


PART III:

Beta-diversity Analysis

PART III:
Beta-diversity Analysis



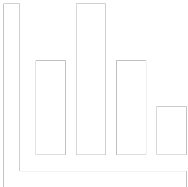


The statistical analysis of microbial metabarcoding sequence data is a rapidly evolving field

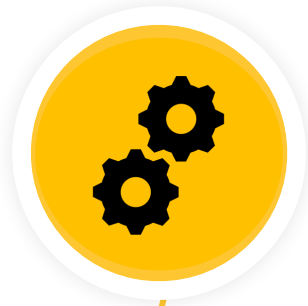
Different solutions (often many) have been proposed to answer the same questions.

Focus on methods that are common in the microbiome literature, well-documented, and reasonably accessible...and a few we think are new and interesting.

Objectives



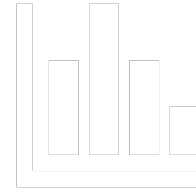
Objectives



01

β -DIVERSITY ANALYSIS

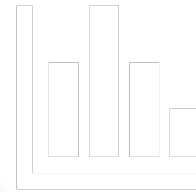
Learn the different steps of the Beta-diversity analysis workflow



Objectives

02 STATISTICS

Know how to choose among the many classification/ordination/statistical methods commonly used with metabarcoding data



01

β -DIVERSITY ANALYSIS

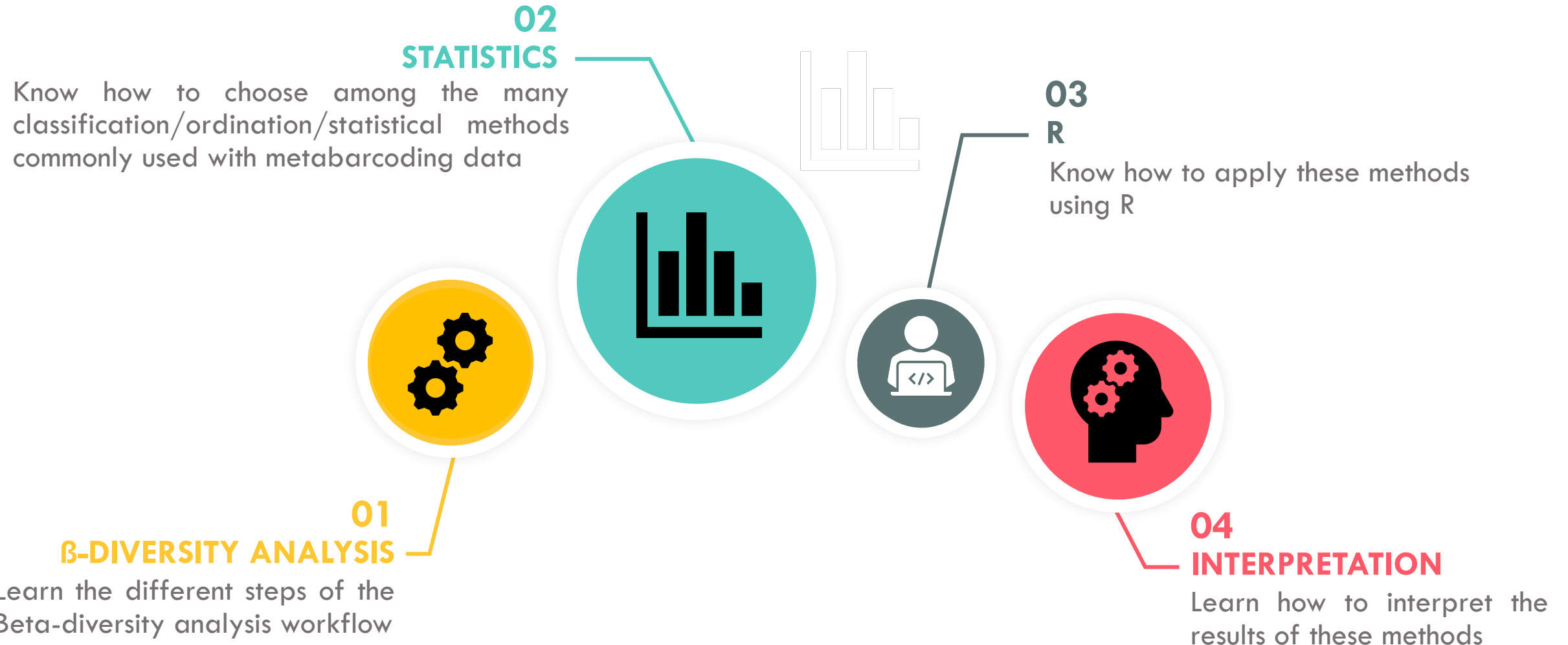
Learn the different steps of the Beta-diversity analysis workflow



Objectives



Objectives



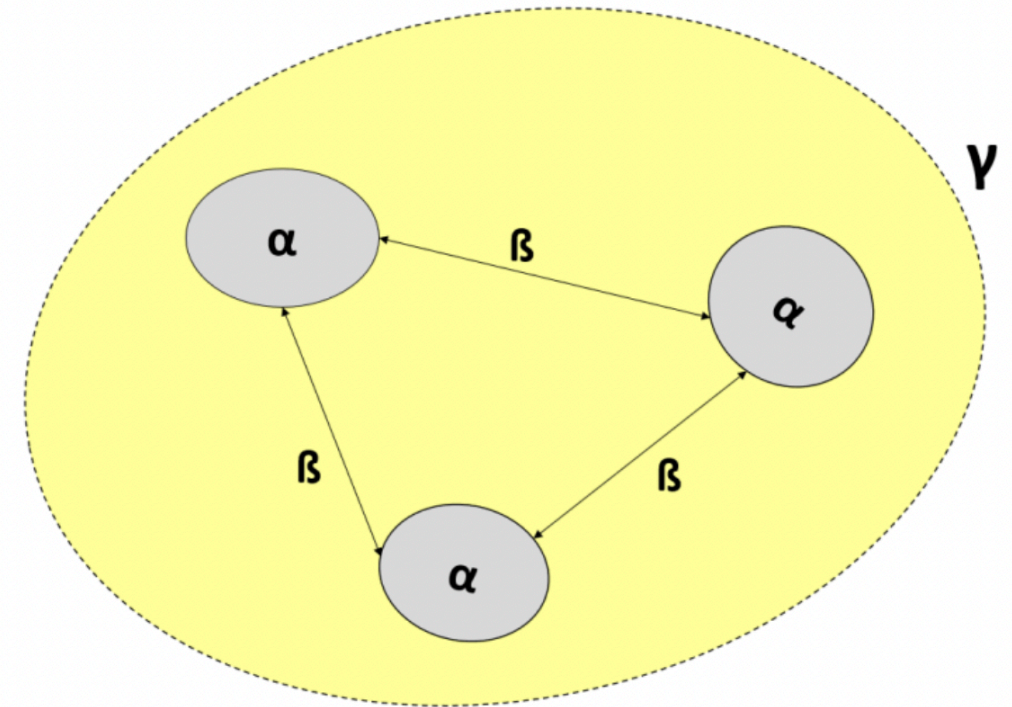
Alpha Diversity

Alpha diversity describes the species diversity *within* a community at a small scale or local scale, generally the size of one ecosystem.

Beta diversity describes the species diversity *between* two communities or ecosystems.

The extent of change in community composition, or degree of community differentiation, in relation to a complex-gradient of environment, or a pattern of environments

Gamma diversity is studied at a very large scale—a biome—where species diversity is compared between many ecosystems. It could range over areas like the entire slope of a mountain, or the entire littoral zone of a sea shore.



Alpha Diversity

Alpha diversity describes the species diversity *within* a community at a small scale or local scale, generally the size of one ecosystem.

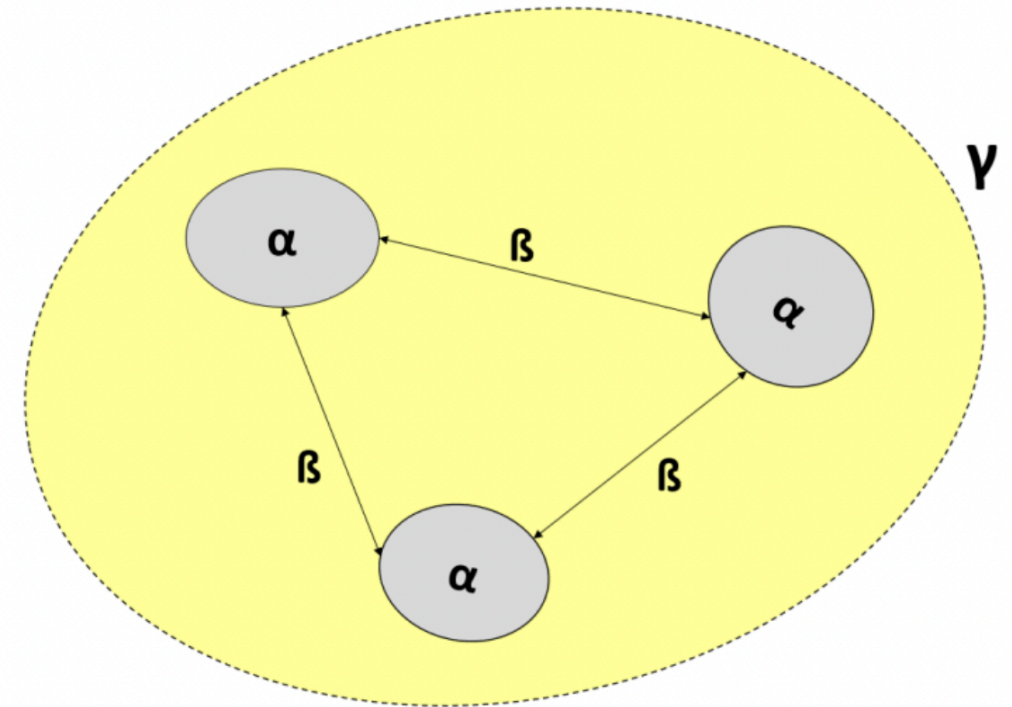
Beta Diversity

Beta diversity describes the species diversity *between* two communities or ecosystems.

The extent of change in community composition, or degree of community differentiation, in relation to a complex-gradient of environment, or a pattern of environments

Gamma Diversity

Gamma diversity is studied at a very large scale—a biome—where species diversity is compared between many ecosystems. It could range over areas like the entire slope of a mountain, or the entire littoral zone of a sea shore.



β Diversity

Inter-sample comparison of the community composition

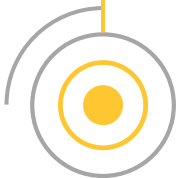
- **Measure of the similarities/dissimilarities between the samples** according to specific criteria of the MEASURE under consideration (e.g. Unifrac, Bray-curtis)
- **Highlight structure** by Ordination Plot in low dimensional space
e.g. PCoA, PCA, Db-RDA, Biplot
- **Test the structure differences & identify main variables/Taxa**
e.g. Permanova, differential abundance



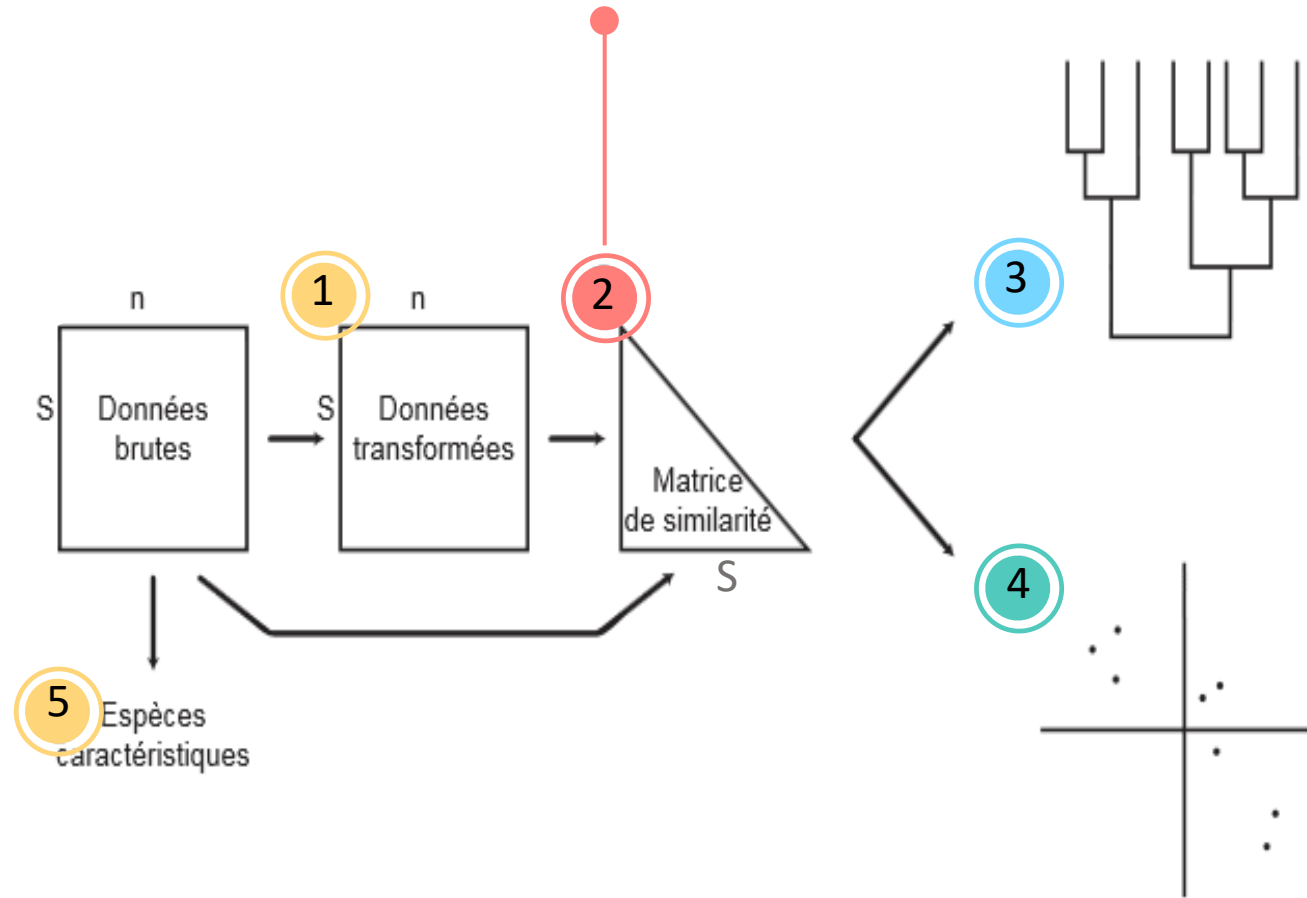
β Diversity

Inter-sample comparison of the community composition

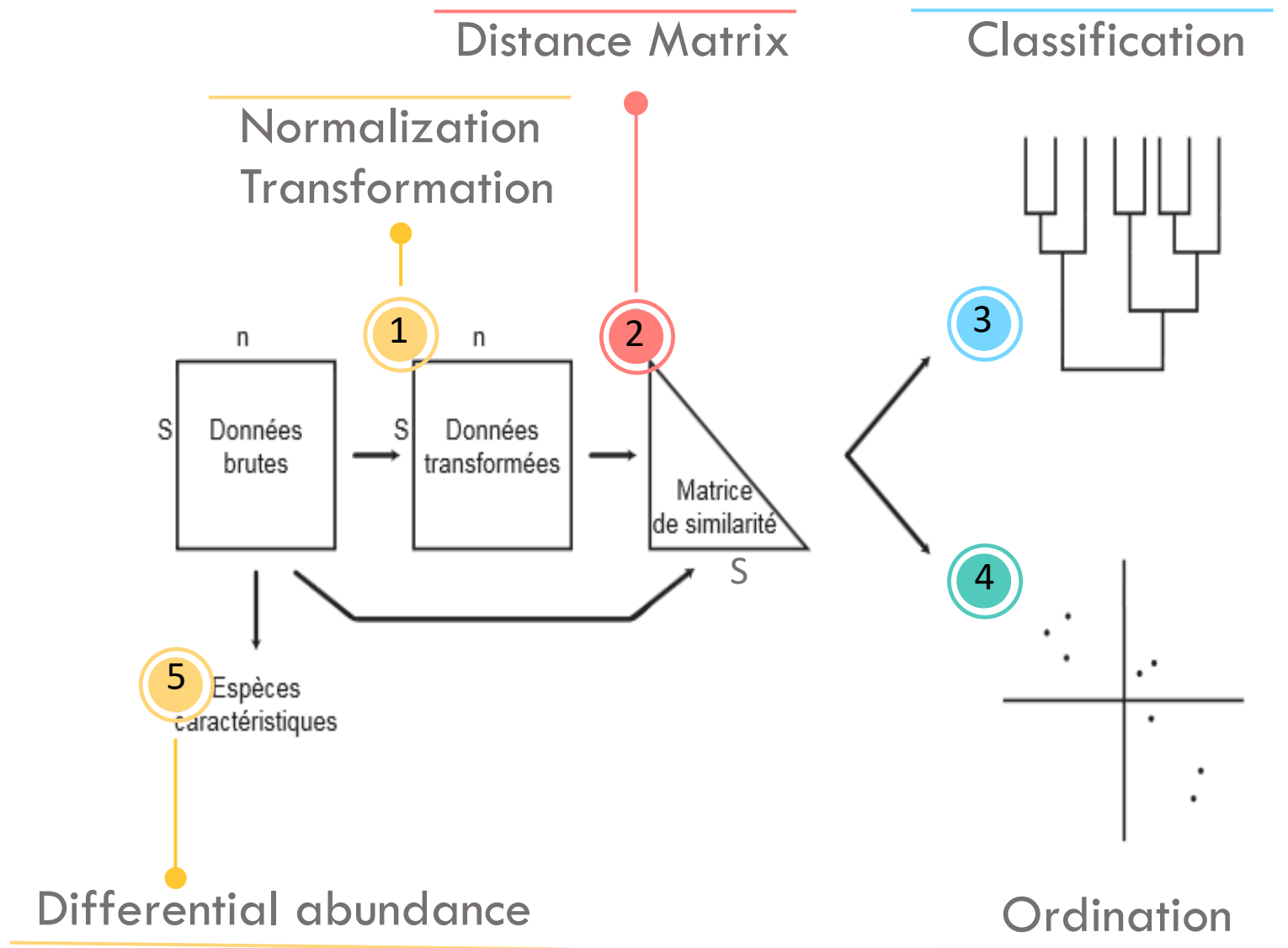
- **Measure of the similarities/dissimilarities between the samples** according to specific criteria of the MEASURE under consideration (e.g. Unifrac, Bray-curtis)
- **Highlight structure** by Ordination Plot in low dimensional space
e.g. PCoA, PCA, Db-RDA, Biplot
- **Test the structure differences & identify main variables/Taxa**
e.g. Permanova, differential abundance



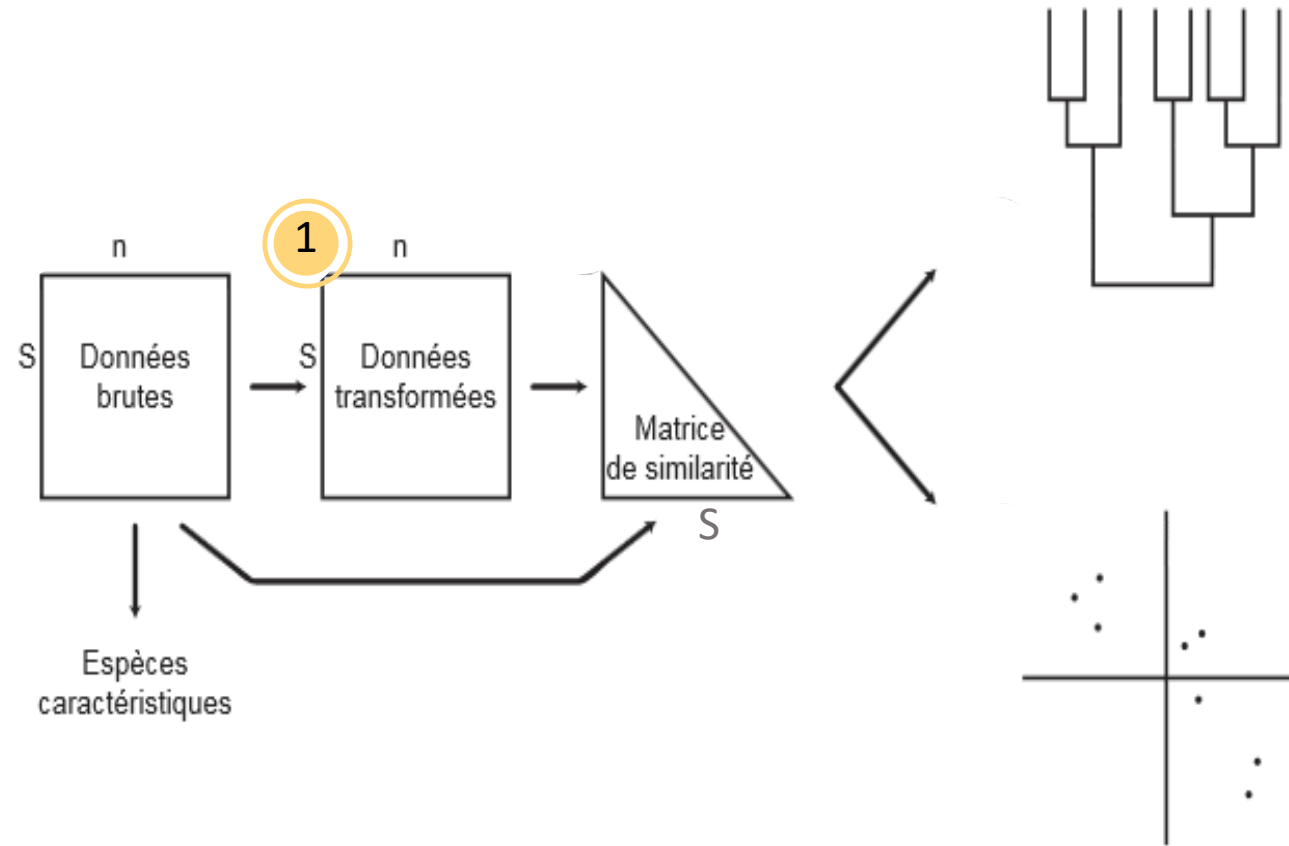
Overview of the Beta-analysis approach



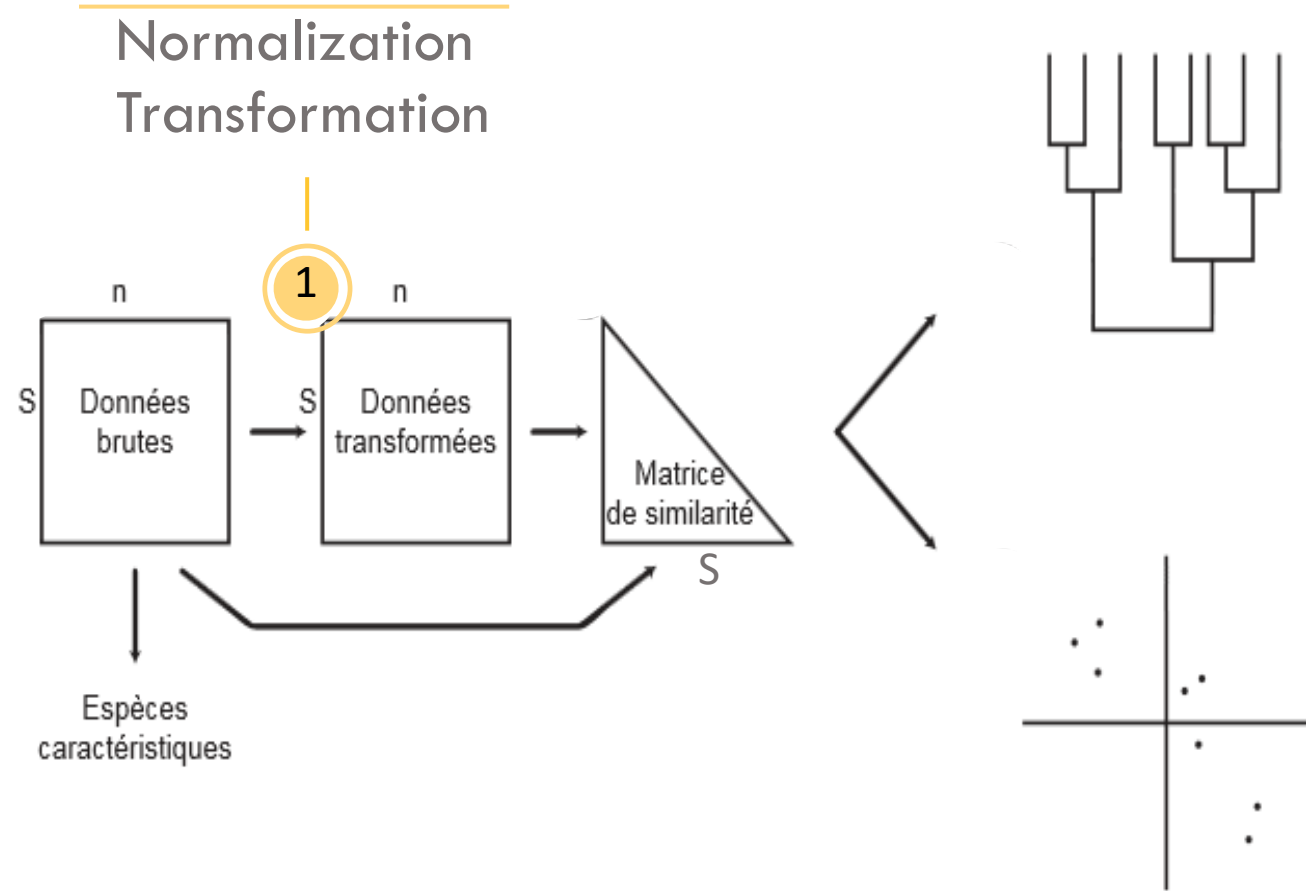
Overview of the Beta-analysis approach



Overview of the Beta-analysis approach



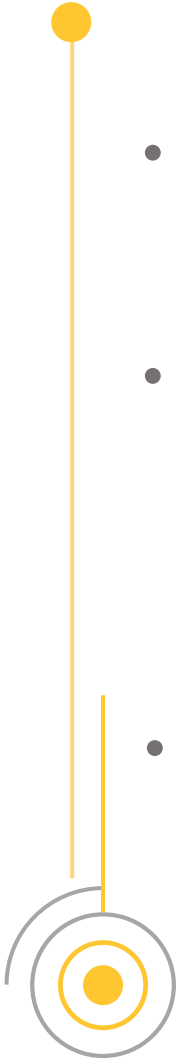
Overview of the Beta-analysis approach



Normalization & transformation

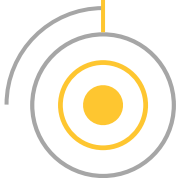
CHARACTERISTICS OF METABARCODING DATA

- The OTU/ASV count matrix is **sparse**, with often between 80 and 95% of the counts being zero
- The library sizes (sum of counts in each sample; also referred to as **sequencing depth**) **vary significantly**, sometimes by several orders of magnitude, making it nonsensical to compare counts directly between samples, since they each represent a different fraction of the composition of a given sample.
- The variances of these count distributions are greater than their means, a phenomenon known as **overdispersion**



