



Atelier 9: Analyses multivariées

Série d'ateliers R

Centre de la Science de la Biodiversité du Québec



À propos de cet atelier

 REPO

 DEV

 WIKI

09

 5

DIAPOS

09

 6

DIAPOS

09

 R

SCRIPT

09

Packages requis

- `ape`
- `gclus`
- `vegan`

```
install.packages(c('ape', 'gclus', 'vegan'))
```

Objectifs d'apprentissage

Utiliser R pour faire des ordinations sans contraintes

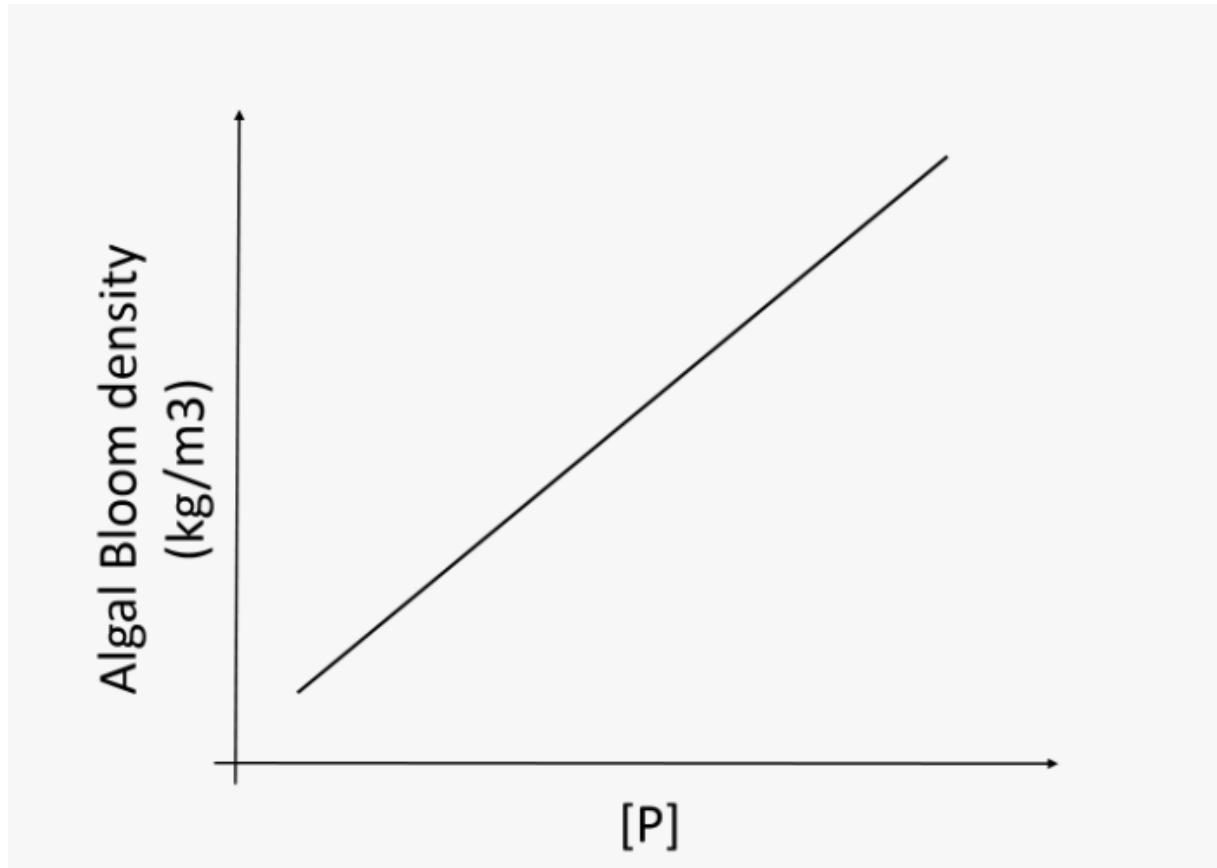
Faire un dendrogramme avec R

1. Introduction

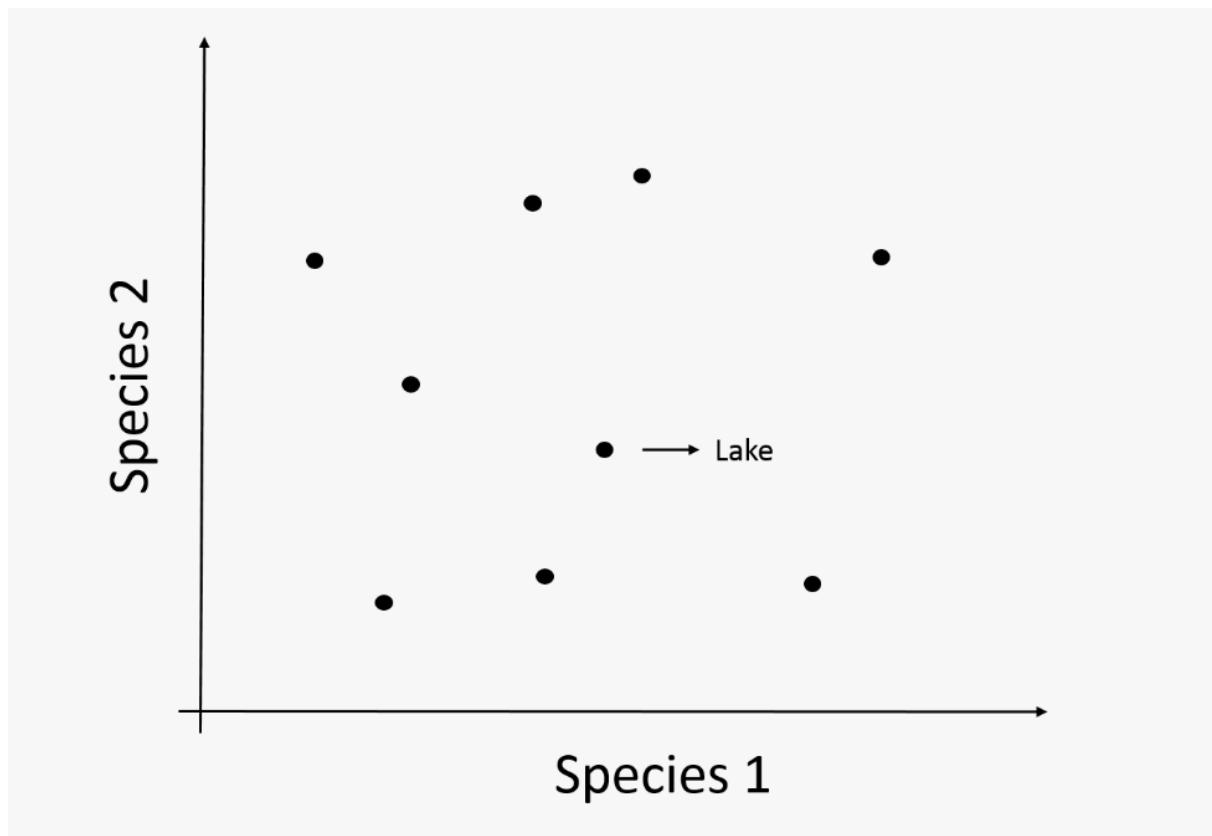
Qu'est-ce que l'ordination?

Une Dimension

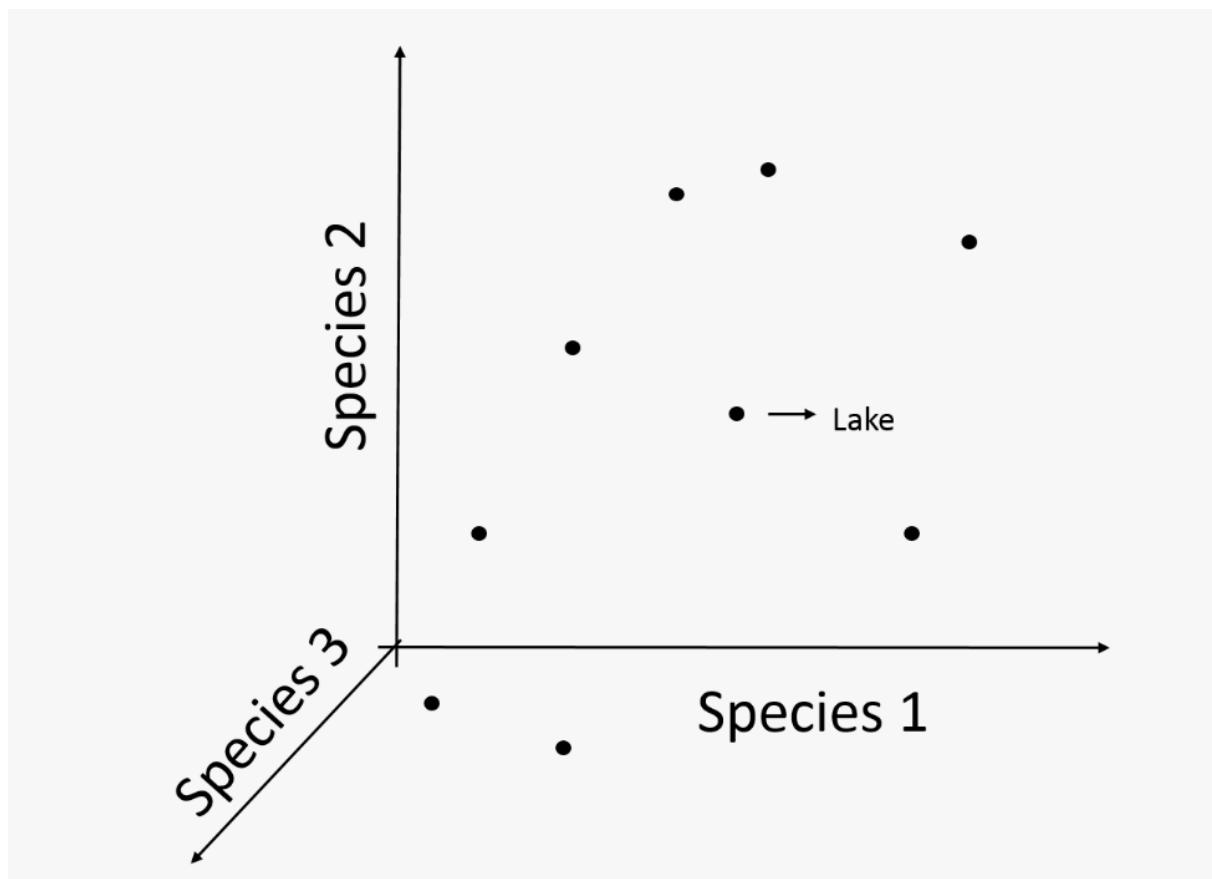
Que se passe-t-il si nous voulons nous intéresser à la réponse de différentes espèces d'algues?



Deux Dimensions



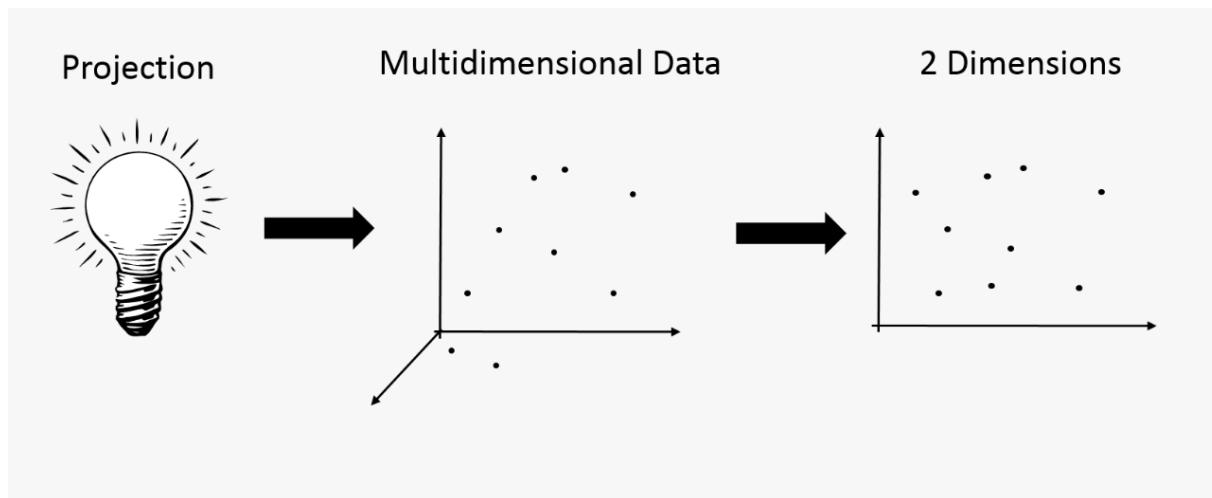
Trois Dimensions



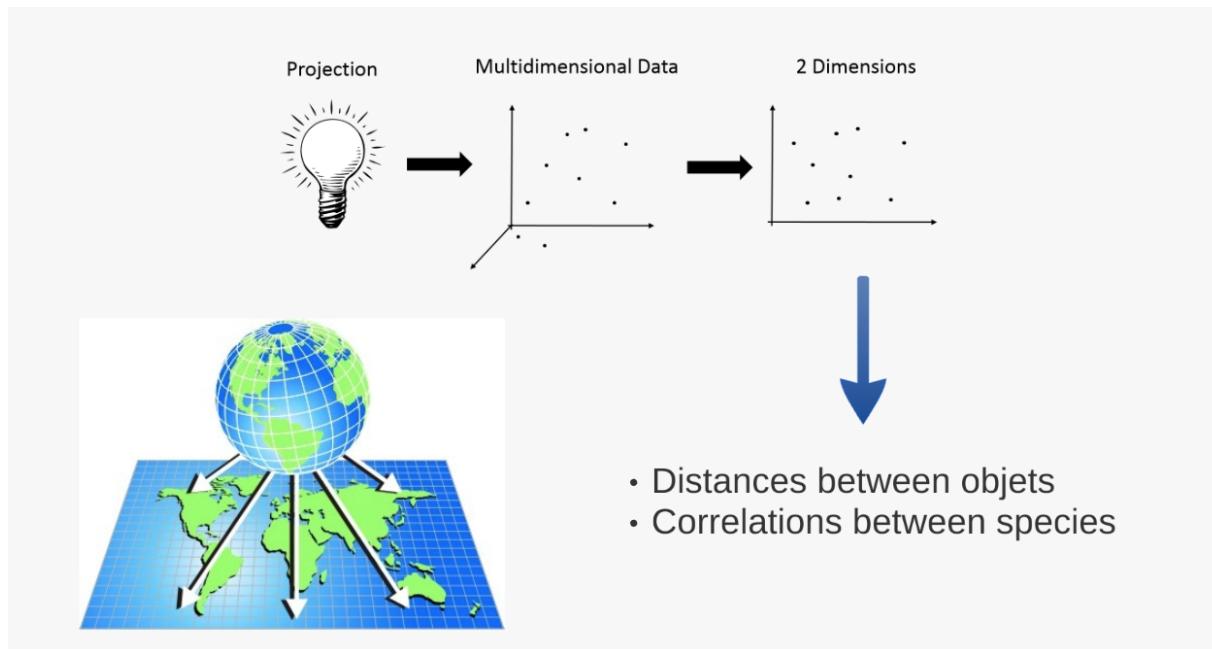
4,5,6, ou plus de Dimensions



Ordination en espace réduit



Ordination en espace réduit



- L'algèbre matricielle est complexe et difficile à comprendre
- Une compréhension générale est suffisante pour utiliser efficacement les méthodes d'ordination

Méthodes pour la recherche scientifique

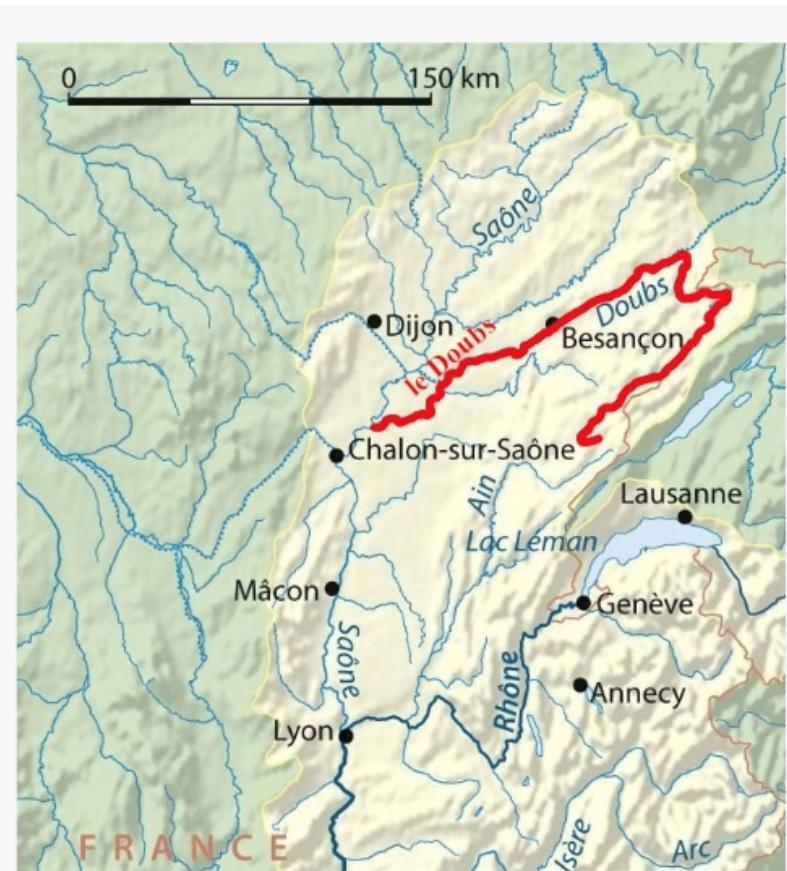
- **Questions / Hypothèses**
- **Design expérimental**
- **Collecte de données**
- **Transformation / Distance**
- **Analyses**
- **Rédaction**
- **Communication**

2. Exploration des données

Données de poissons de la rivière Doubs

Données de Verneaux (1973) :

- caractérisation des communautés de poissons
- 27 espèces
- 30 sites
- 11 variables environnementales



Données de poissons de la rivière Doubs

Chargement des données espèces (`Doubs.Spe.csv`)

```
spe <- read.csv("data/doubsspe.csv", row.names = 1)
spe <- spe[-8,] # supprimer le site vide
```

Chargement des données environnementales (`Doubs.Env.csv`)

```
env <- read.csv("data/doubsenv.csv", row.names = 1)
env <- env[-8,] # remove site with no data
```

Attention, n'exécuter qu'une seule fois

Exploration des données

Explorer le contenu des données espèces :

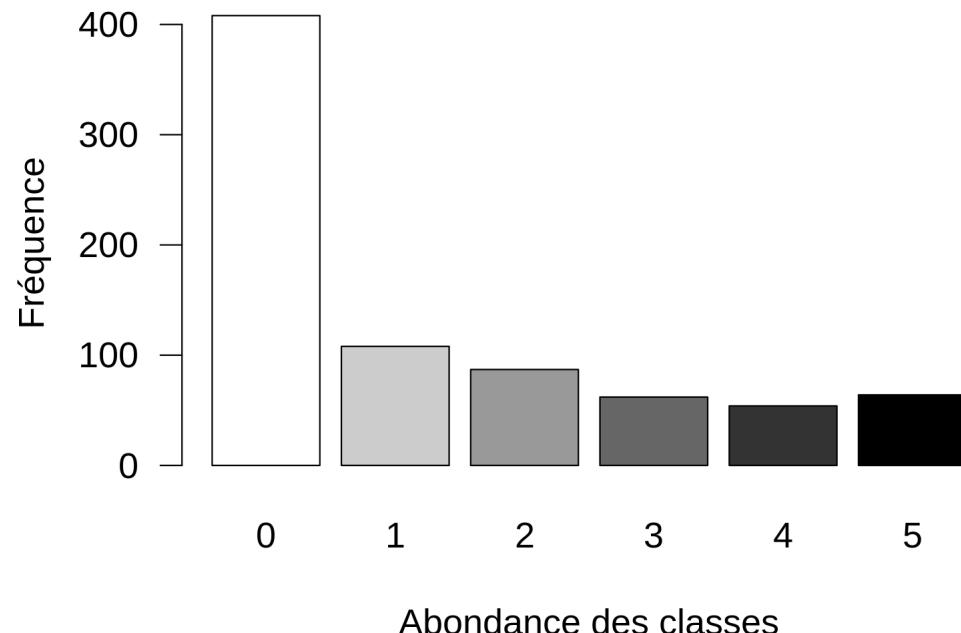
```
names(spe) # noms des objets  
dim(spe) # dimensions  
str(spe) # structure des objets  
summary(spe) # résumé statistique  
head(spe) # 6 premières lignes
```

#	CHA	TRU	VAI	LOC	OMB	BLA	HOT	TOX	VAN	CHE	BAR	SPI	GOU	BRO	PER	BOU	PSO	ROT	CAR
# 1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
# 2	0	5	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
# 3	0	5	5	5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
# 4	0	4	5	5	0	0	0	0	0	0	1	0	0	1	2	2	0	0	0
# 5	0	2	3	2	0	0	0	0	0	5	2	0	0	2	4	4	0	0	2
# 6	0	3	4	5	0	0	0	0	1	2	0	0	1	1	1	0	0	0	0
#	TAN	BCO	PCH	GRE	GAR	BBO	ABL	ANG											
# 1	0	0	0	0	0	0	0	0											
# 2	0	0	0	0	0	0	0	0											
# 3	0	0	0	0	0	0	0	0											
# 4	1	0	0	0	0	0	0	0											
# 5	3	0	0	0	5	0	0	0											
# 6	2	0	0	0	1	0	0	0											

Fréquences des espèces

Observer la distribution de fréquence des espèces :

```
ab <- table(unlist(spe))
barplot(ab, las = 1, col = grey(5:0/5),
        xlab = "Abondance des classes", ylab = "Fréquence")
```



