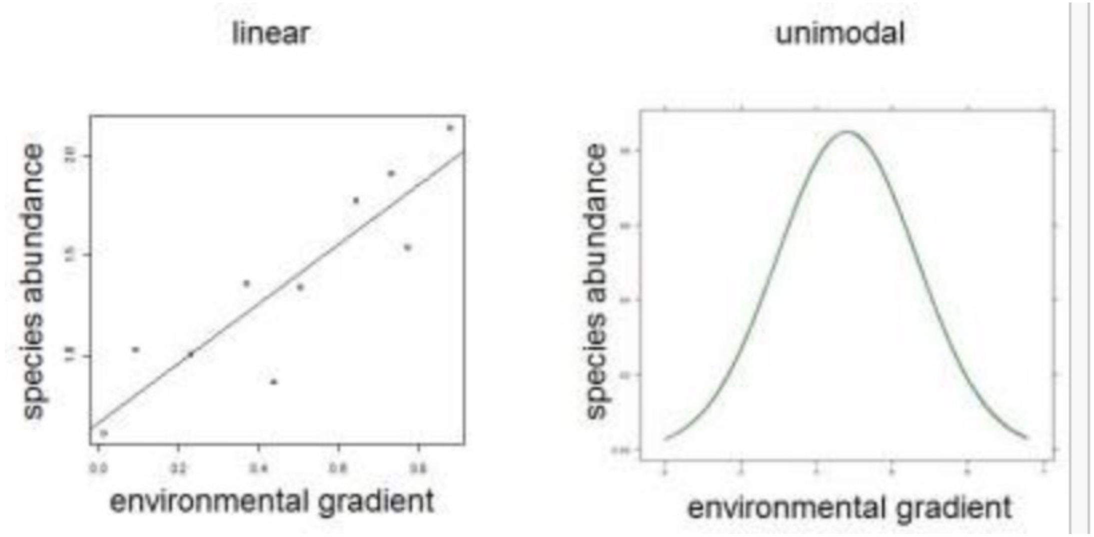
Shepard diag

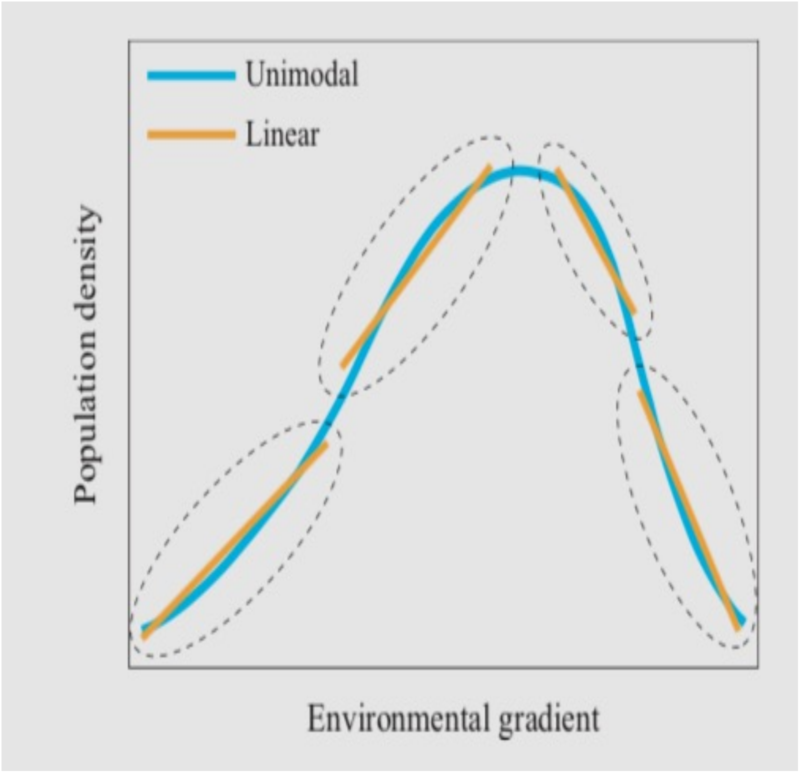
# Models of variable response to environmental gradients

→Ordination type have **specific assumption model** according the **species response along the environmental gradient** = « **variable response model** » (Maths)

Key point:

- **linear relationships** (rarely in nature)
- Non **monotonic relationship** with the environment : **unimodal**





**Gradient : Spatial, temporal, Ph, nutrients, pertubations etc**

# Linear or unimodal ? DCA: Detrended correspondance Analysis

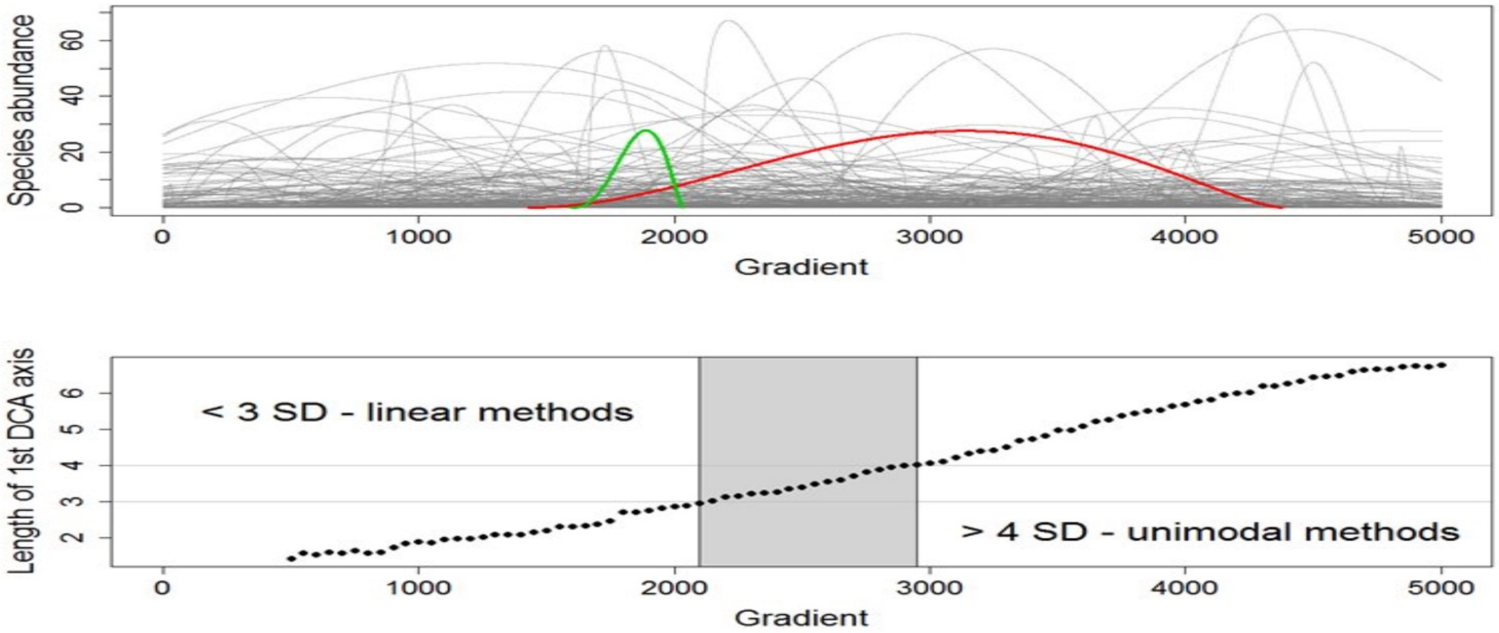
- Apply **linear or unimodal ordination method** on your data? Lepš & Šmilauer 2003)
- Use DCA R package, check the length of the *first* DCA axis