CHARACTERISTICS OF METABARCODING DATA

- The OTU/ASV count matrix is **sparse**, with often between 80 and 95% of the counts being zero
- The library sizes (sum of counts in each sample; also referred to as **sequencing depth**) **vary significantly**, sometimes by several orders of magnitude, making it nonsensical to compare counts directly between samples, since they each represent a different fraction of the composition of a given sample.
- The variances of these count distributions are greater than their means, a phenomenon known as **overdispersion**



Normalization & transformation

Correcting library size, sampling fraction



- **Rarefying : Sub-sampling normalization** (Use rarefaction curves for the minimal library size, remove samples etc)
- **Scaling** : Divide each abundance by a scaling factor to eliminate bias from unequal sampling fraction

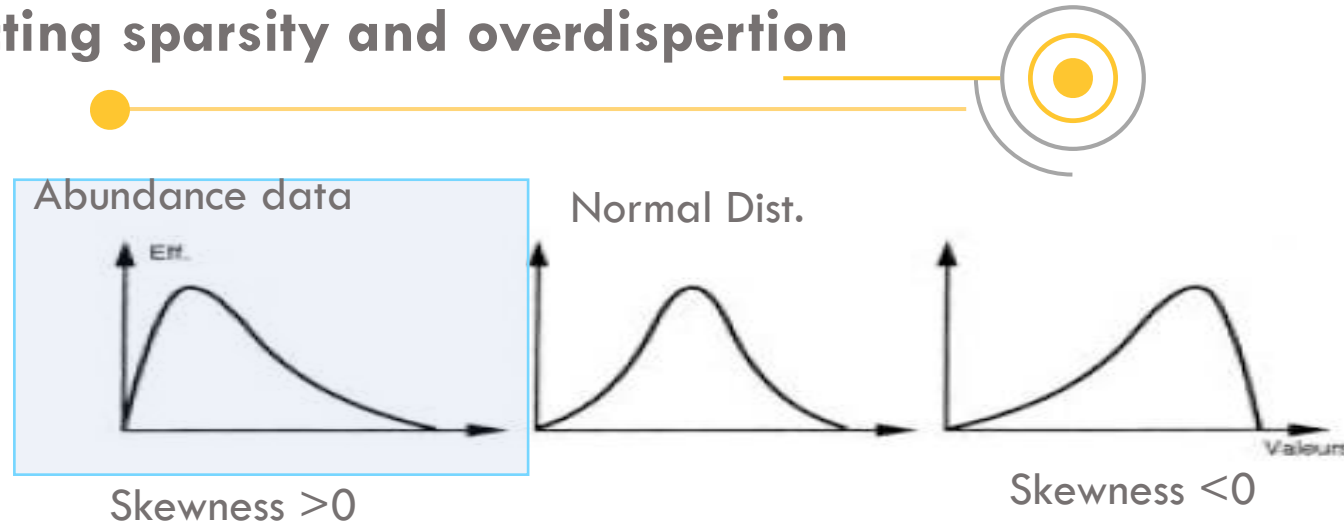
- CSS : Cumulative Sum Scaling (MetagenomeSeq R)
- TMM: Trimmed Mean of M-values (Edge R)
- TSS : Total Sum Scaling = relative abundance

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

Normalization & transformation

Correcting sparsity and overdispersion

Sparse Data
=
contain many Zeros



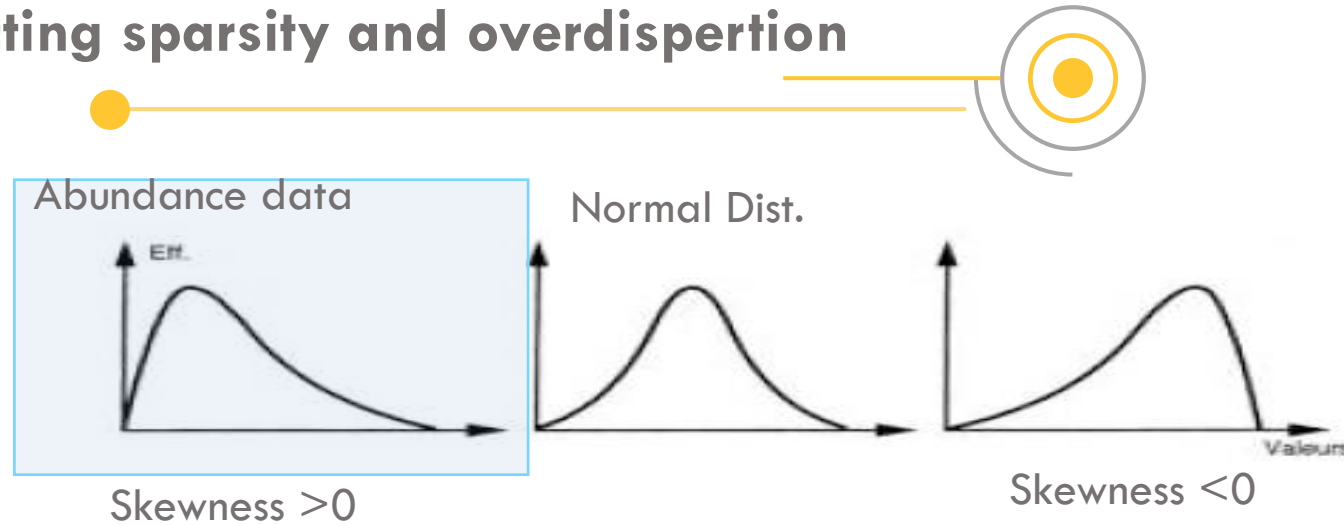
Why transformation ?

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

Normalization & transformation

Correcting sparsity and overdispersion

Sparse Data
=
contain many Zeros



Why transformation ?

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

What kind of Transformations for species abundance data



- Most of the transformation can be perform with `decostand()` from `Vegan`

- Log $x+1$ \rightarrow `(log1p(data))`
- Square root \rightarrow `(sqrt(data))`
- double square \rightarrow `root(sqrt(sqrt(data)))`

Reduction of variation range: $\text{Log} > \text{double sqrt} > \text{sqrt}$

\rightarrow Be careful of the deformation of data with these transformations!

What kind of Transformations for species abundance data



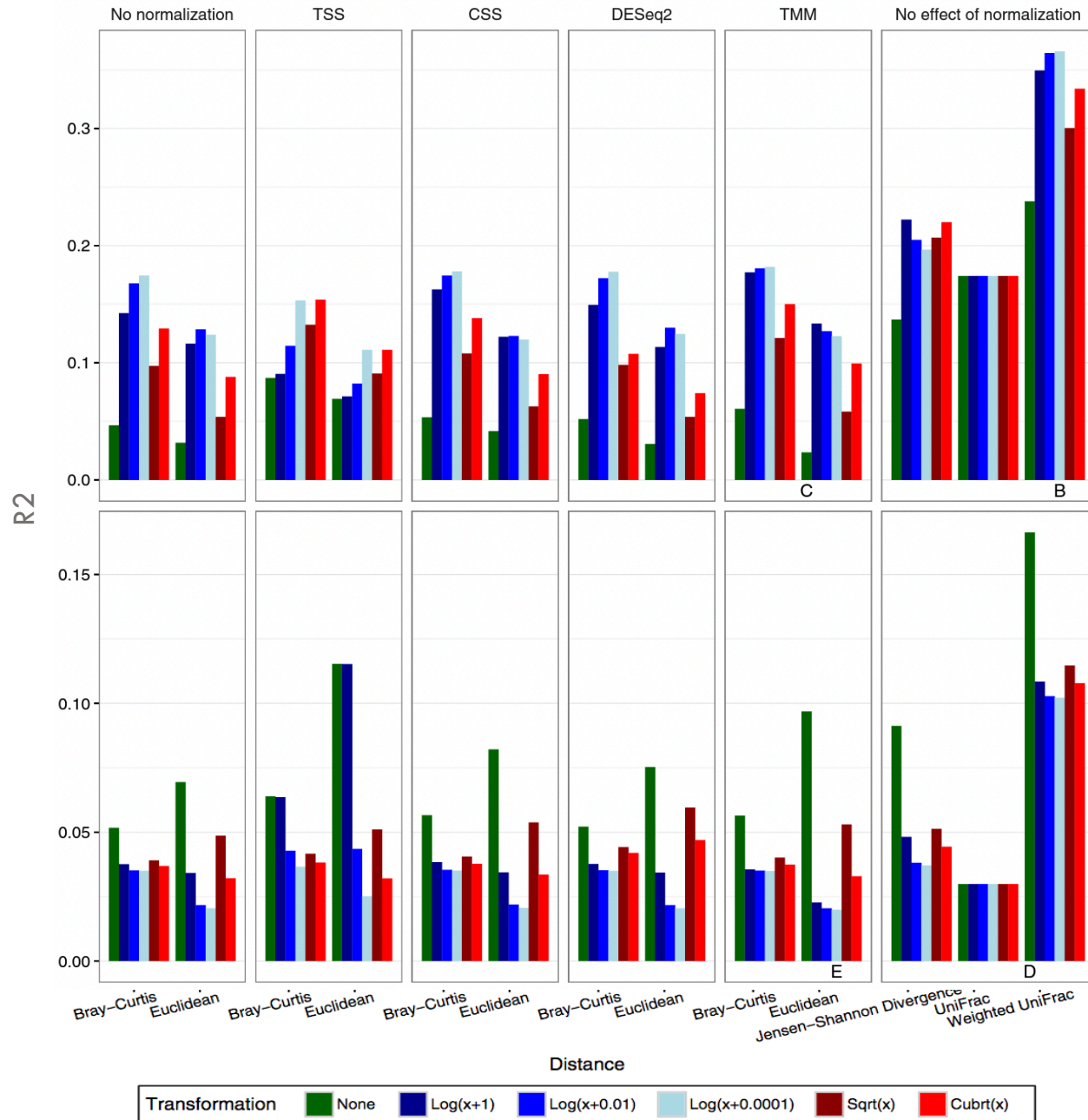
- Most of the transformation can be perform with `decostand()` from `Vegan`

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

Reduction of variation range: `Log > double sqrt > sqrt`

→ Be careful of the deformation of data with these transformations!

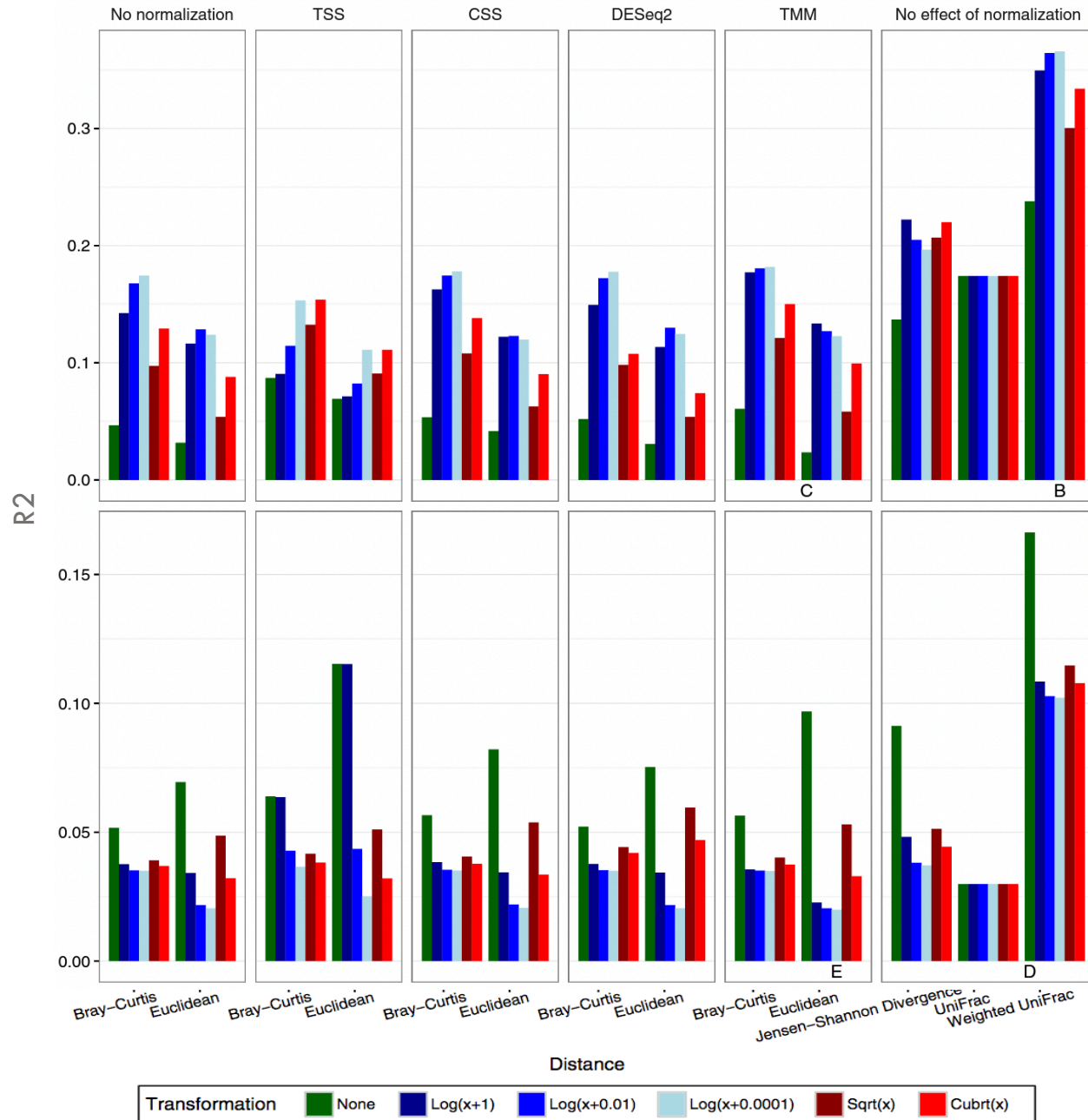
Normalization & transformation



Impact on sample separation

- Normalisation: negligible effect
- Transformation: Log+x was the best transformation (reduced the weight of highly abundant ASV/OTUs / increase the weight of low abundant ASV/OTUs)
- Distance or dissimilarity metric: highest separation effect

Normalization & transformation



Impact on sample separation

- Normalisation: negligible effect
- Transformation: Log+x was the best transformation (reduced the weight of highly abundant ASV/OTUs / increase the weight of low abundant ASV/OTUs)
- Distance or dissimilarity metric: highest separation effect

Normalization & transformation



To the point of view of Compositional data : CoDA

CoDA Aitchison's Log-ratio based-methods :

- Eliminate the sampling fraction effect
- Isometric log-ratio (ILR)
- Centered log-ratio (CLR)
- Additive log-ratio (ALR)
- Phylogenetic Isometric Log-Ratio (phILR)

If you want to test it : zcompositions, composition R packages,
easycoda

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

Normalization & transformation



To the point of view of Compositional data : CoDA

CoDA Aitchison's Log-ratio based-methods :

- Eliminate the sampling fraction effect
- Isometric log-ratio (ILR)
- Centered log-ratio (CLR)
- Additive log-ratio (ALR)
- Phylogenetic Isometric Log-Ratio (phILR)

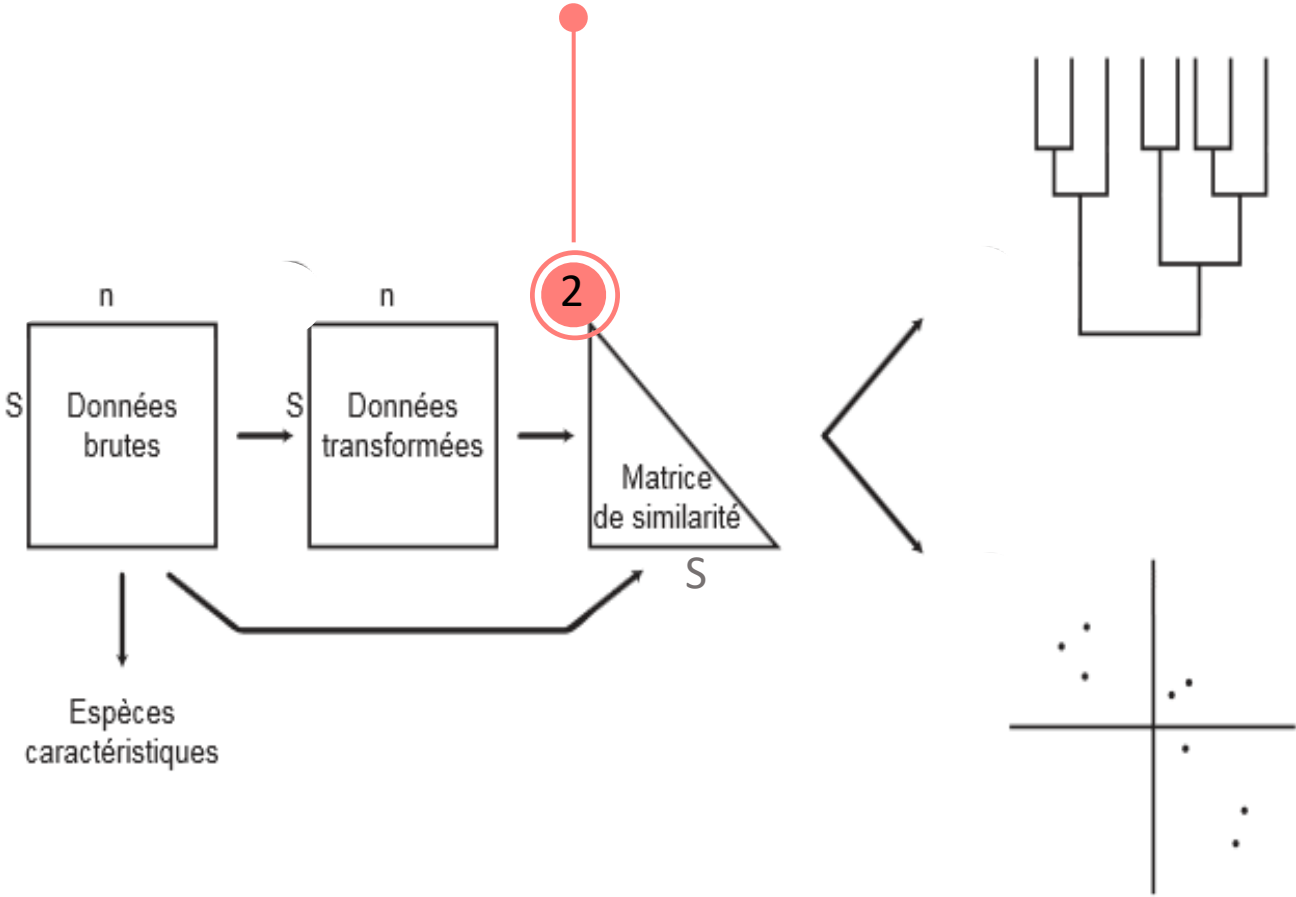
If you want to test it : zcompositions, composition R packages,
easycoda

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

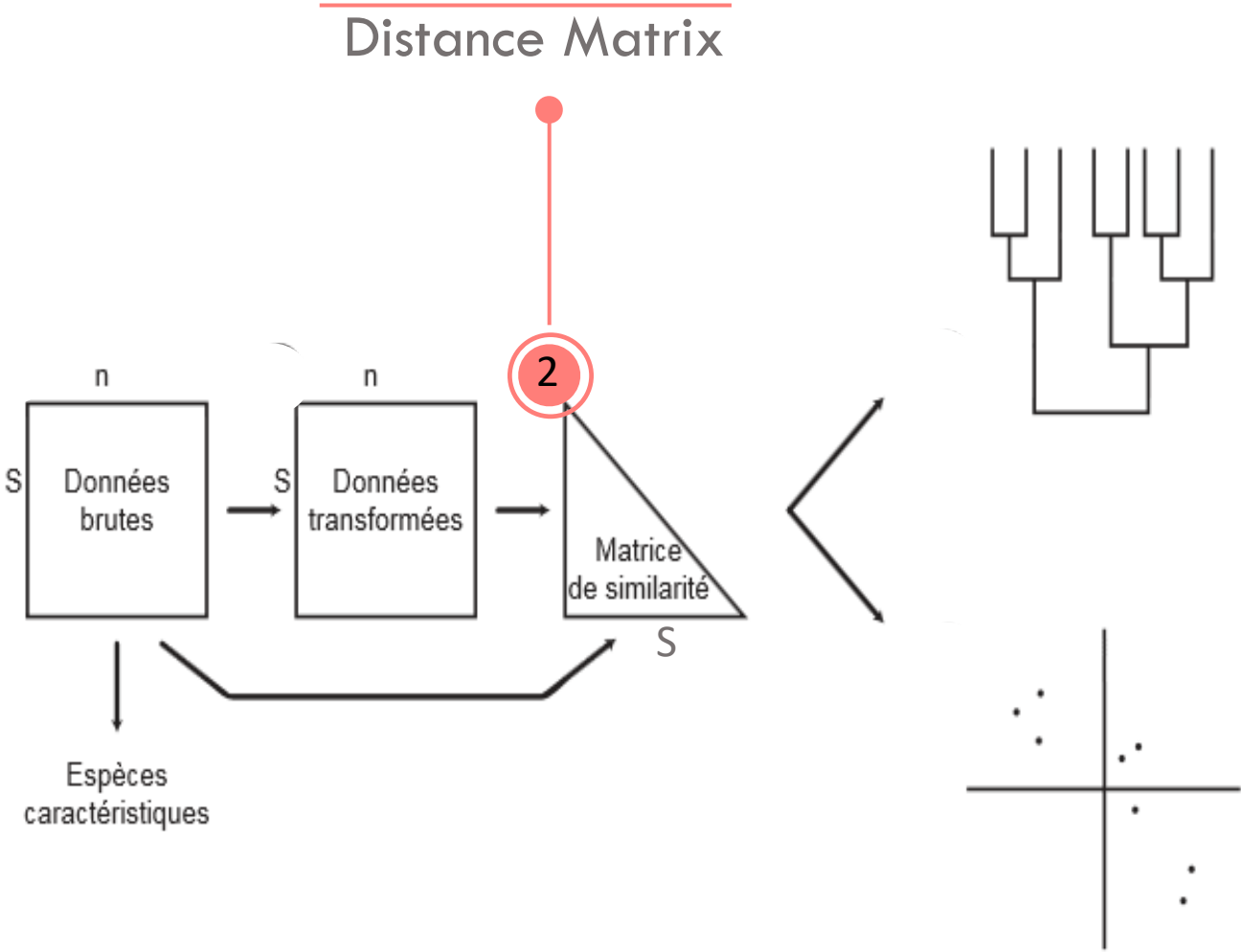


Practice

Overview of the Beta-analysis approach

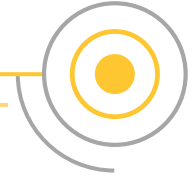


Overview of the Beta-analysis approach



Distance matrix

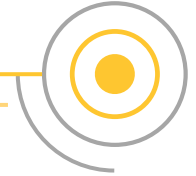
Similarity & Distance: Evaluate the ecological resemblance