Notez la proportion de 0

Fréquences des espèces

Combien de zéros?

```
sum(spe == 0)  
# [1] 408
```

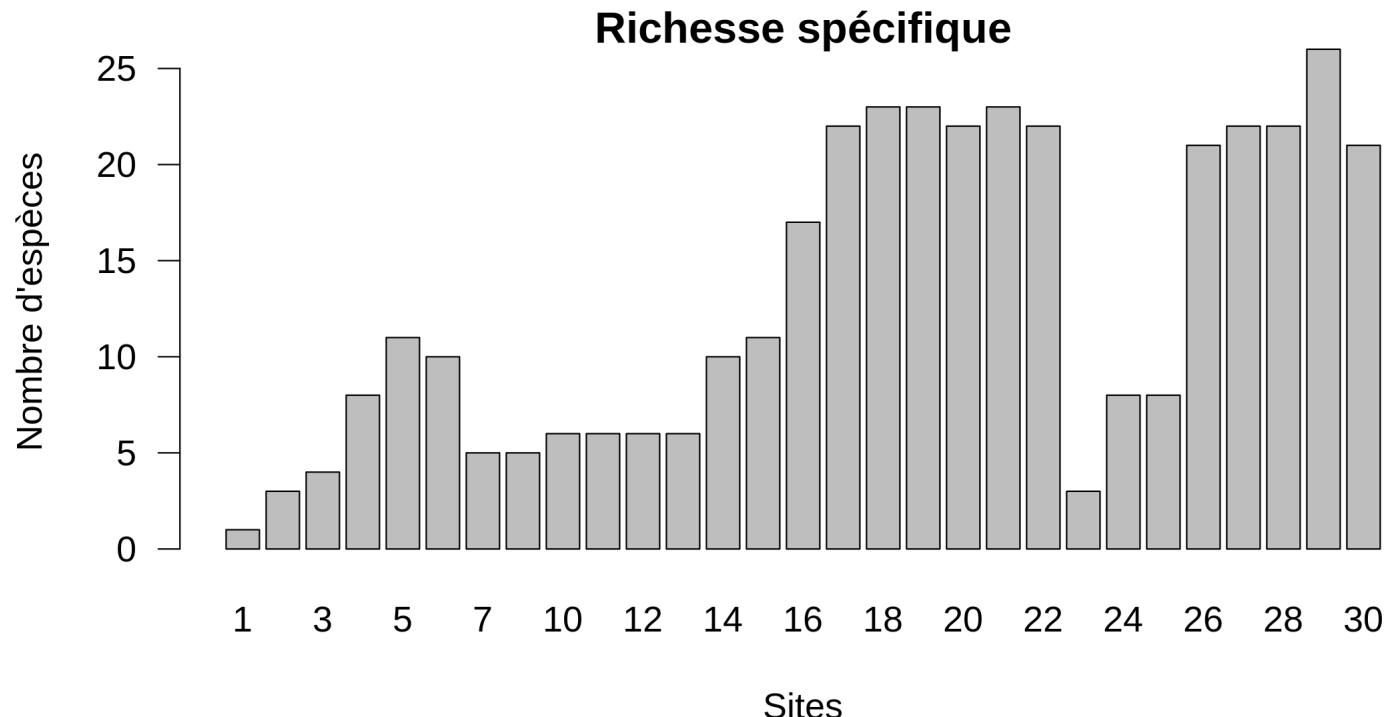
Quelle proportion de zéros?

```
sum(spe == 0)/(nrow(spe)*ncol(spe))  
# [1] 0.5210728
```

Richesse totale en espèce

Observer le nombre d'espèces présentes dans chaque site :

```
site.pre <- rowSums(spe > 0)
barplot(site.pre, main = "Richesse spécifique",
        xlab = "Sites", ylab = "Nombre d'espèces",
        col = "grey ", las = 1)
```



Comprenez vos données!

...pour choisir la transformation et la distance appropriée

- Y-a-t-il beaucoup de zéros?
- Que veulent-ils dire?

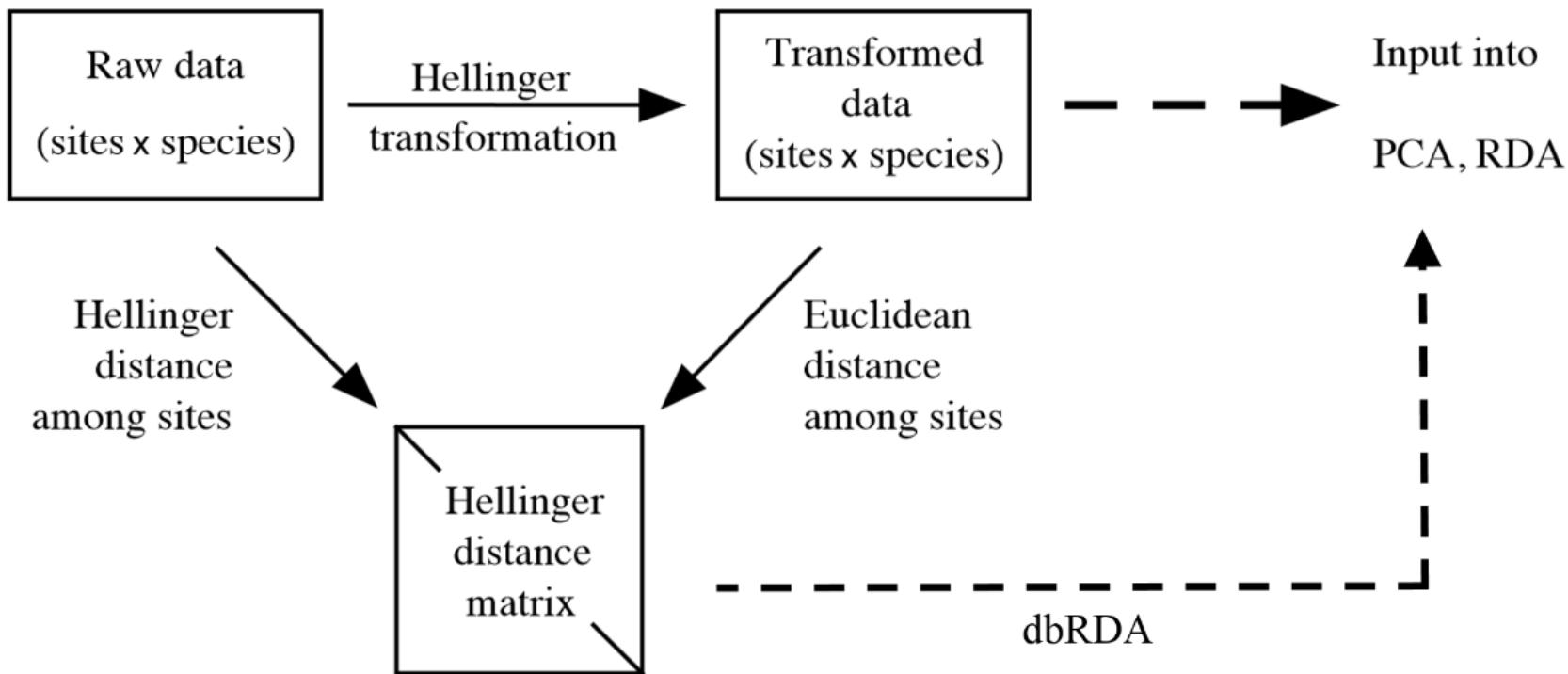
Une mesure de 0 (e.g 0mg/L, 0° C) n'est pas équivalent à un 0 représentant une absence d'observation.

Avant de transformer vos données de composition des communautés...

Considérations importantes:

- abondances/comptes/présence-absence relatives?
- distributions asymétriques ?
- beaucoup d'espèces rares?
- surabondance d'espèces dominantes?
- problème de double Zéro?

Transformer les données de composition des communautés



Modified from Legendre & Gallagher (2001)

Transformer les données de composition des communautés

Exemples

Transformer des comptes en présence - absence

```
library(vegan)
spec.pa <- decostand(spe, method = "pa")
```

Réduire le poids des espèces rares

```
spec.hel <- decostand(spe, method = "hellinger")
spec.chi <- decostand(spe, method = "chi.square")
```

Réduire le poids des espèces abondantes

```
spe.pa <- decostand(spe, method = "log")
```

Données sur l'environnement

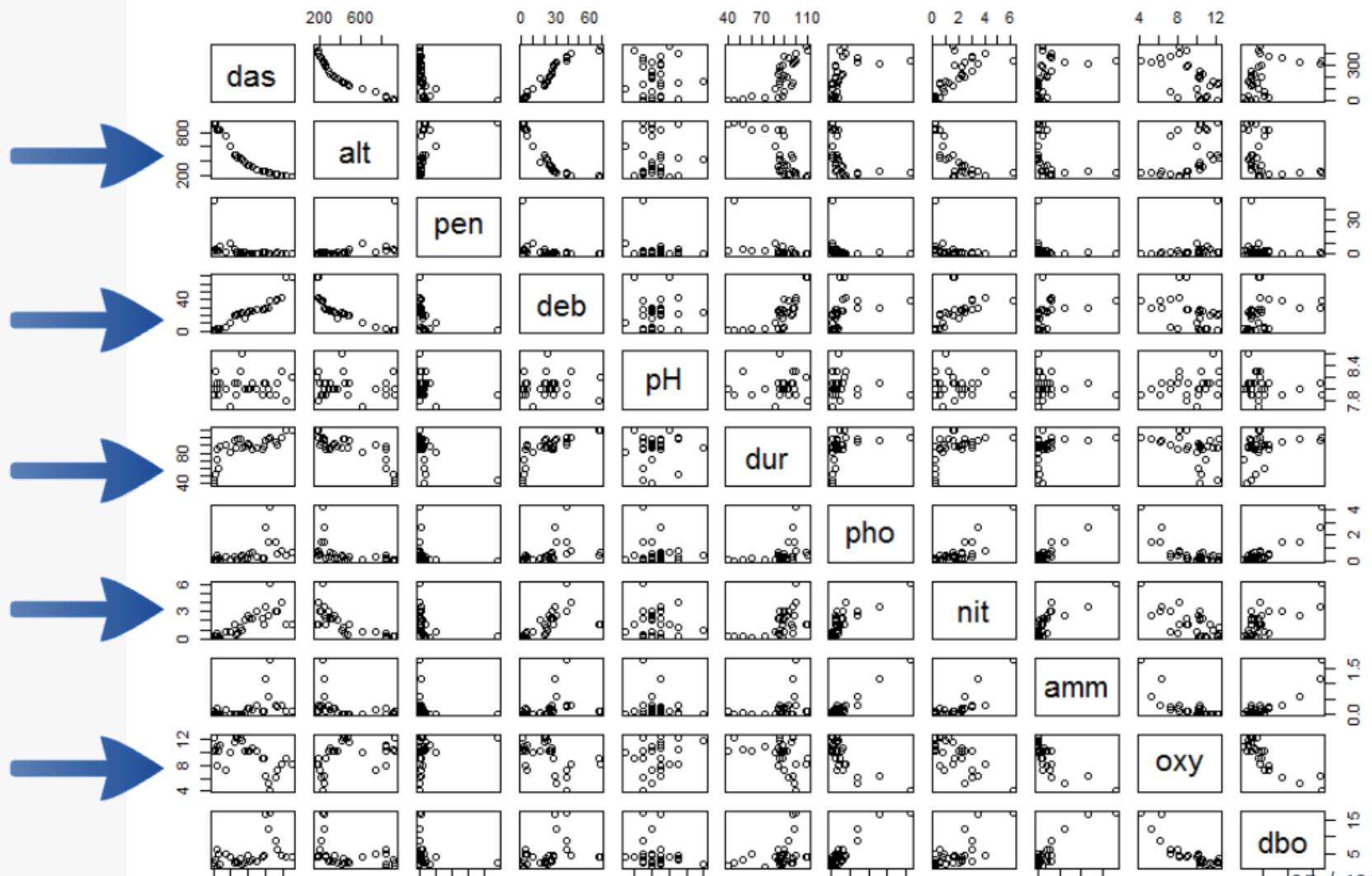
```
names(env) # Names of objects  
dim(env) # dimensions  
str(env) # structure of objects  
summary(env) # summary statistics  
head(env) # first 6 rows  
  
head(env) # first 6 rows  
#   das alt pen deb pH dur pho nit amm oxy dbo  
# 1 0.3 934 48.0 0.84 7.9 45 0.01 0.20 0.00 12.2 2.7  
# 2 2.2 932 3.0 1.00 8.0 40 0.02 0.20 0.10 10.3 1.9  
# 3 10.2 914 3.7 1.80 8.3 52 0.05 0.22 0.05 10.5 3.5  
# 4 18.5 854 3.2 2.53 8.0 72 0.10 0.21 0.00 11.0 1.3  
# 5 21.5 849 2.3 2.64 8.1 84 0.38 0.52 0.20 8.0 6.2  
# 6 32.4 846 3.2 2.86 7.9 60 0.20 0.15 0.00 10.2 5.3
```

Explorer la colinéarité en visualisant les corrélations entre les variables

```
pairs(env, main = "Bivariate Plots of the Environmental Data")
```

Données sur l'environnement

Bivariate Plots of the Environmental Data



Standardisation

Standardiser les variables environnementales est indispensable car il est impossible de comparer des variables d'unités différentes :

```
## ?decostand  
env.z <- decostand(env, method = "standardize")
```

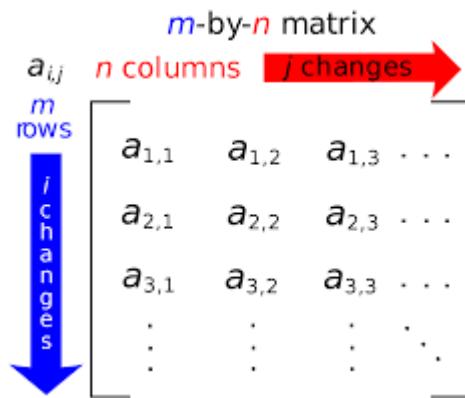
Cette fonction centre-réduit les données pour permettre la fiabilité des analyses :
:

```
apply(env.z, 2, mean)  
# das alt pen deb pH  
# -7.959539e-17 -4.795165e-17 2.494600e-17 -7.323225e-17 -1.730430e-15  
# dur pho nit amm oxy  
# -2.028505e-16 4.445790e-17 2.875893e-17 2.754434e-17 -4.038167e-16  
# dbo  
# 9.829975e-17  
apply(env.z, 2, sd)  
# das alt pen deb pH dur pho nit amm oxy dbo  
# 1 1 1 1 1 1 1 1 1 1 1
```

3. Similarité / Dissimilité

Mesure d'association

L'algébre matricielle est au cœur de plusieurs méthodes d'analyses multivariées

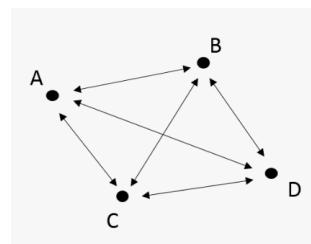


- Explorer différentes mesures de distance entre objets permet de mieux comprendre le fonctionnement de l'ordination

Au-delà de la 1ère dimension

- Les jeux de données écologiques correspondent souvent à de grandes matrices
- L'ordination calcule les relations entre espèces, ou entre objets
- Ces relations peuvent être simplifiées par des mesures de dissimilarités

	sp1	sp2	...
A			
B			
C			
D			
...			

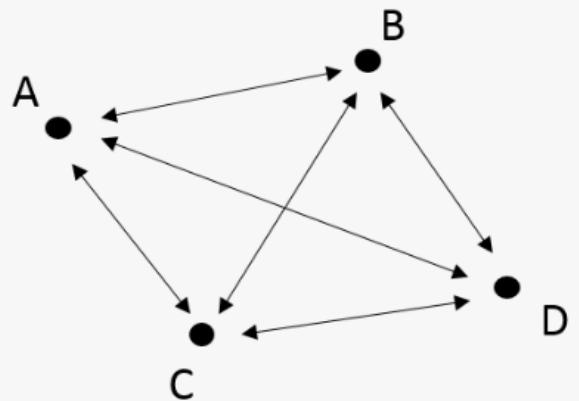


	A	B	C	D
A				
B				
C				
D				

Similarité / Dissimilarité

- Utile pour comprendre vos données
- Certains types d'ordination ou de groupement nécessitent des mesures appropriées

$$\text{Similarité: } S = 1 - D \quad \text{Distance: } D = 1 - S$$



	A	B	C	D
A				
B				
C				
D				

Mesures de distance des communautés

- Euclidienne
- Manhattan
- Corde
- Hellinger
- Chi-carré
- Bray-Curtis

Chaque mesure est utile dans différentes situations

Comparaison des sites de la rivière Doubs

La fonction `vegdist()` comprend les mesures de distances communes :

?vegdist

Comment la composition des communautés diffère-t-elle entre les 30 sites de la rivière Doubs?

```
spe.db.pa <- vegdist(spe, method = "bray")
```

Comparaison des sites

	1	2	3	4	5	6	7	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	
2	0.600																												
3	0.684	0.143																											
4	0.750	0.333	0.189																										
5	0.892	0.696	0.680	0.491																									
6	0.750	0.394	0.297	0.190	0.418																								
7	0.684	0.143	0.125	0.243	0.640	0.243																							
9	1.000	0.692	0.733	0.657	0.583	0.543	0.667																						
10	0.882	0.385	0.400	0.371	0.542	0.257	0.267	0.571																					
11	0.571	0.304	0.407	0.438	0.689	0.438	0.333	0.760	0.440																				
12	0.714	0.200	0.235	0.333	0.692	0.385	0.176	0.688	0.375	0.241																			
13	0.727	0.290	0.314	0.450	0.736	0.550	0.371	0.818	0.576	0.333	0.189																		
14	0.806	0.400	0.318	0.347	0.677	0.429	0.364	0.762	0.476	0.436	0.217	0.191																	
15	0.833	0.511	0.469	0.407	0.552	0.370	0.388	0.660	0.404	0.500	0.333	0.385	0.246																
16	0.860	0.654	0.571	0.475	0.459	0.377	0.536	0.704	0.519	0.647	0.552	0.593	0.441	0.260															
17	0.915	0.679	0.633	0.508	0.513	0.446	0.600	0.690	0.517	0.636	0.581	0.619	0.500	0.403	0.262														
18	0.956	0.741	0.724	0.587	0.500	0.524	0.690	0.643	0.571	0.698	0.667	0.705	0.600	0.467	0.341	0.140													
19	1.000	0.793	0.710	0.612	0.500	0.522	0.677	0.667	0.633	0.825	0.750	0.815	0.676	0.570	0.395	0.311	0.250												
20	1.000	0.912	0.889	0.740	0.489	0.688	0.861	0.686	0.771	0.910	0.892	0.920	0.833	0.708	0.583	0.420	0.327	0.235											
21	1.000	0.946	0.923	0.783	0.500	0.735	0.897	0.763	0.816	0.918	0.925	0.951	0.867	0.768	0.627	0.491	0.404	0.296	0.102										
22	1.000	0.976	0.955	0.828	0.528	0.785	0.932	0.767	0.860	0.952	0.956	0.978	0.900	0.771	0.661	0.552	0.474	0.390	0.188	0.104									
23	1.000	1.000	1.000	0.920	0.895	0.840	0.900	0.778	0.889	0.867	0.909	1.000	0.938	0.946	0.909	0.833	0.826	0.840	0.867	0.879	0.895								
24	1.000	1.000	1.000	0.889	0.796	0.778	0.935	0.724	0.793	0.923	0.939	1.000	0.907	0.875	0.818	0.695	0.649	0.639	0.577	0.610	0.655	0.579							
25	1.000	1.000	0.926	0.812	0.689	0.688	0.852	0.840	0.760	0.909	0.931	1.000	0.846	0.818	0.765	0.745	0.660	0.614	0.672	0.699	0.735	0.467	0.462						
26	1.000	0.964	0.932	0.781	0.558	0.688	0.898	0.719	0.825	0.926	0.934	0.968	0.859	0.763	0.639	0.540	0.459	0.326	0.212	0.200	0.252	0.830	0.483	0.593					
27	1.000	0.973	0.949	0.833	0.567	0.762	0.924	0.766	0.844	0.946	0.951	0.976	0.890	0.771	0.670	0.570	0.486	0.376	0.193	0.136	0.126	0.881	0.615	0.703	0.189				
28	1.000	0.976	0.953	0.824	0.577	0.780	0.930	0.762	0.857	0.951	0.955	0.978	0.898	0.786	0.691	0.579	0.500	0.414	0.222	0.167	0.127	0.892	0.647	0.728	0.239	0.098			
29	0.978	0.939	0.922	0.815	0.537	0.778	0.903	0.782	0.842	0.898	0.905	0.906	0.843	0.733	0.654	0.511	0.442	0.414	0.245	0.181	0.119	0.912	0.706	0.776	0.338	0.187	0.146		
30	1.000	1.000	0.981	0.873	0.593	0.836	0.962	0.845	0.903	0.980	0.981	1.000	0.932	0.820	0.721	0.579	0.527	0.481	0.297	0.232	0.180	0.914	0.712	0.780	0.364	0.197	0.157	0.148	