The length of first DCA axis > 4 SD.

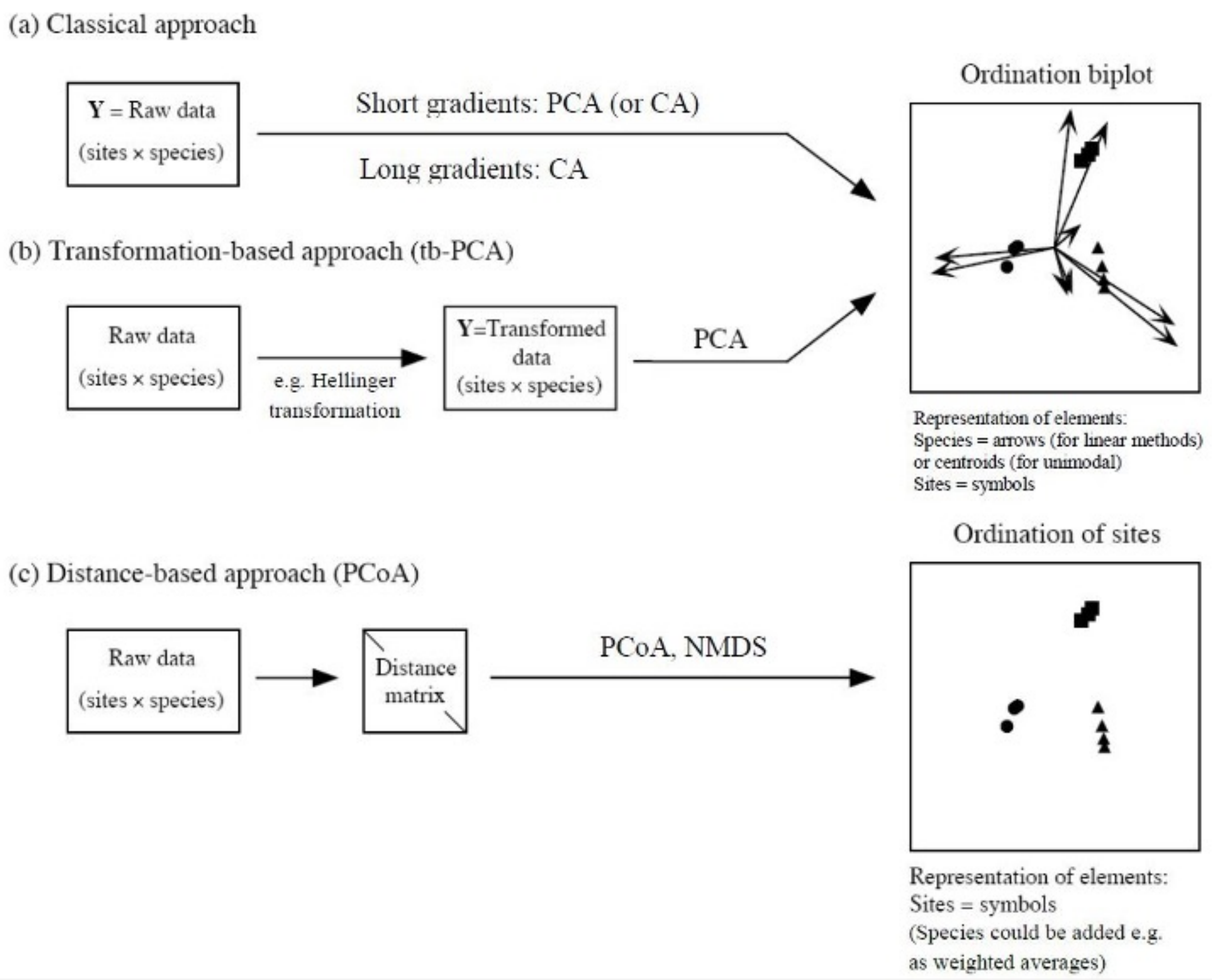
→ heterogeneous dataset on which unimodal methods should be used

The length of first DCA axis < 3 SD.

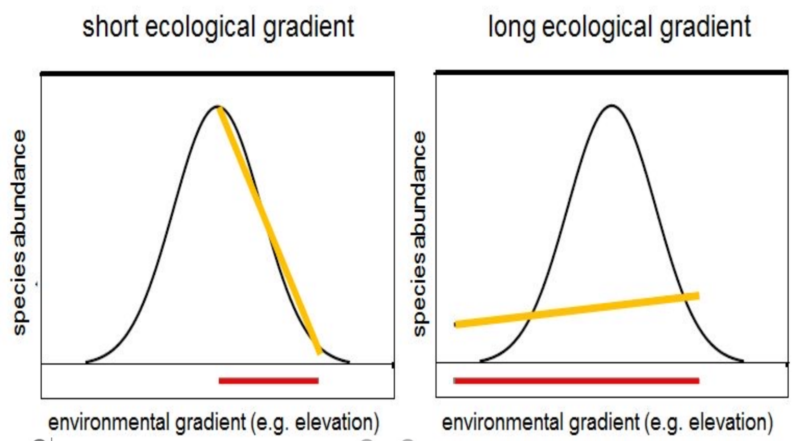
→ homogeneous dataset for which linear methods are suitable