- Quantifying ecological **resemblances between samples**, including **similarities (S)** and **dissimilarities** (or distances), is the basic approach of handling **multivariate ecological data**
- Two samples, which contain the same species with the same abundances, have the highest similarity, the similarity decreases with the differences in species composition
- Ordination methods operate with **distances or dissimilarities** between samples (e.g. $1-S$)

Distance matrix

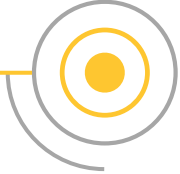
Similarity & Distance: Evaluate the ecological resemblance



- Quantifying ecological **resemblances between samples**, including **similarities (S)** and **dissimilarities** (or distances), is the basic approach of handling **multivariate ecological data**
- Two samples, which contain the same species with the same abundances, have the highest similarity, the similarity decreases with the differences in species composition
- Ordination methods operate with **distances or dissimilarities** between samples (e.g. $1-S$)

Distance matrix

The process: ASV/OUT Abundance to Distance to Ordination of samples

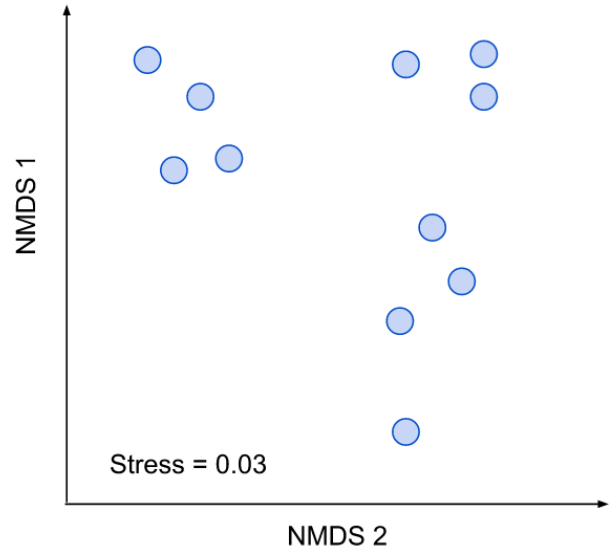


		Variables			
		X1	X2	X3	X4
Samples	S1	14	2	14	14
	S2	10	14	0	8
	S3	0	5	0	2
	S4	0	0	1	0

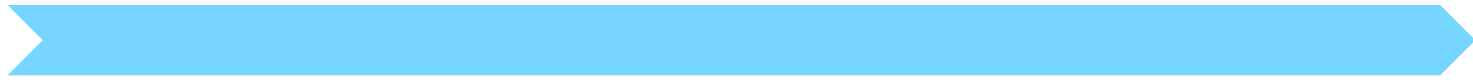
Abundance Matrix
Contingency table
OUT/ASV table

		Samples			
		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
matrix



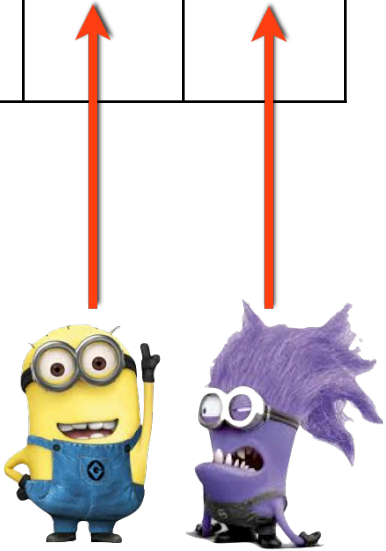
Ordination plot in a reduced
dimensional space



Distance matrix

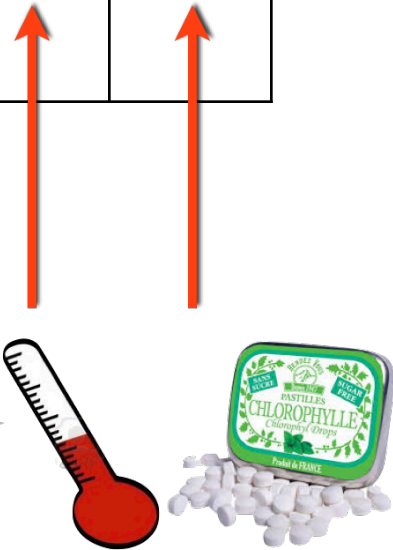
Contingency table
ASV/OTU Abundance Table

	Var1	Var2
Obj1		
Obj2		
Obj3		



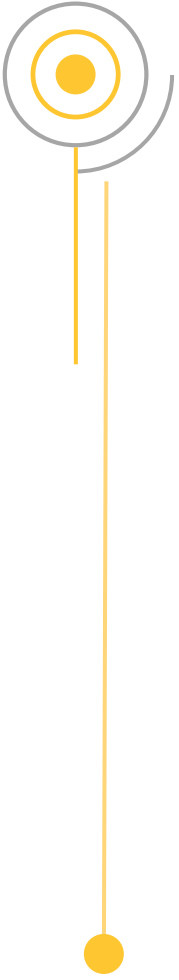
Environmental table
(Sample_data in phyloseq)

	Var1	Var2
Obj1		
Obj2		
Obj3		



Distance matrix

ASV/OTU Abundance Table



Q mode

	Obj1	Obj2
Obj1		
Obj2		

Association measure
=
Distance
Dissimilarity
Similarity

R mode

	Var1	Var2
Var1		
Var2		

Association measure
=
Correlation
Covariance

When pairs of objects are compared, we talk about Q mode. We talk about R mode when variables are compared.

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 samples for which 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 Consider them as missing data!

Distance matrix

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



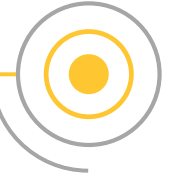
- Species composition data are **sparse matrix**, which means that it contains lot of zeros, double zeros
- Double zero” is a situation when **certain species are missing** in both compared community samples for which similarity/distance will be next calculated!

	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 Consider them as missing data!

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)

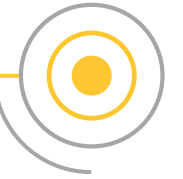
- Recommendation is to use **dissimilarity indices or distance-based** method that do **not take into account the double zero as a resemblance!!!**

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

Questions you should ask yourself before choosing a dissimilarity/distance metric



- Do I compare variables or objects? (**R vs Q modes**)
- Do I use ASV/OTU/species variables or another type (e.g.; physico-chemical)? **Asymmetrical Vs symmetrical dissimilarity or distance index**
- What type of data do I have? **Binary Vs quantitative Vs multifactor**



Distance matrix

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

- for binary data (presence/absence)
- for quantitative data
- 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 Distance chord Distance d'Hellinger	Non Normalisation de Chord Transformation d'Hellinger	vegdist decostand puis dist decostand puis dist
		Binaire	Dissimilarité de Jaccard Dissimilarité de Sorensen Dissimilarité de Ochiai	/ / /	dist.binary
		Multifacteur	/	/	/
R	Asymétrique	Quantitative	Corrélation de Pearson Corrélation de Spearman Distance du Chi carré	/ / Transformation du Chi carré	cor cor decostand puis dist
		Binaire	Dissimilarité de Jaccard Dissimilarité de Sorensen Dissimilarité de Ochiai	/ / /	dist.binary
	Symétrique	Binaire	Corrélation de Pearson	/	cor
		Multifacteur	Corrélation de Pearson	/	cor

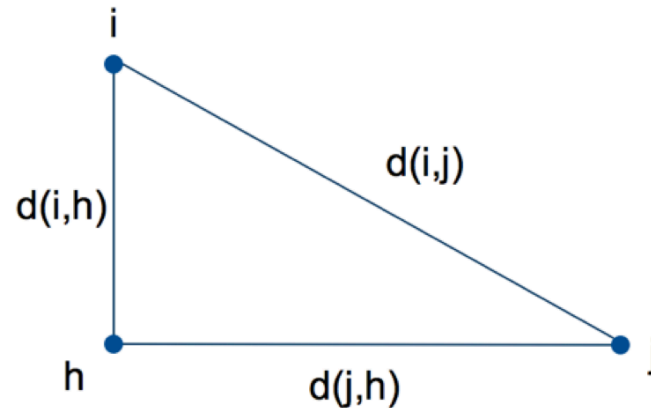
Distance matrix

Most common dissimilarities/distance used for species data

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

Properties of distance measures

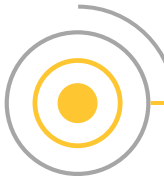
- D1: $d(i,j) \geq 0$
- D2: $d(i,i) = 0$
- D3: $d(i,j) = d(j,i)$
- D4: $d(i,j) \leq d(i,h) + d(h,j)$ (triangle inequality)



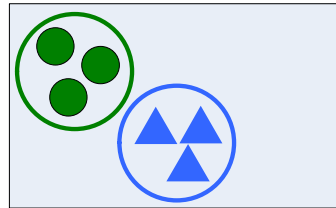
Distance matrix

Euclidean distance

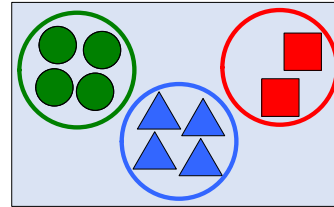
$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$



Community 1



Community 2



$$(x_{i1} - x_{j1})^2 = (3 - 4)^2 = 1$$

$$(x_{i2} - x_{j2})^2 = (3 - 4)^2 = 1$$

$$(x_{i3} - x_{j3})^2 = (0 - 2)^2 = 4$$

$$D(i, j) = \sqrt{1 + 1 + 4} = 2.45$$



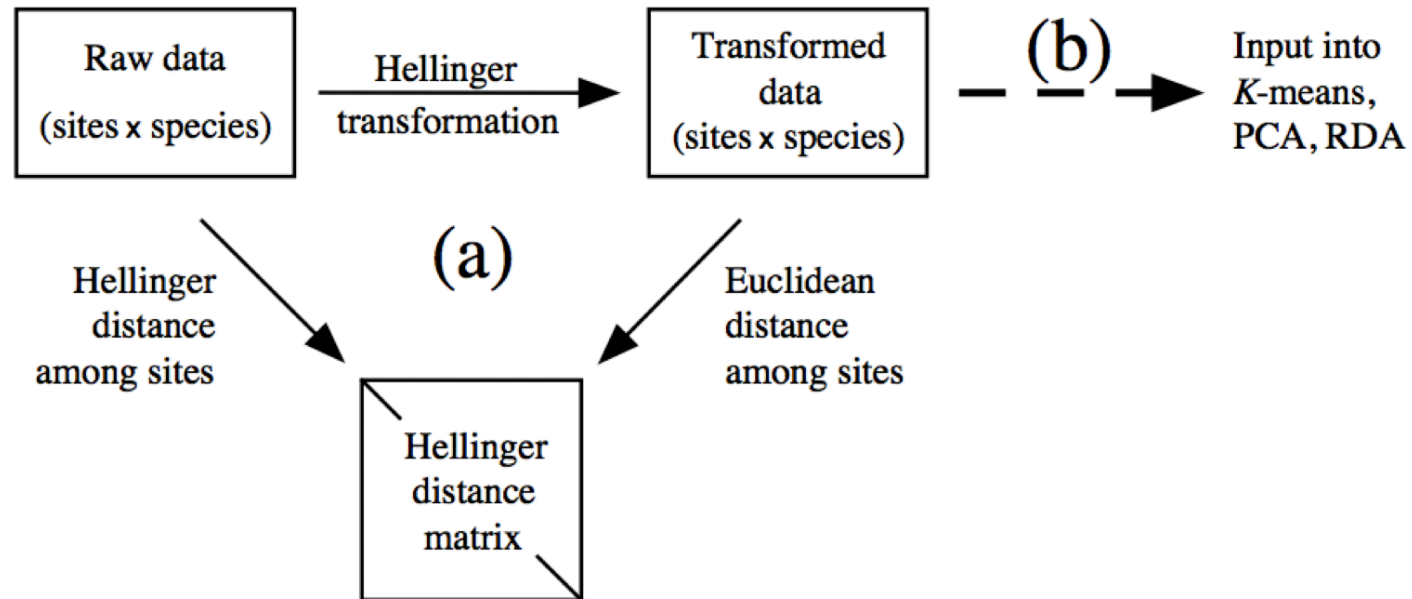
- Highly impacted by the unit or scale of the descriptor
- Standardize or not standardize?????
- Standardise if descriptors have not the same units

Distance matrix

Hellinger distance

Two steps calculation:

- Hellinger Transformation
- Euclidian distance calculation



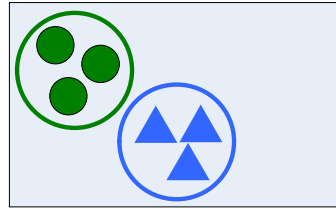
Distance matrix

Hellinger distance

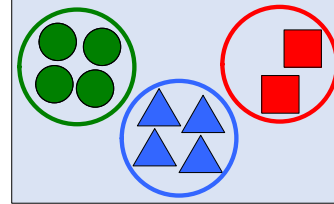
$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}}$$

Hellinger transformation

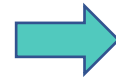
Community 1



Community 2



	Sp1	Sp2	Sp3
Com1	3	3	0
Com2	4	4	2



	Sp1	Sp2	Sp3
Com1	0.7	0.7	0
Com2	0.6	0.6	0.4



	Com1	Com2
Com1	0	0.42
Com2	0.42	0

Hellinger transformation

Euclidean distance



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.

Dissimilarities

- More flexible than distances

D1: $d(i,j) \geq 0$

D2: $d(i,i) = 0$

D3: $d(i,j) = d(j,i)$

	M	P	H
M	10	1	8
P		10	5
H			10

- Example: What do you think, how different are the topics Mathematics, Physics, History on a scale from 0 to 10 (very different)?
- Could also work with “Similarities” (e.g. 1-Dissimilarity)

Dissimilarities

- More flexible than distances

D1: $d(i,j) \geq 0$

D2: $d(i,i) = 0$

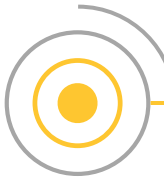
D3: $d(i,j) = d(j,i)$

	M	P	H
M	10 0	1	8
P		10 0	5
H			10 0

- Example: What do you think, how different are the topics Mathematics, Physics, History on a scale from 0 to 10 (very different)?
- Could also work with “Similarities” (e.g. 1-Dissimilarity)

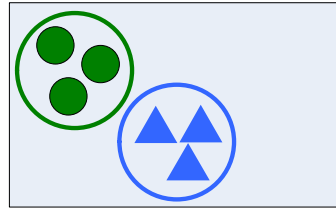
Distance matrix

Bray-Curtis

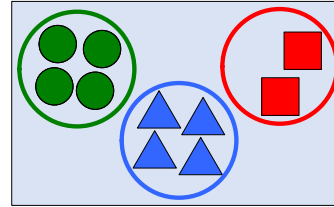


$$BC_{jk} = 1 - \frac{2 \sum_{i=1}^p \min(N_{ij}, N_{ik})}{\sum_{i=1}^p (N_{ij} + N_{ik})}$$

Community 1



Community 2



$\text{Min}(N_{ij}, N_{ik}) = 3 \text{ green} + 3 \text{ blue} = 6$

$\text{Sum}(N_{ij} + N_{ik}) = 6 \text{ (community 1)} + 10 \text{ (community 2)} = 16$

$\text{BC} = 1 - (2 \times 6) / 16 = 0.25$



- Values range from 0 (maximum of similarity) to 1
- Same sampling depth in each sample

Distance matrix

Choose the right distance/dissimilarity

Species abundance paradox data \Rightarrow
(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

Euclidian
distance

$$D_1(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^p (y_{1j} - y_{2j})^2}$$

Chord
Distance

$$D_3(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^p \left(\frac{y_{1j}}{\sqrt{\sum_{j=1}^p y_{1j}^2}} - \frac{y_{2j}}{\sqrt{\sum_{j=1}^p y_{2j}^2}} \right)^2}$$

Chi-Square
Distance

$$D_{18}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^p \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

Hellinger
Distance

$$D_{17}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^p \left[\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2}$$

Transformations

\Downarrow

$$y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^p y_{ij}^2}}$$

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

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

$$\mathbf{D}_1 = \begin{bmatrix} 0.0000 & 7.6158 & 9.0000 \\ 7.6158 & 0.0000 & 1.7321 \\ 9.0000 & 1.7321 & 0.0000 \end{bmatrix}$$

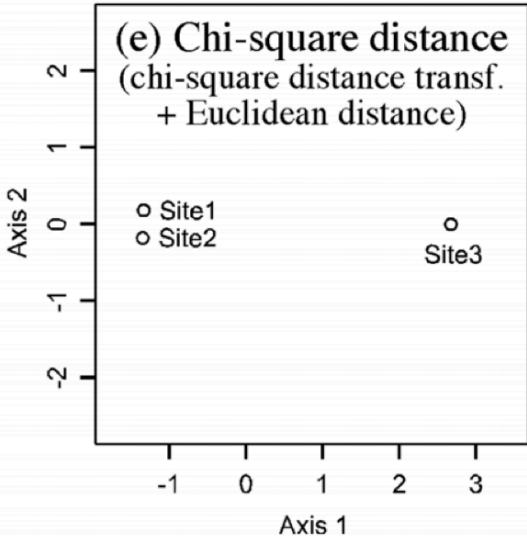
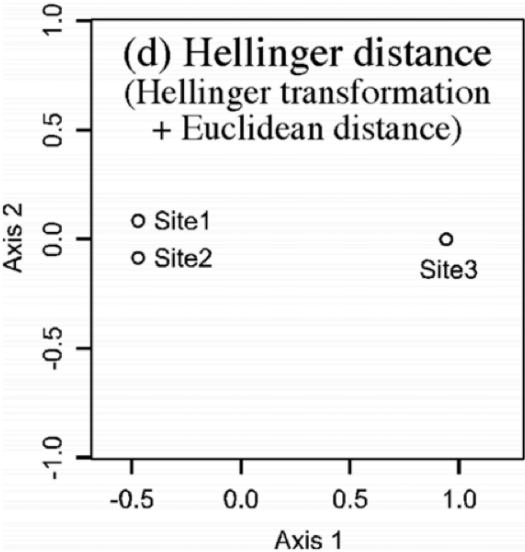
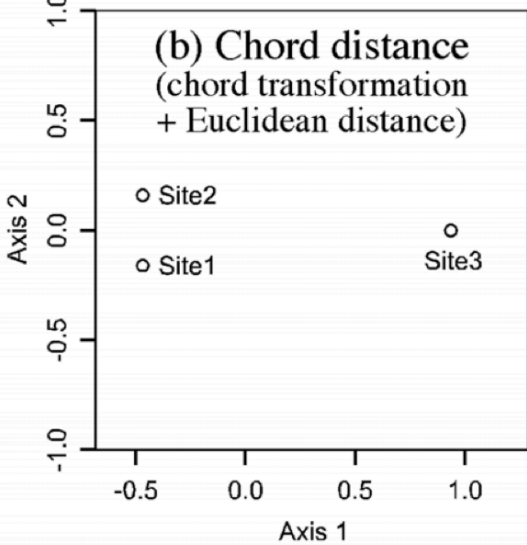
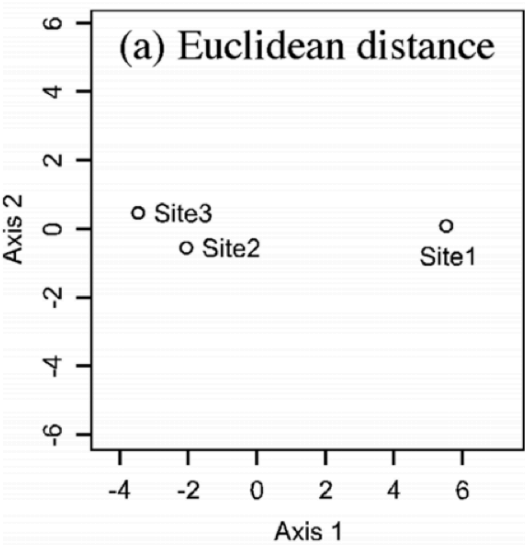
$$\mathbf{D}_3 = \begin{bmatrix} 0.0000 & 0.3204 & 1.4142 \\ 0.3204 & 0.0000 & 1.4142 \\ 1.4142 & 1.4142 & 0.0000 \end{bmatrix}$$

$$\mathbf{D}_{18} = \begin{bmatrix} 0.0000 & 0.2357 & 1.2472 \\ 0.2357 & 0.0000 & 1.2247 \\ 1.2472 & 1.2247 & 0.0000 \end{bmatrix}$$

$$\mathbf{D}_{17} = \begin{bmatrix} 0.0000 & 0.1697 & 1.4142 \\ 0.1697 & 0.0000 & 1.4142 \\ 1.4142 & 1.4142 & 0.0000 \end{bmatrix}$$

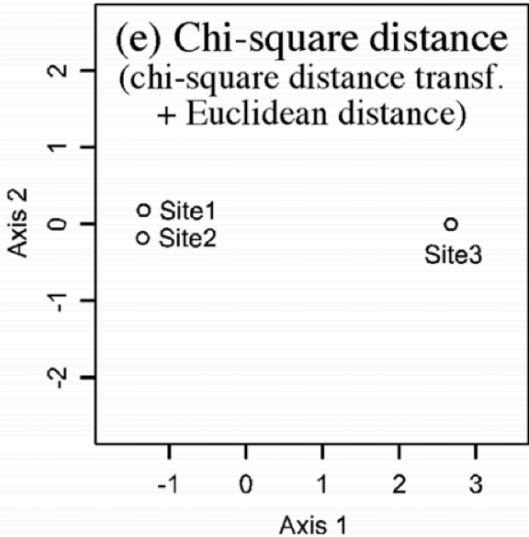
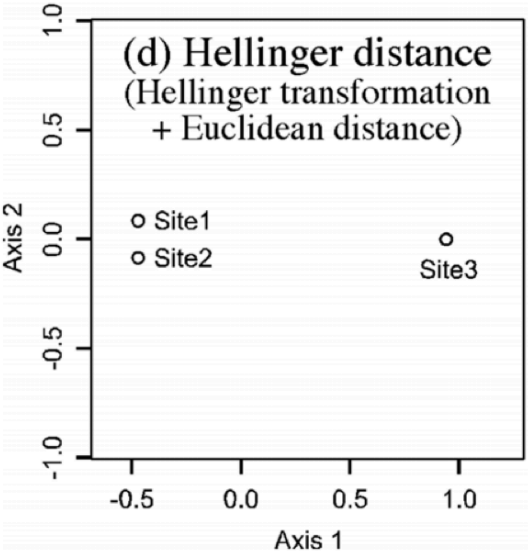
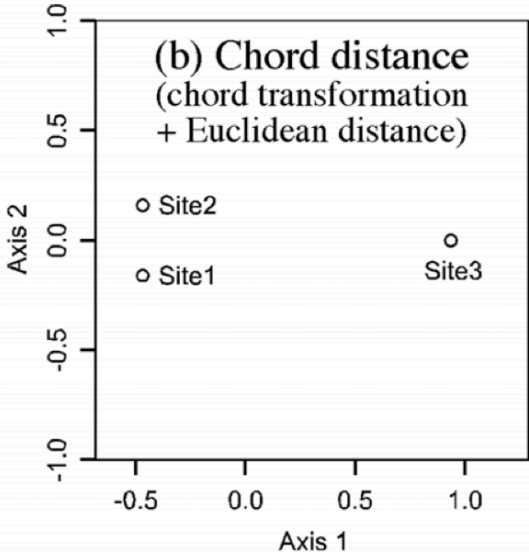
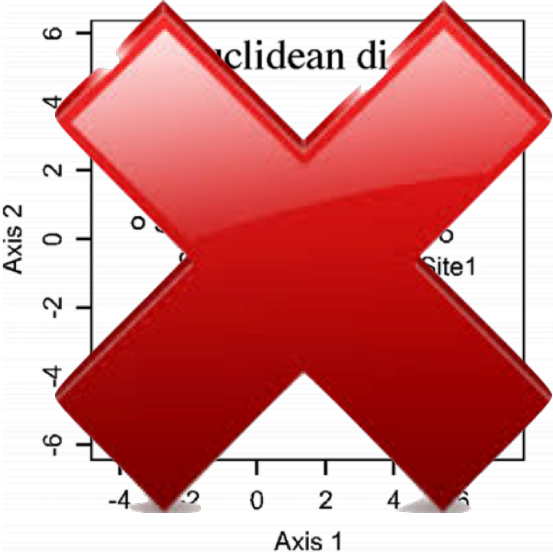
Distance matrix