Comparaison des sites

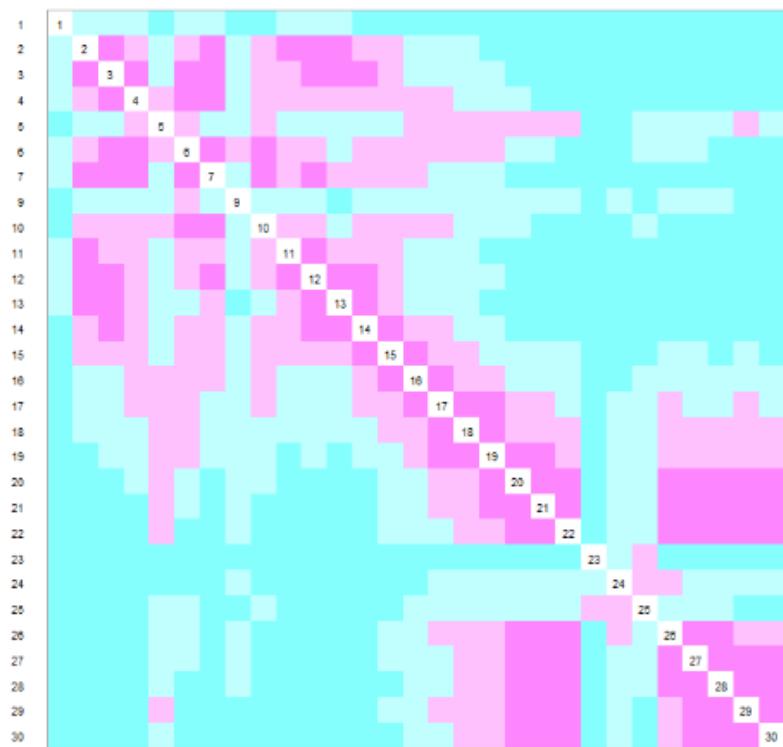
	1	2	3	4	5	6	7	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29		
2	0.600																													
3	0.684	0.143																												
4	0.750	0.333	0.189																											
5	0.892	0.696	0.680	0.491																										
6	0.750	0.394	0.297	0.190	0.118																									
7	0.684	0.143	0.125	0.243	0.081	0.243																								
8	1.000	0.692	0.733	0.657	0.583	0.543	0.667																							
9	0.882	0.385	0.400	0.371	0.542	0.457	0.267	0.571																						
10	0.571	0.304	0.407	0.438	0.689	0.471	0.333	0.760	0.440																					
11	0.714	0.200	0.235	0.333	0.692	0.383	0.176	0.688	0.375	0.241																				
12	0.727	0.290	0.314	0.450	0.736	0.550	0.471	0.818	0.576	0.333	0.189																			
13	0.806	0.400	0.318	0.347	0.677	0.429	0.311	0.762	0.476	0.436	0.217	0.191																		
14	0.833	0.511	0.469	0.407	0.552	0.370	0.388	0.660	0.404	0.500	0.333	0.385	0.246																	
15	0.860	0.654	0.571	0.475	0.459	0.377	0.536	0.704	0.519	0.647	0.552	0.593	0.441	0.260																
16	0.915	0.679	0.633	0.508	0.513	0.446	0.600	0.181	0.517	0.636	0.581	0.619	0.500	0.403	0.262															
17	0.956	0.741	0.724	0.587	0.500	0.524	0.690	0.643	0.571	0.698	0.667	0.705	0.600	0.467	0.341	0.140														
18	1.000	0.793	0.710	0.612	0.500	0.522	0.677	0.667	0.633	0.825	0.750	0.815	0.676	0.570	0.395	0.311	0.250													
19	1.000	0.912	0.889	0.740	0.489	0.688	0.861	0.686	0.701	0.910	0.892	0.920	0.833	0.708	0.583	0.420	0.327	0.235												
20	1.000	0.946	0.923	0.783	0.500	0.735	0.897	0.763	0.816	0.918	0.925	0.951	0.867	0.768	0.627	0.491	0.404	0.296	0.102											
21	1.000	0.976	0.955	0.828	0.528	0.785	0.932	0.767	0.860	0.952	0.956	0.978	0.900	0.771	0.661	0.552	0.474	0.390	0.188	0.104										
22	1.000	1.000	0.920	0.895	0.840	0.900	0.778	0.889	0.883	0.909	1.000	0.938	0.946	0.909	0.833	0.876	0.840	0.867	0.879	0.895										
23	1.000	1.000	1.000	0.920	0.895	0.840	0.900	0.778	0.889	0.883	0.909	1.000	0.938	0.946	0.909	0.833	0.876	0.840	0.867	0.879	0.895									
24	1.000	1.000	1.000	0.889	0.796	0.778	0.935	0.724	0.793	0.923	0.939	1.000	0.907	0.875	0.818	0.695	0.649	0.639	0.577	0.610	0.655	0.579								
25	1.000	1.000	0.926	0.812	0.689	0.688	0.852	0.840	0.760	0.909	0.931	1.000	0.846	0.818	0.765	0.745	0.660	0.614	0.672	0.699	0.735	0.467	0.462							
26	1.000	0.964	0.932	0.781	0.558	0.688	0.898	0.719	0.825	0.926	0.951	0.968	0.859	0.763	0.639	0.540	0.459	0.326	0.212	0.200	0.252	0.830	0.483	0.593						
27	1.000	0.973	0.949	0.833	0.567	0.762	0.924	0.766	0.844	0.946	0.951	0.976	0.890	0.771	0.670	0.570	0.486	0.376	0.193	0.136	0.126	0.881	0.615	0.703	0.189					
28	1.000	0.976	0.953	0.824	0.577	0.780	0.930	0.762	0.857	0.951	0.955	0.978	0.898	0.786	0.691	0.579	0.500	0.414	0.222	0.167	0.127	0.892	0.647	0.728	0.239	0.098				
29	1.000	0.939	0.922	0.815	0.537	0.778	0.903	0.782	0.842	0.898	0.905	0.931	0.843	0.733	0.654	0.511	0.442	0.414	0.245	0.181	0.119	0.912	0.706	0.776	0.338	0.187	0.146			
30	1.000	1.000	0.981	0.873	0.593	0.836	0.962	0.845	0.903	0.980	0.981	1.000	0.932	0.820	0.721	0.579	0.527	0.481	0.297	0.232	0.180	0.914	0.712	0.780	0.364	0.197	0.157	0.148		

	1	2	3
1	0.000	0.600	0.684
2	0.600	0.000	0.143
3	0.684	0.143	0.000

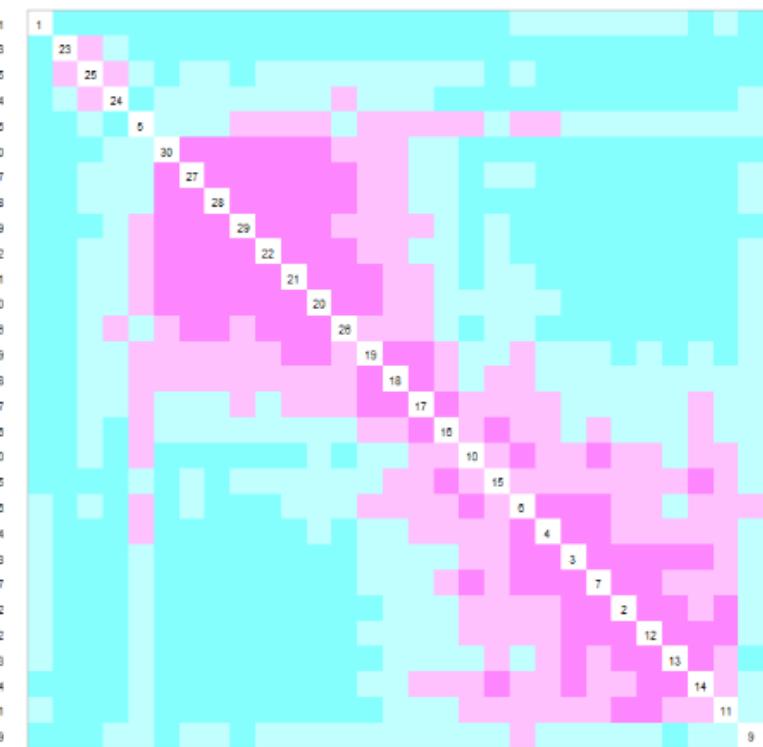
- Diagonal is zero
- Site 2 and 3 most similar
- Site 1 and 3 most different

Visualisation d'une matrice de distances

Dissimilarity Matrix



Ordered Dissimilarity Matrix



Défi #1



Discuter avec votre voisin:

Comment savoir si deux objets caractérisés par des données multidimensionnelles sont similaires?

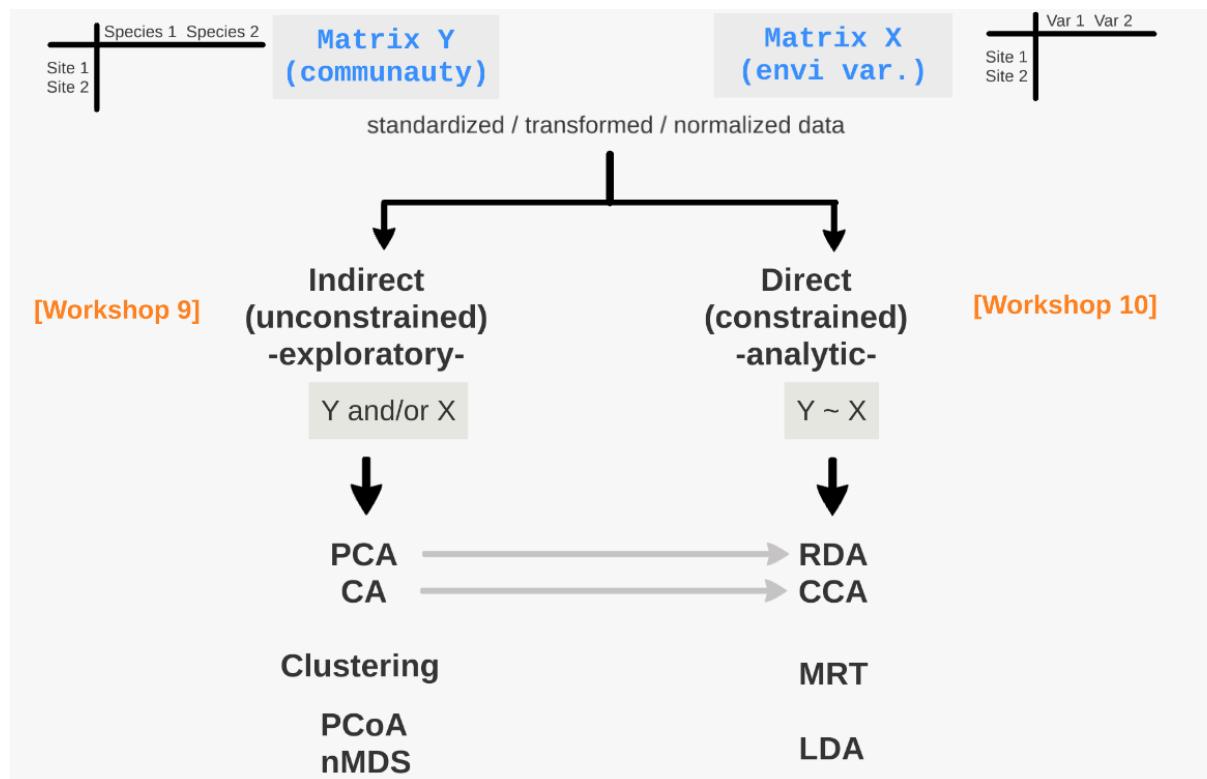
- Faites une liste de vos suggestions

Et qu'en est-il de l'ordination?

Avec des méthodes d'ordination, nous ordonnons vos objets (sites) en fonction de leur similarité

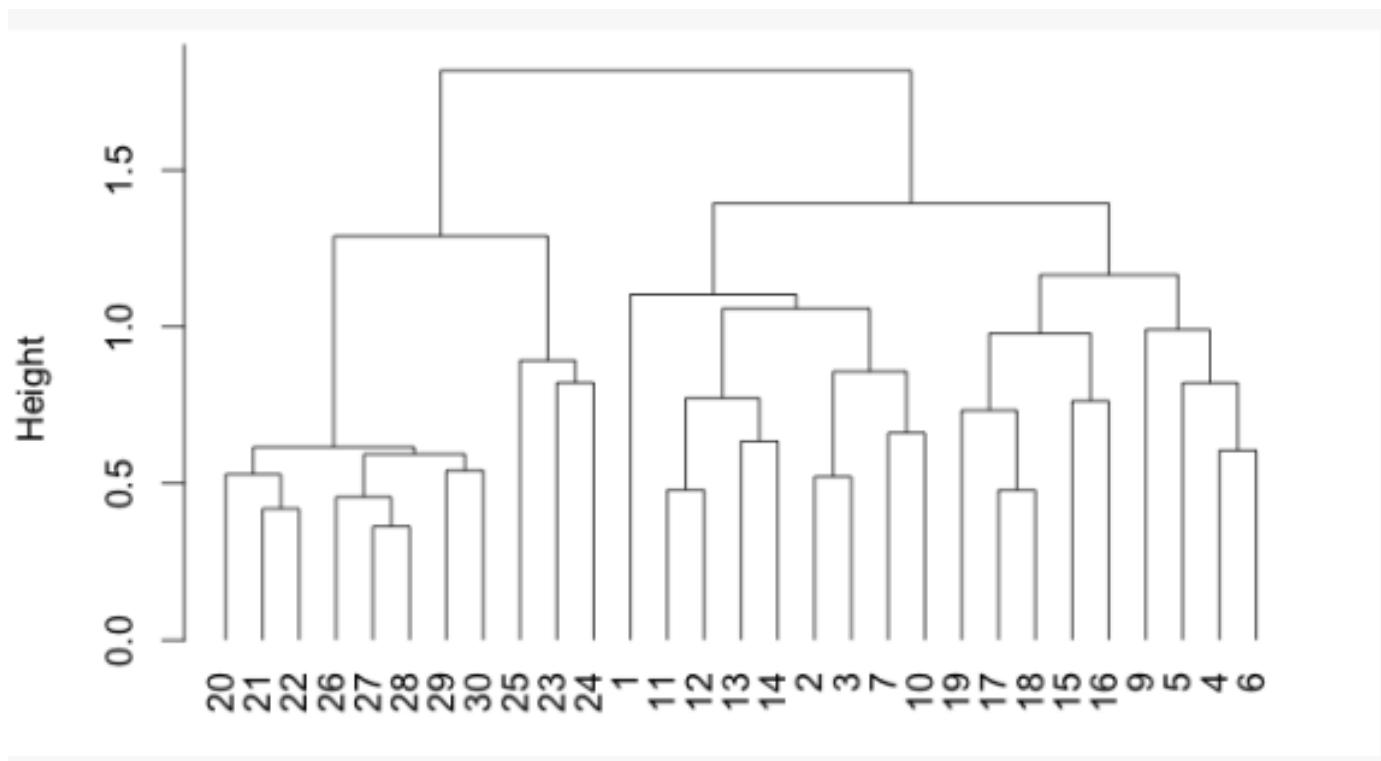
- Plus les sites sont similaires, plus ils sont proches dans l'espace d'ordination (plus petites distances)
- En écologie, on calcule habituellement la similarité entre sites en fonction de leur composition en espèces ou de leur conditions environnementales.

Analyse schématique des analyses multivariées



Groupement

- Permet de mettre en lumière des structures dans les données en partitionnant les objets
- Les résultats sont représentés sous forme de dendrogramme (arbre)
- Pas une méthode statistique!



Aperçu de 3 méthodes hiérarchiques

- Groupement agglomératif à liens simples
- Groupement agglomératif à liens complets
- Groupement de Ward
- Les éléments de petits ensembles se regroupent en groupes plus vastes de rang supérieur
 - (e.g. espèces, genres, familles, ordres...)

Groupement hiérarchique

À partir d'une matrice de distances, on classe les objets en ordre croissant

	2	3	4	5
1	0.10	0.15	0.40	0.80
2		0.60	0.35	0.70
3			0.30	0.65
4				0.75

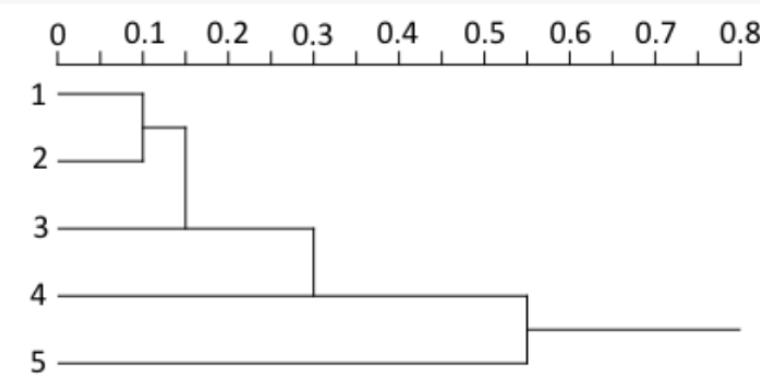


Dist	Pair
0.10	1-2
0.15	1-3
0.30	3-4
0.35	2-4
0.40	1-4
0.60	2-3
0.65	3-5
0.70	2-5
0.75	4-5
0.80	1-5

Groupement à liens simples

Dist	Pair
0.10	1-2
0.15	1-3
0.30	3-4
0.35	2-4
0.40	1-4
0.60	2-3
0.65	3-5
0.70	2-5
0.75	4-5
0.80	1-5

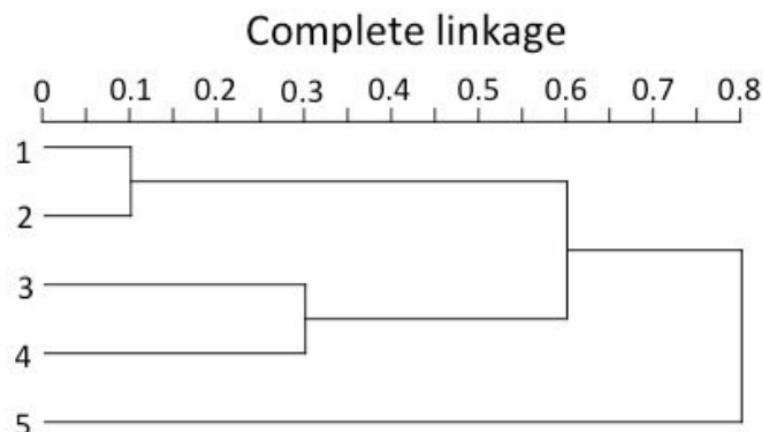
- Les deux objets les plus proches se regroupent
- Ensuite les deux objets les plus proches suivants
- et ainsi de suite.



Groupement à liens complets

Dist	Pair
0.10	1-2
0.15	1-3
0.30	3-4
0.35	2-4
0.40	1-4
0.60	2-3
0.65	3-5
0.70	2-5
0.75	4-5
0.80	1-5

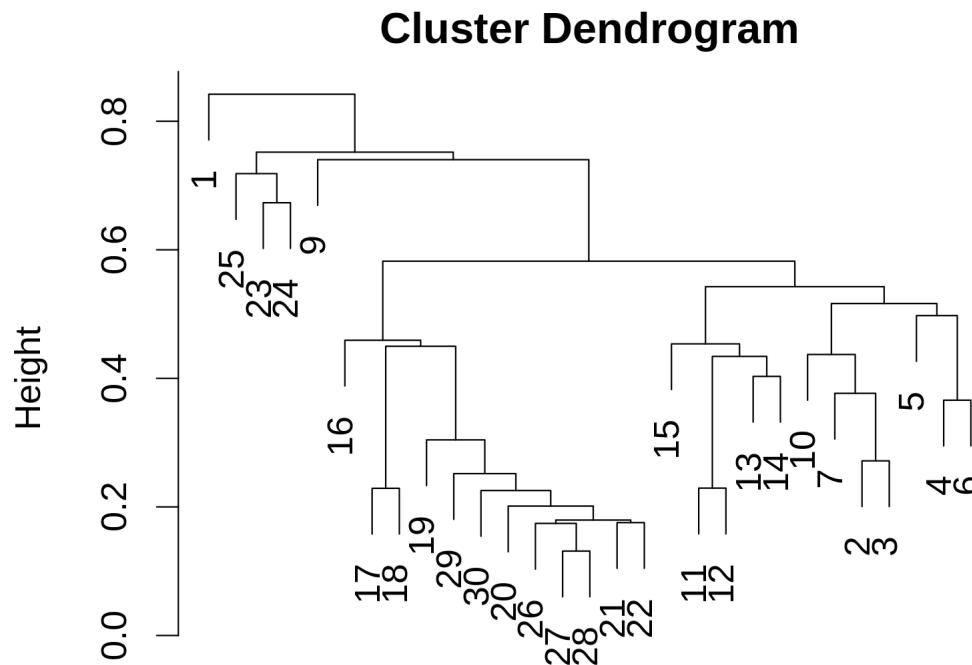
- Les deux objets les plus proches se regroupent
- Ensuite les groupes se lient à la distance à laquelle les objets qu'ils contiennent sont tous liés



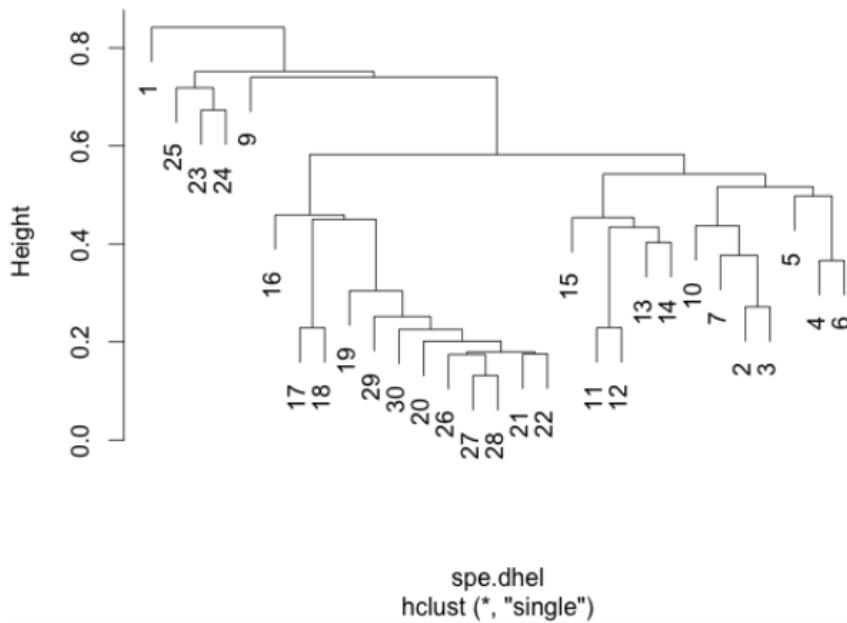
Comparaison

Créer une matrice de distance à partir des données de la rivière Doubs transformées Hellinger et faire le groupement à liens simples :

```
spe.dhe1 <- vegdist(spec.hel, method = "euclidean")
spe.dhe1.single <- hclust(spe.dhe1, method = "single")
plot(spe.dhe1.single)
```