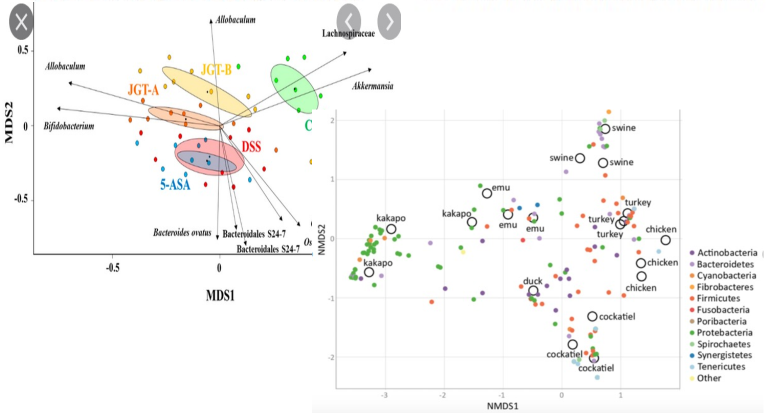
# To summarize :Unconstrained Ordination



## DCA package R



short ecological gradient
long ecological gradient
species abundance
species abundance
environmental gradient (e.g. elevation)
environmental gradient (e.g. elevation)

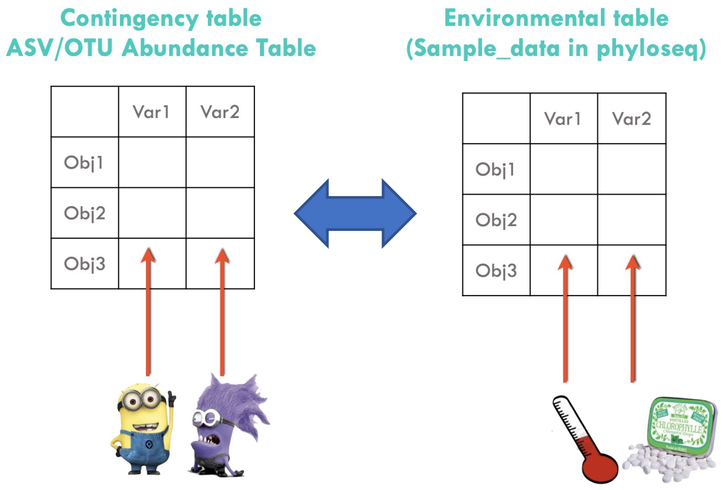


- **Linear relationship** : Principal Component Analysis, PCoA, Tb-PCA
- **Unimodal relationship**: CA = correspondance Analysis
- **Not based on specific underlying model of variable response**:  NMDS

*Legendre & legendre: Numerical ecology*

# Constrained Ordination : The gradient is imposed! NOT exploratory)

- **Objective** : Attempt to explain differences in species **composition between sites** by the **environmental gradient**



Contingency table
ASV/OTU Abundance Table

| | Var1 | Var2 |
|---|---|---|
| Obj1 | | |
| Obj2 | | |
| Obj3 | | |

Environmental table
(Sample_data in phyloseq)

| | Var1 | Var2 |
|---|---|---|
| Obj1 | | |
| Obj2 | | |
| Obj3 | | |

- **Key points**

- Computes axes that are **linear combinations of the explanatory variables** (e.g ph, T°C, …)

- It is constrained because you are **directly** testing the **influence of CHOSEN explanatory variables**

- Consequence : probably **only a fraction of the variance** from data is explained by explanatory variables, that means you can miss/not see patterns in your data!

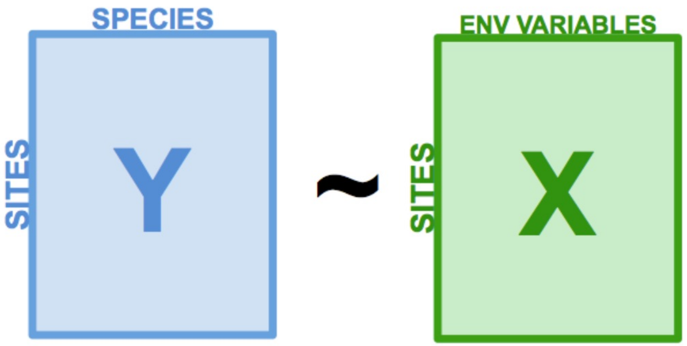
# Constrained Ordination (Direct Gradient Analysis)

## Redundancy Analysis (RDA)

Conceptually, RDA is an extension of **multiple linear regression**

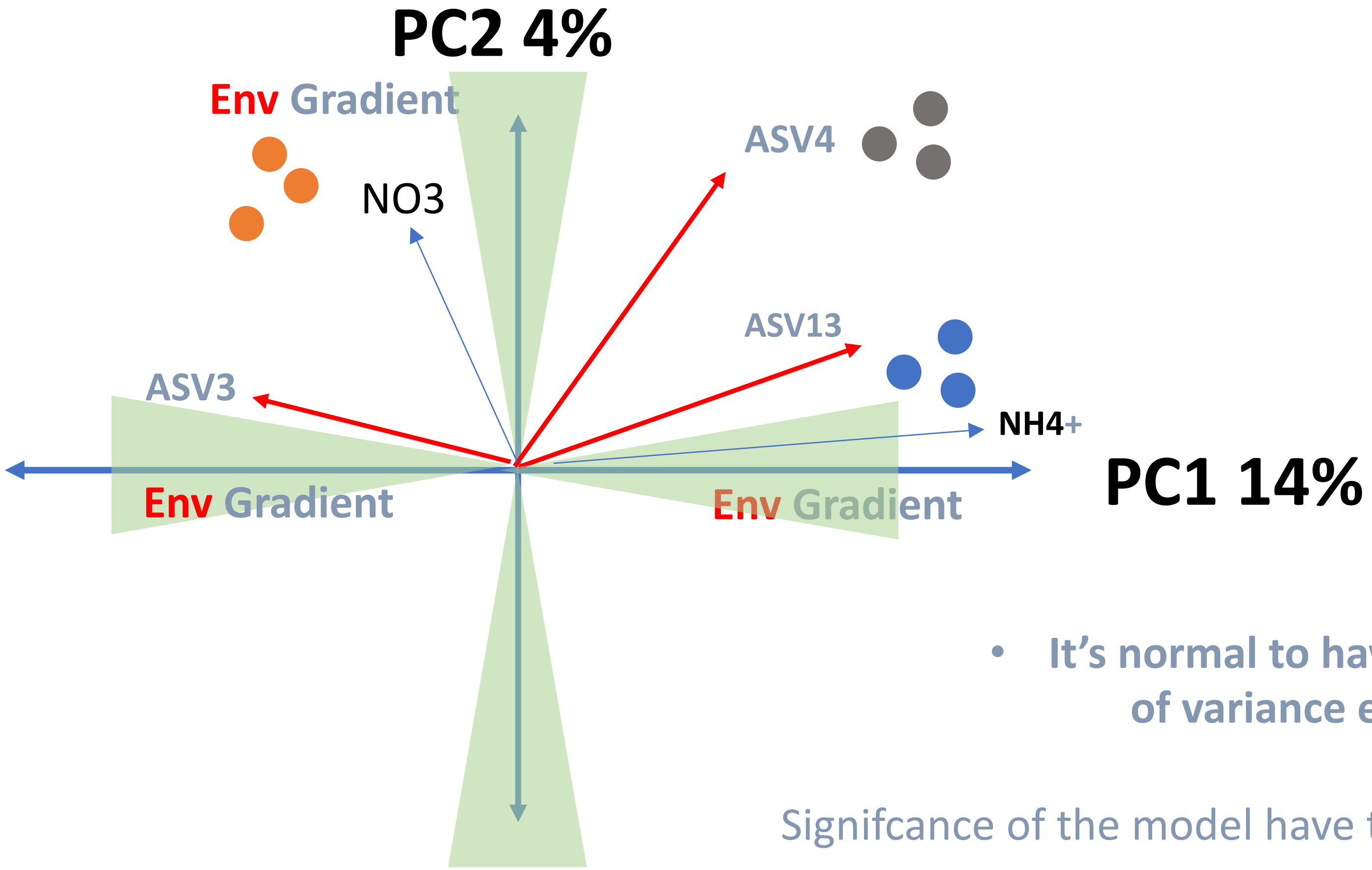
**RDA models the effect of an explicative matrix X (env data) on a response matrix Y (community data)**