Choose the right distance/dissimilarity



Distance matrix

Choose the right distance/dissimilarity



Distance matrix

Most common dissimilarities/distance used for species data

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

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

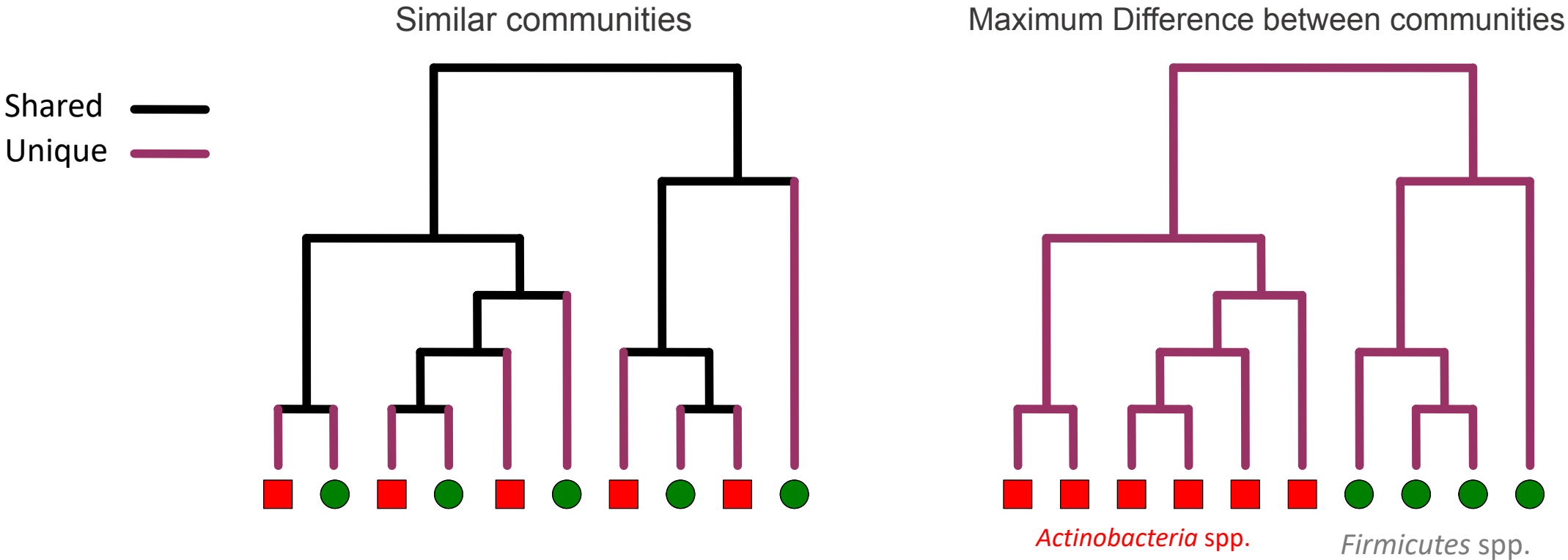
Two modes :

Unweighted Unifrac

Weighted Unifrac (take into account the relative abundance of taxa)

Distance matrix

Unweighted Unifrac



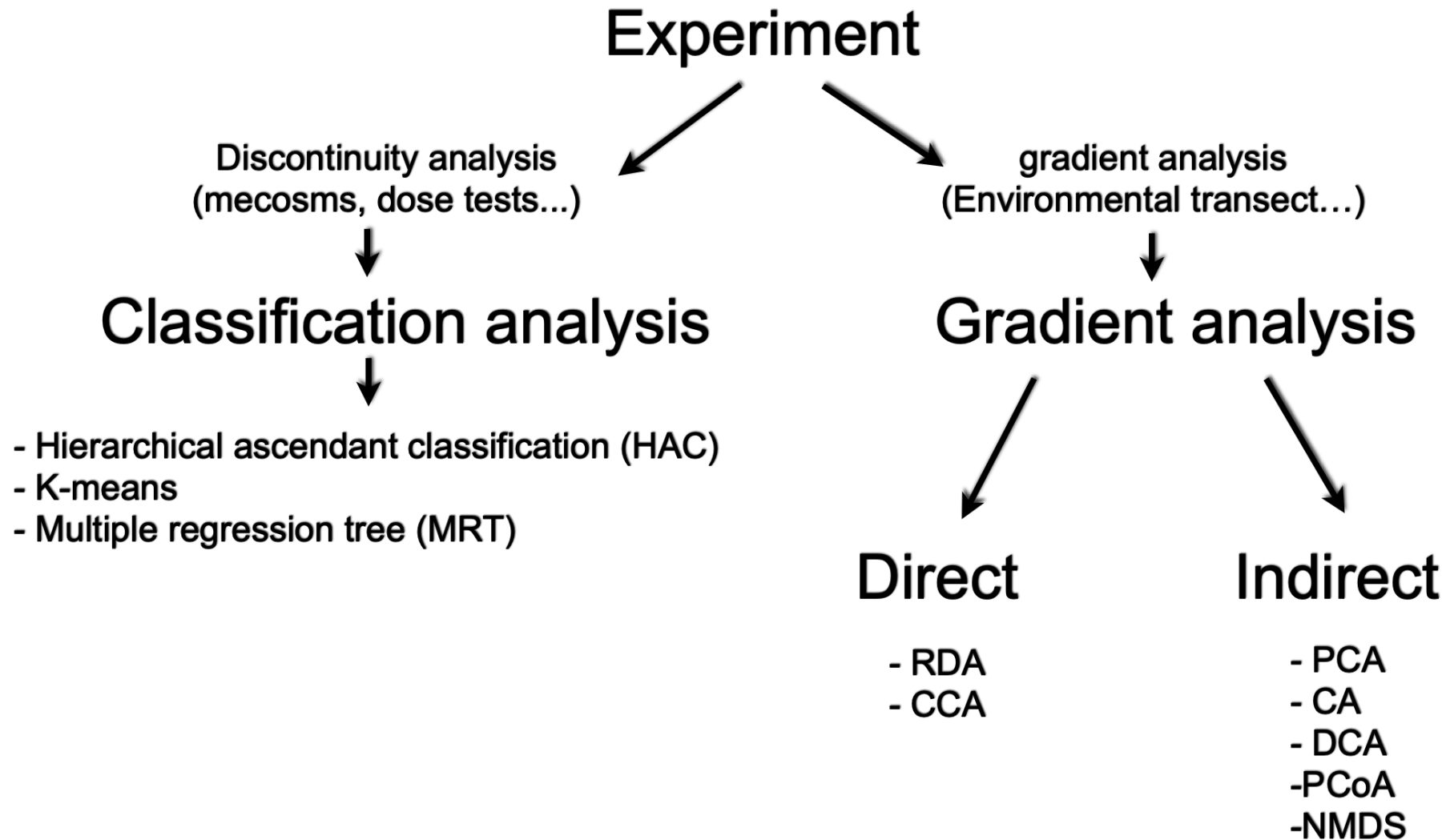
$$\text{Distance Measure of UniFrac} = \frac{\text{—————}}{\text{—————} + \text{—————}}$$

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

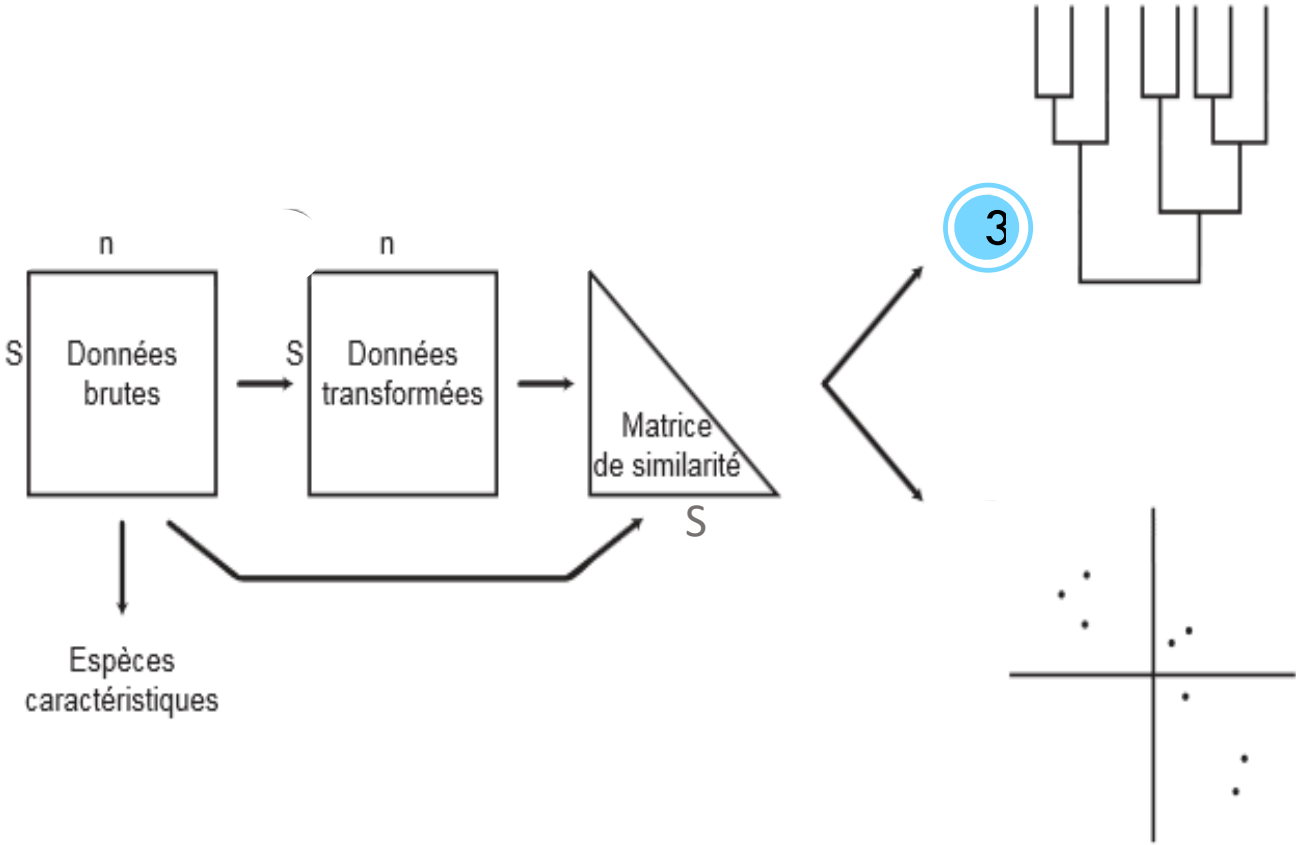


Practice

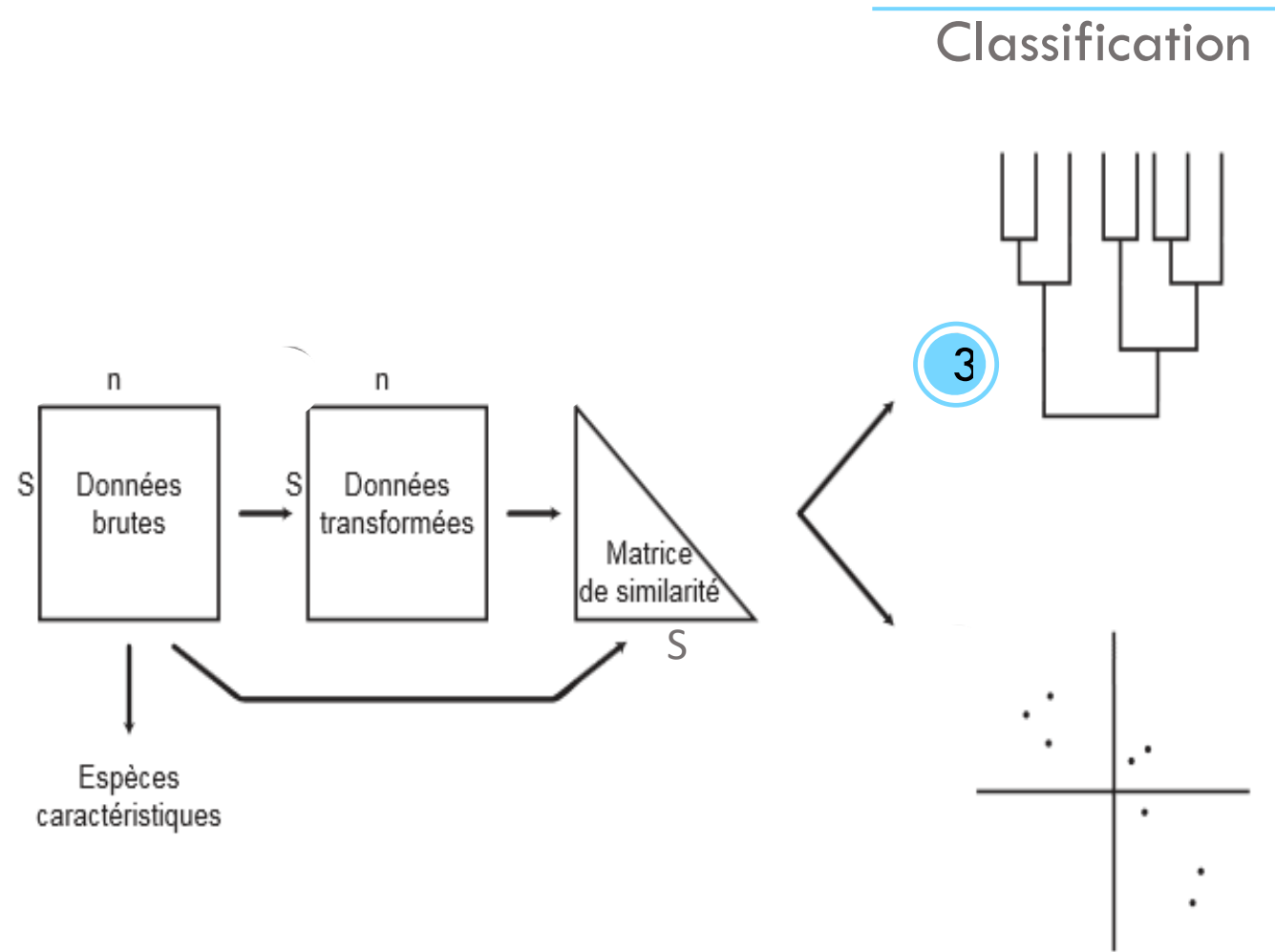
Classification and ordination summarize community data by producing a low-dimensional ordination space in which similar samples are plotted close together, and dissimilar samples are placed far apart. Ideally and typically, dimensions of this low dimensional space will represent important and interpretable environmental gradients.



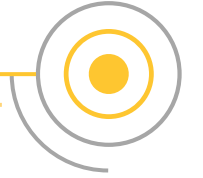
Overview of the Beta-analysis approach



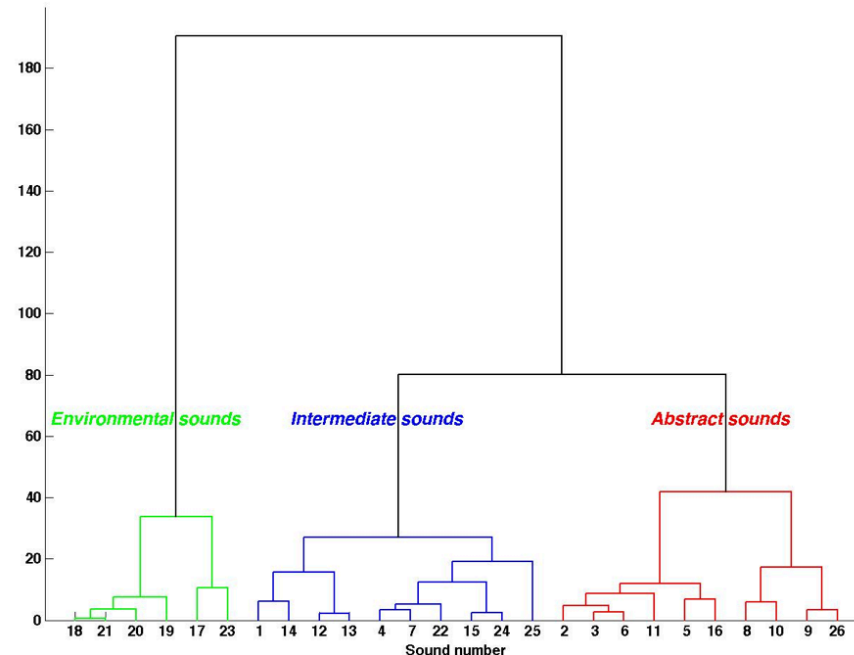
Overview of the Beta-analysis approach



Classification methods (cluster 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 and 2) the aggregation criterion chosen for the partitions formed



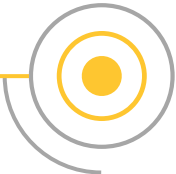
Classification methods (cluster analysis)



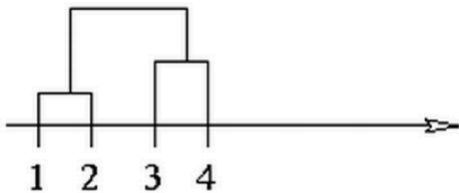
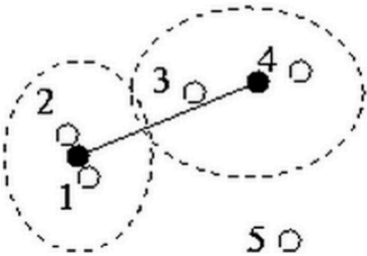
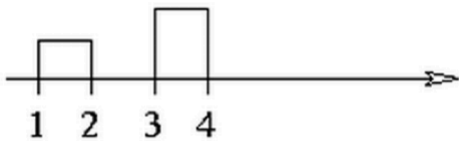
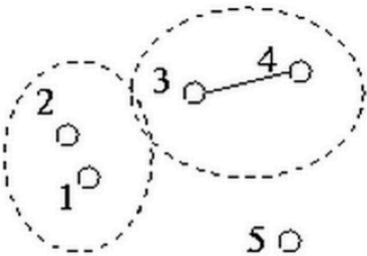
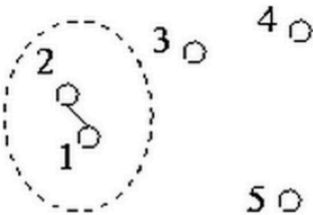
- Not a classical statistical methods in that no hypothesis is formulated
- The user interpret if the final topology has an ecological explanation

-
- Hierarchical ascendant classification (HAC)
 - K-means
 - Multiple regression tree (MRT)

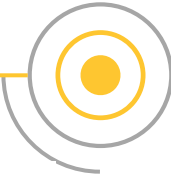
Hierarchical ascendant classification (HAC)



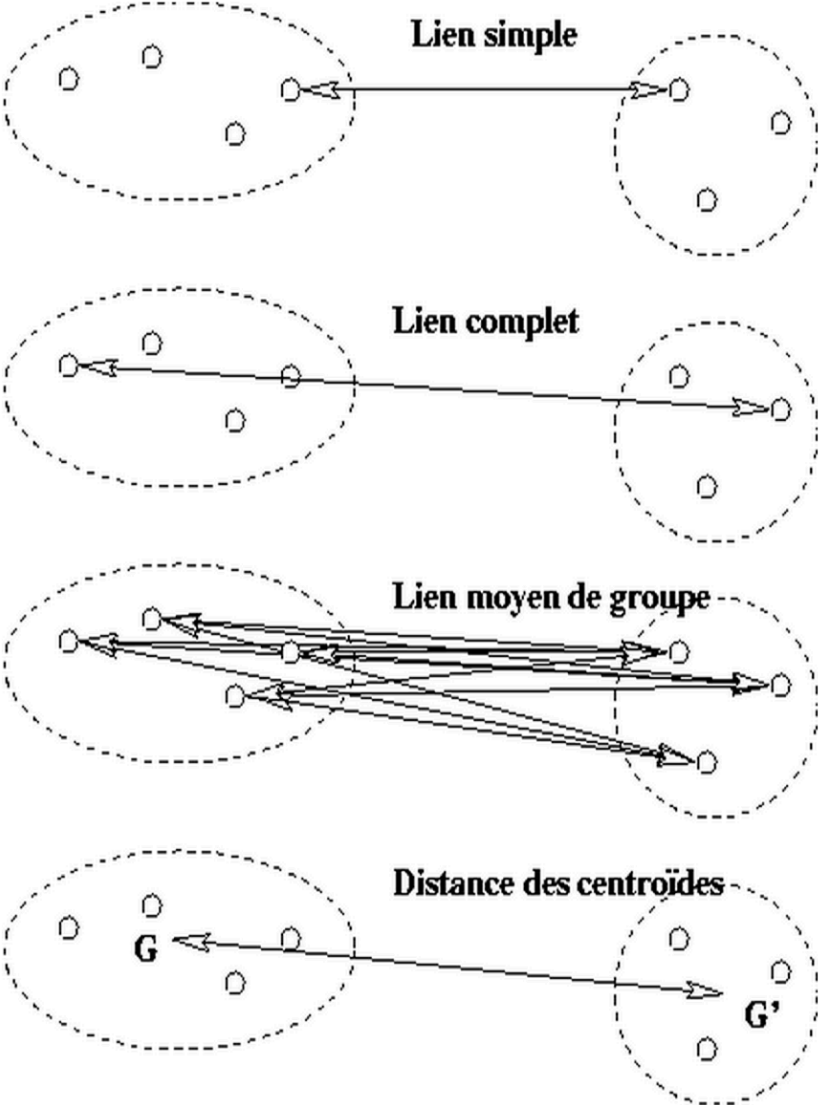
Select the closest objects and cluster them



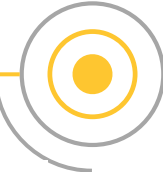
Aggregation criteria used in HAC



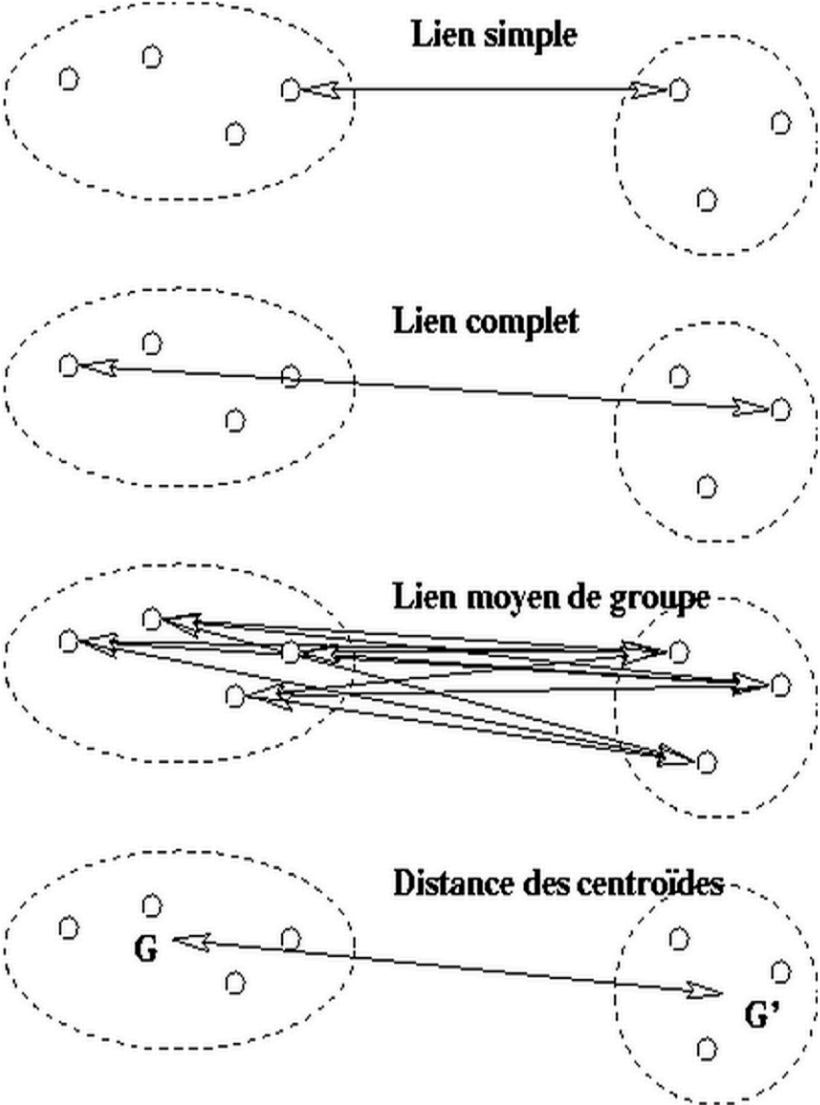
Agglutinates at each stage the two clusters having the smallest distance between their nearest neighbors.