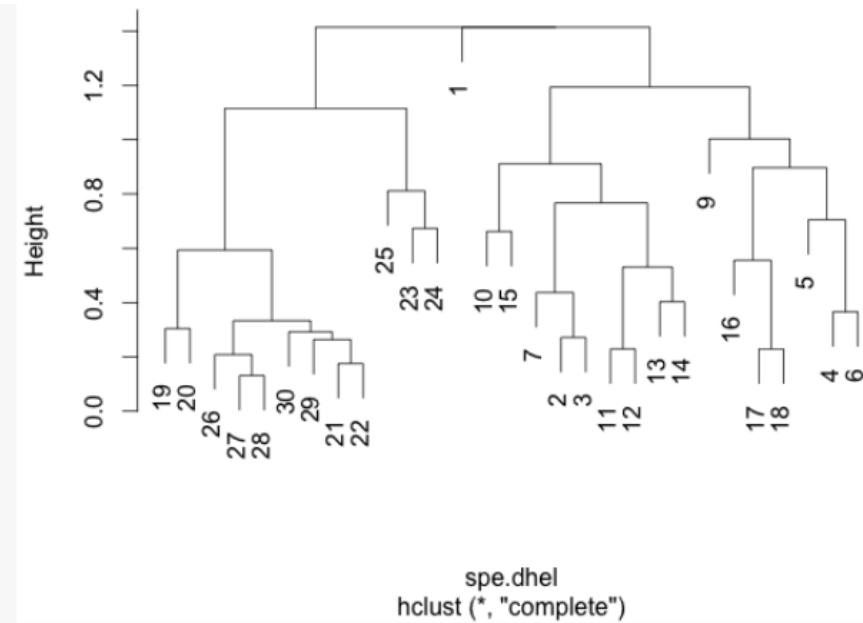


Comparaison



Liens simples :

Les objets ont tendance à s'enchaîner (e.g. 19,29,30,26)



Liens complets : Les groupes sont plus distincts

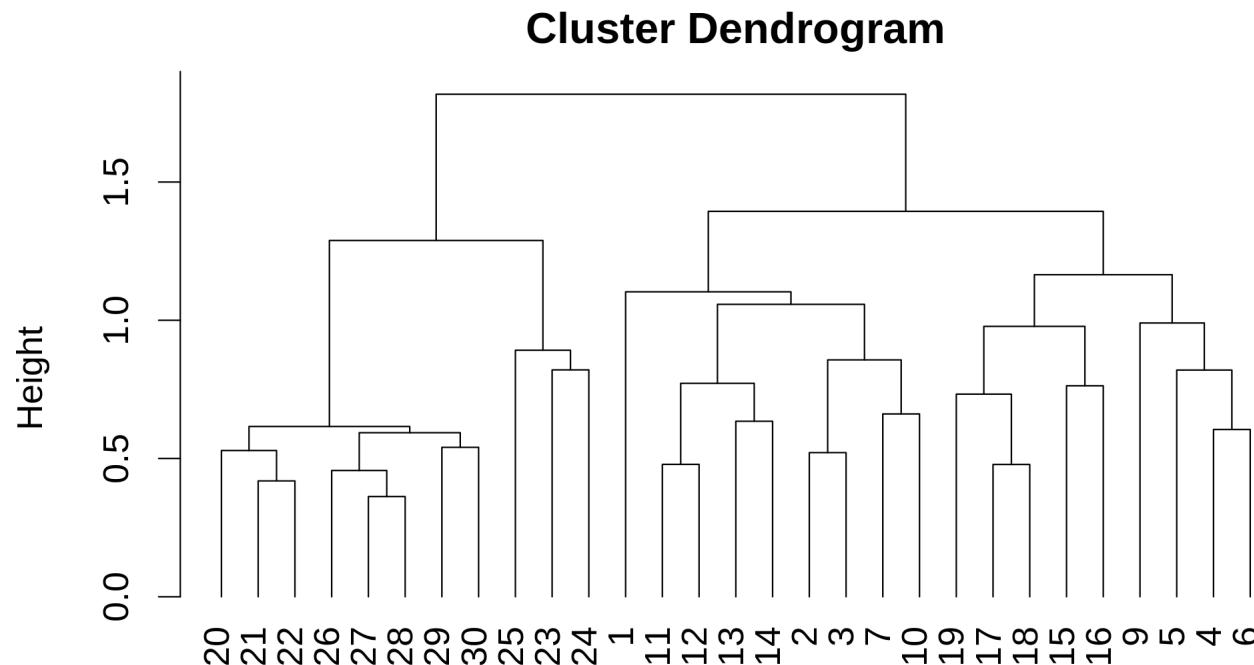
Groupement de Ward

- Utilise la méthode des moindres carrés pour lier les objets
 - les groupes fusionnent de façon à minimiser la variance intragroupe
 - à chaque étape, la paire de groupes à fusionner est celle qui résulte à la plus petite augmentation de la somme des carrés des écarts intra-groupes

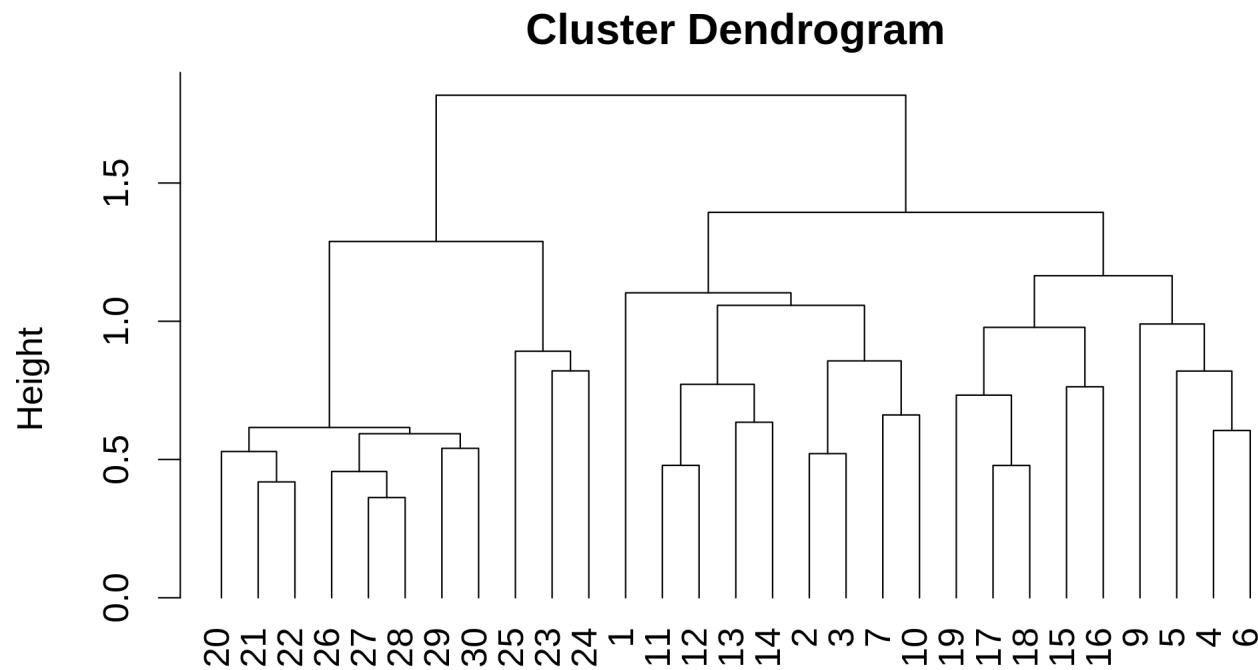
Groupement de Ward

Faire le groupement de Ward et dessiner le dendrogramme en utilisant la racine carrée des distances :

```
spe.dhel.ward <- hclust(spe.dhe1, method = "ward.D2")
spe.dhel.ward$height <- sqrt(spe.dhel.ward$height)
plot(spe.dhel.ward, hang = -1) # hang = -1 aligns objects at the same level
```



Groupement de Ward



Les objets ont tendance à former des groupes plus sphériques et homogènes

Comment choisir la bonne méthode ?

- Dépend de votre objectif
 - démontrer des gradients? des contrastes?
- Si plus d'une méthode semble adéquate, comparer les dendrogrammes
- Encore une fois : ceci **n'est pas** une méthode statistique Mais! il est possible de:
 - déterminer le nombre de groupe optimal
 - faire des tests statistiques sur les résultats
 - combiner le groupement à l'ordination pour distinguer des groupes de sites

4. Ordination non contrainte

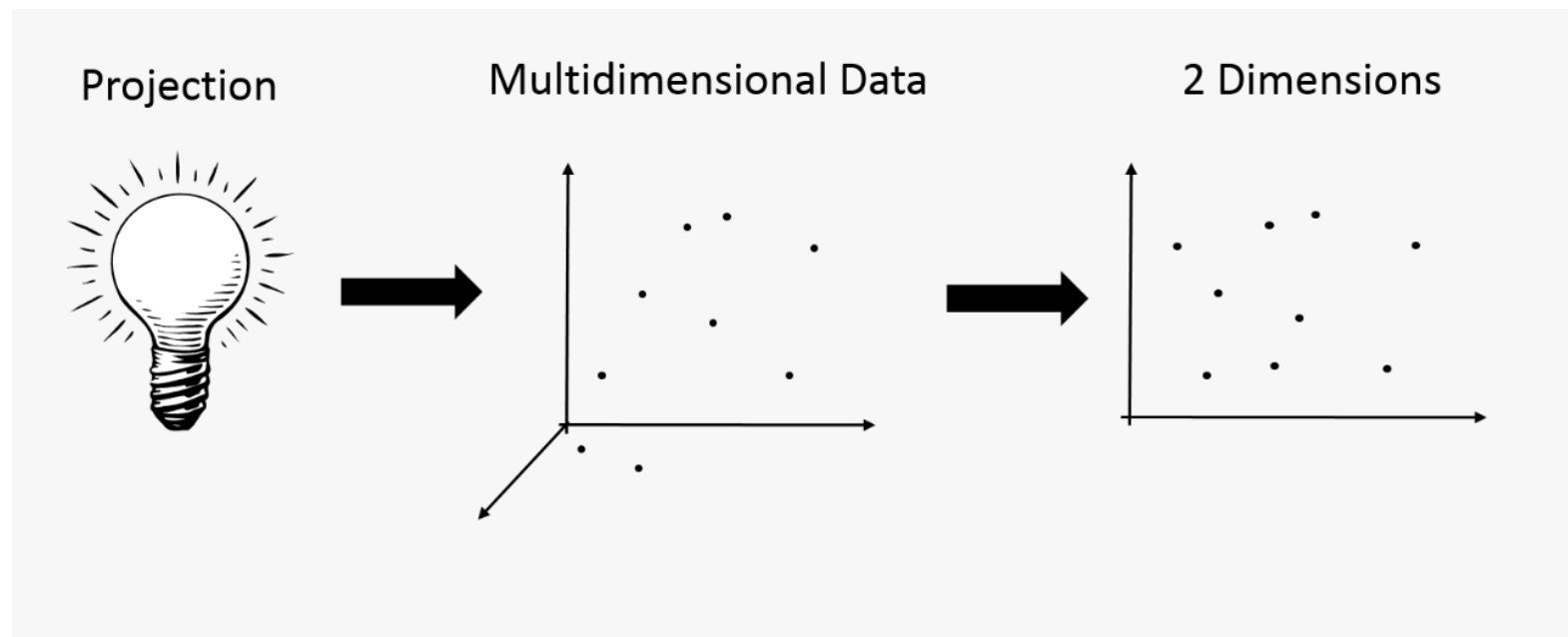
Définitions

- **Variance:** mesure de la dispersion d'une variable y_j de sa moyenne
- **Co-variance:** mesure de co-dispersion des variables y_j et y_i de leur moyenne
- **Corrélation:** mesure de la force du lien entre 2 variables : $r_{ij} = (d_{ij}/d_j x d_k)$
- **Valeurs propres:** proportion de variance (dispersion) représentée par un axe d'ordination
- **Orthogonalité:** angle droit entre 2 axes ou 2 flèches, ce qui veut dire qu'ils sont indépendants = non corrélés
- **Score:** position d'un point sur un axe. Tous les scores d'un point donnent ses coordonnées dans l'espace multidimensionnel. Ils peuvent être utilisés pour d'autres analyses (e.g combinaison linéaire de variables mesurées)
- **Dispersion** (inertie): Mesure de la variabilité totale du diagramme de dispersion de l'espace multidimensionnel en fonction de son centre de gravité

Ordination non contrainte

- Évalue la relation **dans** un ensemble de variables (espèces ou variables environnementales, et non **parmi** les ensembles, i.e analyse sous contraintes)
- Trouve les composants clés de la variation entre échantillons, sites, espèces, etc...
- Réduit le nombre de dimensions dans les données multivariées sans perte d'informations considérables
- Créer de nouvelles variables pour des analyses ultérieures (comme la régression)

4.1. Analyse en Composantes Principales (ACP ou PCA)



- Préserve, en 2 dimensions, le maximum de variation des données
- Il en résulte des variables synthétiques orthogonales entre elles (et donc non corrélées)

PCA - Ce qu'il vous faut

- Un jeu de données correspondant à des variables réponses (eg. composition de communautés) OU à des variables explicatives (e.g variables environnementales)

PAS LES DEUX!

- Échantillons correspondant à des mesures du même jeu de variables
- Généralement, un jeu de données plus long que large est préféré

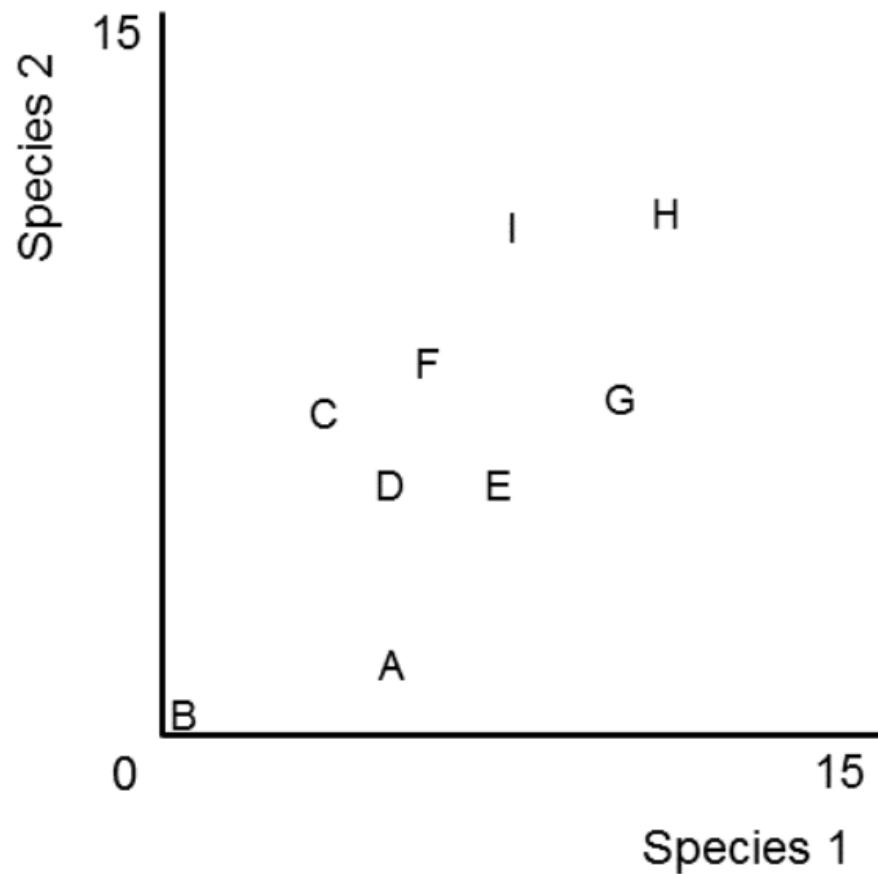
	sp1	sp2	...
A			
B			
C			
D			
...			

PCA - Principles

Site	Species 1	Species 2
A	7	3
B	4	3
C	12	10
D	23	11
E	13	13
F	15	16
G	18	14

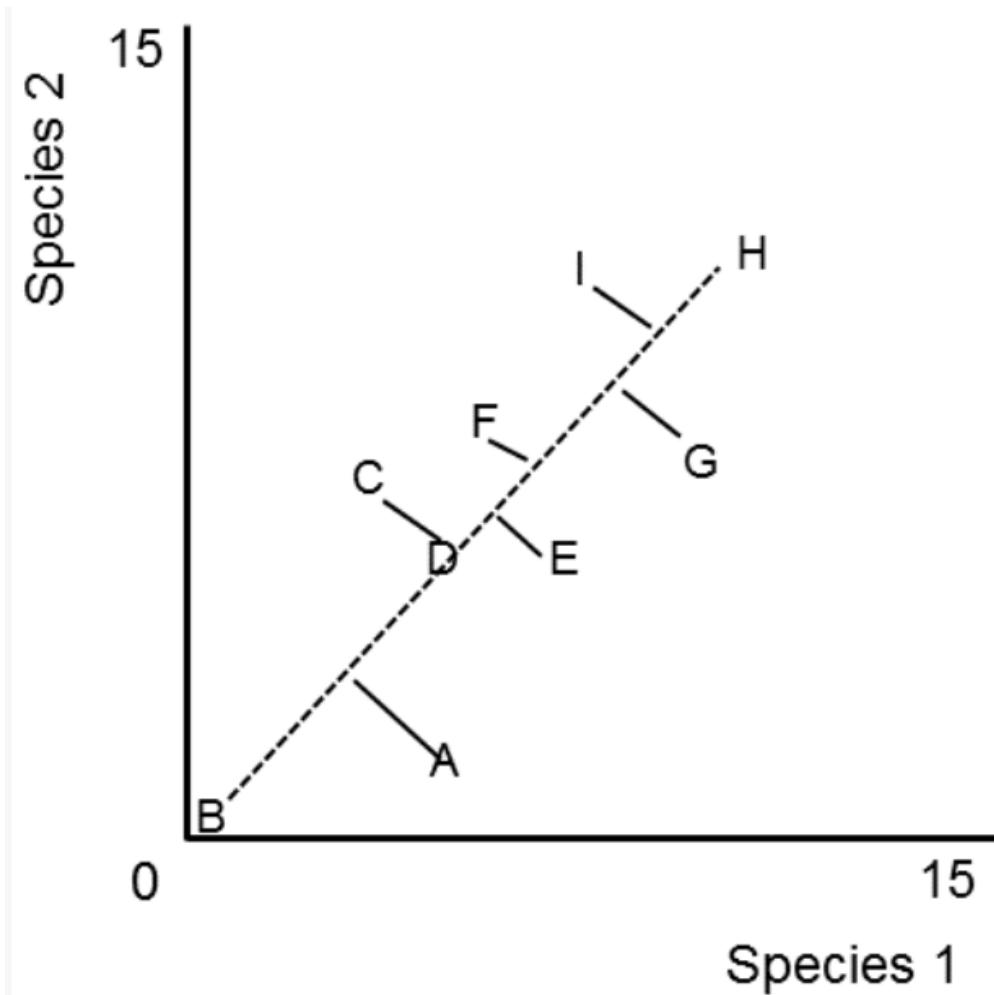
Un exemple simplifié

PCA - Principes



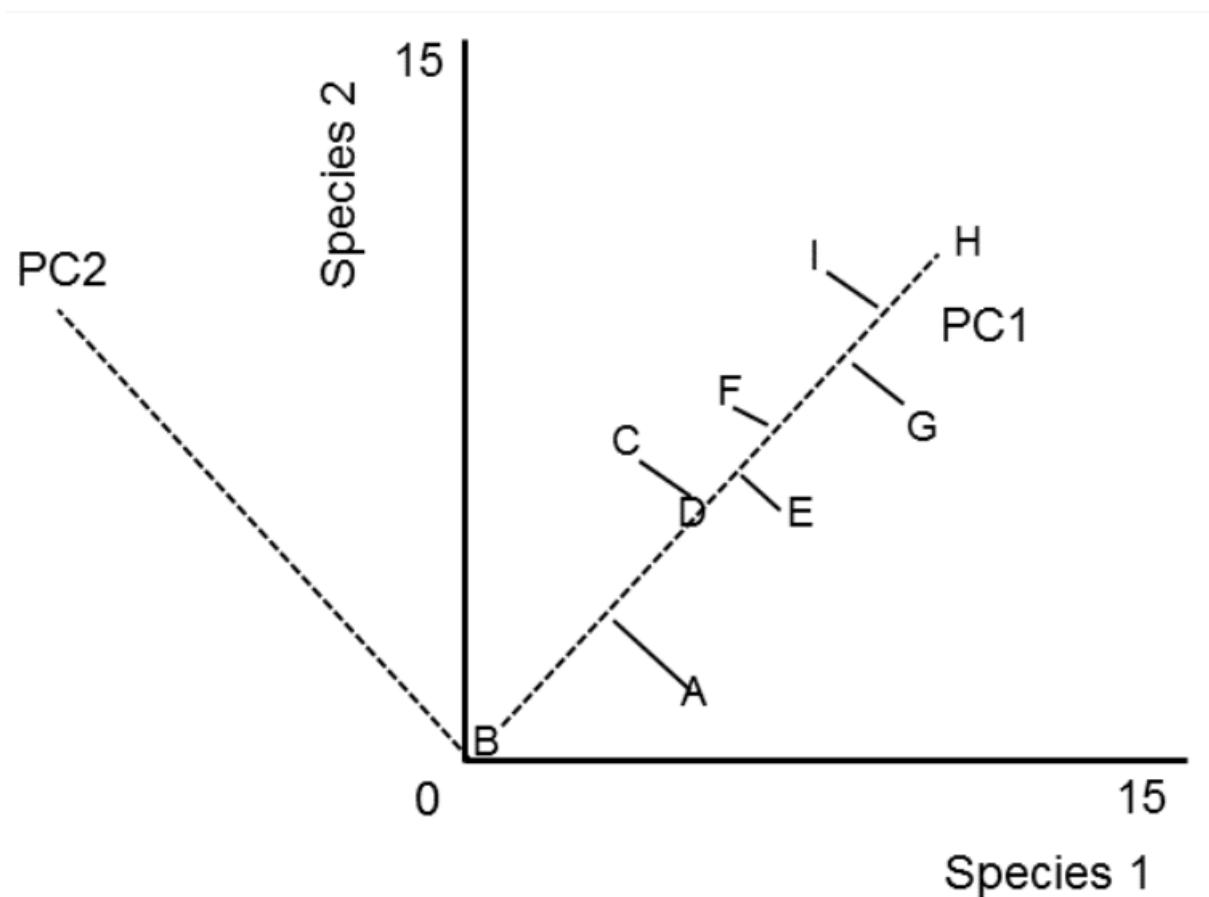
En 2D, les sites seraient disposés de cette façon... Notez la dispersion dans le diagramme de dispersion