→ **Effects of environmental factors on assemblages**



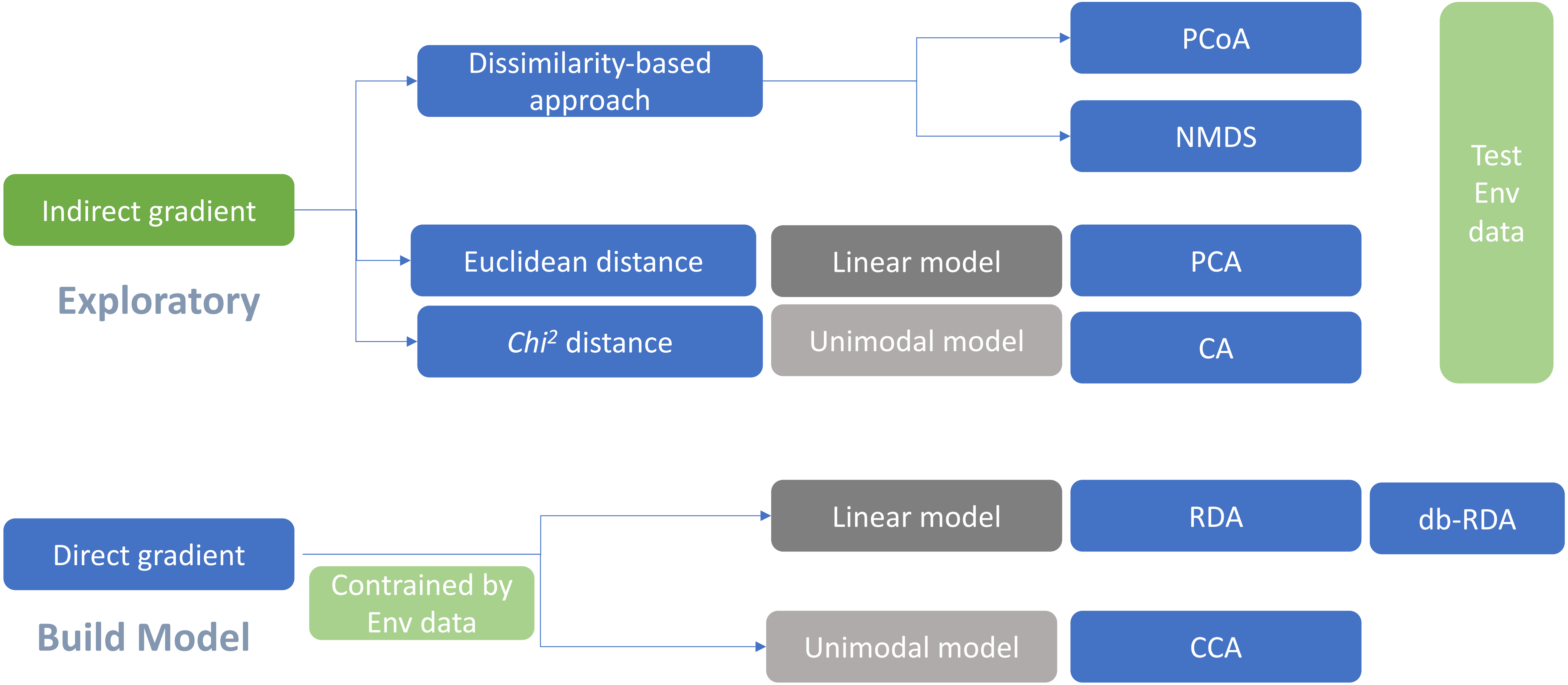
**RDA** = raw data
**Tb-RDA**= transformed raw data
**Db-RDA** = distance-based

1- Multiple linear regression (Y~X)
2- PCA (dimension reduction)



- It's normal to have a low score of variance explained!

Signifcance of the model have to best tested (Axis)



Indirect gradient

Exploratory

Dissimilarity-based approach

PCoA

NMDS

Euclidean distance

Linear model

PCA

*Chi*$^2$ distance

Unimodal model

CA

Test Env data

Direct gradient

Build Model

Contrained by Env data

Linear model

RDA

db-RDA

Unimodal model

CCA

# Selection of environmental factors/variables

**You Need to find the best combinaison of variables for the model**

→The combinaison **which best explains** the abundance/species composition variations

**But too much environmental variables (noise)… How to deal with?**

- **Ecological meaningful : Removing or keep some variables according to your expertise!!**