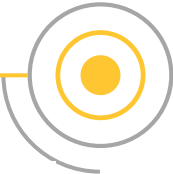
Aggregation criteria used in HAC



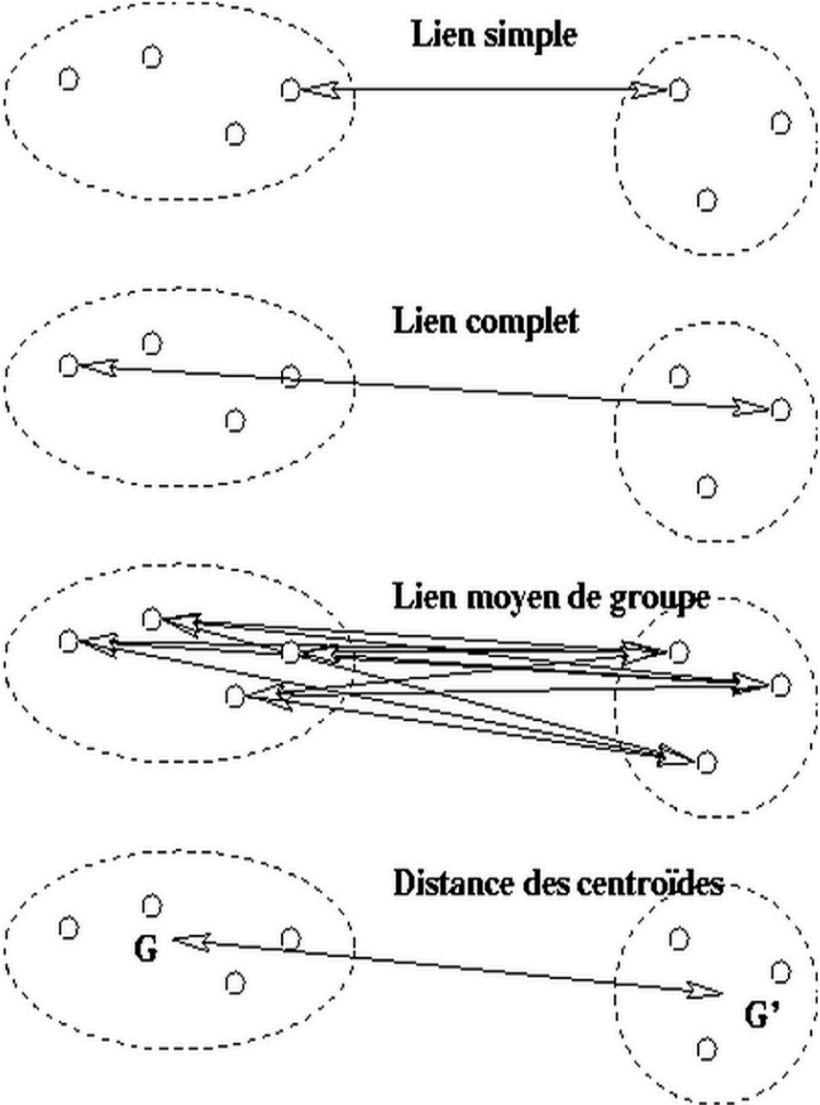
Tends to quickly build large clusters and poorly isolates clusters that are poorly separated



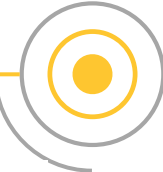
Aggregation criteria used in HAC



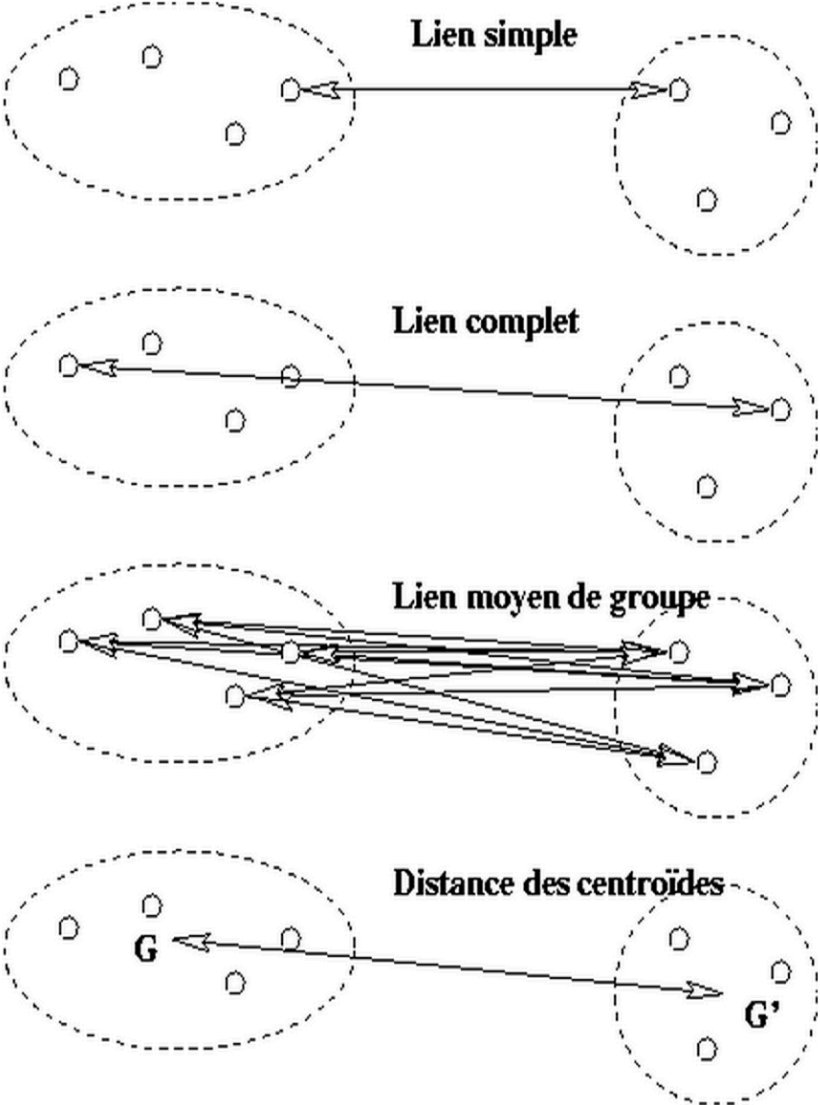
Agglutinates at each stage the two clusters having the smallest distance between their most distant neighbors.



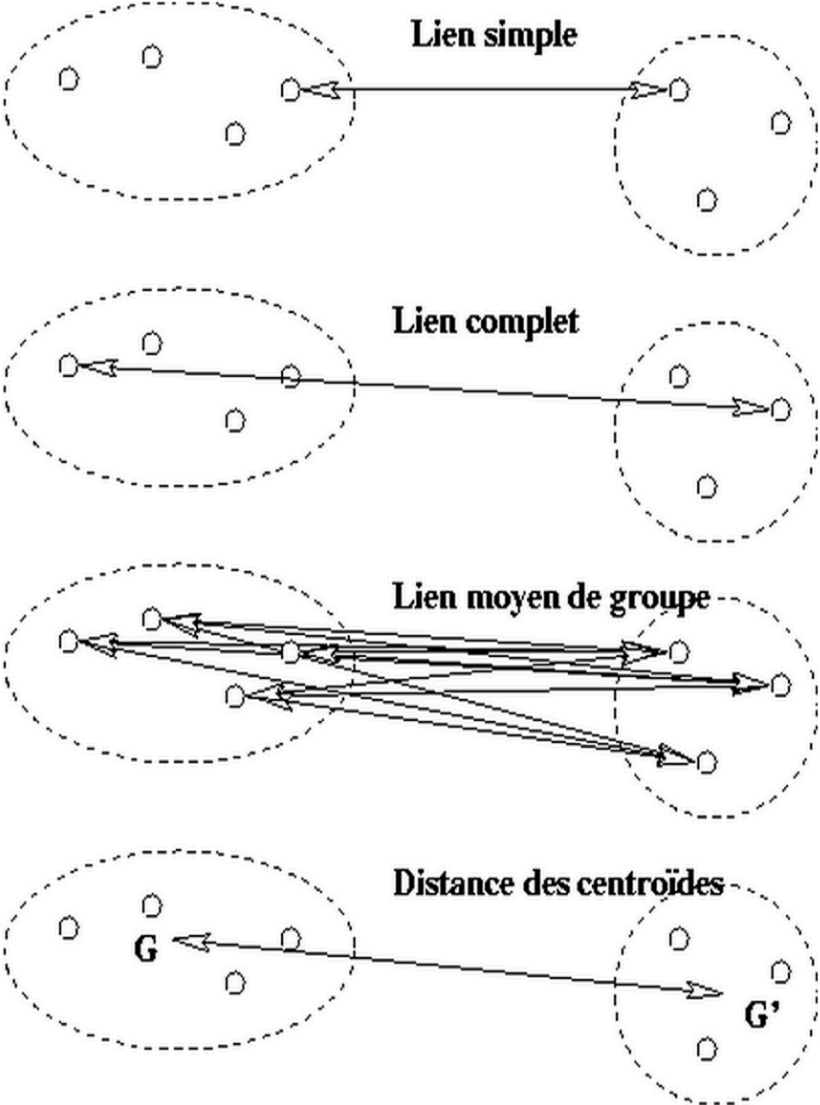
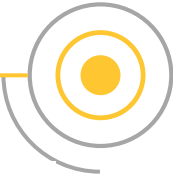
Aggregation criteria used in HAC



Tends to form small compact clusters.

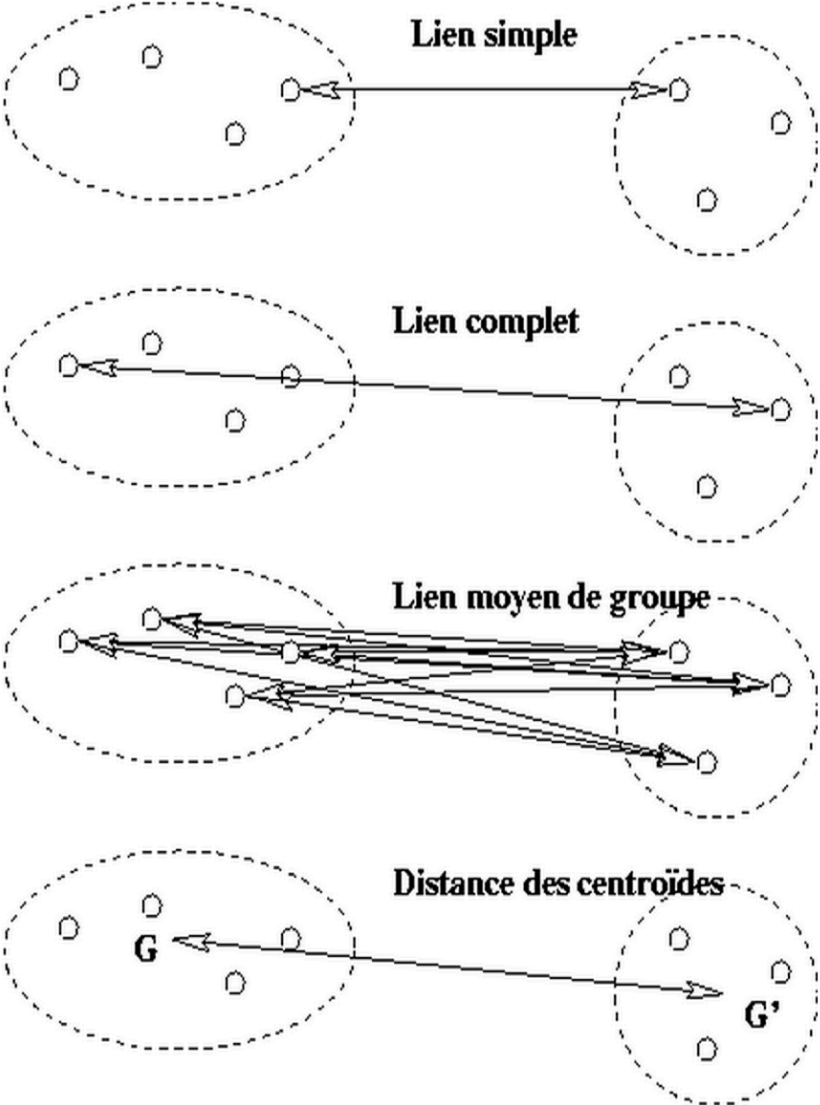
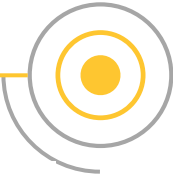


Aggregation criteria used in HAC



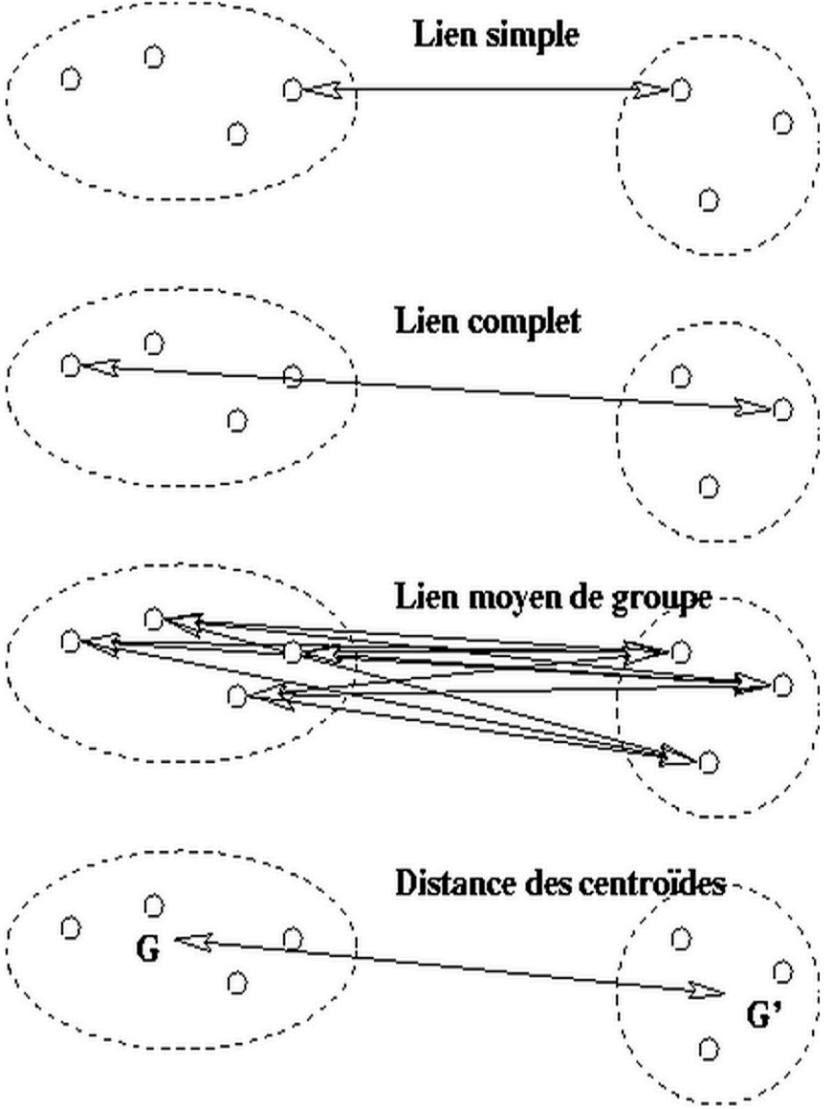
Agglutinates at each stage the two clusters whose means of distances between neighbors are the weakest.

Aggregation criteria used in HAC



Produces clusters whose size is intermediate between clusters produced by the two previous methods.

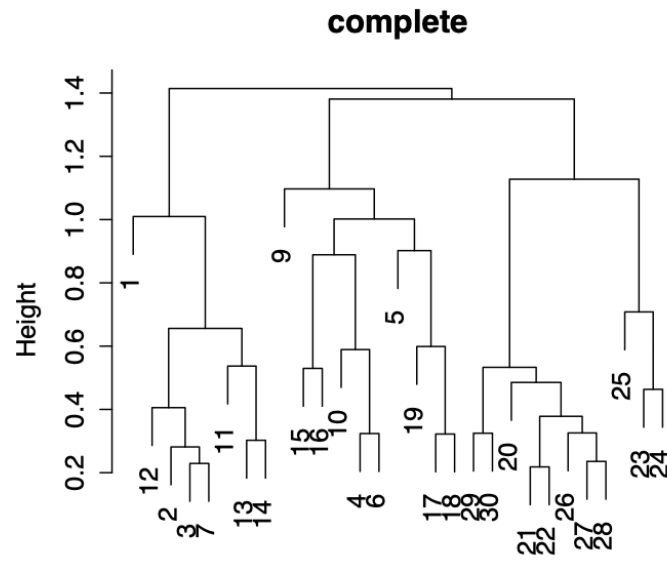
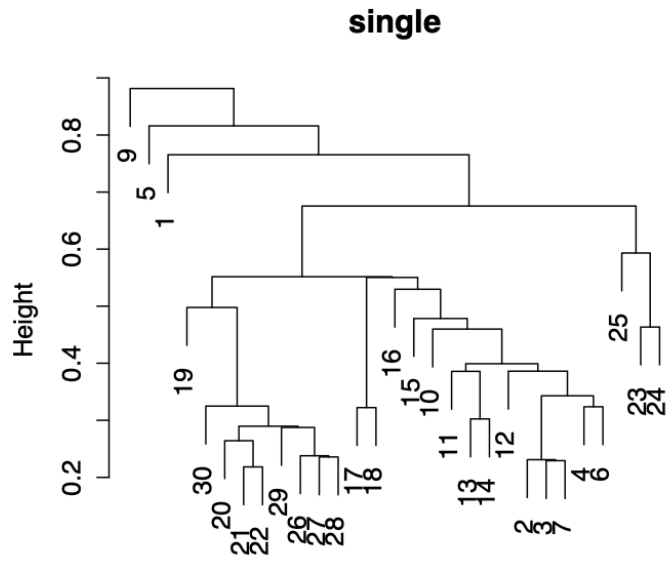
Aggregation criteria used in HAC



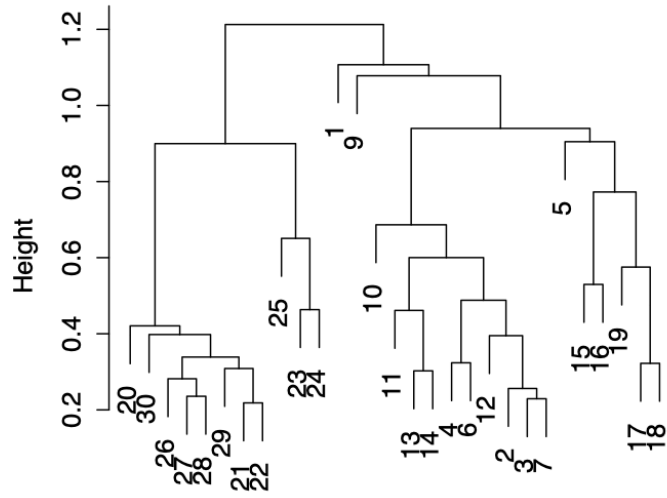
Minimum variance method that agglutinates at each step the two clusters whose junction minimizes the sum of squares of internal errors (Euclidean distances to centroids).



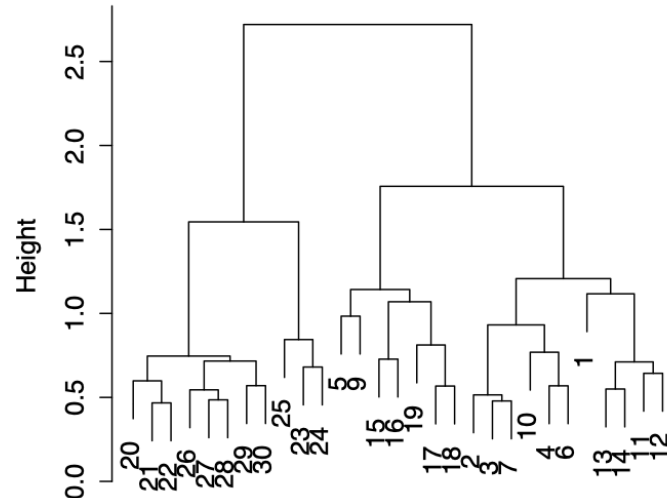
Practice



Choice of the aggregation criteria?

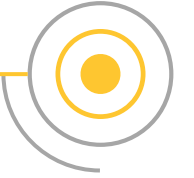


Chorddist
hclust(*, "average")



Chorddist
hclust(*, "ward")

Interprete and compare HAC results



Classification is an heuristic method not a statistical test

Classification methods modify the original distances

Cophenetic distance matrix

	Obj1	Obj2
Obj1		
Obj2		

VS
Corrélation
Pearson

Original distance matrix

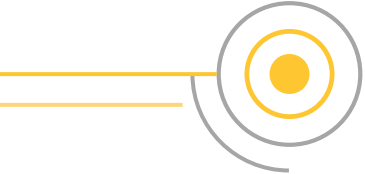
	Obj1	Obj2
Obj1		
Obj2		



Practice

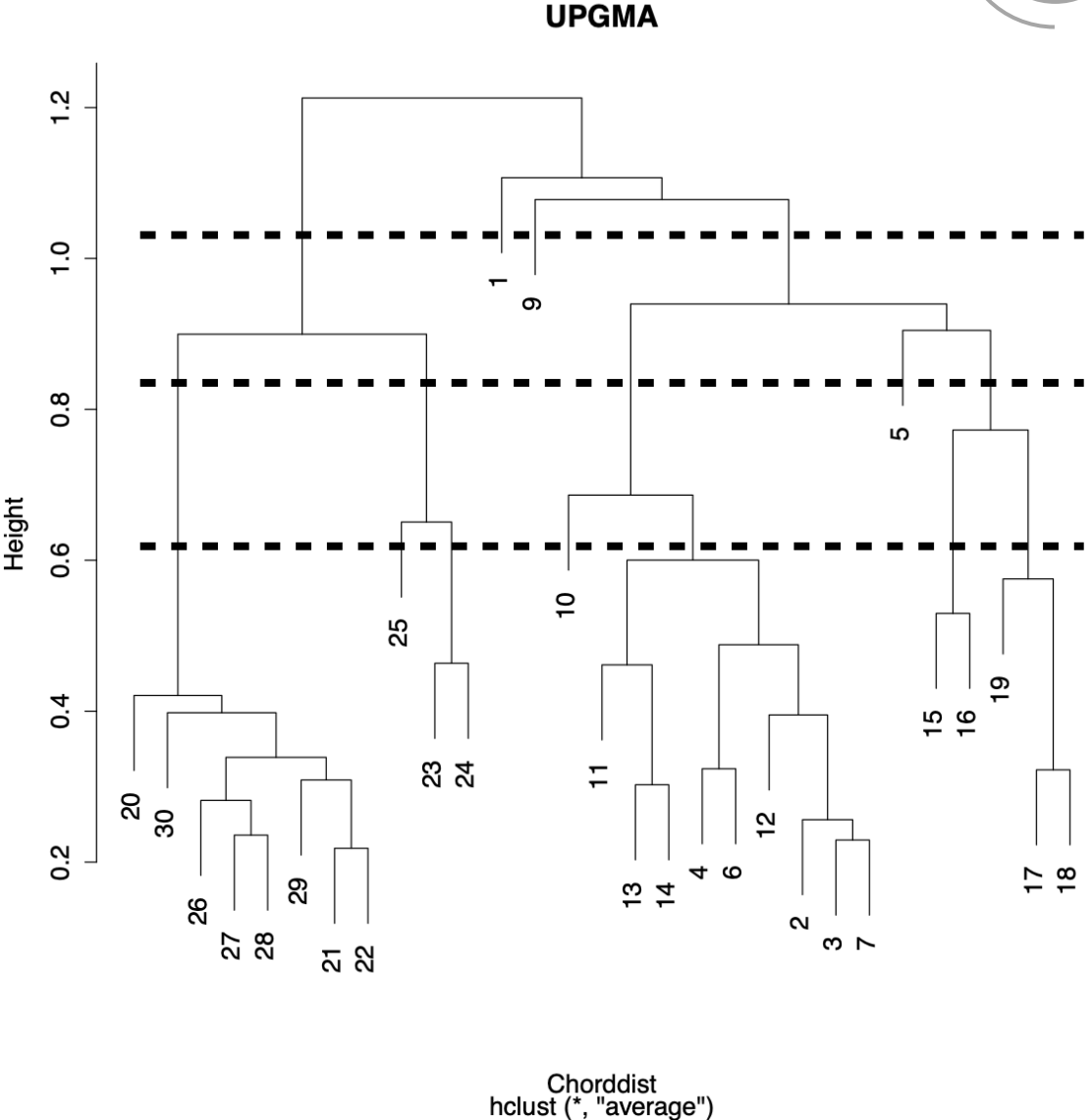
Classification

Looking for Interpretable Clusters



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

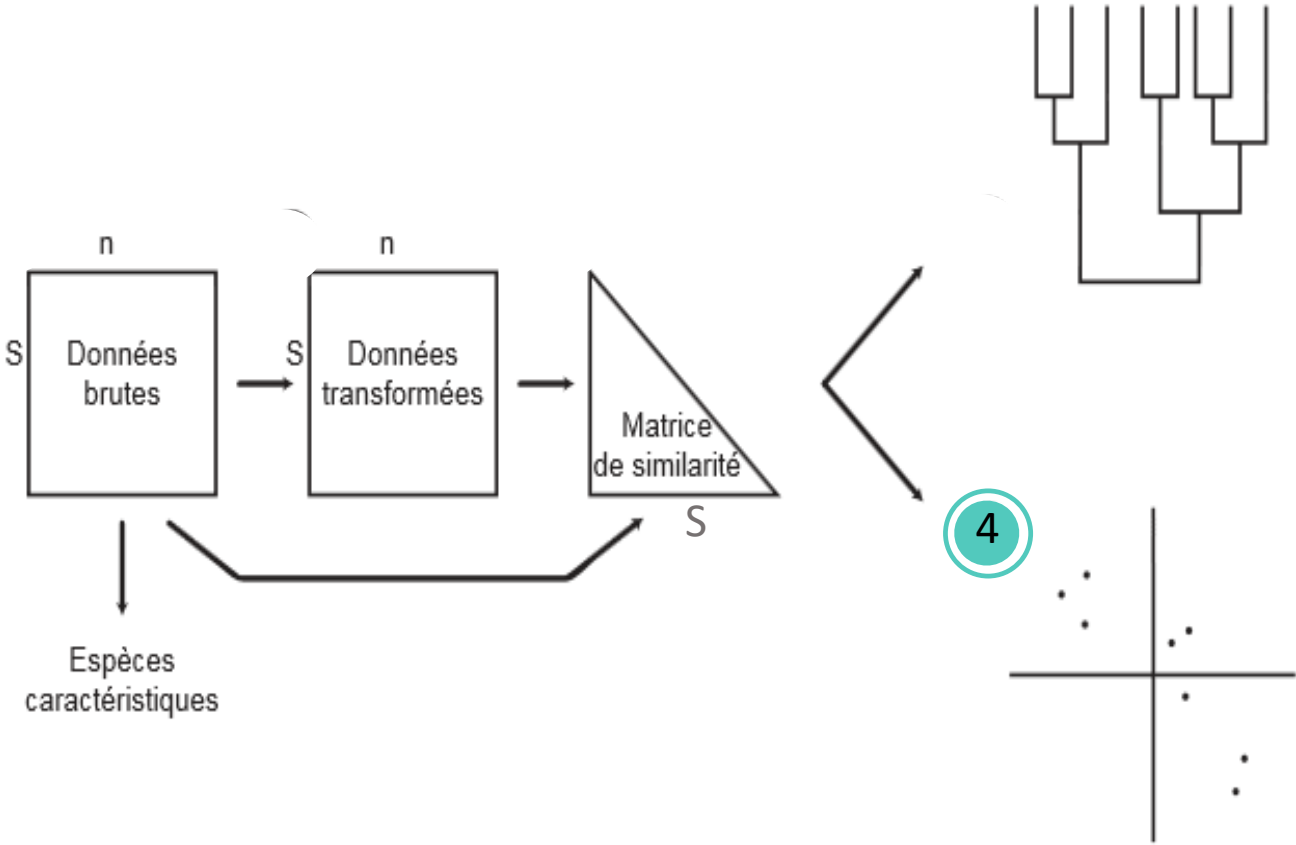
Many indices (more than 30) has been published in the literature for finding the right number of clusters in a dataset.