PCA - Principes



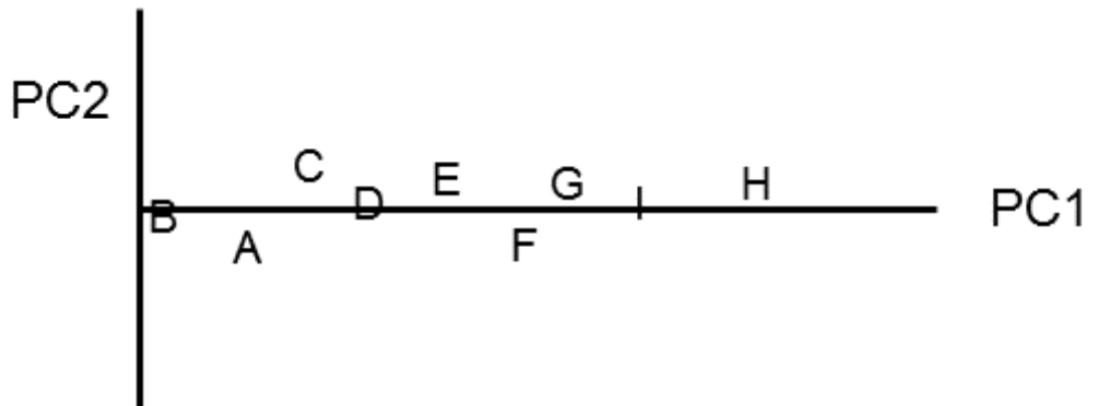
La première composante principale est celle qui maximise la variation observée... la meilleure droite entre les sites

PCA - Principes



La seconde composante principale est ajoutée perpendiculairement au premier axe

PCA - Principles



Le graphique final subit une rotation afin que les deux axes correspondent aux composantes principales (et non plus aux espèces)

PCA - Cas multidimensionnel

- **PC1** --> axe qui maximise la variance des points projetés perpendiculairement sur les axes.
- **PC2** --> doit être orthogonal à PC1, mais sa direction maximise la variance des points projetés.
- **PC3** --> et ainsi de suite : orthogonale à PC1 et PC2...

Quand il y a plus de deux dimensions, la PCA produit un nouvel espace dans lequel tous les axes sont orthogonaux (i.e. la corrélation entre les axes =0) et où les axes sont ordonnés selon le pourcentage de variation des données brutes qu'ils représentent (valeur propre).

PCA - Essayons sur les données Poissons!

- La PCA (tout comme la RDA) est implémentée par la fonction `rda()` de la librairie vegan
- Effectuer une PCA sur les abondances de poissons transformées Hellinger

```
spe.h.pca <- rda(spec.hel)

summary(spe.h.pca)
#
# Call:
# rda(X = spec.hel)
#
# Partitioning of variance:
#                  Inertia Proportion
# Total          0.5025        1
# Unconstrained 0.5025        1
#
# Eigenvalues, and their contribution to the variance
#
# Importance of components:
#                   PC1     PC2     PC3     PC4     PC5     PC6     PC7
# Eigenvalue      0.2580  0.06424  0.04632  0.03850  0.02197  0.01675  0.01472
# Proportion Explained 0.5133  0.12784  0.09218  0.07662  0.04371  0.03334  0.02930
```

Fonction `rda()`

- RDA en 2 étapes :
 - régressions multiples
 - PCA sur les valeurs régressées
- Si on donne seulement un tableau à la fonction `rda()`, la fonction roule une PCA sans faire les régressions

`rda(Y~X) → RDA`

`rda(Y) ou rda(X) → PCA`

PCA - Interprétation des sorties

Partitioning of variance:

	Inertia	Proportion
Total	0.5025	1
Unconstrained	0.5025	1

- Total de variance capturée par les descripteurs (ici les espèces de poissons)
- Dans une PCA, la proportion "Total" et "Unconstrained" de variance capturée est identique

PCA - Interprétation des sorties

	PC1	PC2	PC3	PC4	PC5
Eigenvalue	0.2579605	0.06424089	0.04632294	0.03850244	0.02196526
Proportion Explained	0.5133400	0.12784000	0.09218000	0.07662000	0.04371000
Cumulative Proportion	0.5133400	0.64118000	0.73337000	0.80999000	0.85370000

- Liste des valeurs propres associées à chaque Composantes Principales (ici 27 PCs sont identifiées, soit le nombre de dimensions des données)

La valeur propre est la valeur du changement dans la longueur d'un vecteur, et ici représente la quantité de variation capturée par chaque Composante Principale.

$$0.258 + 0.064 + \dots = 0.5025 \text{ Variance totale capturée}$$

PCA - Interprétation des sorties

Partitioning of variance:

	Inertia	Proportion
--	---------	------------

Total	0.5025	1
Unconstrained	0.5025	1

	PC1	PC2	PC3	PC4	PC5
Eigenvalue	0.2579605	0.06424089	0.04632294	0.03850244	0.02196526
Proportion Explained	0.5133400	0.12784000	0.09218000	0.07662000	0.04371000
Cumulative Proportion	0.5133400	0.64118000	0.73337000	0.80999000	0.85370000

- Liste de la proportion de variance capturée par chaque Composante Principale (et la proportion cumulée)

51.3% de 0.5025 égal 0.258

PCA - Interprétation des sorties

Scaling 2 for species and site scores

- * Species are scaled proportional to eigenvalues
- * Sites are unscaled: weighted dispersion equal on all dimensions
- * General scaling constant of scores: 1.93676

- Il existe deux façons principales de représenter une ordination en 2D... ici la sortie R nous informe que le cadrage utilisé par défaut est de type 2...

À suivre!

PCA - Interprétation des sorties

Species scores

	PC1	PC2	PC3	PC4	PC5	PC6
CHA	0.17336	0.08295	-0.064963	0.2539861	-0.0285801	0.019057
TRU	0.64860	0.01162	-0.261994	-0.1606020	-0.0745819	-0.088616
VAI	0.51810	0.14773	0.165304	0.0241017	0.1012928	0.104748
LOC	0.38606	0.16615	0.242995	-0.0275216	0.1258011	0.048299
OMB	0.16893	0.06274	-0.096143	0.2426514	0.0140574	0.062117

- *Species* réfère aux colonnes de votre jeu de données, ici différentes espèces de poissons
- Les scores correspondent aux coordonnées de chaque espèce le long de chaque PC

PCA - Interprétation des sorties

site scores (weighted sums of species scores)

	PC1	PC2	PC3	PC4	PC5	PC6
1	0.367401	-0.39935	-1.08857	-0.63304	-0.512027	-0.858378
2	0.503582	-0.05683	-0.19259	-0.43441	0.389533	0.069451
3	0.461709	0.02262	-0.06522	-0.49798	0.309425	0.270577
4	0.298336	0.15130	0.26748	-0.53196	0.003088	0.184821
5	-0.002222	0.07631	0.54769	-0.50936	-0.780261	-0.169353
6	0.212816	0.08345	0.55091	-0.42210	-0.139518	-0.104278

- Site réfère aux lignes de votre jeu de données, ici différentes stations d'échantillonnage le long de la rivière,
- Les scores correspondent aux coordonnées de chaque site le long de chaque PC

Accéder à une partie de la sortie R

La sortie R est très dense, mais vous pouvez accéder au besoin à des informations spécifiques. Par exemple, vous pouvez extraire les valeurs propres et leur contribution à la variance capturée :

```
summary(spe.h.pca, display = NULL)
#
# Call:
# rda(X = spec.hel)
#
# Partitioning of variance:
#                  Inertia Proportion
# Total          0.5025      1
# Unconstrained 0.5025      1
#
# Eigenvalues, and their contribution to the variance
#
# Importance of components:
#                 PC1     PC2     PC3     PC4     PC5     PC6     PC7
# Eigenvalue    0.2580  0.06424  0.04632  0.03850  0.02197  0.01675  0.01472
# Proportion Explained 0.5133  0.12784  0.09218  0.07662  0.04371  0.03334  0.02930
# Cumulative Proportion 0.5133  0.64118  0.73337  0.80999  0.85370  0.88704  0.91634
#                 PC8     PC9     PC10    PC11    PC12    PC13
# Eigenvalue   0.01156  0.006936  0.006019  0.004412  0.002982  0.002713
```

Accéder à une partie de la sortie R

Vous pouvez calculer les valeurs propres :

```
eigen(cov(spec.hel))  
# eigen() decomposition  
# $values  
# [1] 2.579605e-01 6.424089e-02 4.632294e-02 3.850244e-02 2.196526e-02  
# [6] 1.675463e-02 1.472430e-02 1.155759e-02 6.936149e-03 6.019271e-03  
# [11] 4.412388e-03 2.982309e-03 2.713021e-03 1.834874e-03 1.454670e-03  
# [16] 1.117858e-03 8.308832e-04 5.415301e-04 4.755244e-04 3.680458e-04  
# [21] 2.765106e-04 2.252760e-04 1.429425e-04 7.618319e-05 4.989831e-05  
# [26] 1.525627e-05 9.117507e-06  
  
#  
# $vectors  
# [,1] [,2] [,3] [,4] [,5]  
# [1,] -0.12492725 -0.11979088 -0.11047444 0.4737644443 0.070581708  
# [2,] -0.46740781 -0.01678206 -0.44554311 -0.2995735541 0.184188349  
# [3,] -0.37336215 -0.21333150 0.28111355 0.0449572376 -0.250153773  
# [4,] -0.27821421 -0.23994030 0.41323337 -0.0513364598 -0.310679865  
# [5,] -0.12173642 -0.09059800 -0.16349912 0.4526216196 -0.034716207  
# [6,] -0.05610722 -0.21147318 -0.05340233 0.4363710457 0.254947303  
# [7,] 0.13325245 -0.07077305 -0.07670861 0.0371901204 -0.169737613  
# [8,] 0.10553143 -0.25754282 -0.01860002 0.1212372044 0.001538404  
# [9,] 0.08240964 -0.22633305 0.24186207 -0.0237391101 0.491305413  
# [10,] 0.06977391 0.22309480 0.41314626 0.2097008258 -0.057746857
```

Accéder à une partie de la sortie R

Vous pouvez extraire les scores (des sites ou des espèces) pour réaliser des graphiques ou les utiliser dans de nouvelles analyses :

- Extraire le score des espèces pour PC1 et PC2 :

```
spe.scores <- scores(spe.h.pca,  
                      display = "species",  
                      choices = c(1,2))
```

- Extraire le score des sites pour PC1 et PC2 :

```
site.scores <- scores(spe.h.pca,  
                      display = "sites",  
                      choices = c(1,2))
```

Sélection des PCs significatives

- La force de la PCA est de condenser la variance contenue dans un grand jeu de données en jeu de données synthétiques moins nombreuses
- Dans notre cas, 27 PCs sont identifiées mais seules les premières contribuent de façon importante à la variance capturée tandis que les autres représentent le bruit des données et peuvent être écartées...

Comment choisir les PCs les plus importantes?

Critère de Kaiser - Guttman

Sélectionner les PCs qui capturent plus de variance que la moyenne de tous les PCs

- Extraire les valeurs propres de chaque PCs :

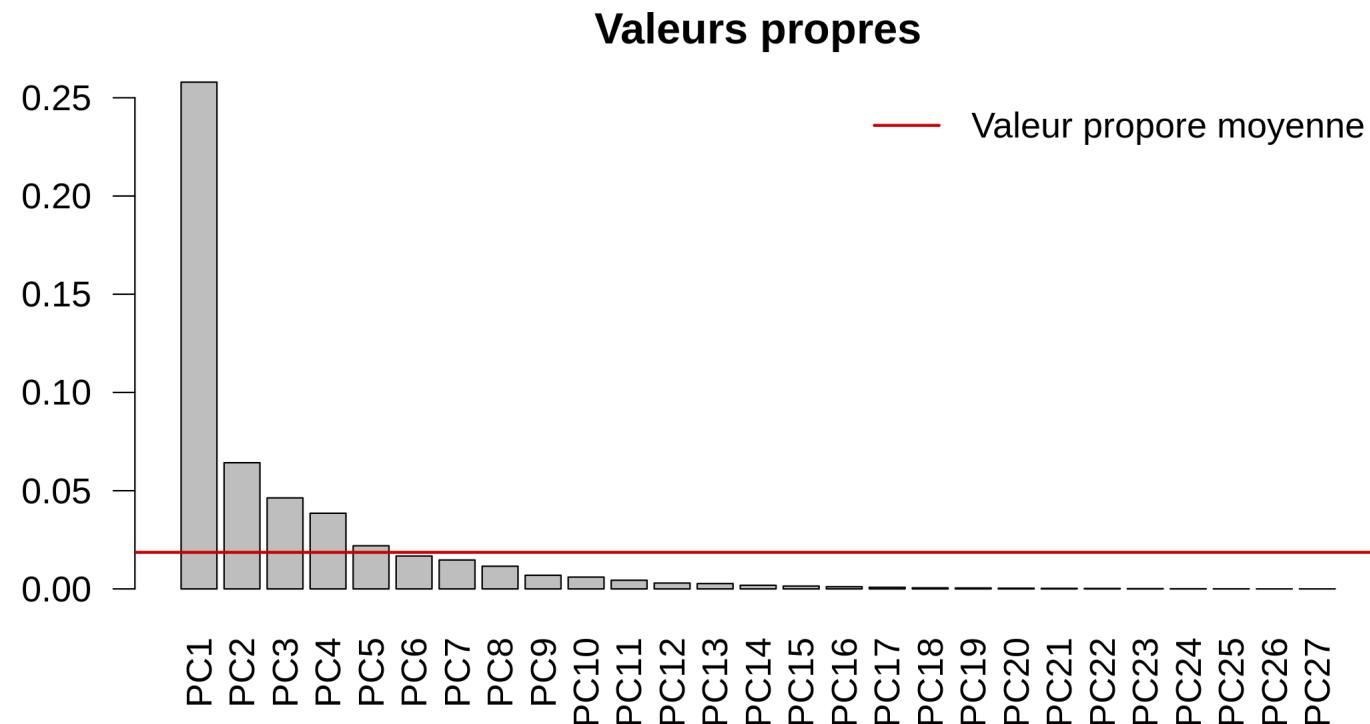
```
ev <- spe.h.pca$CA$eig
```

- Sélectionner les valeurs propres supérieures à la moyenne :

```
ev[ev>mean(ev)]  
#          PC1          PC2          PC3          PC4          PC5  
# 0.25796049 0.06424089 0.04632294 0.03850244 0.02196526
```

Critère de Kaiser - Guttman (illustration)

```
n <- length(ev)
barplot(ev, main = "Valeurs propres", col = "grey", las = 2)
abline(h = mean(ev), col = "red3", lwd = 2)
legend("topright", "Valeur propre moyenne",
       lwd = 2, col = "red3" , bty = "n")
```



PCA - variables environnementales

Une PCA peut aussi être effectuée sur les variables environnementales standardisées pour comparer les sites ou évaluer les corrélations entre variables...

- Effectuer une PCA sur les variables environnementales standardisées

```
env.pca <- rda(env.z)
summary(env.pca, scaling = 2) # default
#
# Call:
# rda(X = env.z)
#
# Partitioning of variance:
#           Inertia Proportion
# Total          11      1
# Unconstrained 11      1
#
# Eigenvalues, and their contribution to the variance
#
# Importance of components:
#           PC1    PC2    PC3    PC4    PC5    PC6    PC7
# Eigenvalue   6.0980 2.1671 1.03760 0.70351 0.35185 0.31913 0.16455
# Proportion Explained 0.5544 0.1970 0.09433 0.06396 0.03199 0.02901 0.01496
# Cumulative Proportion 0.5544 0.7514 0.84570 0.90966 0.94164 0.97066 0.98561
```

PCA - variables environnementales

- Extraire les valeurs propres de chaque PC :

```
ev <- env.pca$CA$eig
```

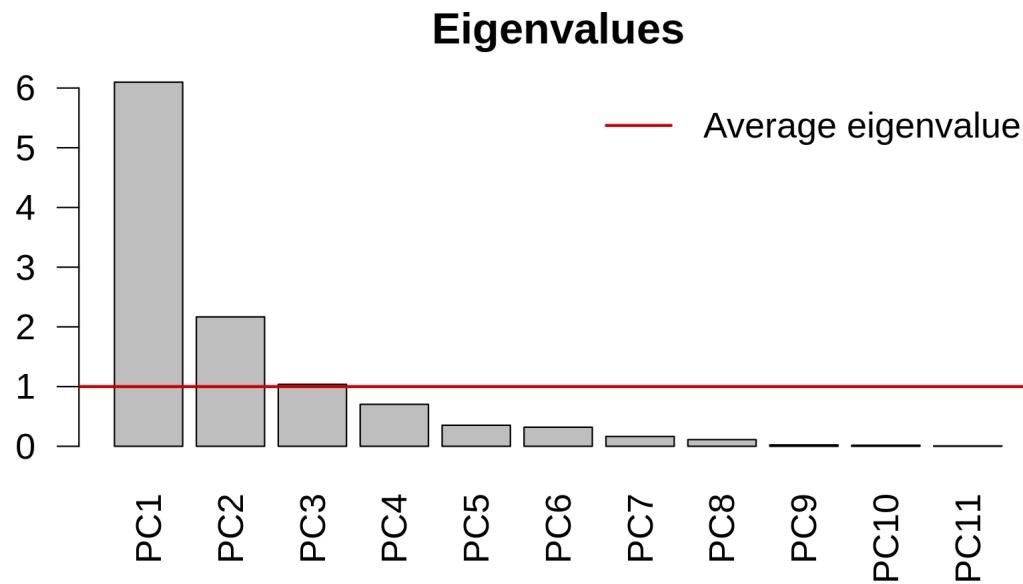
- Sélectionner les valeurs propres supérieures à la moyenne :

```
ev[ev>mean(ev)]  
#      PC1      PC2      PC3  
# 6.097995 2.167126 1.037603
```

PCA - variables environnementales

- Créer le graphique :

```
n <- length(ev)
barplot(ev, main = "Eigenvalues", col = "grey", las = 2)
abline(h = mean(ev), col = "red3", lwd = 2)
legend("topright", "Average eigenvalue",
      lwd = 2, col = "red3" , bty = "n")
```