- **Progressive strategy : Remove non significant env variables**

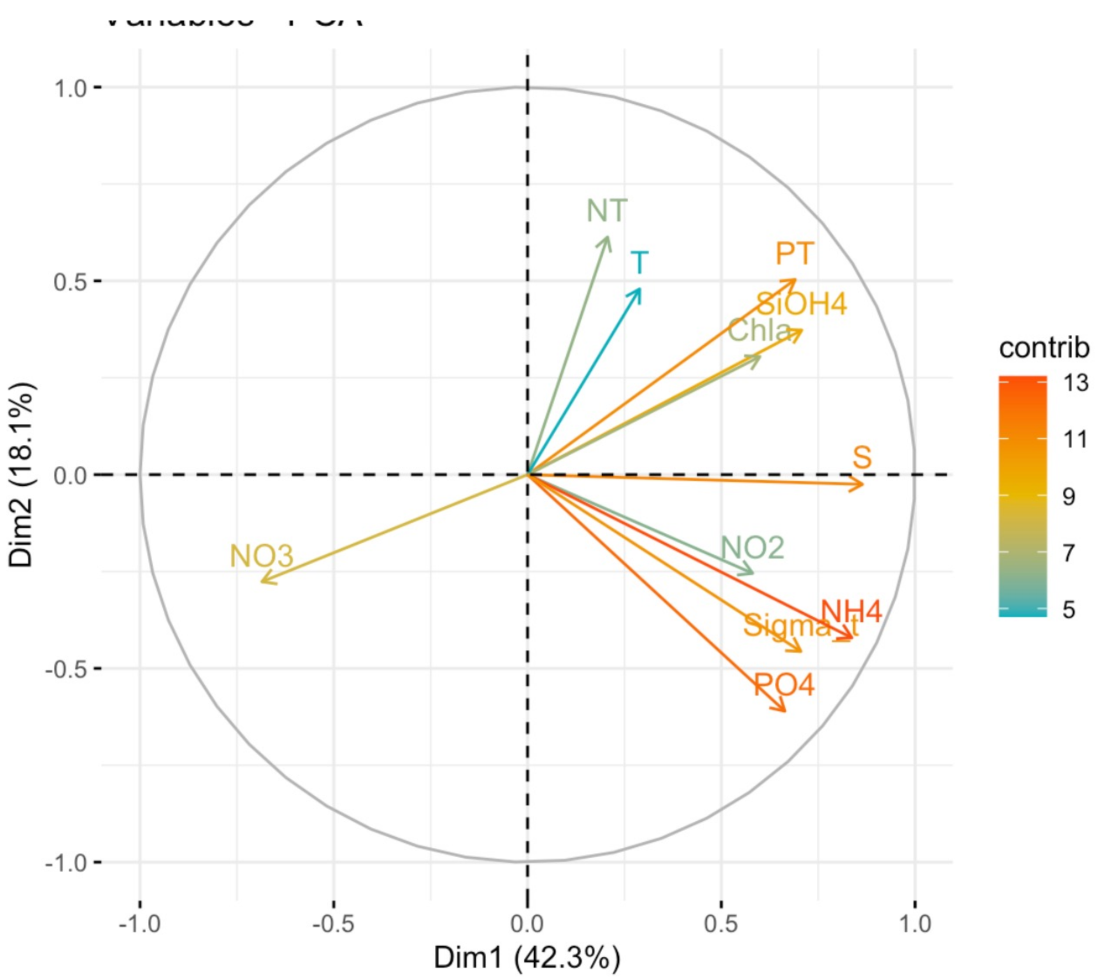
→Add one by one the environmental factor & evaluate if the model is better (**$R^2$** score, **BIC** and **AIC** criteria)

Tools :

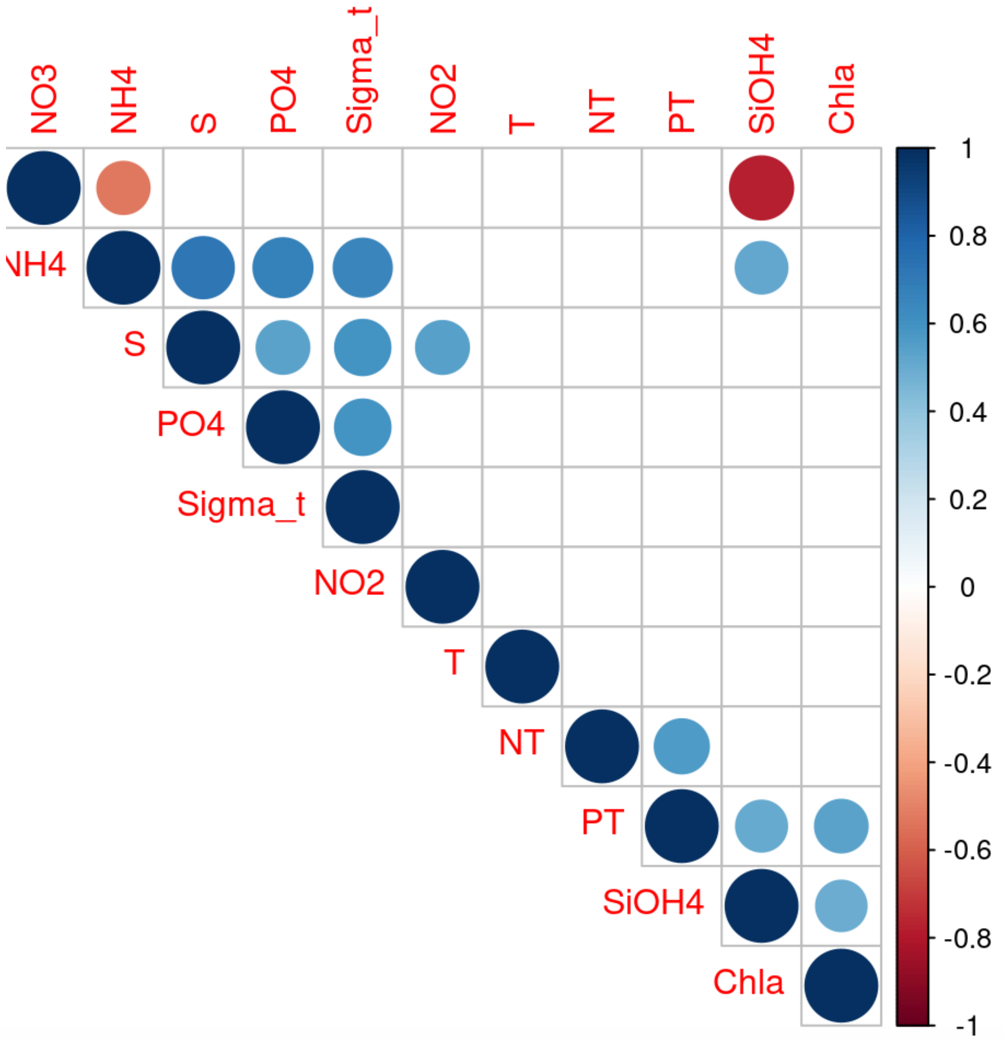
- **Ordistep, forward.sel** R functions
- **Evaluate model with BIC, VIF**