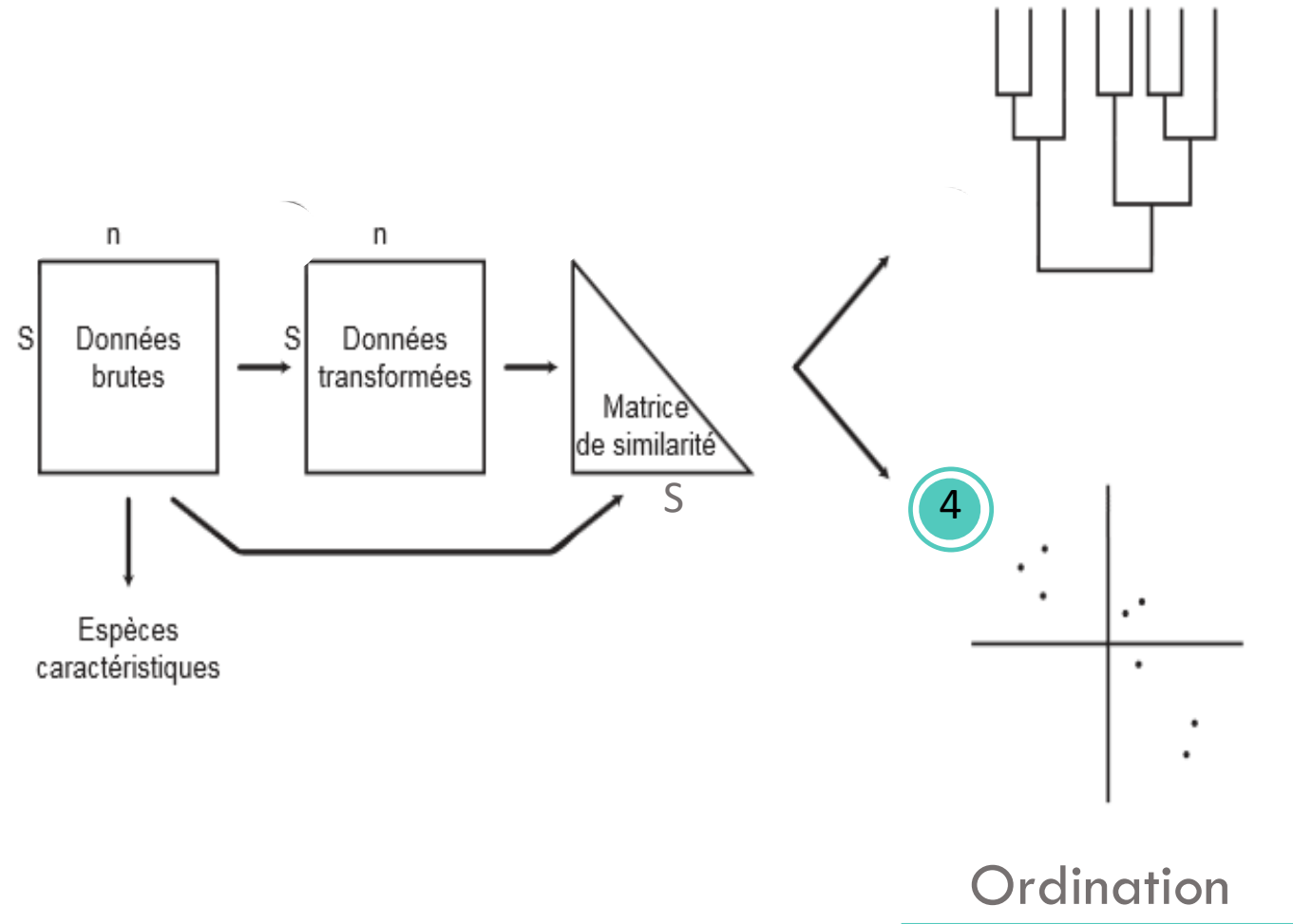


Practice

Overview of the Beta-analysis approach



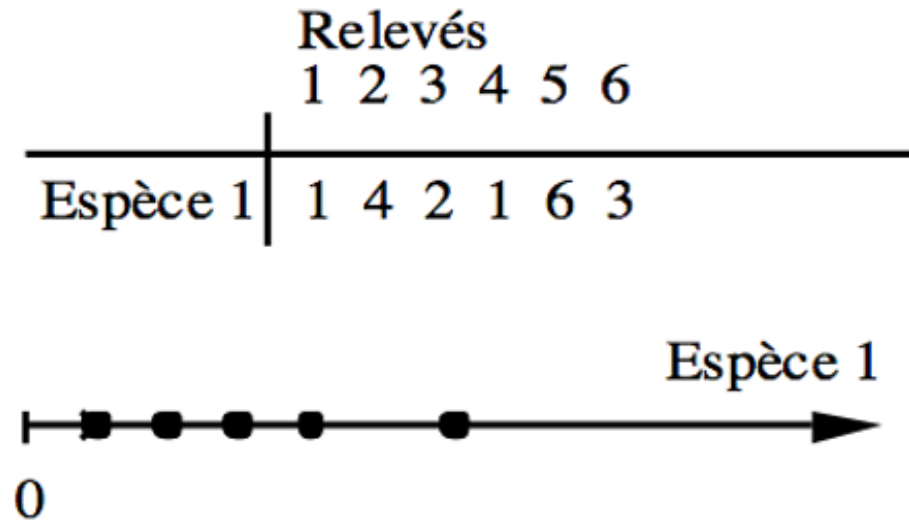
Overview of the Beta-analysis approach



Ordination

Objectifs: représenter les relations entre les objets et les variables dans un espace de faible dimension

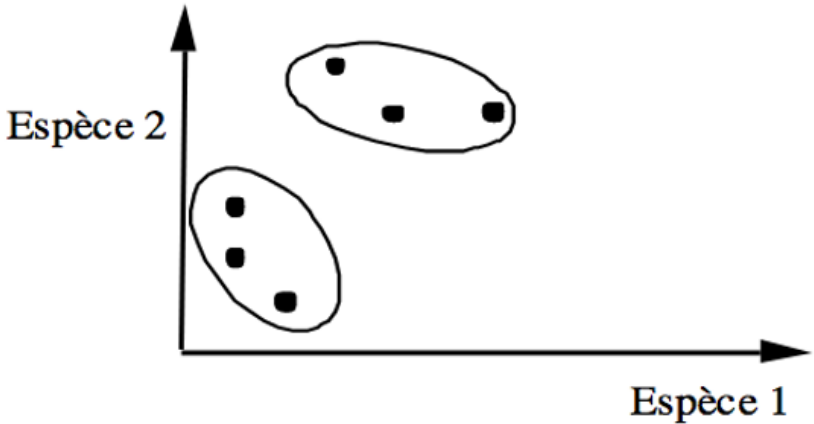
Unidimensional Data



Ordination

Bidimensional Data

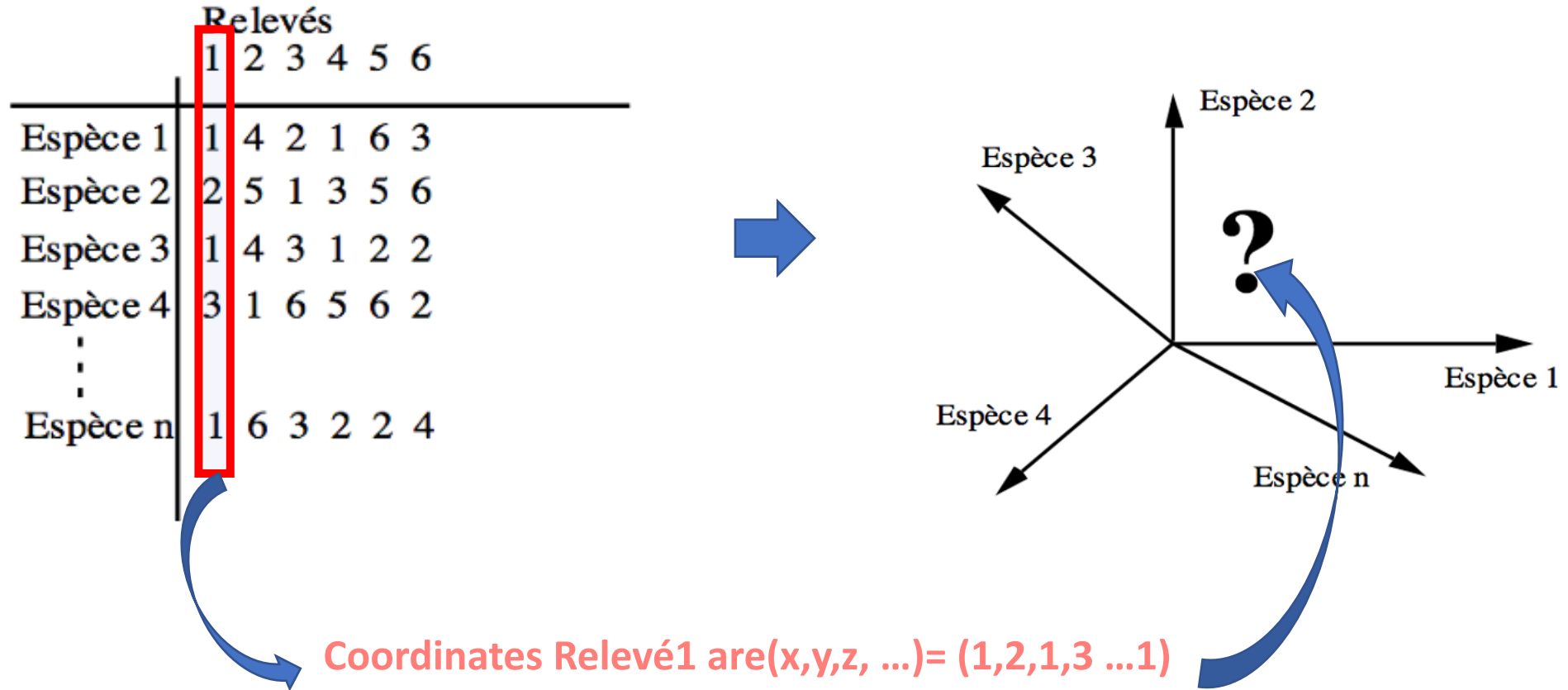
	Relevés					
	1	2	3	4	5	6
Espèce 1	1	4	2	1	6	3
Espèce 2	2	5	1	3	5	6



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

Ordination

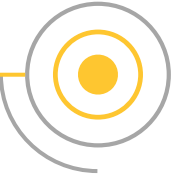
Multidimensional Data (e.g. Metabarcoding)



Impossible to graphically display all the axes!

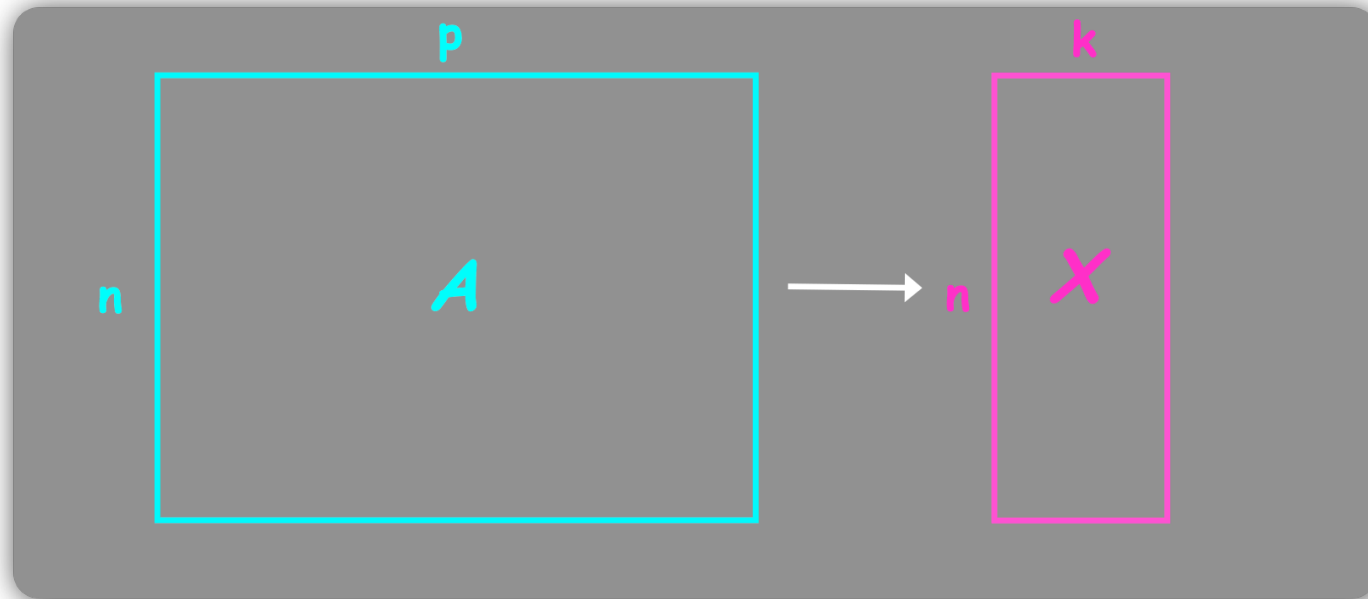
Ordination

How to visualize data in more than 3 dimensions ??



The **ordination methods** respond to this problem by summarizing the data with many (p) variables by a smaller set of (k) derived (synthetic, composite) variables.

Balancing act between clarity of representation, **ease of understanding**, oversimplification
loss of important or relevant information



bénéfices

- représenter les gradients environnementaux les plus importants et interprétables
- réduire le bruit en mettant l'accent sur un espace de faible dimension
- efficacité statistique : une analyse globale vs de multiples analyses univariées

limitations

- analyse exploratoire, pas de test statistique facile à utiliser
- chaque méthode a ses propres limitations
- bonne compréhension de la logique mathématique sous-jacente à chaque méthode
 - pour choisir la méthode appropriée
 - pour faire des interprétations pertinentes

Unconstrained Ordination

In unconstrained methods, the ordination procedure itself is not influenced by external variables. The data matrix express the relationships among objects and variables without constraint. This can be tested after the computation of the ordination. This is an exploratory, descriptive approach.

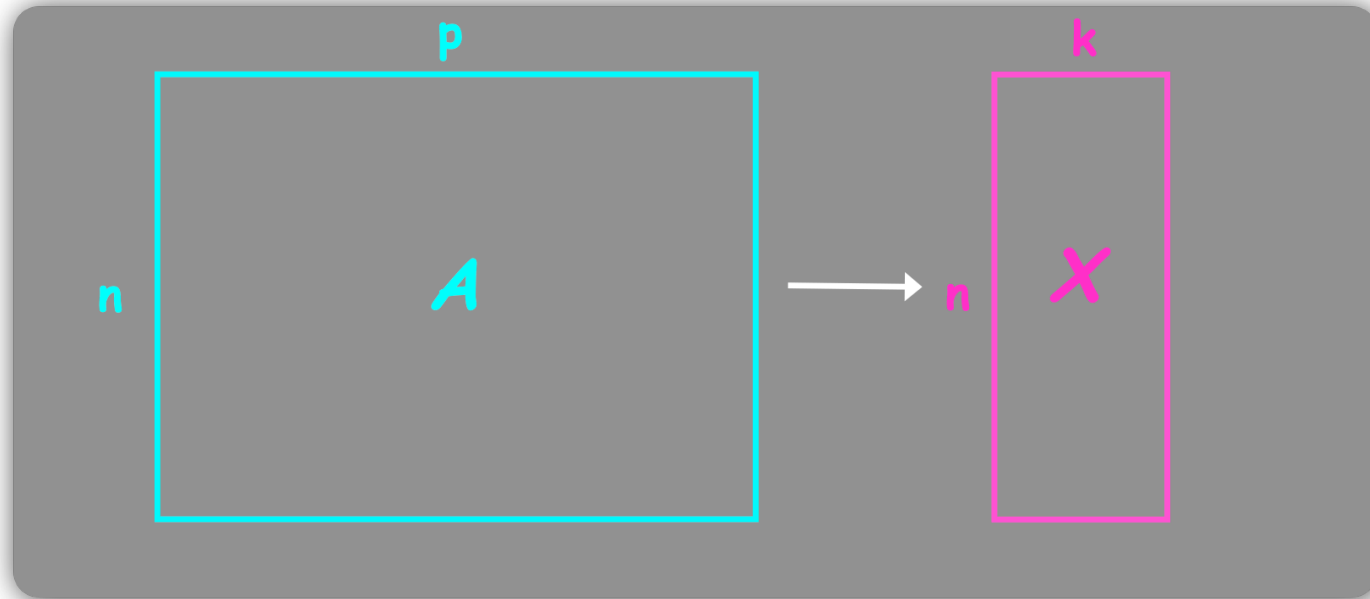
Méthodes	basées sur	gradient	type de données
PO	dist	-	-
PCoA	dist	linéaire	-
NMDS	dist	-	-
PCA	valeurs propres	linéaire	quantitative
CA	valeurs propres	unimodal	tableau de contingence ou au moins positives
DCA	valeurs propres	unimodal	tableau de contingence ou au moins positives

Unconstrained Ordination

Principal Component Analysis (PCA) (ACP en français)

Takes a data matrix of n objects by p variables, which may be correlated, and summarizes it by **uncorrelated axes (principal components or principal axes)** that are **linear combinations** of the original p variables

The first k components display **as much as possible of the variation** among objects.

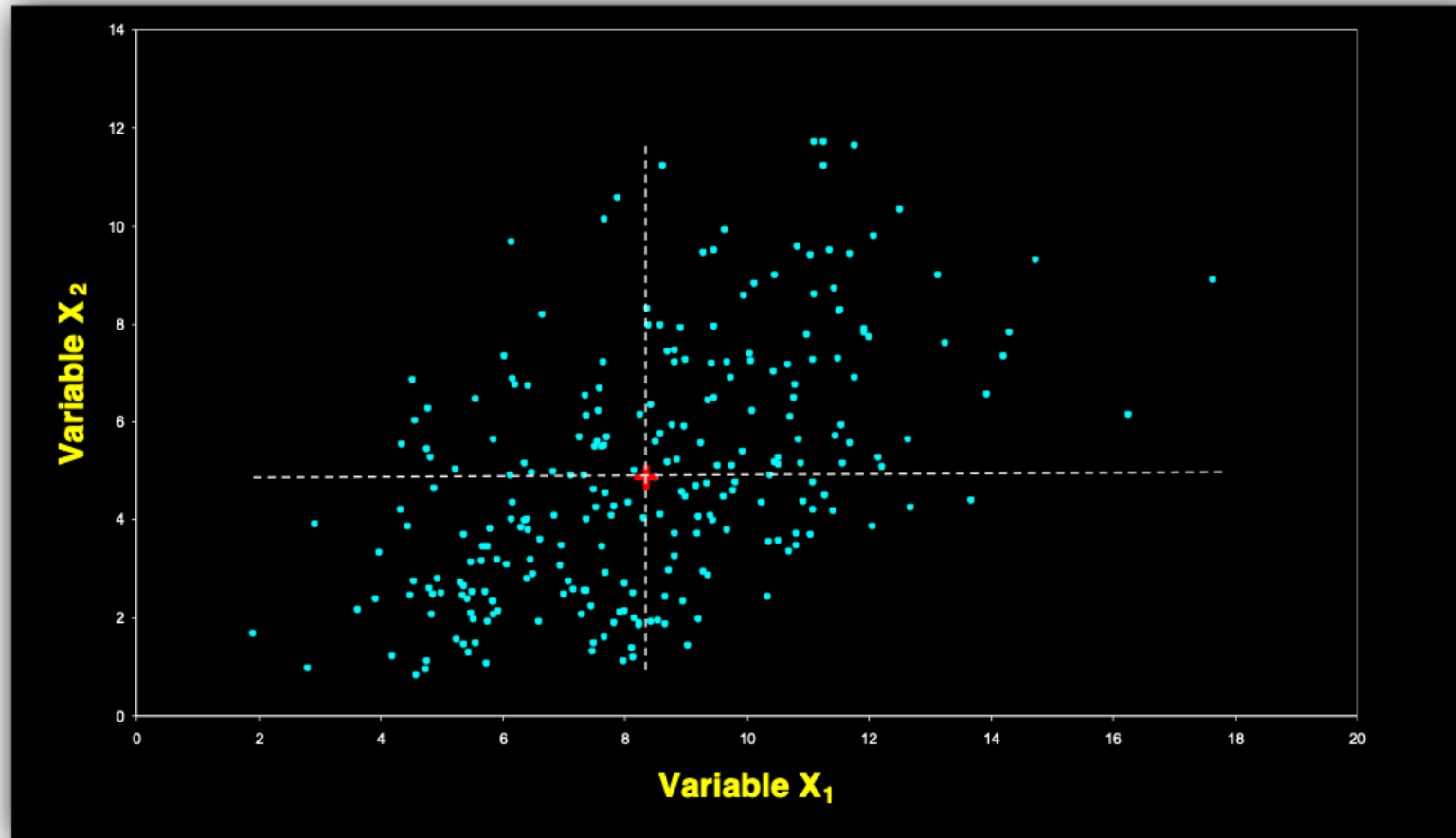


Unconstrained Ordination

PCA principle

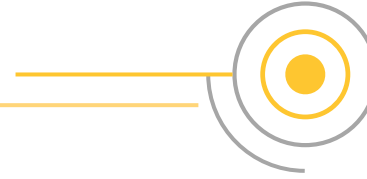


Objects are represented as a cloud of n points in a multidimensional space with an axis for each of the p variables