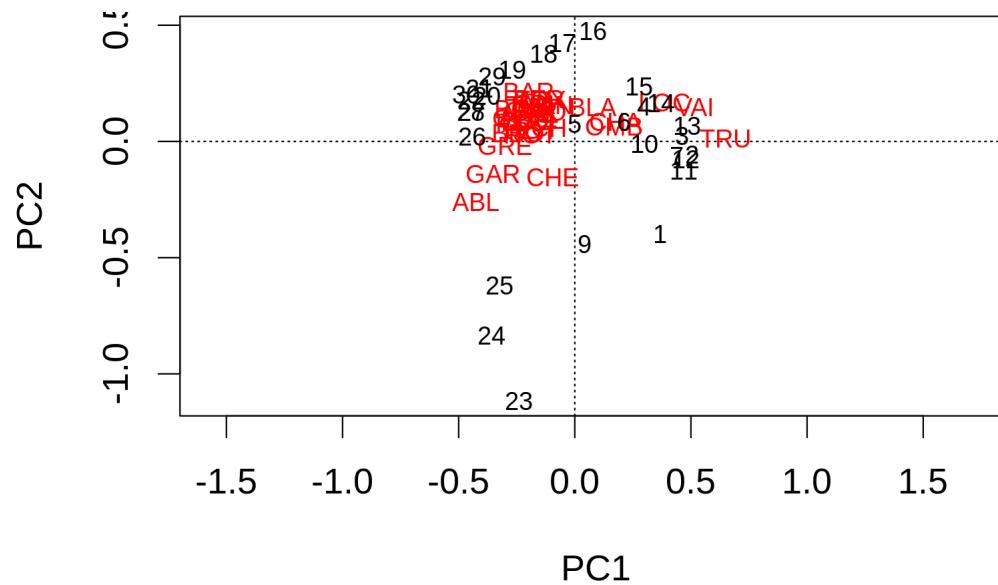


PCA - Illustration graphique

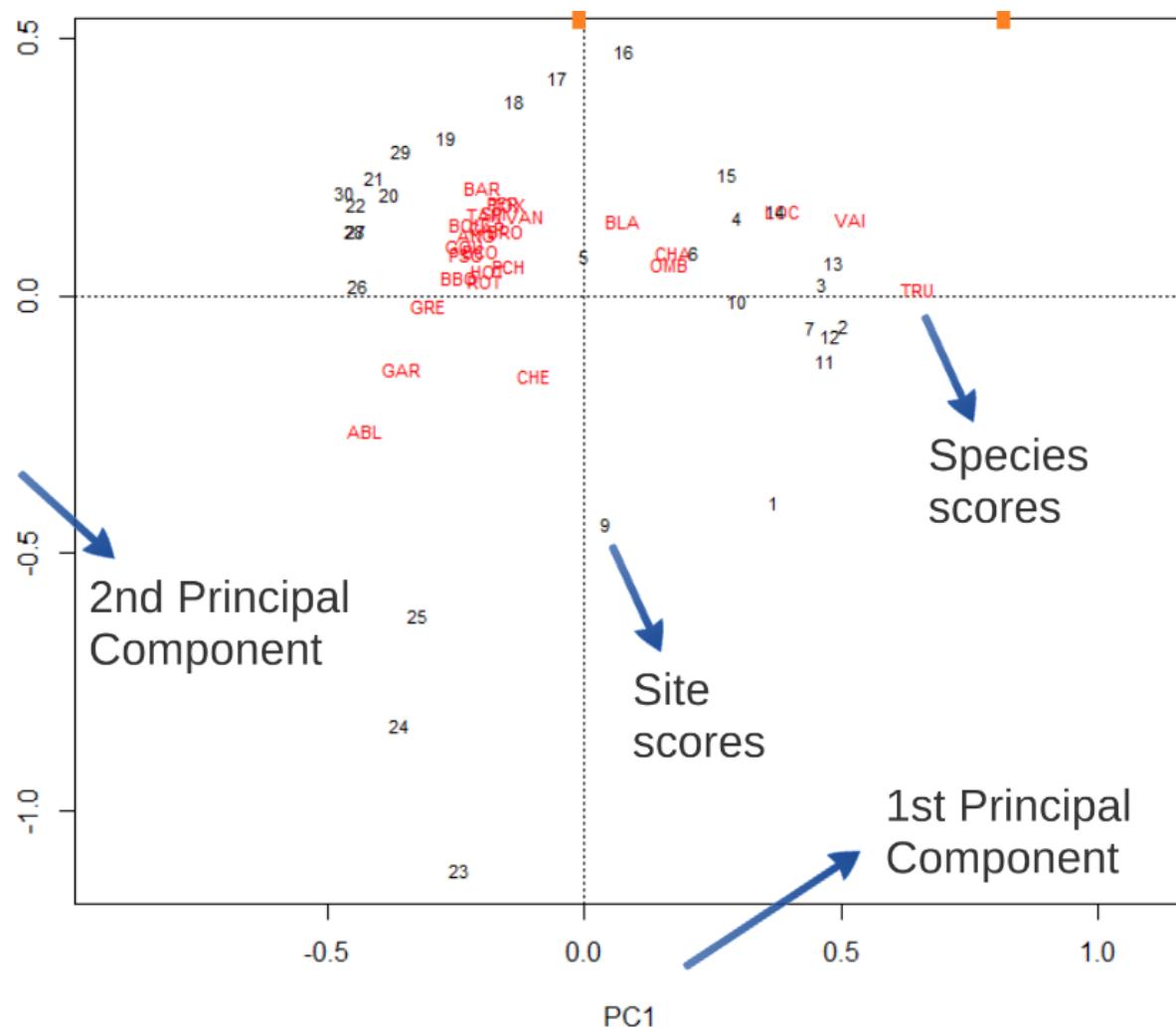
L'abondante information produite par une PCA est plus facile à comprendre et à interpréter à l'aide de biplots permettant de visualiser les patrons présents dans les données.

- Un biplot peut être rapidement créé via la fonction `plot()`

```
plot(spe.h.pca)
```



Biplot de PCA avec plot()

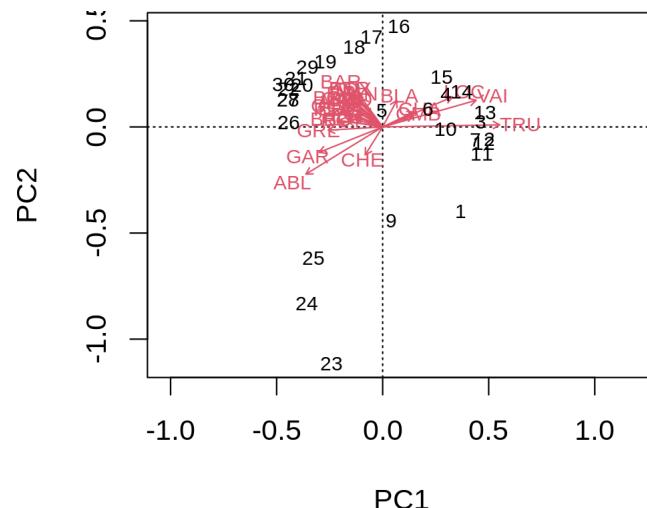


`plot()` est rapide mais il est difficile d'interpréter les angles entre espèces

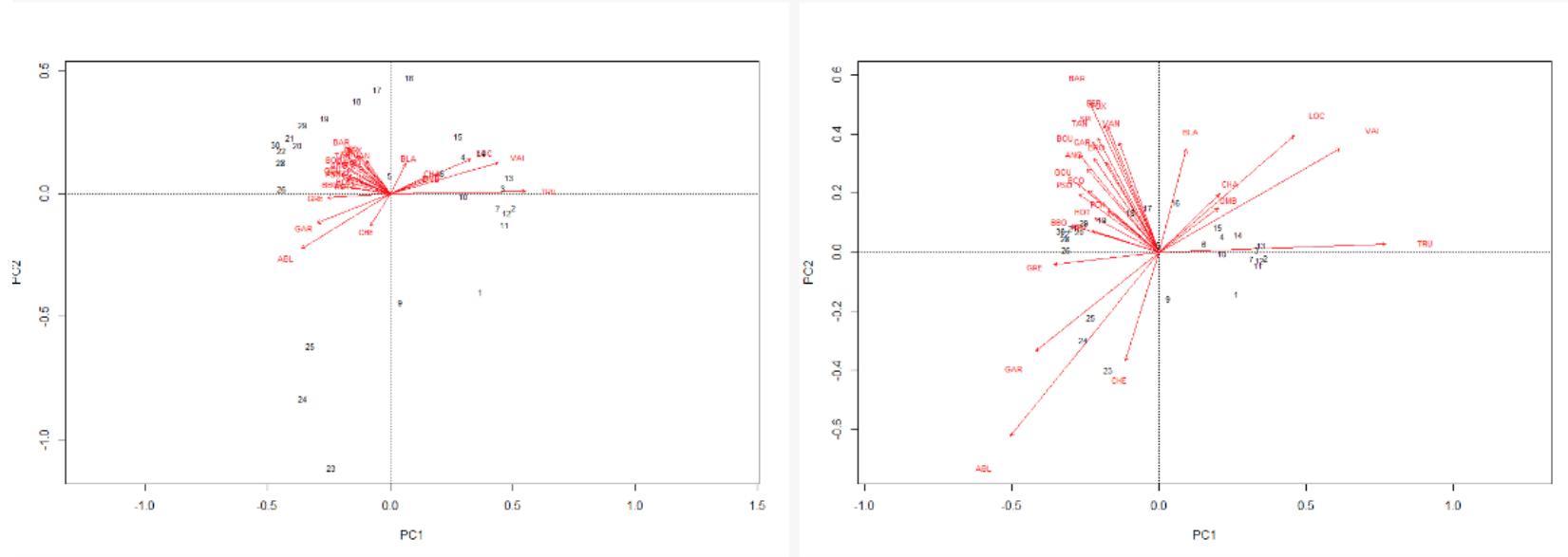
Basique biplot() de PCA

- Avec la fonction `biplot()` de base R, des flèches sont tracées pour montrer les directions et les angles entre descripteurs dans l'ordination
- **Descripteurs distants de 180 degrés : corrélation négative**
- **Descripteurs distants de 90 degrés : pas de corrélation**
- **Descripteurs distants de 0 degré : corrélation positive**

```
biplot(spe.h.pca)
```



Type de scaling



Cadrage 2 (DEFAULT): les distances entre objets ne sont pas des approximations de leurs distances euclidiennes, mais les angles entre descripteurs reflètent leurs corrélations.

Meilleur cadrage pour interpréter les relations entre descripteurs (espèces) !

Cadrage 1 : préserve au maximum la distance euclidienne (dans l'espace d'ordination) entre objets (ex. sites); les angles entre descripteurs (ex. espèces) ne sont pas informatifs.

Meilleur cadrage pour interpréter les relations entre objects (sites)!

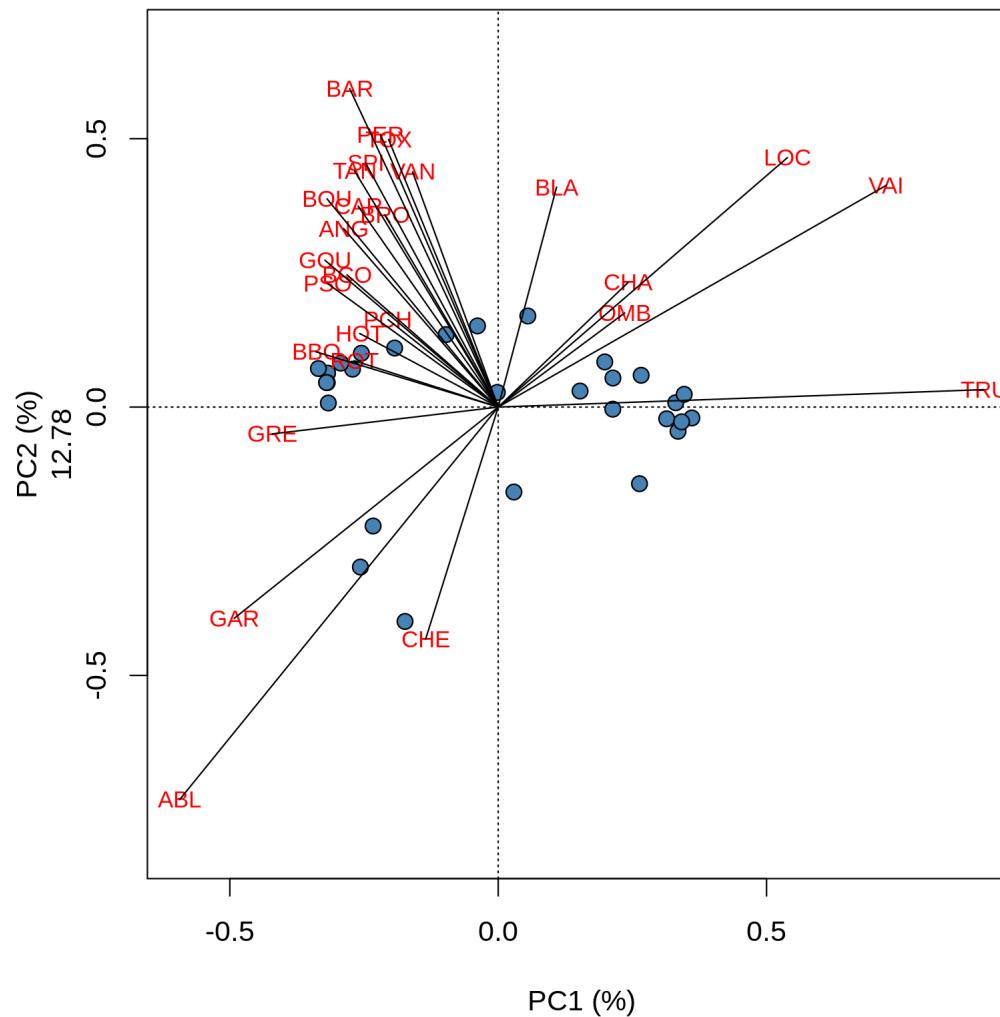
"Biplot" avancés

- En extrayant certaines parties de la sortie R, il est possible de créer des biplots plus détaillés et clairs :

```
plot(spe.h.pca, scaling = 1, type = "none",
      xlab = c("PC1 (%)", round(spe.h.pca$CA$eig[1]/sum(spe.h.pca$CA$eig)*100,2)),
      ylab = c("PC2 (%)", round(spe.h.pca$CA$eig[2]/sum(spe.h.pca$CA$eig)*100,2)))
points(scores(spe.h.pca, display = "sites", choices = c(1,2), scaling = 1),
       pch=21, col = "black", bg = "steelblue" , cex = 1.2)
text(scores(spe.h.pca, display = "species", choices = 1, scaling = 1),
      scores(spe.h.pca, display = "species", choices = 2, scaling = 1),
      labels = rownames(scores(spe.h.pca, display = "species", scaling = 1)),
      col = "red", cex = 0.8)
spe.cs <- scores(spe.h.pca, choices = 1:2, scaling = 1 , display = "sp")
arrows(0, 0, spe.cs[,1], spe.cs[,2], length = 0)
```

voir la fonction `arrows()` de `graphics` pour ajouter des vecteurs

"Biplot" avancés



Autres options graphiques : ggvegan

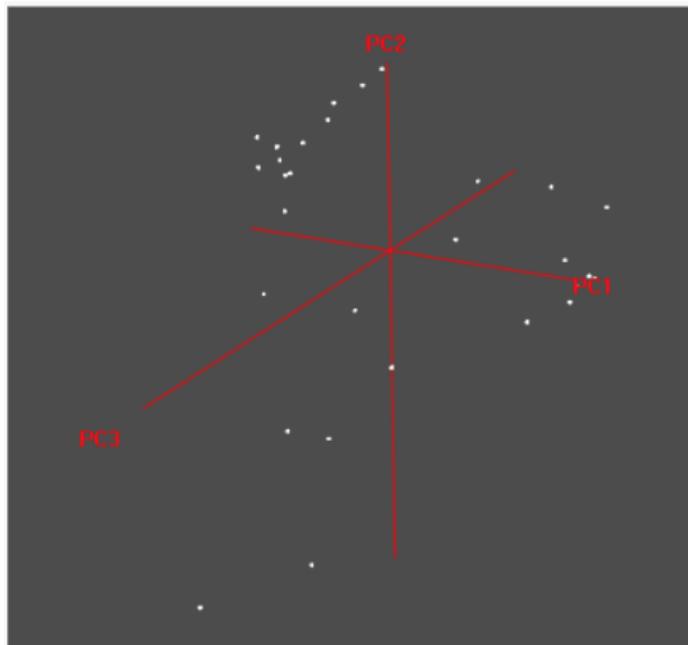
- Un ensemble d'outils pour créer des biplots avec ggplot2 :

```
install.packages("devtools")
require("devtools")
install_github("ggvegan", "gavinsimpson")
require("ggvegan")
autoplot()
```

Autres options graphiques : rgl et vegan3d

Biplot interactif avec rgl

```
require(rgl)
require(vegan3d)
ordirgl(spe.h.pca)
```





Défi # 3

Exécuter une PCA sur les données d'abondance d'acariens :

```
data(mite)
```

- Quels sont les axes significatifs?
- Quels groupes de sites pouvez-vous identifier?
- Quelles espèces caractérisent ces groupes de sites?

Solution #3

- Transformation Hellinger des données :

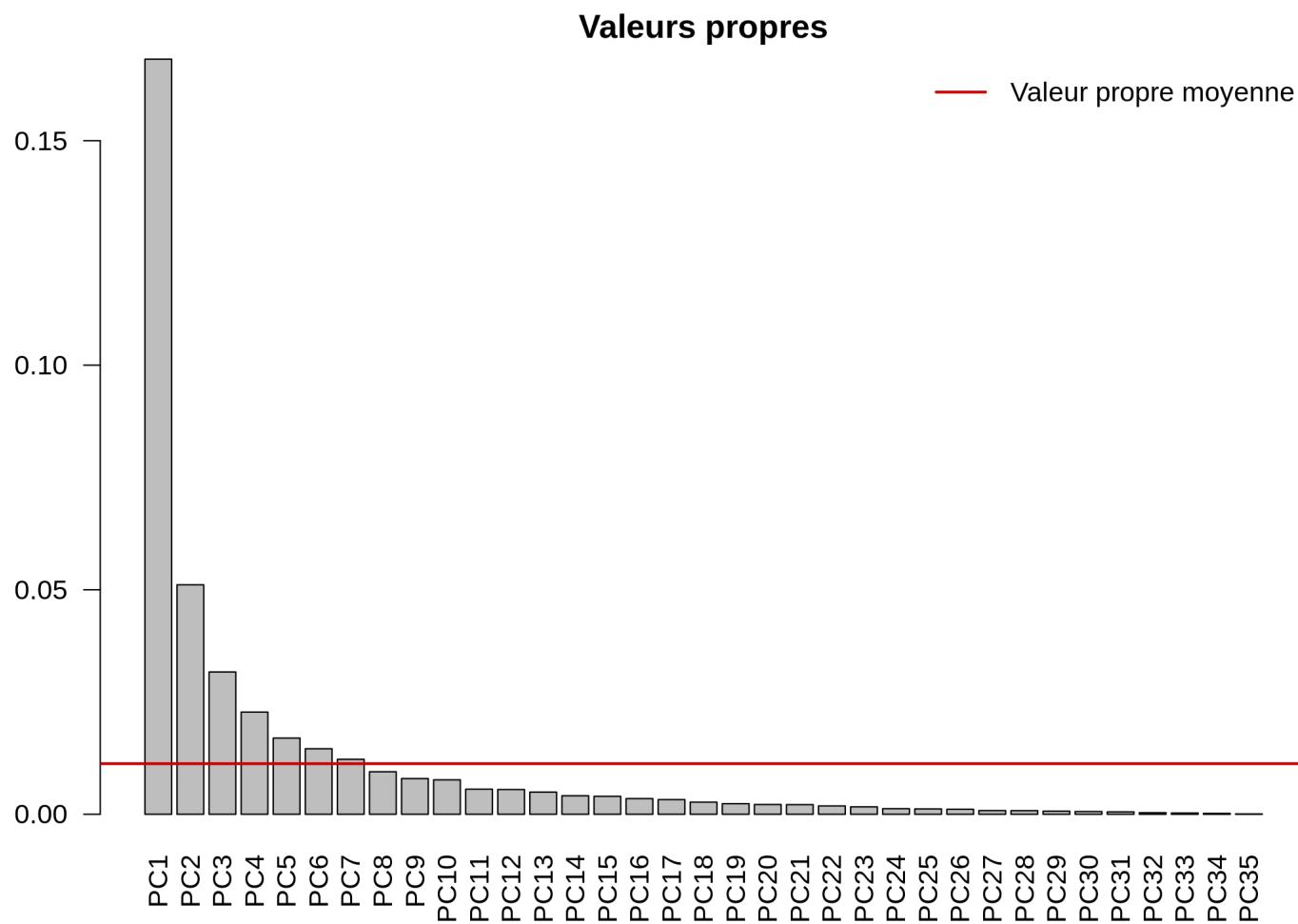
```
mite.spe.hel <- decostand(mite, method = "hellinger")
```

```
mite.spe.h.pca <- rda(mite.spe.hel)
```

- Recherche des axes significatifs par critère de Gutman-Kaiser :

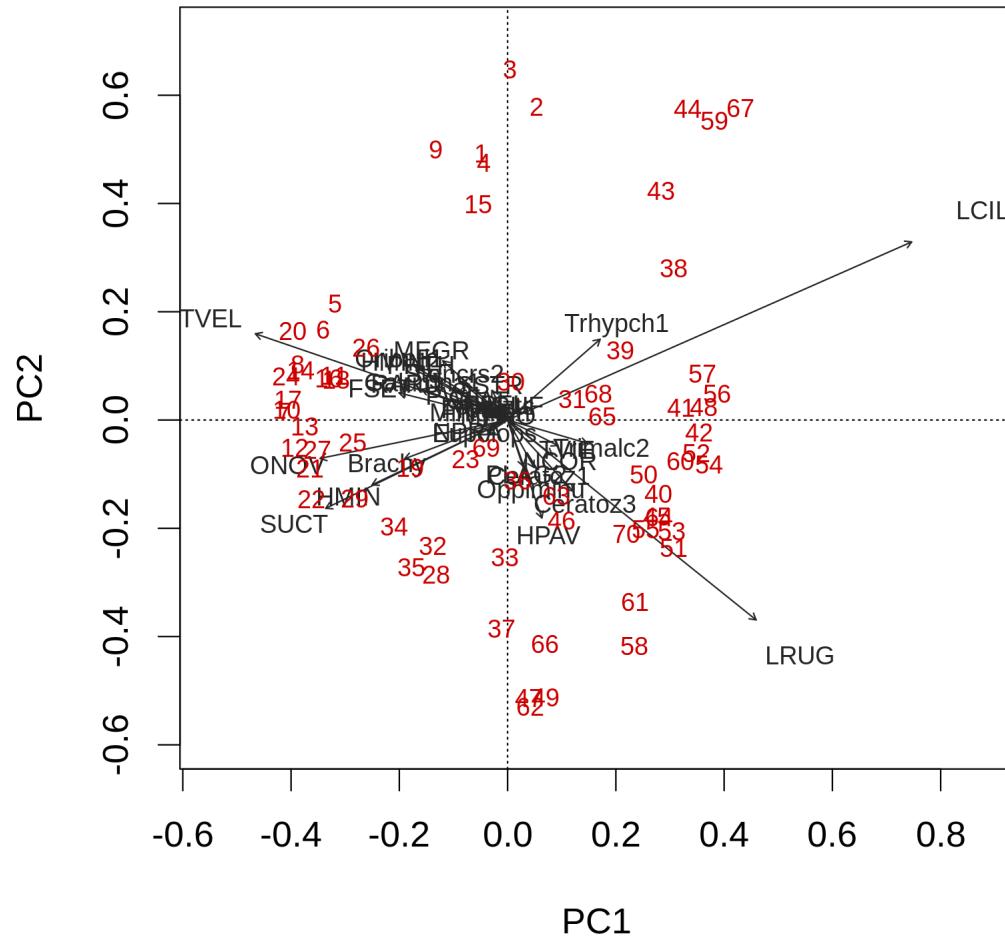
```
ev <- mite.spe.h.pca$CA$eig
ev[ev > mean(ev)]
n <- length(ev)
barplot(ev, main = "Valeurs propres", col = "grey", las = 2)
abline(h = mean(ev), col = "red3", lwd = 2)
legend("topright", "Valeur propre moyenne", lwd = 2, col = "red3", bty = "n")
```

Solution #3



Solution #3

```
biplot(mite.spe.h.pca, col = c("red3", "grey15"))
```



Attention

- La PCA est une méthode linéaire basée sur quelques hypothèses clefs :
 - distribution multinormale des données (seulement pour faire des inférences)
 - nombre limité de zéros
 - le gradient d'intérêt doit causer la majorité de la variance dans le jeu de données

Le non-respect de ces hypothèses peut causer une forme de fer à cheval sur les biplots (*horseshoe effect*), sur lesquels les extrémités du fer à cheval sont proches mais représentent en réalité des conditions opposées du gradient

Attention

- Certains de ces problèmes peuvent être réglés en utilisant des transformations appropriées des données avant d'effectuer une PCA
- Dans certains cas, tels que les études couvrant de longs gradients environnementaux, il est préférable d'utiliser d'autres méthodes d'ordination non-constraines (ex. CA)

4.1. Analyse des Correspondances (CA)

Distances euclidiennes vs distances de Chi²

- La PCA préserves les **distances euclidiennes** entre objets et postule une relation linéaire entre descripteurs
- ...mais dans certains cas (ex. gradients longs), **les espèces présentent une réponse unimodale** aux gradients environnementaux

Principes de la CA

- Dans de tels cas, la CA devrait être préférée à la PCA car elle préserve les **distances de Chi2 entre objets**... et représente donc mieux les réponses unimodales

Comment effectuer une CA?