→ The lowest BIC value correspond to the model that best fits the data

See ANF website

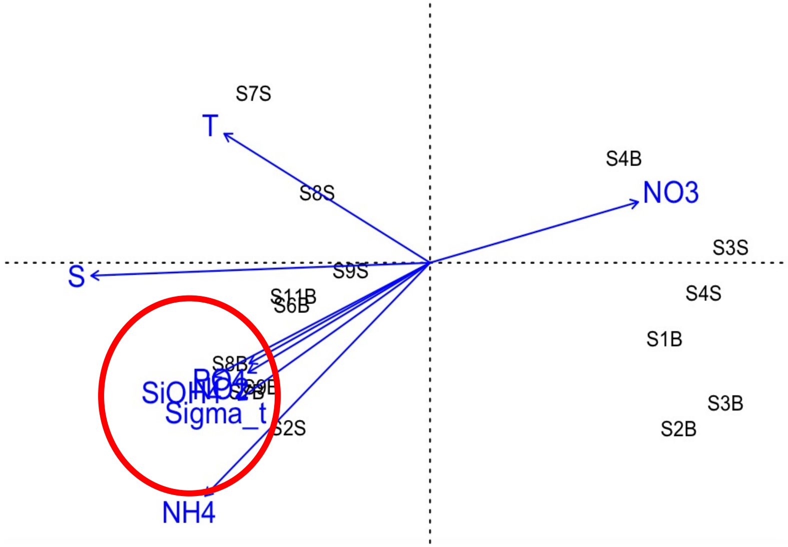


Correlation between variables by PCA



Correlation between variables Spearman/Pearson

# Multicollinarity issue: Remove redundancy : make choices!



Where collinearity exists between variables there is **redundancy** between **predictor variables** (= env. variables)

→ The solution of the **model becomes unstable**

**How to evaluate collinearity between env. variables?**

- **Use VIF (Variance Inflation Factor, vif.cca with R )**

→ VIF > 10 indicates collinearity problems with that variable

# RDA Statistics

## Model meaningful

- **Explicative power of the included env variables?**
- **Are the relations observed are significant ?**

- **$R^2$**, strength of the relation between Y and X thanks to the % of variation of species matrix explained by ENV

- **Adjusted $R^2$**, **Apply a correction** : taking into account the explicative variable number!! This is this value that you must report!

- **F statistics** :

  - Global test of the model significance
  - **Test of each RDA axis**

# RDA triPLot

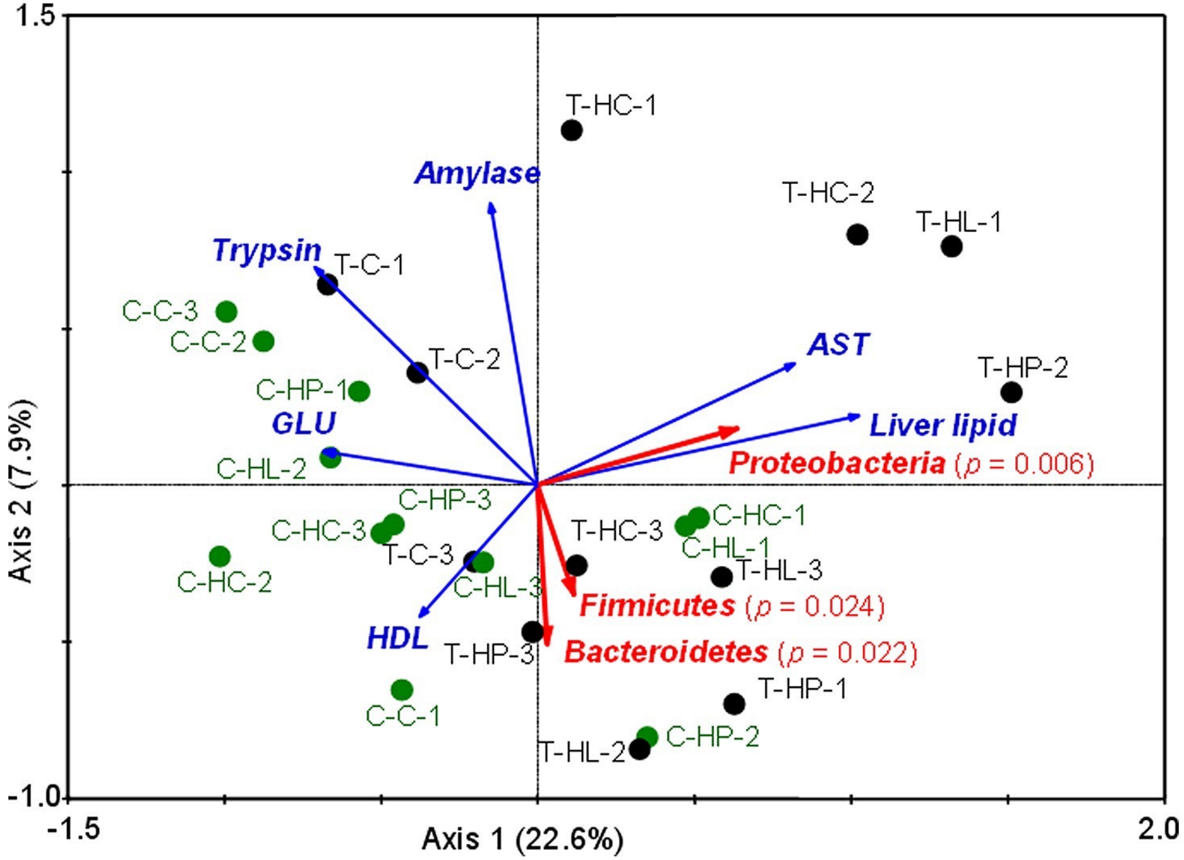
There are three different entities in the plot: **sites, response variables and explanatory variables**