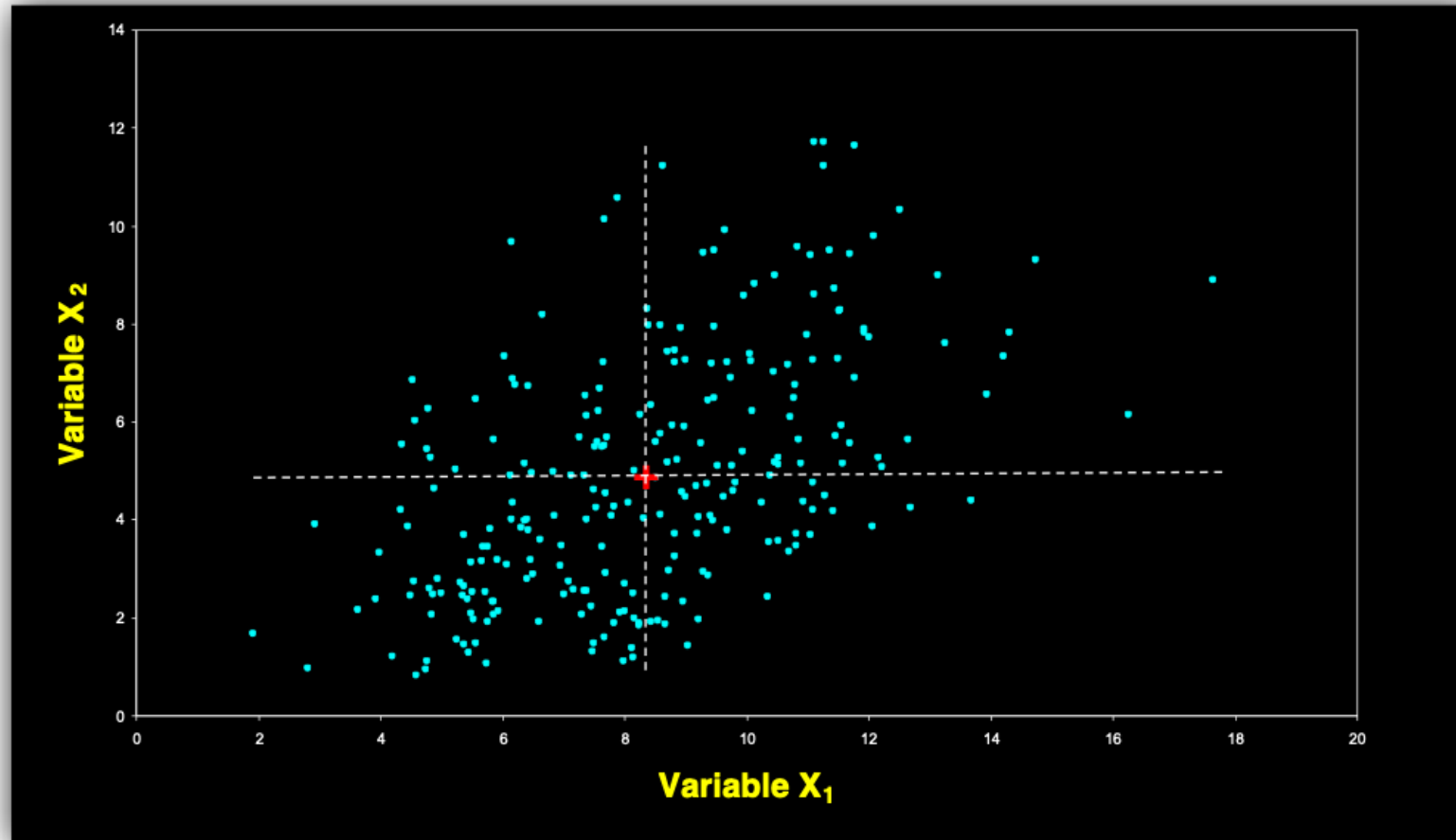


Unconstrained Ordination

PCA principle

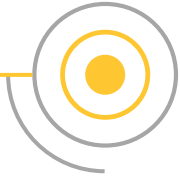


Variables X_1 and X_2 are centered first by subtracting the mean from each value



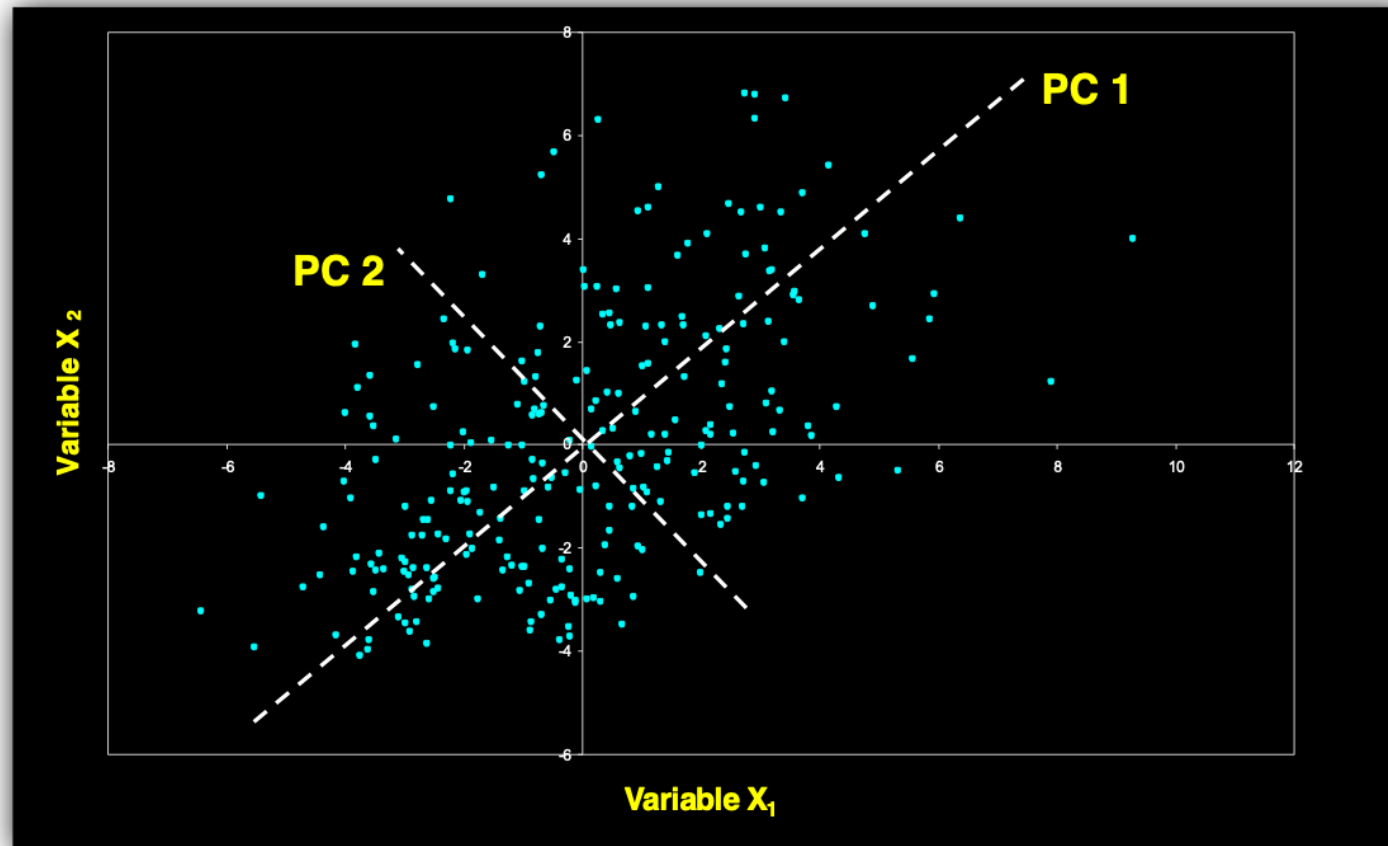
Unconstrained Ordination

Principal Components are Computed



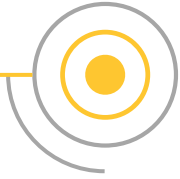
PC axes are a **rigid rotation** of the original variables

PC 1 is simultaneously the direction of **maximum variance** and a least-squares “line of best fit” (squared distances of points away from PC 1 are minimized)

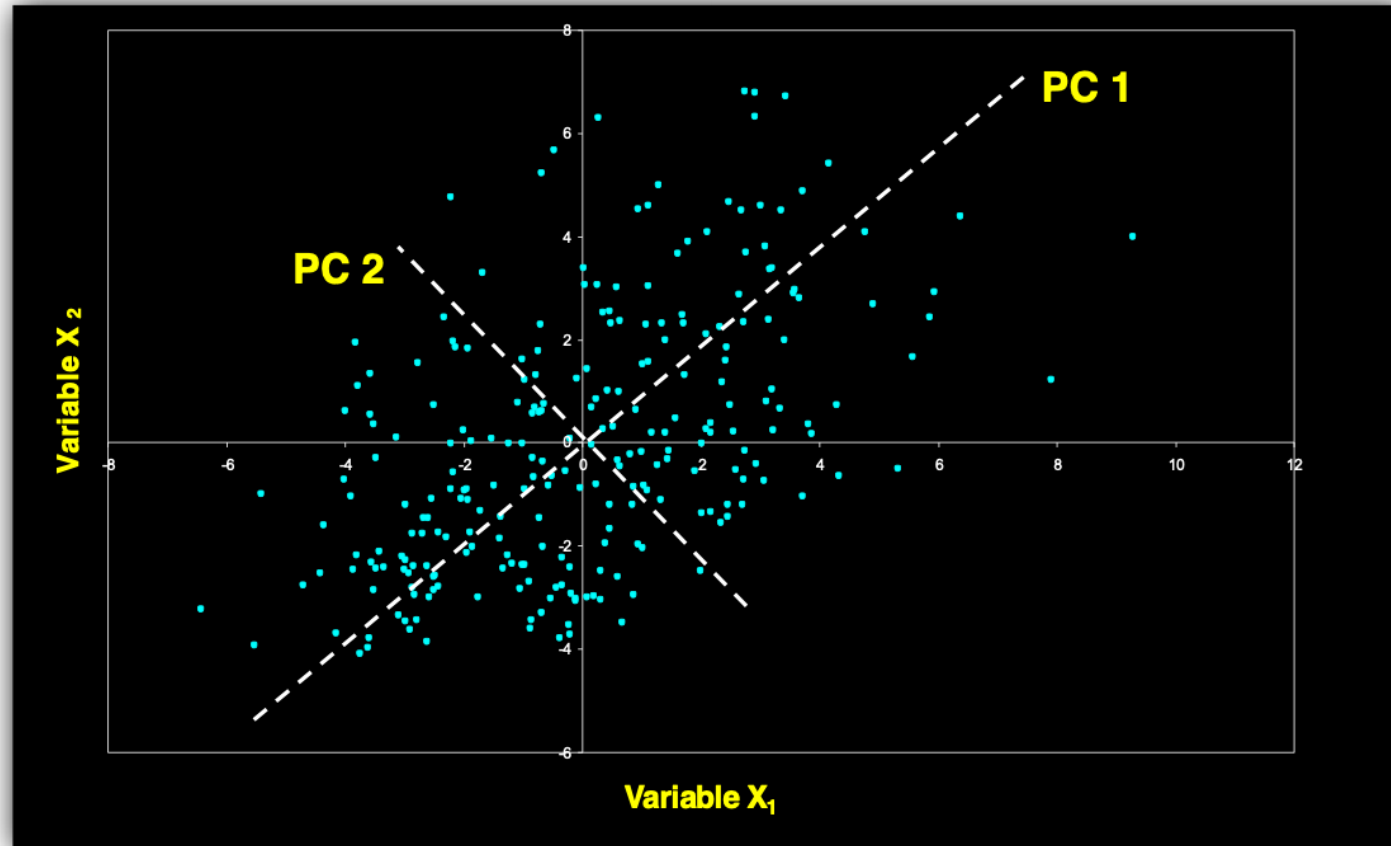


Unconstrained Ordination

Principal Components are Computed

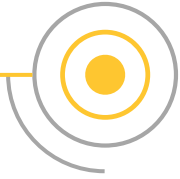


Covariance among each pair of the PCs is zero (the **PCs are uncorrelated**)
Each PC is a **linear combination of the original variables**



Unconstrained Ordination

Generalization to p-dimensions



PCA uses **Euclidean Distance** calculated from the p variables as the measure of dissimilarity among the n objects

PC 1 is the direction of maximum variance in the p -dimensional cloud of points

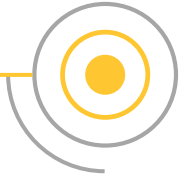
PC 2 is in the direction of the next highest variance, subject to the constraint that it has zero covariance with PC 1

PC 3 is in the direction of the next highest variance, subject to the constraint that it has zero covariance with both PC 1 and PC 2

and so on... up to PC p

Unconstrained Ordination

Eigenvalues, eigenvector, scores



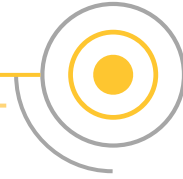
The **eigenvalue** represents the **variance** displayed (“explained” or “extracted”) by the k^{th} axis

Each **eigenvector** consists of p values which represent the “contribution” of each variable to the principal component axis

Coordinates of each object i on the k^{th} principal axis are known as the **scores** on PC

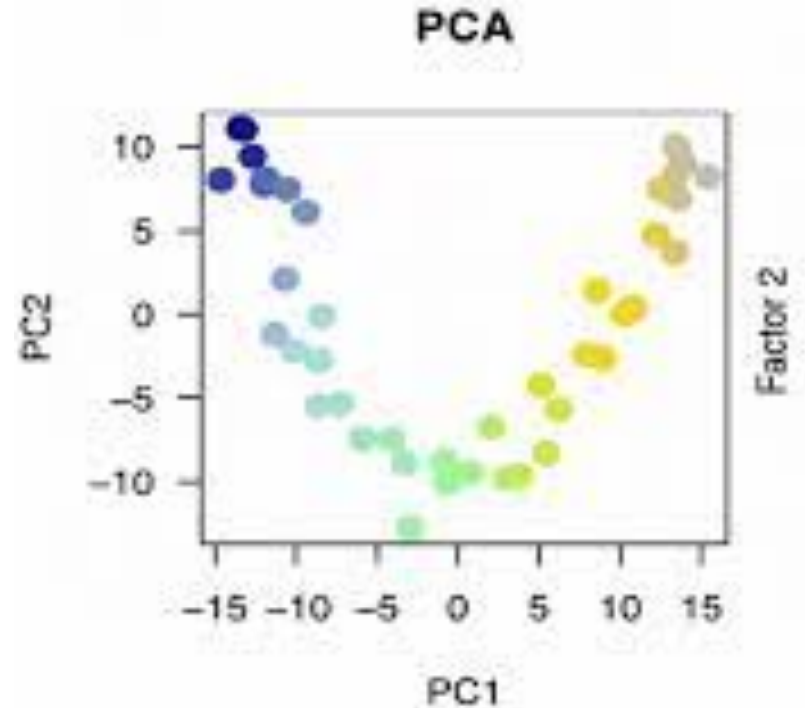
Unconstrained Ordination

What are the assumptions of PCA?



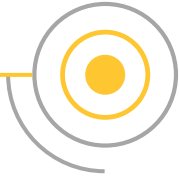
Assumes relationships among variables are **LINEAR**

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



Unconstrained Ordination

When should PCA be used?



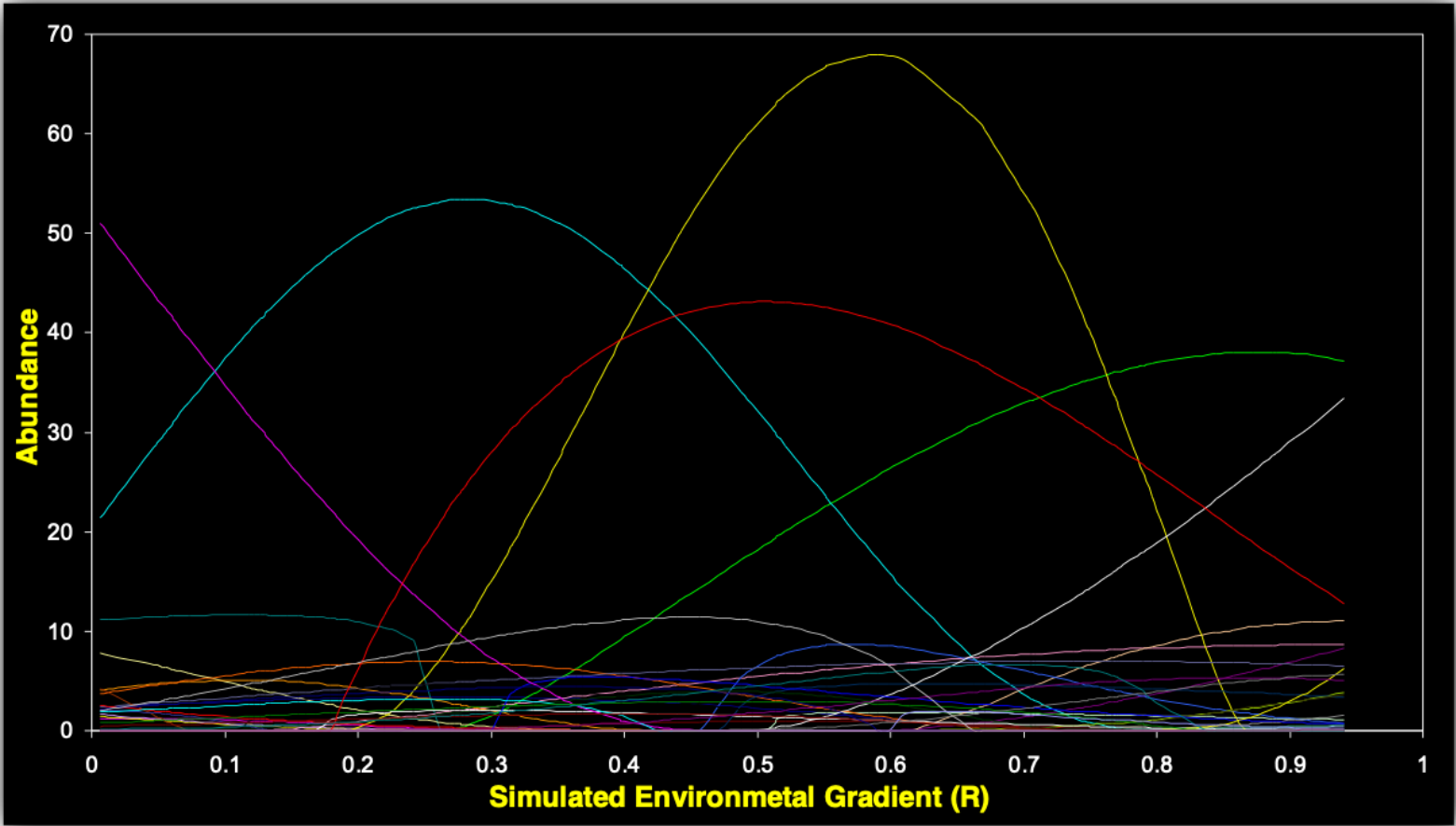
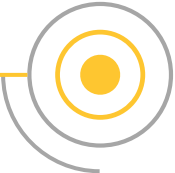
In community ecology, PCA is useful for summarizing variables whose relationships are approximately linear or at **least monotonic**
e.g. A PCA of many soil properties might be used to extract a few components that summarize main dimensions of soil variation

PCA is not always useful for ordinating community data

Why? Because relationships among species are **highly nonlinear**.

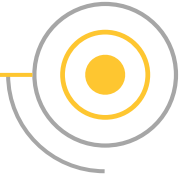
Unconstrain Ordination

When should PCA be used?



Unconstrained Ordination

The “Horseshoe” or Arch Effect



Community trends along environmental gradients appear as “**horseshoes**” in PCA ordinations

None of the PC axes effectively summarizes the trend in species composition along the gradient

Recommendation is to use an ordination that does not assume a linear relationship with environmental gradient such as NMDS



Practice

Unconstrained Ordination based on distances

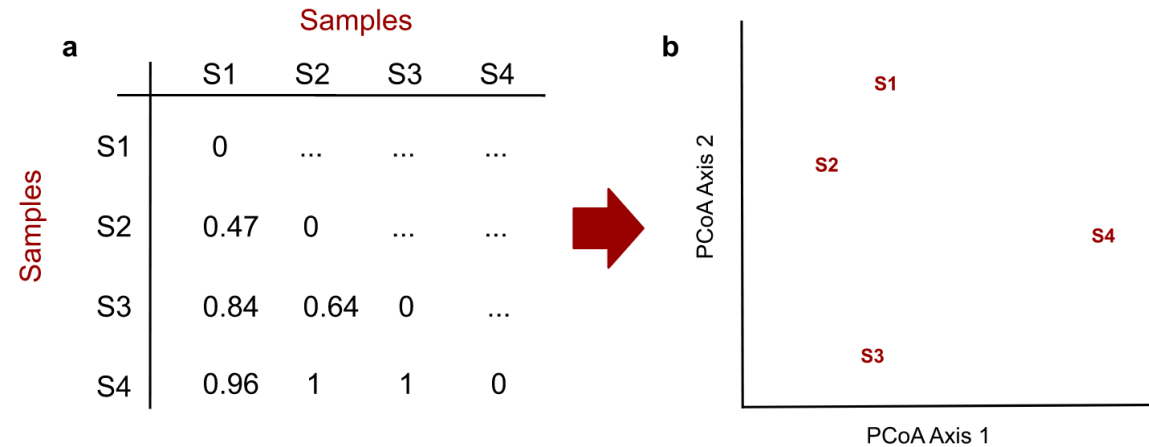
- ces méthodes font référence à une matrice de distance carrée, symétrique, aussi appelée matrice de similarité.
- à l'inverse des méthodes basées sur les valeurs propres, ces méthodes ne donnent pas les scores des espèces et des sites simultanément.
- certaines méthodes 'valeurs propres' sont des cas spéciaux de méthodes 'distance', où la distance est basée sur une distribution du χ^2 .
- **mais:** la philosophie des méthodes 'valeurs propres' est différente: elles ont pour objectif de positionner fidèlement les espèces sur un gradient (soit inféré soit mesuré), et pas de positionner les sites en fonction de leur similarité.

Unconstrained Ordination based on distances

Principal Coordinate Analysis (PCoA or MDS)

It provides a Euclidean representation (distances are preserved) of a set of objects whose relationships are measured by any similarity or distance measure chosen by the user

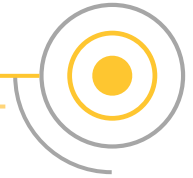
Does not use original data (e.g. PCA)...



Like PCA and CA, PCoA produces a set of orthogonal axes **which maximize the correlation between the dissimilarity matrix and the Euclidian distance among samples in ordination space.** Their importance is measured by eigenvalues.

Unconstrained Ordination based on distances

Principal Coordinate Analysis (PCoA or MDS)



If it is necessary to project variables, e.g. species, on a PCoA ordination of the objects, the variables can be related *a posteriori* to the ordination axes using correlations or weighted averages and drawn on the ordination plot.

The **most common** ordination used in microbial ecology with NMDS



Practice

Unconstrained Ordination based on distances

Non Metric Multidimensional Scaling (NMDS)

NMDS attempts to represent the pairwise dissimilarity between objects in a low-dimensional space. **Any dissimilarity coefficient or distance measure** may be used to build the distance matrix used as input.

NMDS is an iterative algorithm. NMDS routines often begin by random placement of data objects in ordination space. The algorithm then begins to refine this placement by an iterative process, attempting to find an ordination in which ordinated object distances closely match the order of object dissimilarities in the original distance matrix.

The **stress value** reflects how well the ordination summarizes the observed distances among the samples. 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)

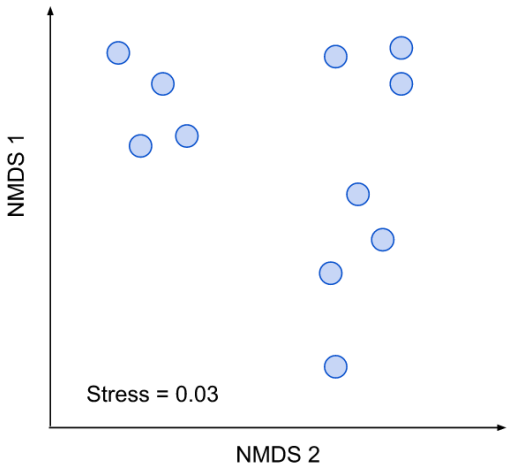
NMDS is a rank-based approach. This means that the original distance data is substituted with ranks. While information about the magnitude of distances is lost, rank-based methods are generally more robust to data which do not have an identifiable distribution.