- La CA est implémentée dans la librairie `vegan` par la fonction `cca()`:

```
spe.ca <- cca(spe[ -8, ])  
# prend seulement les colonnes dont rowsums est > à 0.
```

- CA effectuée sur les abondances de poissons

CA: sortie de R

- Les résultats d'une CA sont présentés de la même manière qu'une PCA et peuvent être appelés par :

```
summary(spe.ca)
#
# Call:
# cca(X = spe[-8, ])
#
# Partitioning of scaled Chi-square:
#           Inertia Proportion
# Total      1.128      1
# Unconstrained 1.128      1
#
# Eigenvalues, and their contribution to the scaled Chi-square
#
# Importance of components:
#           CA1     CA2     CA3     CA4     CA5     CA6     CA7
# Eigenvalue 0.6062 0.1423 0.10251 0.07319 0.04912 0.03909 0.03341
# Proportion Explained 0.5374 0.1262 0.09087 0.06488 0.04354 0.03465 0.02962
# Cumulative Proportion 0.5374 0.6635 0.75437 0.81925 0.86279 0.89745 0.92706
#           CA8     CA9     CA10    CA11    CA12    CA13
# Eigenvalue 0.01709 0.01302 0.010765 0.008141 0.007533 0.005820
# Proportion Explained 0.01515 0.01154 0.009543 0.007217 0.006678 0.005159
# Cumulative Proportion 0.94221 0.95375 0.963294 0.970511 0.977188 0.982347
```

CA: Interprétation des résultats

```
Call:  
cca(X = spe)
```

```
Partitioning of mean squared contingency coefficient:
```

	Inertia	Proportion
Total	1.167	1
Unconstrained	1.167	1

```
Eigenvalues, and their contribution to the mean squared contingency coefficient
```

```
Importance of components:
```

	CA1	CA2	CA3	CA4	CA5	CA6	CA7	CA8
Eigenvalue	0.601	0.1444	0.10729	0.08337	0.05158	0.04185	0.03389	0.02883
Proportion Explained	0.515	0.1237	0.09195	0.07145	0.04420	0.03586	0.02904	0.02470
Cumulative Proportion	0.515	0.6388	0.73069	0.80214	0.84634	0.88220	0.91124	0.93594
	CA9	CA10	CA11	CA12	CA13	CA14	CA15	
Eigenvalue	0.01684	0.01083	0.01014	0.007886	0.006123	0.004867	0.004606	
Proportion Explained	0.01443	0.00928	0.00869	0.006760	0.005250	0.004170	0.003950	
Cumulative Proportion	0.95038	0.95965	0.96835	0.975100	0.980350	0.984520	0.988470	
	CA16	CA17	CA18	CA19	CA20	CA21		
Eigenvalue	0.003844	0.003067	0.001823	0.001642	0.001295	0.0008775		
Proportion Explained	0.003290	0.002630	0.001560	0.001410	0.001110	0.0007500		
Cumulative Proportion	0.991760	0.994390	0.995950	0.997360	0.998470	0.9992200		
	CA22	CA23	CA24	CA25	CA26			
Eigenvalue	0.0004217	0.0002149	0.0001528	8.949e-05	2.695e-05			
Proportion Explained	0.0003600	0.0001800	0.0001300	8.000e-05	2.000e-05			
Cumulative Proportion	0.9995900	0.9997700	0.9999000	1.000e+00	1.000e+00			

```
Scaling 2 for species and site scores
```

```
* Species are scaled proportional to eigenvalues
```

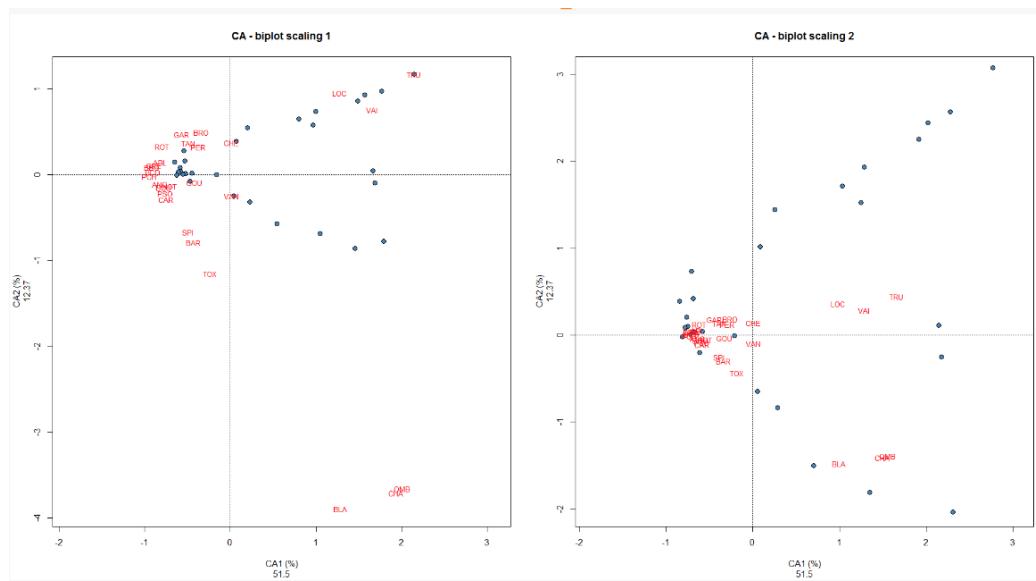
```
* Sites are unscaled: weighted dispersion equal on all dimensions
```

26 axes CA identifiés

% CA1 = 51.50%

% CA2 = 12.37%

CA: biplots



Ces biplots montrent qu'un groupe de sites (à gauche) possède des communautés similaires de poissons caractérisées par de nombreuses espèces dont *GAR*, *TAN*, *PER*, *ROT*, *PSO* et *CAR*

Dans le coin supérieur droit, un second groupe de sites se caractérise par les espèces *LOC*, *VAI* et *TRU*

Le dernier groupe de sites dans le coin inférieur droit montre des communautés abondantes en *BLA*, *CHA* et *OMB*



Défi #4

Exécuter une CA sur les données d'abondance des espèces d'acariens

```
mite.spe <- mite
```

- Quels sont les axes importants?
- Quels groupes de sites se distinguent?
- Quelles espèces caractérisent chaque groupe de site?

Solution #4

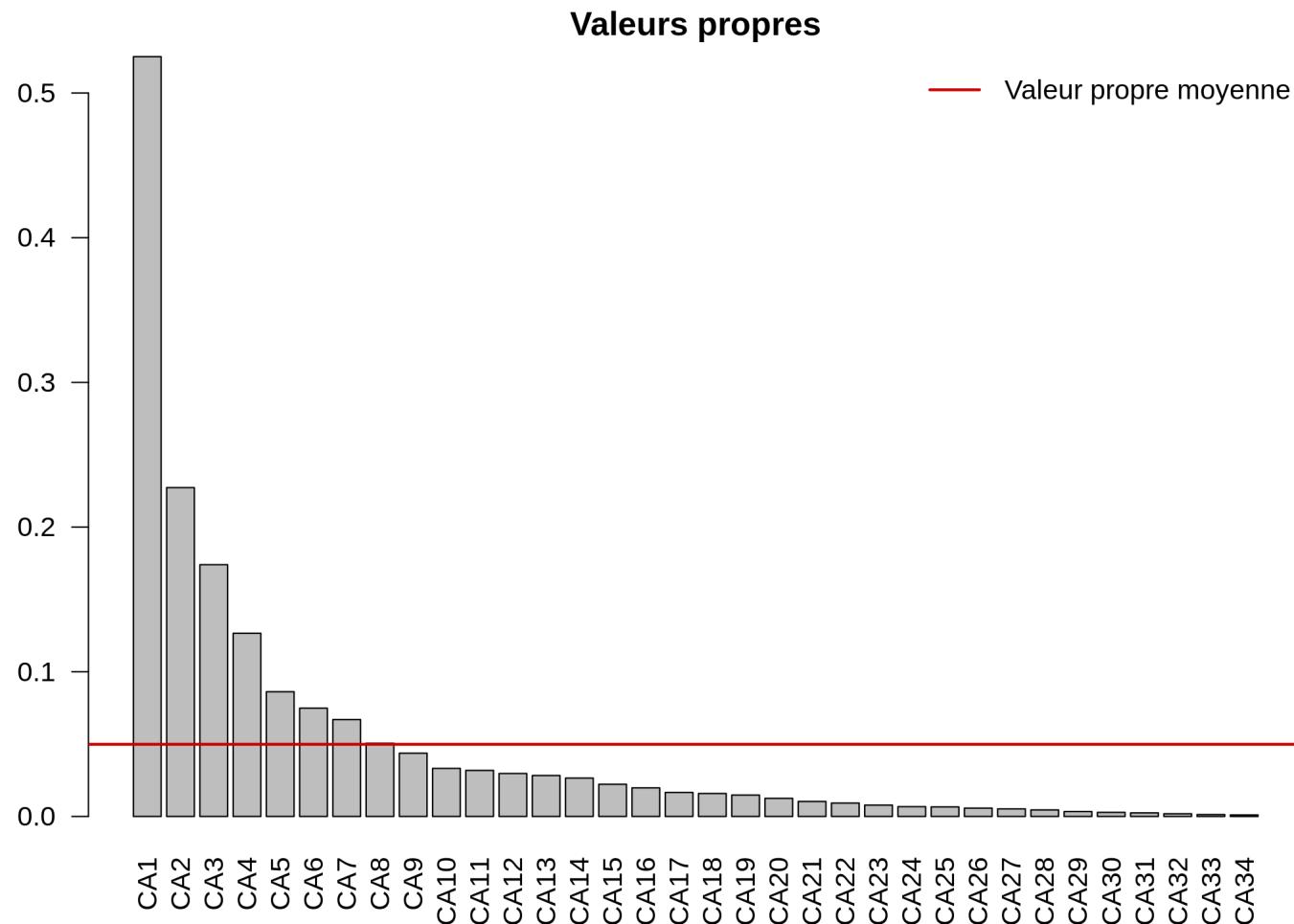
- Calcule de la CA:

```
mite.spe.ca <- cca(mite.spe)
```

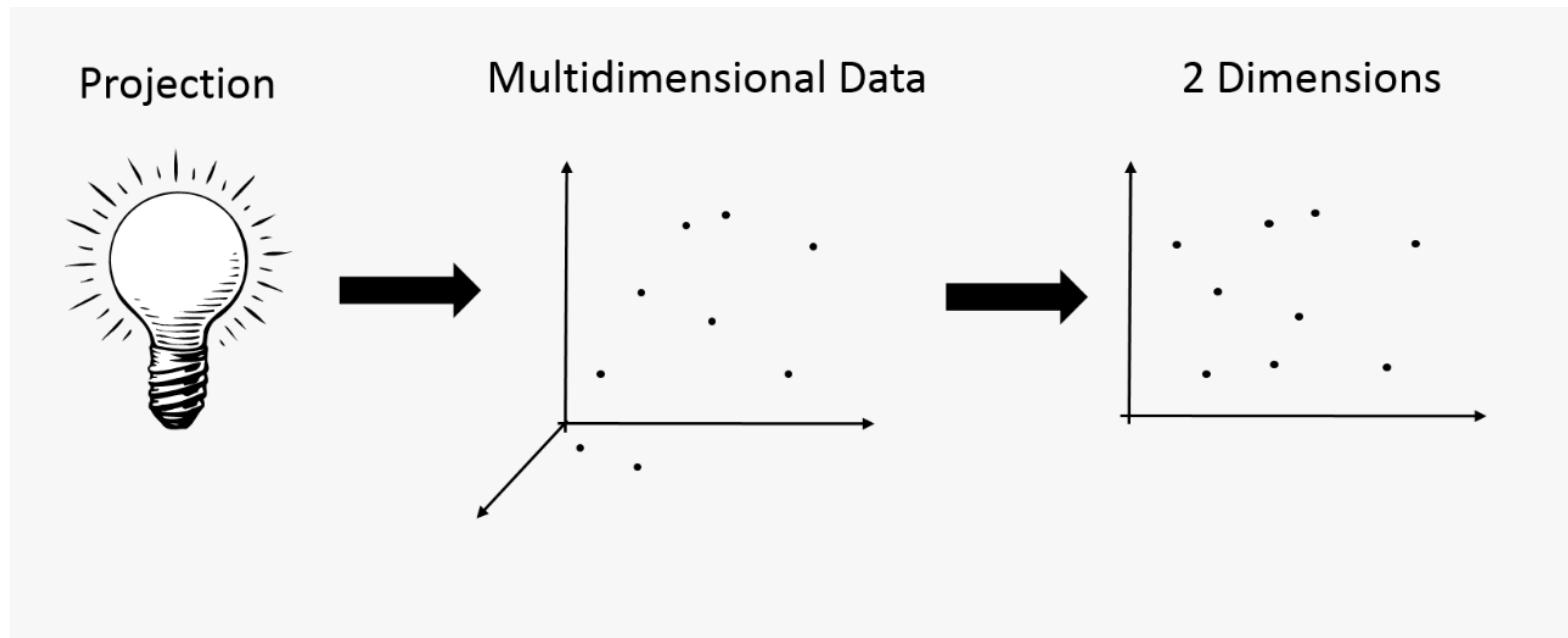
- Recherche des axes significatifs par critère de Guttman-Kaiser :

```
ev <- mite.spe.ca$CA$eig
ev[ev > mean(ev)]
n <- length(ev)
barplot(ev, main = "Valeurs propres", col = "grey", las = 2)
abline(h = mean(ev), col = "red3", lwd = 2)
legend("topright", "Valeur propre moyenne", lwd = 2, col = red3, bty = "n")
```

Solution #4



4.3. Analyse en Coordonnées Principales



- En PCA, le maximum de la variation des données est préservée
- En PCoA, la distance entre objets est préservée autant que possible dans un espace multidimensionnel

La PCoA est particulièrement consilie pour des jeux de données plus larges que longs (problème typique en génétique)

PCoA - Essayons avec les données Poissons!

- Les fonctions `cmdscale()` et `pcoa()`, des librairies **stats** et **ape** permettent d'effectuer une PCoA:

```
?cmdscale  
library(ape)  
?pcoa
```

- Effectuer une PCoA sur les abondances de poissons transformées Hellinger

```
spe.h.pcoa <- pcoa(dist(spec.hel))  
summary(spe.h.pcoa)  
#           Length Class    Mode  
# correction     2   -none- character  
# note          1   -none- character  
# values         5   data.frame list  
# vectors       783   -none- numeric  
# trace          1   -none- numeric
```

PCoA - Interprétation des sorties R

```
$values
```

	Eigenvalues	Relative_eig	Broken_stick	Cumul_eig	Cumul_br_stick
1	7.2228938501	5.133437e-01	0.144128028	0.5133437	0.1441280
2	1.7987448715	1.278400e-01	0.107090991	0.6411837	0.2512190
3	1.2970422885	9.218307e-02	0.088572472	0.7333668	0.3397915
4	1.0780684157	7.662021e-02	0.076226793	0.8099870	0.4160183
5	0.6150272794	4.371107e-02	0.066967534	0.8536980	0.4829858

- Valeurs propres
- Valeurs propres relatives
- Modèle broken stick: évalue les axes significatifs
- Valeurs propres cumulées: cumul des valeurs propres relatives
- Broken stick cumulés: cumul des valeurs du modèle broken stick

PCoA - Interprétation des sorties R

	Axis.1	Axis.2	Axis.3	Axis.4	Axis.5	Axis.6
1	-0.509824403	-0.276543720	0.64011383	-0.339373399	0.207330880	0.303563377
2	-0.698794880	-0.039355856	0.11324989	-0.232885899	-0.157730682	-0.024561130
3	-0.640690642	0.015667069	0.03835044	-0.266970577	-0.125293094	-0.095688892
4	-0.413985947	0.104770836	-0.15728486	-0.285182806	-0.001250382	-0.065361378
5	0.003083242	0.052843104	-0.32206098	-0.273069271	0.315944703	0.059891284
6	-0.295314224	0.057788054	-0.32395301	-0.226290236	0.056493824	0.036877698

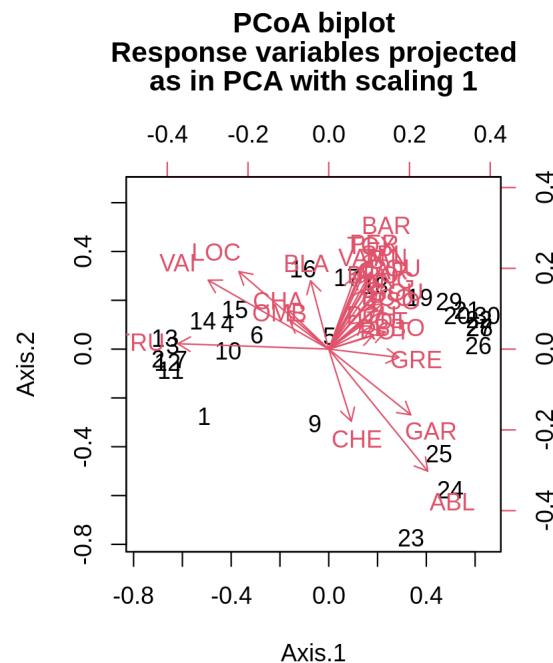
-Vecteurs : Vecteurs propres associés à chaque valeur propre contenant les coordonnées de chaque site dans l'espace euclidien.

Ce sont les résultats les plus utiles pour des analyses subséquentes puisqu'ils capturent fidèlement la distance entre objets.

Biplot de PCoA avec `biplot.pcoa()`

La fonction `biplot.pcoa()` permet de visualiser en 2D les distances entre sites, et les espèces associées à chaque site

```
biplot.pcoa(spe.h.pcoa, spec.hel)
```



PCoA et distances non-métriques

- La PCoA peut aussi être utilisée pour capturer de l'information à partir de distance non-métriques, telles que la distance de Bray-Curtis. Essayons :

```
spe.bray.pcoa <- pcoa(spe.db.pa)
# spe.bray.pcoa
```

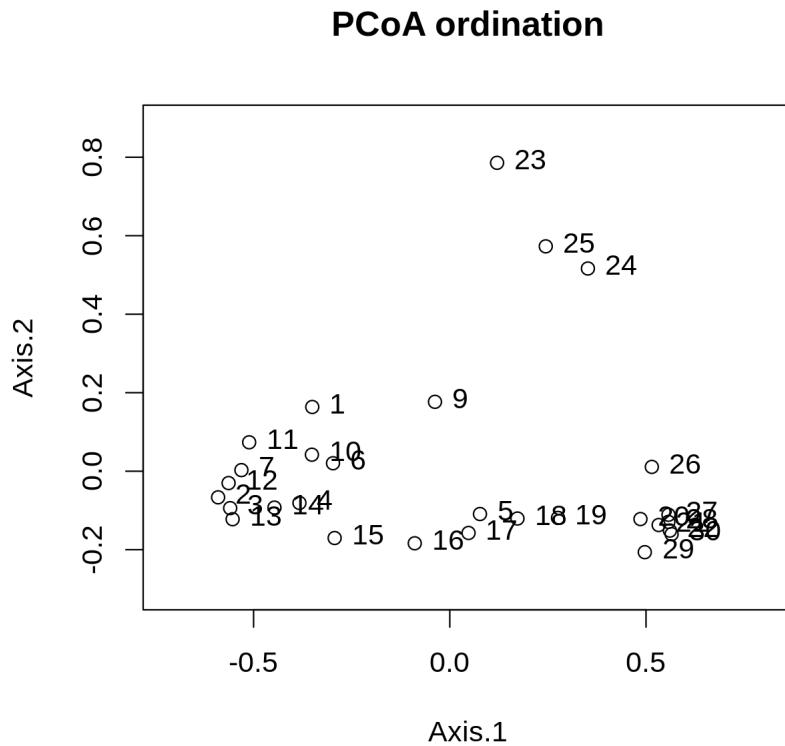
- Observer la sortie R et noter la présence de valeurs propres négatives. Elles sont liées à l'impossibilité de représenter des distances non-métriques dans un espace euclidien sans corrections (*voir Legendre & Legendre 2012*) :

```
spe.bray.pcoa <- pcoa(spe.db.pa, correction = "cailliez")
# spe.bray.pcoa
```

PCoA et distances non-métriques

- Construisons maintenant le biplot (sans espèces)

```
biplot.pcoa(spe.bray.pcoa)
```



Défi #5



Exécuter une PCoA sur les données d'abondances des espèces d'acariens transformées Hellinger.