**Samples (sites):** distances between points approximate compositional dissimilarity among samples

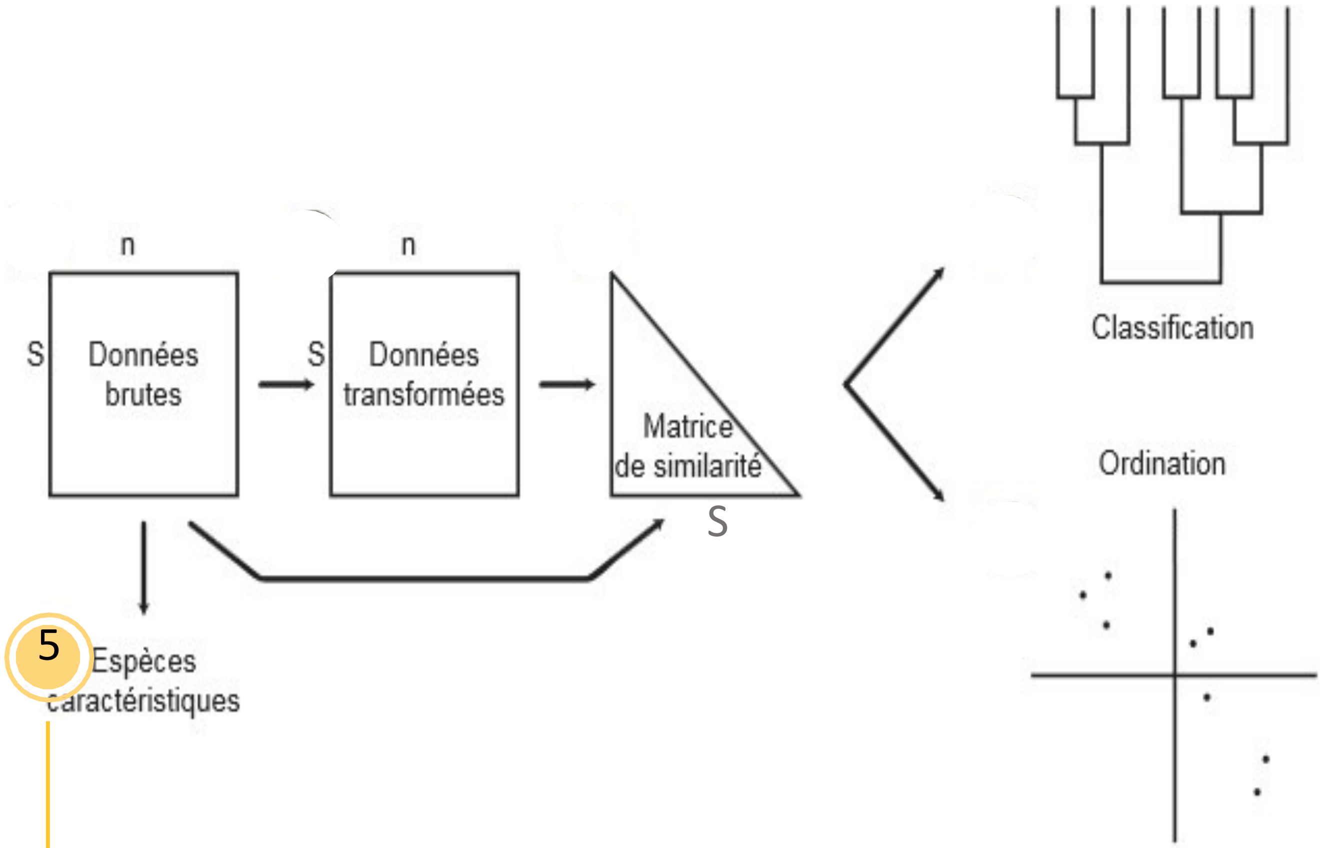
**The distance between site and species** position on the triplot is indicative of the abundance of the species for the site

**The angle between variables and species** reflects their correlations

**Environmental variables :** arrows indicate in which direction the value of environmental variable increases



# Overview of the Beta-analysis approach



Differential abundance

# Differential abundance analysis (DAA)

The goal of differential abundance testing is to identify specific taxa associated with metadata variables of interest. **This is a difficult task (Compositional data)**

This is related to concerns that normalization and testing approaches have generally **failed to control false discovery rates**

Nearing et al. (2022) compared all the methods across 38 different datasets and showed that ALDEx2 and ANCOM-BC produce the most consistent results across studies.

→ **Log ratio transformation**

# Differential Abundance Analysis

**Choose an appropriate analysis unit (e.g. ASV, genus, family level)**

- ASV level (finest resolution) can **lack of detection power** (due to sparse counts, unassigned sequence) for an effective comparison

- Increasing power by **aggregating ASVs to upper taxonomic** rank (i.e genera, family etc) BUT at the cost of a **coarser resolution**

# Differential abundance

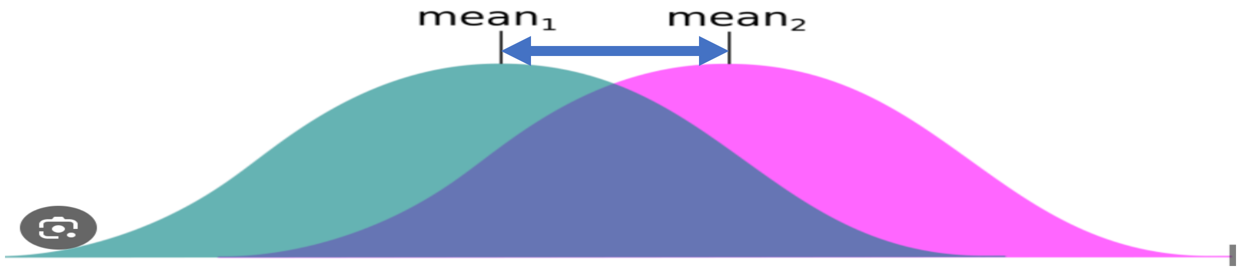
## Linear Discriminant Analysis Effect Size (LEFse), Segata et al. (2011)

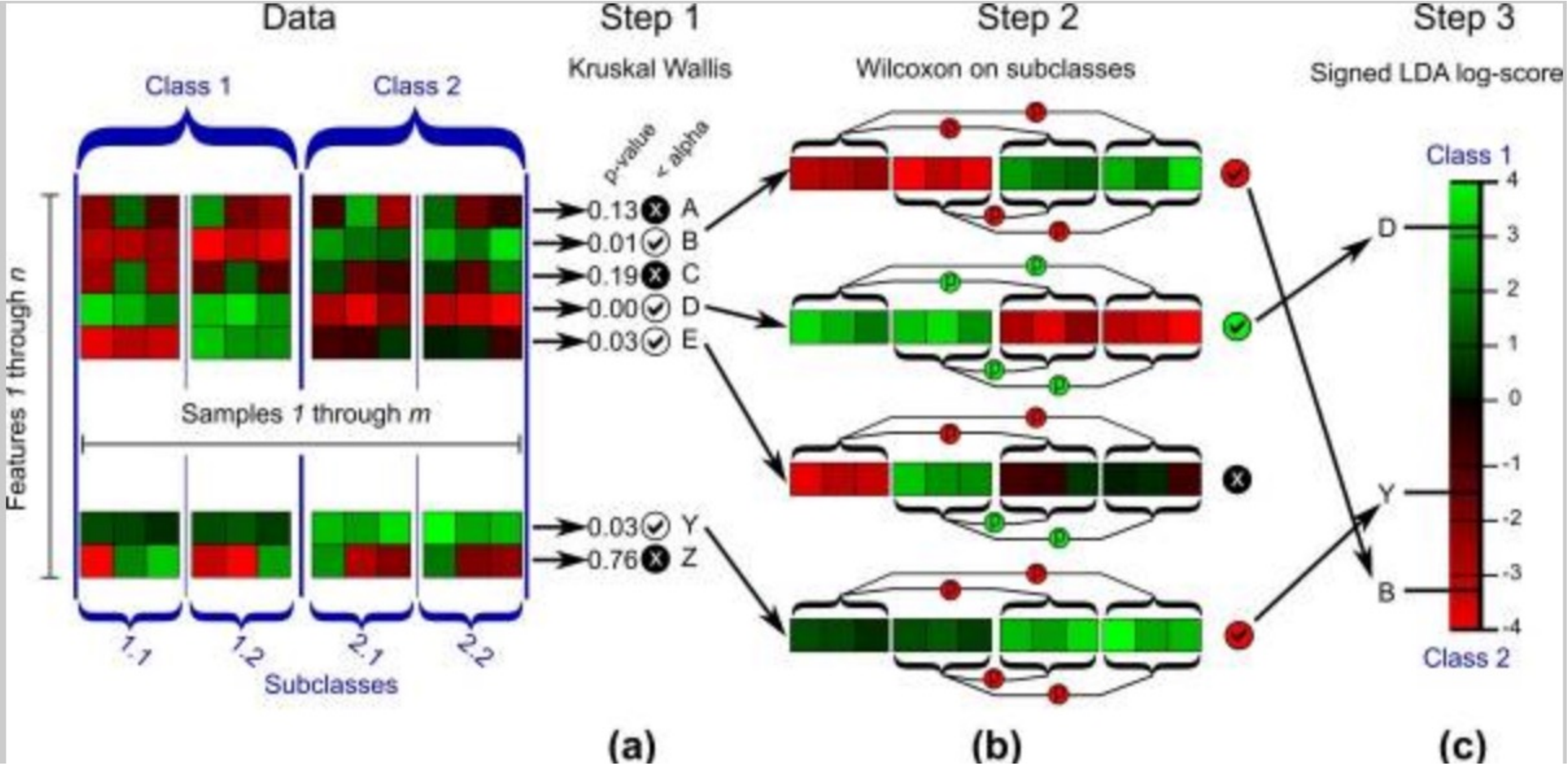
LEFse uses non-parametric tests :