		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

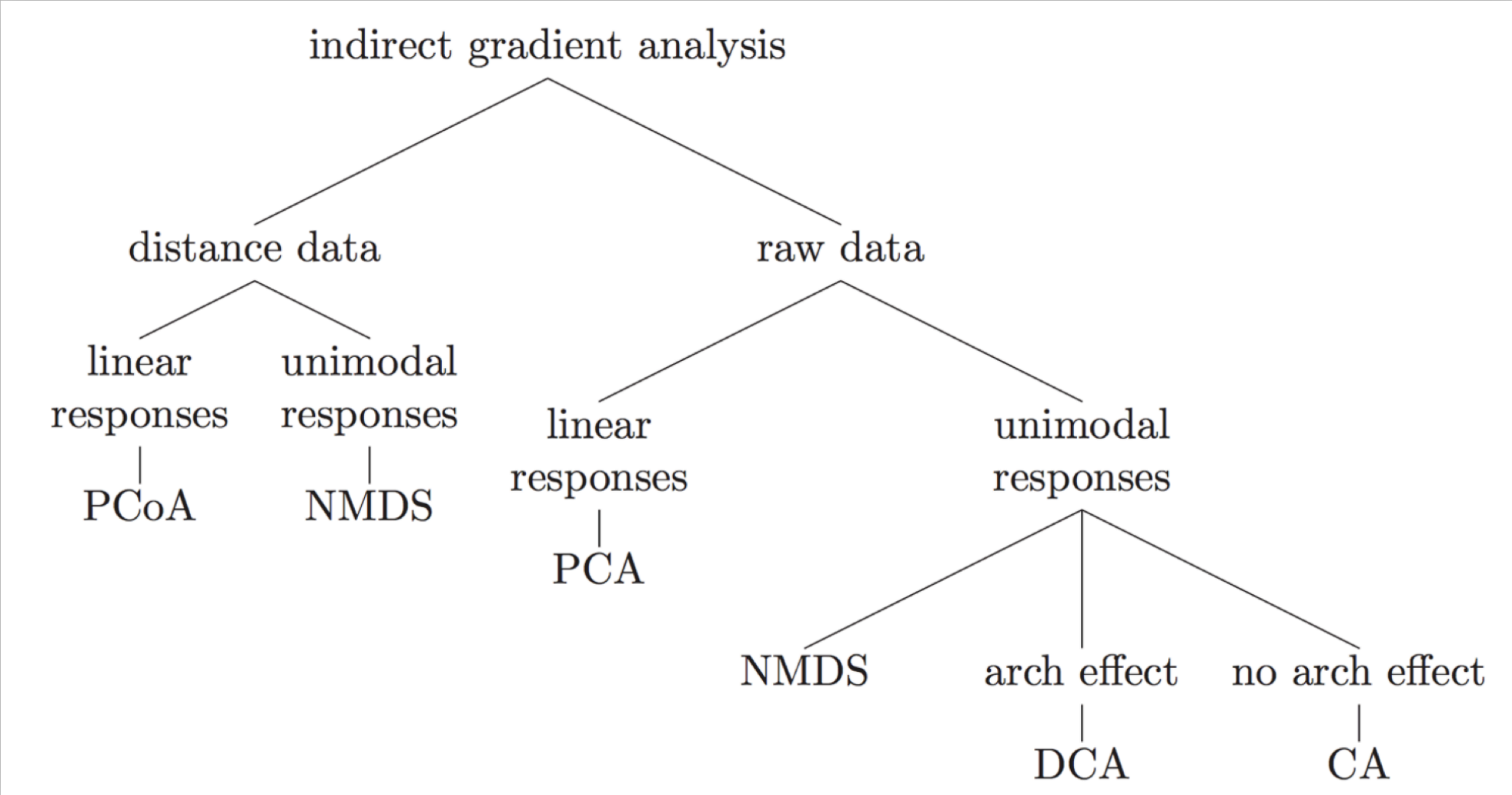
		Samples			
		S1	S2	S3	S4
Samples	S1	0
	S2	1	0
	S3	3	2	0	...
	S4	4	5.5	5.5	0

Rank calcul



NMDS
Axes are arbitrary
No % of inertia/
variance

Unconstrained Ordination



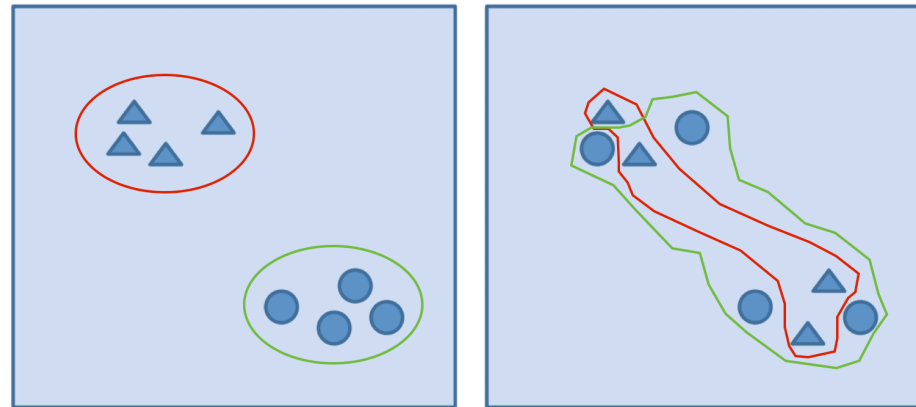


Practice

Hypotheses testing

- ANOSIM (Analysis of Similarity)
- PERMANOVA (Permutational MANOVA)

Multivariate analysis of variance based on distance matrices and permutation. They do this by partitioning the sums of squares for the **within- and between-cluster components**

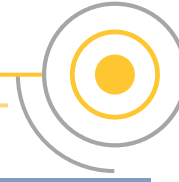


Pseudo-F
(observed)

Pseudo-F
(after permutation)

$$P = \frac{(\text{No. of } F' \geq F) + 1}{(\text{Total no. of } F') + 1}$$

perMANOVA Vs ANOSIM



- PERMANOVA , uses **actual Bray-Curtis** coefficients where ANOSIM uses only **ranks of Bray-Curtis**, therefore preserving more information
- PERMANOVA allows for **partitioning of variability**, similar to ANOVA, therefore allowing for more complex designs (multiple factors, nested factors, interactions, covariates)



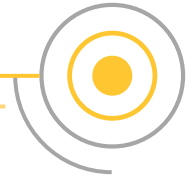
Practice

Constrained Ordination

Constrain ordination **explicitly** explores the relationships between two or more matrices: a response matrix (often your species matrix) and one or more explanatory matrices (often your environmental matrices).

- **Objective** : Attempt to explain differences in species composition between sites by differences in environmental variables
- **Key points**
 - Computes axes that are **combinations of the explanatory variables**(e.g ph, T°C, ...) in order to explain the most variation of the species matrix
 - It is constrained because you are **directly** testing the **influence of explanatory variables**
 - Consequence : probably **only a fraction of the variance** from data is correlated to explanatory variables

Redundant Analysis (RDA)



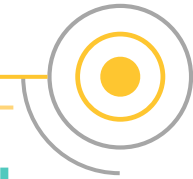
Conceptually, RDA is a **multivariate (meaning *multiresponse*) multiple linear regression followed by a PCA of the table of fitted values.**

It works as follows, on a matrix **Y of centred response data** and a **matrix X of centred (or, more generally, standardized) explanatory variables:**

- **Regress** each (centred) **y** variable on explanatory table **X** and compute the fitted values of **y**. Assemble all vectors into a matrix of fitted values \hat{Y} .
- **Compute a PCA** of the matrix of fitted values \hat{Y} ; this analysis produces a vector of canonical eigenvalues and a matrix **U** of *canonical eigenvectors*.
- **Use matrix U** to compute *the ordination site scores*

Constrained Ordination

Redundant Analysis (RDA)



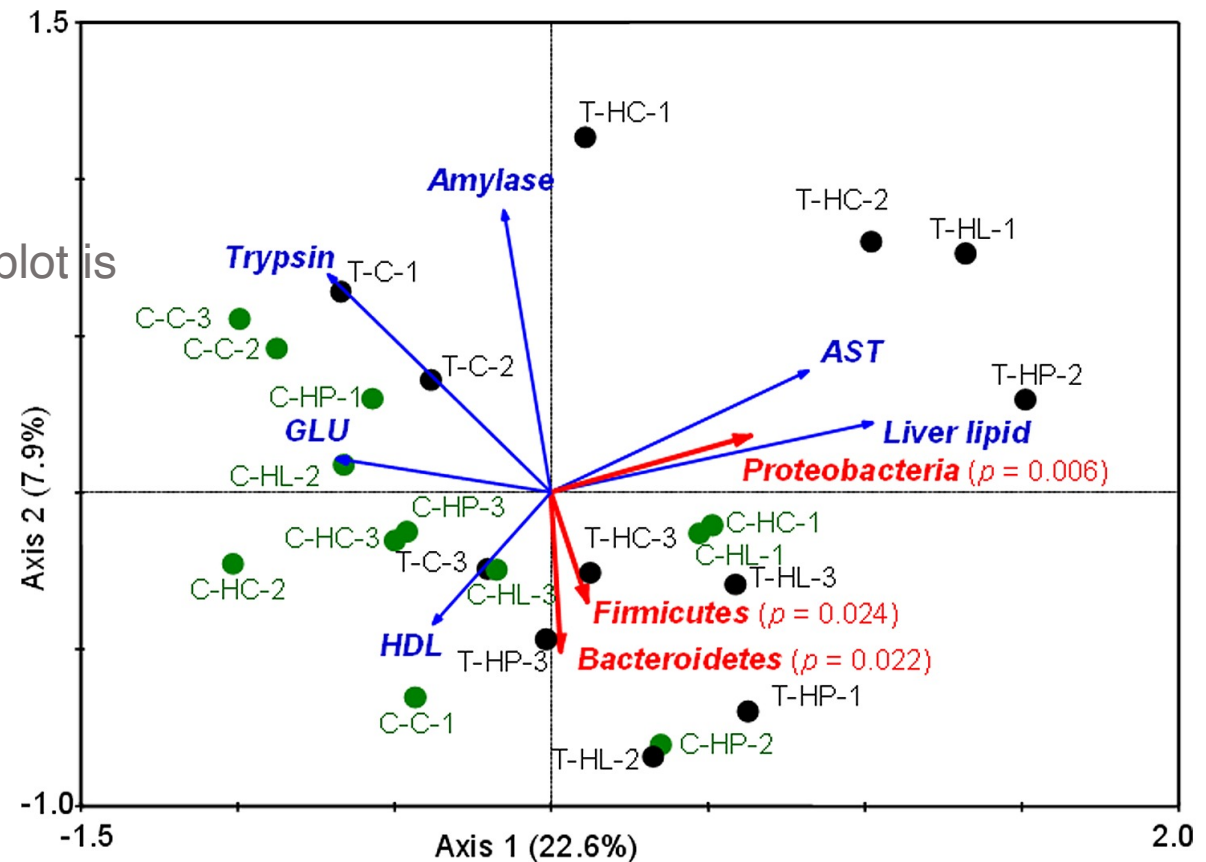
There are three different entities in the plot: **sites**, **response variables** and **explanatory variables**.

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

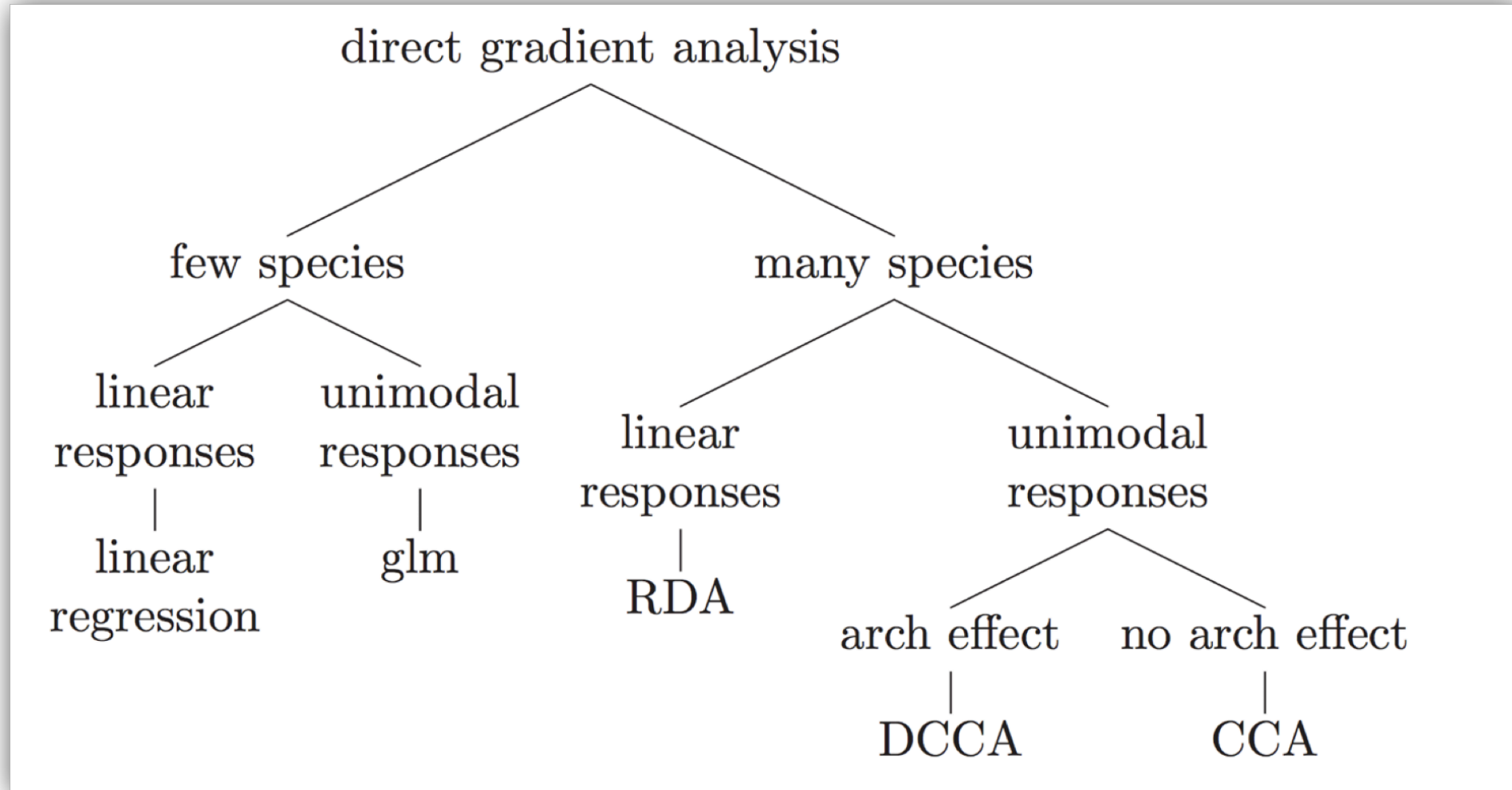
The distance between site and species position on the triplot is indicative of the abundance of the species for the site

The angle between variables and species reflects their correlations

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



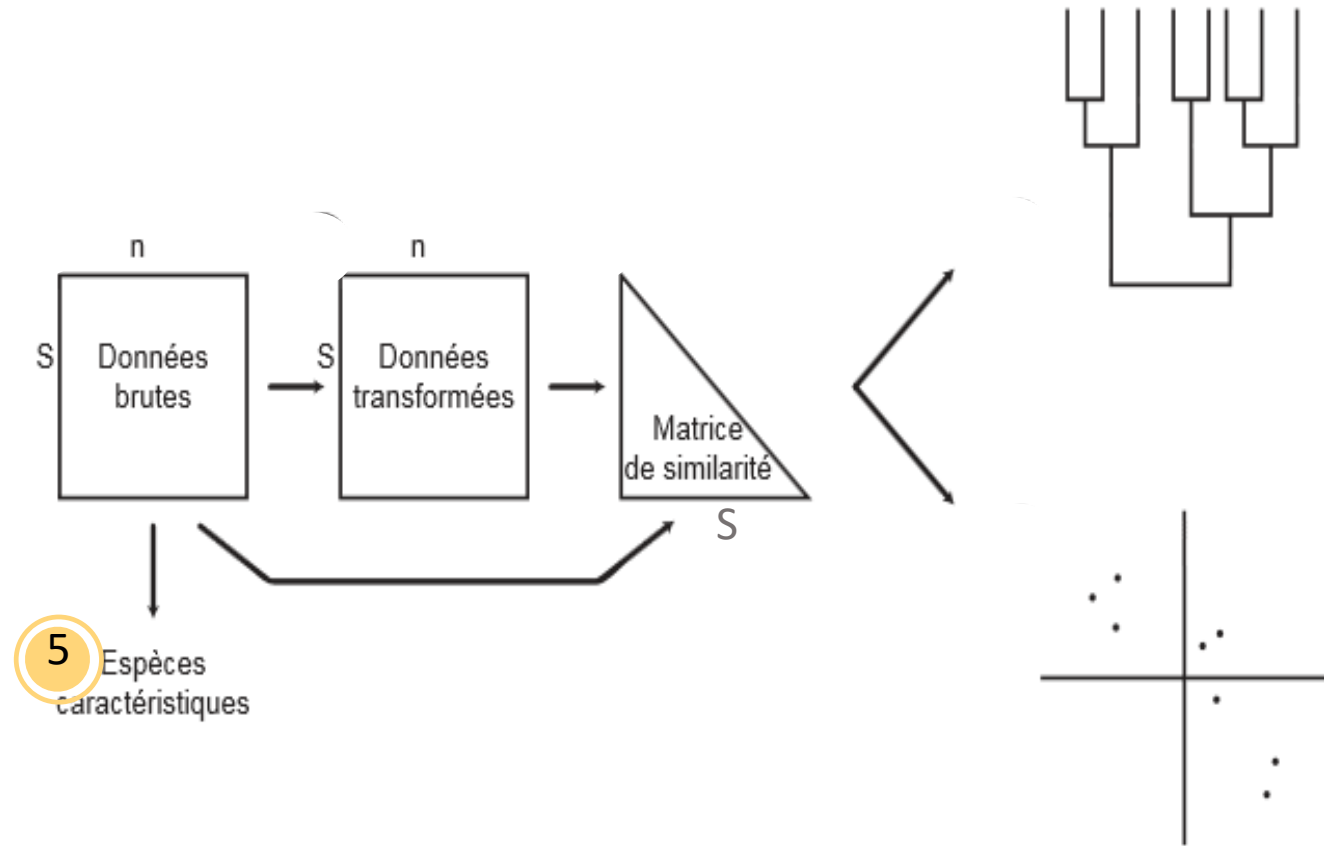
Constrained Ordination



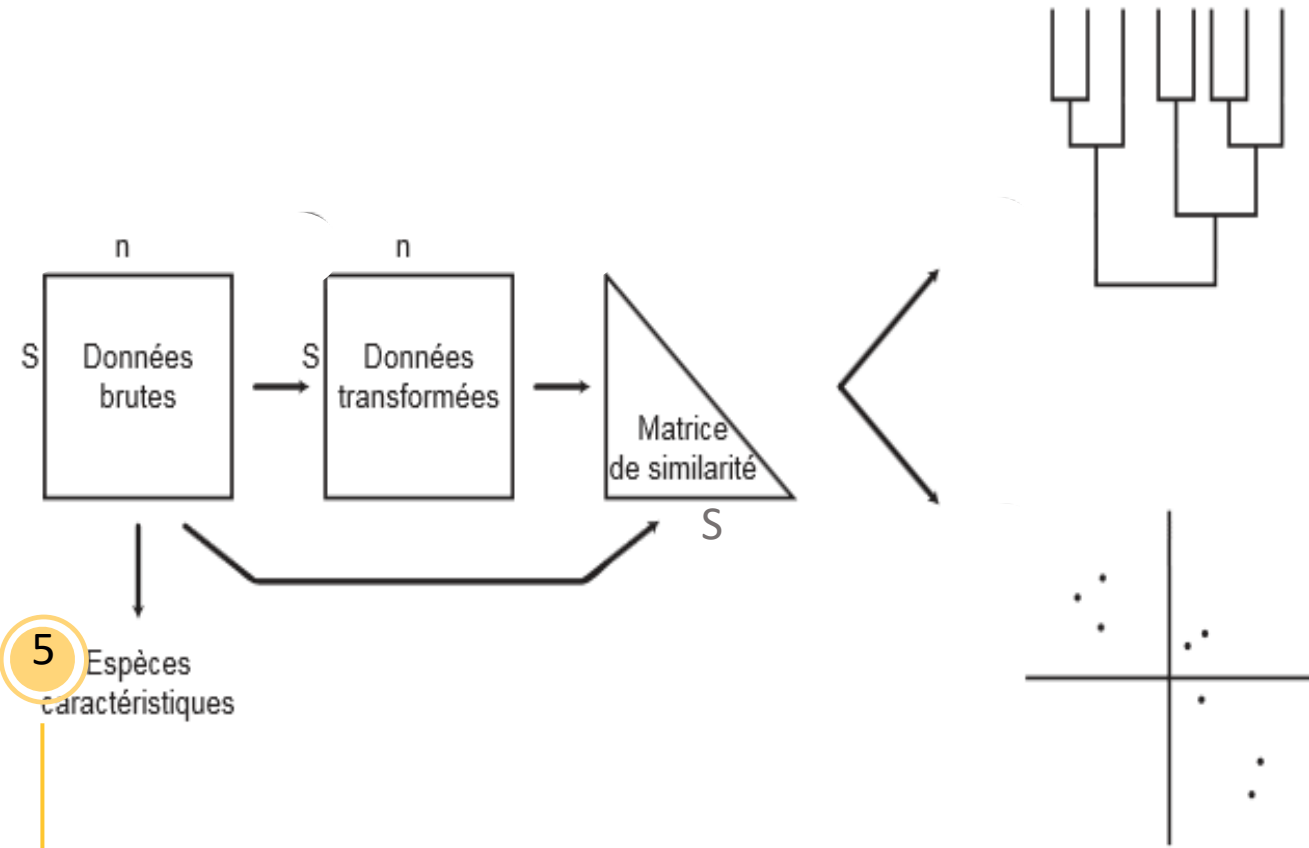


Practice

Overview of the Beta-analysis approach



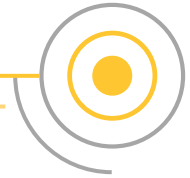
Overview of the Beta-analysis approach



Differential abundance

Differential abundance

Differential abundance analysis (DAA)



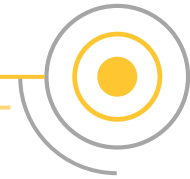
The goal of differential abundance testing is to identify specific taxa associated with metadata variables of interest. **This is a difficult task.**

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.

Differential abundance

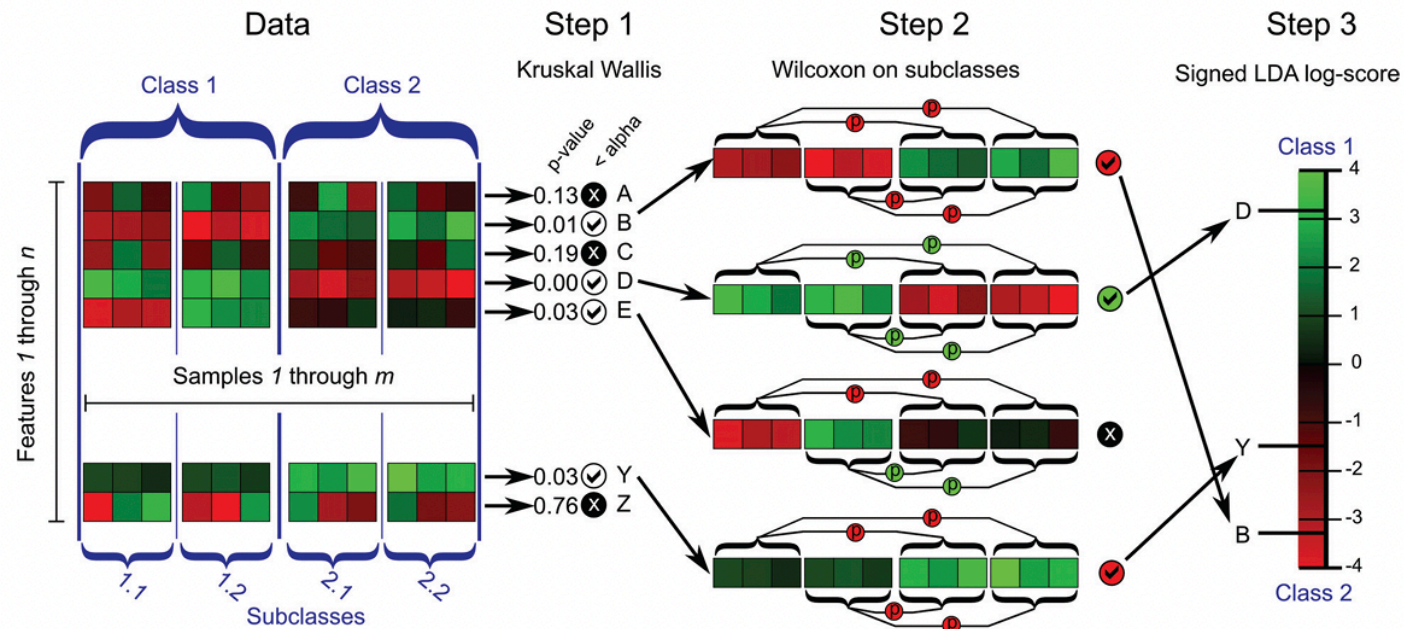
Linear discriminant analysis Effect Size (LEFse)



LEFse first use the **non-parametric factorial Kruskal-Wallis (KW) sum-rank test** to detect features with significant differential abundance

Biological consistency is subsequently investigated using a set of **pairwise tests among subclasses using the (unpaired) Wilcoxon rank-sum test**.

As a last step, LEFse uses **LDA** to estimate the effect size of each differentially abundant features.



Differential abundance

CoDA methods (ALDEx2, ANCOM-BC)



Sequencing data are **compositional**, meaning that sequencing only provides information on the relative abundance of features and that each feature's observed abundance is dependent on the observed abundances of all other features.

Compositional data analysis (CoDa) methods circumvent this issue by reframing the focus of analysis to **ratios of read counts** between different taxa within a sample.

The difference among CoDa methods considered is what abundance value is used as **the denominator**, or the reference, for the transformation.

CoDA Aitchison's Log-ratio based-methods :

- Centered log-ratio (CLR) -> ALDEx2
- Additive log-ratio (ALR) -> ANCOM-BC