- Quels sont les axes importants?
- Quels groupes de sites pouvez-vous identifier?
- Quelles espèces sont liées à chaque groupe de sites?
- Comment les résultats de cette PCoA se comparent-ils avec ceux de la PCA?

Solution #5

- Transformation Hellinger pour les données espèces

```
mite.spe.hel <- decostand(mite.spe, method = "hellinger")
```

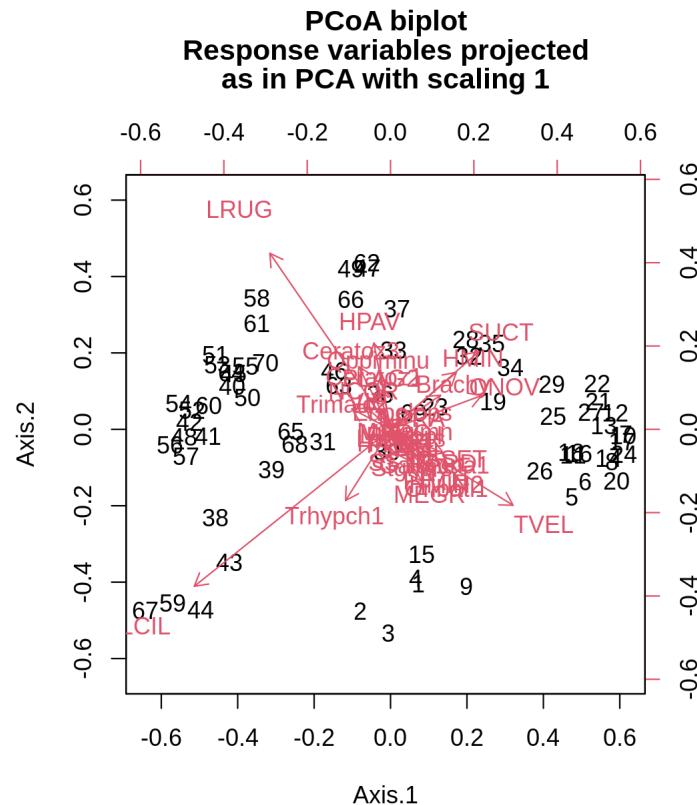
- Calcul de la PCoA

```
mite.spe.h.pcoa <- pcoa(dist(mite.spe.hel))
```

Solution #5

- Biplot pour visualiser les données:

```
biplot.pcoa(mite.spe.h.pcoa, mite.spe.hel)
```



Positionnement multidimensionnel non-métrique (NMDS)

- En PCA, CA et PCoA, les objets sont ordonnés dans un petit nombre de dimensions (i.e. axes) généralement > 2
- En conséquence, les biplots 2D ne représentent pas toute la variation présente dans les données.
- Parfois, l'objectif est cependant de représenter les données dans un nombre défini de dimensions.
- Comment effectuer une ordination pour illustrer l'ensemble de la variation des données ?

Principe du NMDS

- NMDS
 - équivalent non-métrique de la PCoA
 - basé sur un algorithme itératif d'optimisation pour identifier la meilleure représentation possible des données dans l'espace d'ordination de plus en plus populaire
- Dans un nMDS, l'utilisateur peut ainsi spécifier:
 - le nombre de dimensions
 - la mesure de distance

Effectuer un NMDS

- La fonction `metaMDS()` de la librairie `vegan` permet de réaliser un NMDS
 - *distance* spécifie la mesure de distance choisie
 - *k* spécifie le nombre de dimensions

```
spe.nmds <- metaMDS(spe, distance = 'bray', k = 2)
# Run 0 stress 0.07478058
# Run 1 stress 0.1124391
# Run 2 stress 0.1127638
# Run 3 stress 0.120581
# Run 4 stress 0.07477835
# ... New best solution
# ... Procrustes: rmse 0.001073915 max resid 0.005160002
# ... Similar to previous best
# Run 5 stress 0.07429471
# ... New best solution
# ... Procrustes: rmse 0.02379294 max resid 0.09228435
# Run 6 stress 0.1209567
# Run 7 stress 0.1124437
# Run 8 stress 0.08927212
# Run 9 stress 0.1121906
# Run 10 stress 0.1127533
# Run 11 stress 0.1111093
# Run 12 stress 0.07506894
```

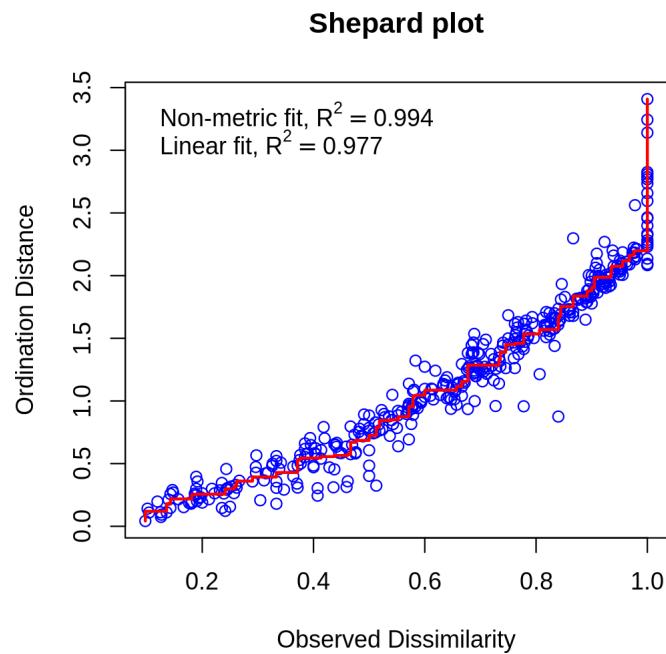
NMDS : qualité de l'ajustement

- Le NMDS applique une procédure itérative qui vise à positionner les objets dans le nombre spécifié de dimensions de façon à minimiser une fonction de stress (variant de 0 à 1) qui mesure la qualité de l'ajustement de la distance entre objets dans l'espace d'ordination.
- Ainsi, plus la valeur du stress sera faible, plus la représentation des objets dans l'espace d'ordination sera exacte.

NMDS : qualité de l'ajustement

- La valeur de stress et le diagramme de Shepard peuvent être obtenus avec :

```
spe.nmds$stress  
# [1] 0.07429471  
stressplot(spe.nmds, main = "Shepard plot")
```

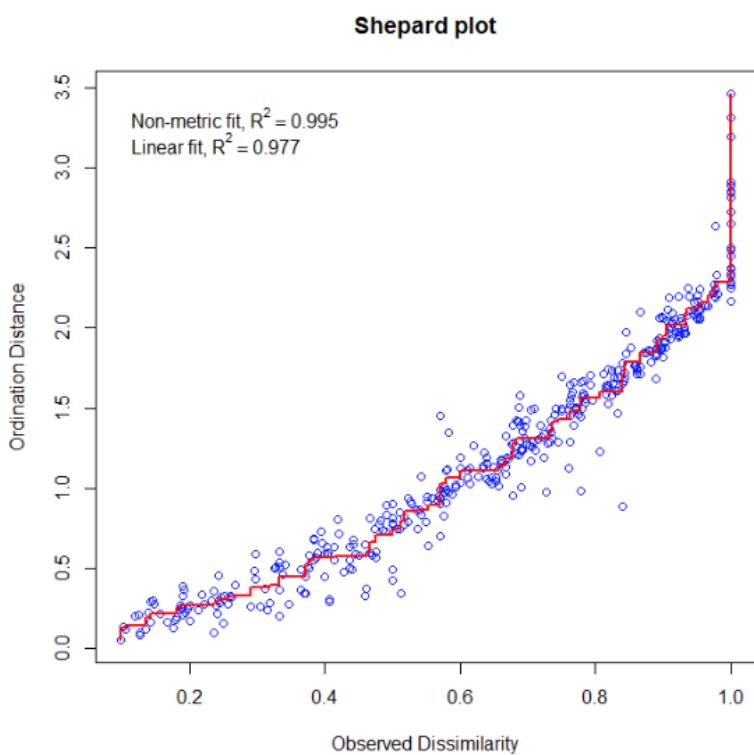


NMDS sur les données Poissons

- Effectuer le NMDS et vérifier la qualité de l'ajustement

```
spe.nmds <- metaMDS(spe, distance = 'bray', k = 2)
spe.nmds$stress
stressplot(spe.nmds, main = "Shepard plot")
```

NMDS sur les données Poissons



- Le diagramme de Shepard identifie une forte corrélation entre les distances observées et les distances de l'ordination ($R^2 > 0.95$), et donc une bonne qualité de l'ajustement du NMDS.

NMDS sur les données Poissons

- Construction du biplot

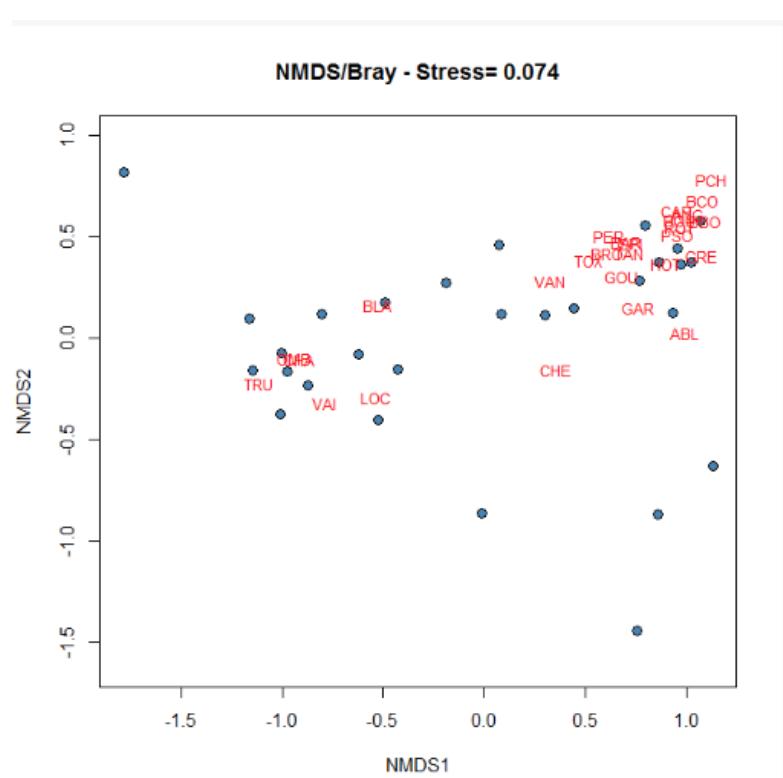
```
plot(spe.nmds, type = "none",
      main = paste("NMDS/Bray - Stress =",
                  round(spe.nmds$stress, 3)),
      xlab = c("NMDS1"), ylab = "NMDS2")

points(scores(spe.nmds, display = "sites",
              choices = c(1,2),
              pch = 21,
              col = "black",
              g = "steelblue",
              cex = 1.2))

text(scores(spe.nmds, display = "species", choices = c(1)),
      scores(spe.nmds, display = "species", choices = c(2)),
      labels = rownames(scores(spe.nmds, display = "species")),
      col = "red", cex = 0.8)
```

NMDS sur les données Poissons

Le biplot du nMDS identifie un groupe de sites caractérisées par les espèces BLA, TRU, VAI, LOC, CHA et OMB, tandis que les autres espèces caractérisent un groupe de sites situés dans le coin supérieur droit du biplot.

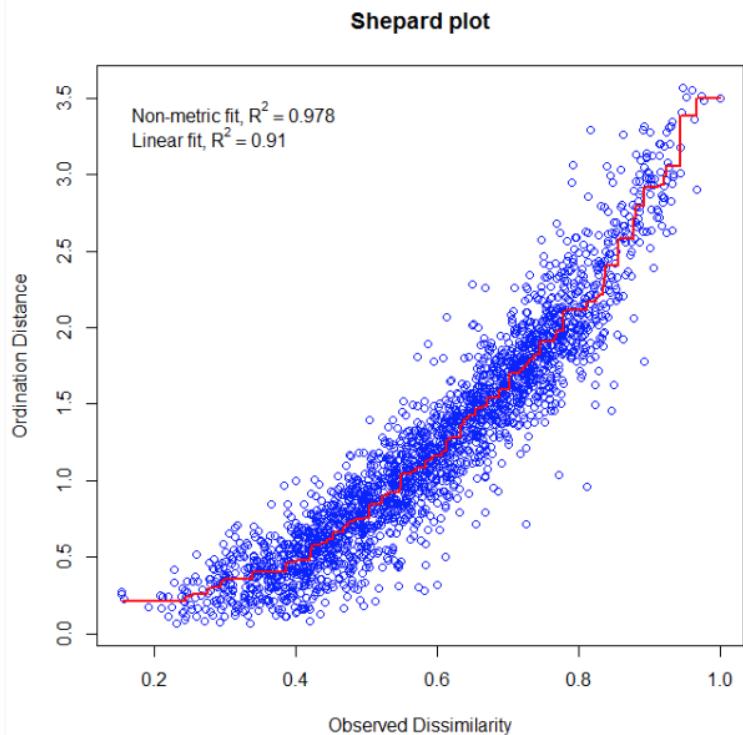


Défi #6



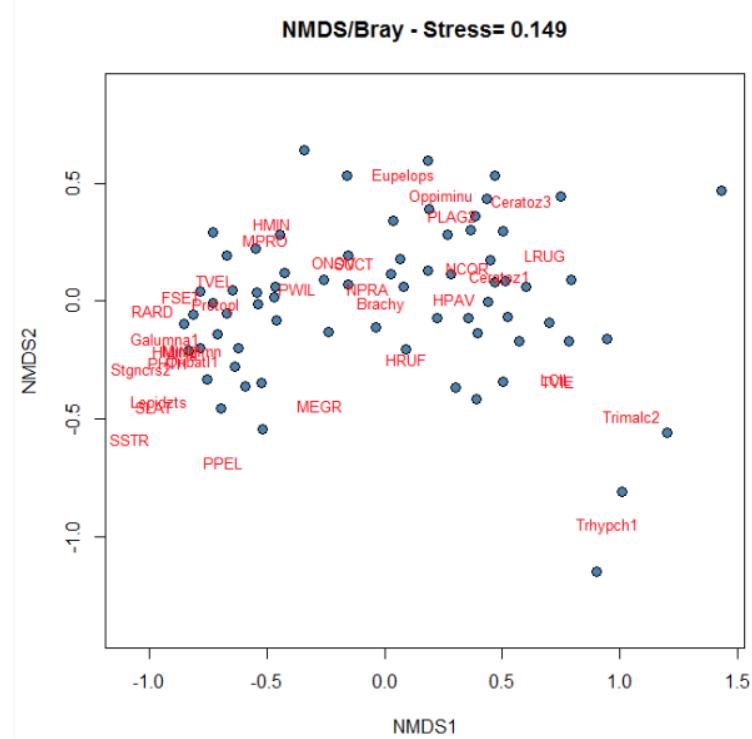
- Exécuter un NMDS sur les données d'abondance des espèces d'acariens en deux dimensions à partir de distances de Bray-Curtis.
- Évaluer la qualité de l'ajustement et interpréter le biplot

Solution #6



La corrélation entre distance observée et distance d'ordination ($R^2 > 0.91$) et la valeur de stress relativement faible identifient une bonne qualité de l'ajustement du NMDS.

Solution #6



Aucun groupe de sites ne peut être précisément identifié à partir du biplot, ce qui montre que la plupart des espèces sont présentes dans la plupart des sites.

Conclusion

Beaucoup de méthodes d'ordination existent mais leurs spécificités doivent guider le choix de la méthode à utiliser :

	Distance préservée	Variables	Nombre maximum d'axes
PCA	Euclidienne	Données quantitatives, relation linéaires	p
CA	Chi2	Non-négatives, données quantitatives dimensionnellement homogènes, ou données binaires	p-1
PCoA	Définie par l'utilisateur	Quantitatives, semi-quantitatives ou mixtes	p-1
NMDS	Définie par l'utilisateur	Quantitatives, semi-quantitatives ou mixtes	Définie par l'utilisateur

C'est l'heure du quiz !

Que signifie PCA?

Principal Component Analysis

Laquelle de ces méthodes est la meilleure pour visualiser les *distances* entre composition des communautés de différents sites?

Principal Coordinate Analysis (PCoA)

Que représente une valeur propre dans une PCA ?

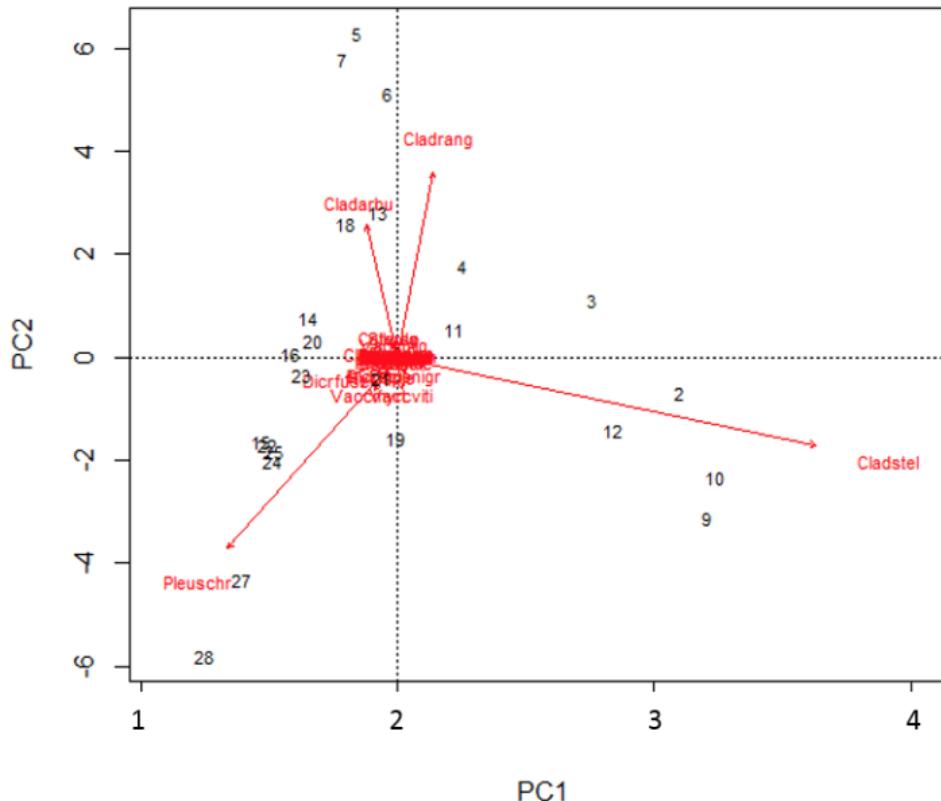
La proportion de variance capturée par une composante principale

C'est l'heure du quiz !

Trouvez l'erreur!



You !



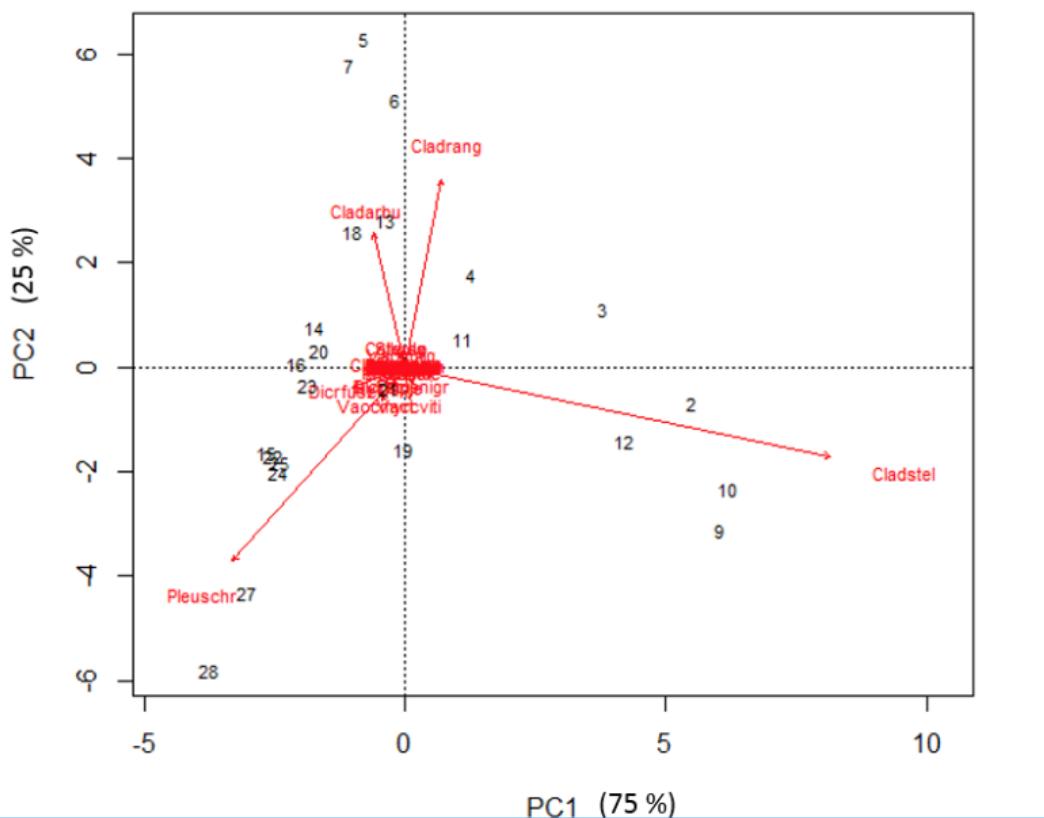
- Données non centrées, Beurk!

C'est l'heure du quiz !

Trouvez l'erreur!



You !



- Les 2 premiers axes capturent 100% de la variation

Live Long and Ordinate

Merci pour votre participation à cet atelier!