1- Kruskal-Wallis sum-rank test : detect differential abundance within class of interest = GROUPS

2- Biological consistency is tested using among subclasses (=Sub groups) with Wilcoxon rank-sum test

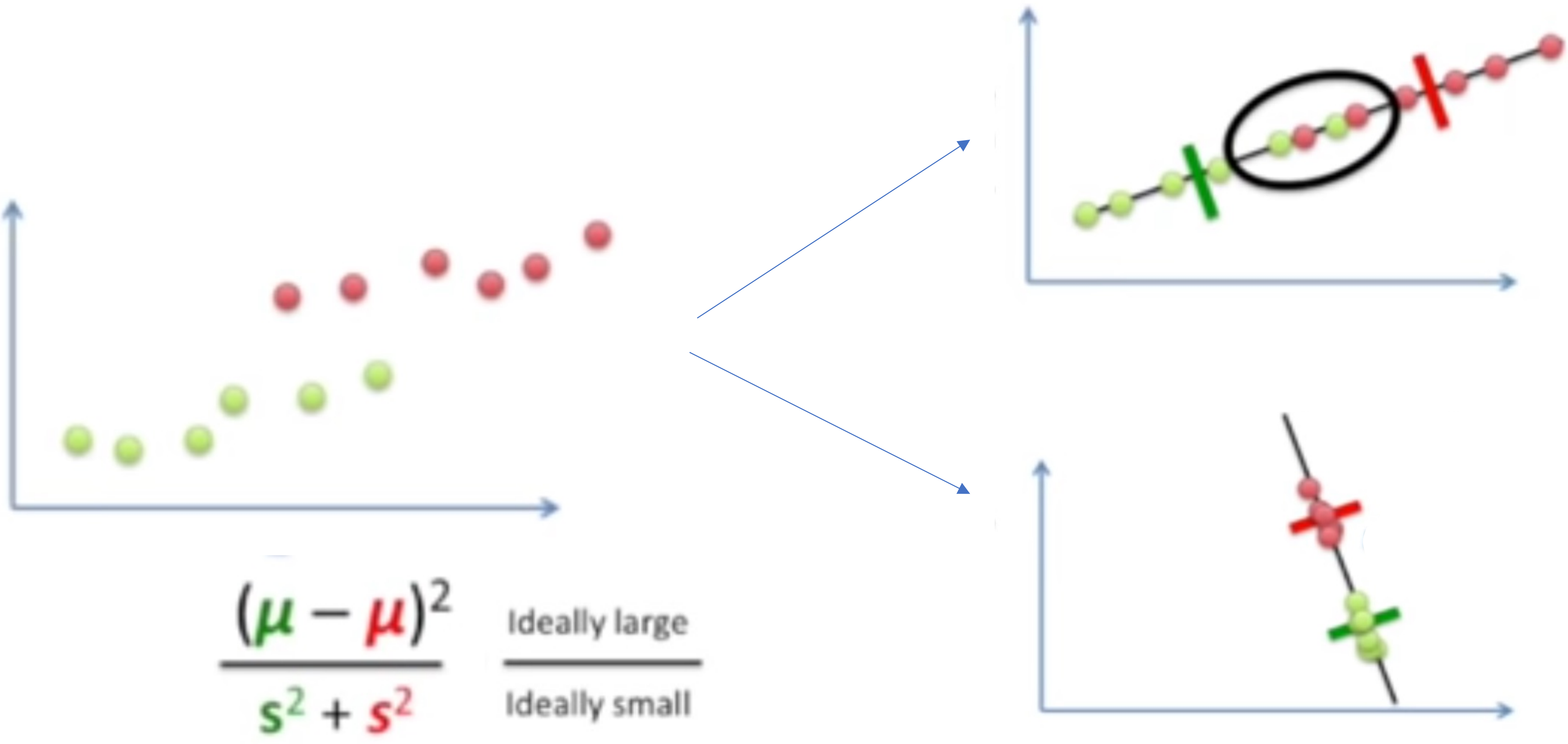
3- LEfSe uses LDA to estimate the effect size of each differentially abundant features




Data
Step 1
Kruskal Wallis
Step 2
Wilcoxon on subclasses
Step 3
Signed LDA log-score
Class 1
Class 2
p-value
< alpha
0.13 A
0.01 B
0.19 C
0.00 D
0.03 E
0.03 Y
0.76 Z
Features 1 through n
Samples 1 through m
1.1
1.2
2.1
2.2
Subclasses
p
Class 1
4
3
2
1
0
-1
-2
-3
-4
Class 2
D
Y
B
(a)
(b)
(c)

# Linear Discriminant Analysis

## How to separate groups using an new axis than maximize the distance (mean, effect size) AND minimize the dispersion



$$\frac{(\mu - \mu)^2}{s^2 + s^2}$$