



Le ReJMiC présente :



Méthodes d'analyse non supervisées

Journée d'initiation à la bio-informatique n°2
24 juin 2022



Plan

Partie I : L'Analyse en Composantes Principales (ACP)

Partie II : Clustering



Plan

Partie I : L'Analyse en Composantes Principales (ACP)

Partie II : Clustering



Introduction/ La PCA

PCA = Principal Component Analysis

ACP = Analyse en Composantes Principales

- Analyse **multivariée** (> 2 variables qui caractérisent un individu)
- Jeu de données contenant des **individus** décrits par **plusieurs variables quantitatives**
- Méthode de **visualisation sans a priori** → **méthode non supervisée**

Résumé des informations (tableau de données × individus) :

→ **On parle de réduction de dimension**

→ **En biologie :** données transcriptomiques, protéomiques...

Les données

Exemple données transcriptomiques

Individus (gènes)

Conditions expérimentales = variables

Fragment des données transcriptomiques brutes

| | day1_2 | day1_3 | day2_1 | day2_2 | day2_3 | day3_1 |
|--------------------|-------------|-------------|-------------|-------------|------------|------------|
| ENSMUSG00000000567 | 599.648075 | 304.093177 | 1052.689447 | 106.995584 | 347.13842 | 479.59911 |
| ENSMUSG00000000568 | 1008.026306 | 1349.126716 | 818.257157 | 116.136417 | 3406.45030 | 766.09722 |
| ENSMUSG00000000579 | 599.832586 | 500.735597 | 473.399472 | 36.258005 | 410.57787 | 347.05054 |
| ENSMUSG00000000581 | 942.511039 | 744.646735 | 546.260344 | 87.788535 | 319.12679 | 461.45732 |
| ENSMUSG00000000594 | 3063.845705 | 2743.404374 | 2283.051270 | 1115.588250 | 1491.61384 | 1576.68451 |
| ENSMUSG00000000600 | 3341.854530 | 3561.805180 | 3674.188108 | 589.840068 | 1399.10751 | 3446.01856 |
| ENSMUSG00000000605 | 791.312561 | 558.710943 | 489.579657 | 77.008213 | 256.00200 | 282.80077 |
| ENSMUSG00000000606 | 4.785252 | 6.566374 | 3.970399 | 0.000000 | 138.48677 | 0.00000 |
| ENSMUSG00000000617 | 15.839486 | 29.138902 | 30.228506 | 6.670500 | 18.27493 | 13.41152 |
| ENSMUSG00000000627 | 36.673777 | 35.967747 | 46.297517 | 2.878275 | 17.03638 | 15.45854 |

Variables quantitatives

Rappel

Variance : mesure de la dispersion des variables

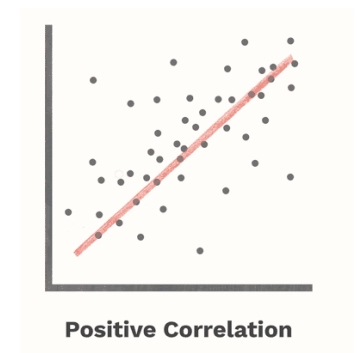
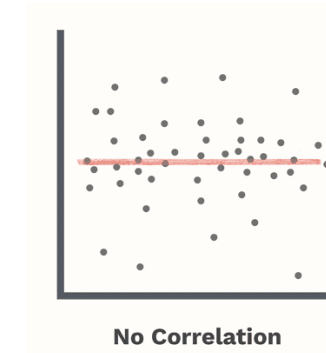
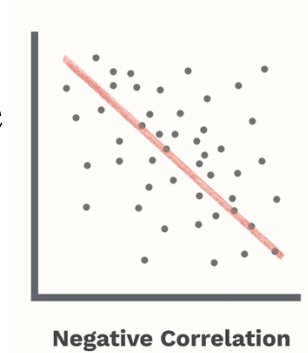
→ variance élevée = points éloignés autour de la moyenne

Covariance : mesure de la liaison entre deux variables

→ covariance élevée = relation forte entre deux variables

Corrélation : mesure standardisée de la covariance

→ la corrélation varie entre -1 et 1



La variance = information

Pourquoi la PCA ?

- Former des groupes d'individus semblables → ressemblance
- Former des groupes de variables liées entre elles → liaison - corrélation



Caractérisation des groupes d'individus par les variables

Quelles (groupes de) variables expliquent le plus la variabilité inter-individus ?

En biologie (exemple) :

Est-ce que la variance contenue dans les variables permet d'expliquer mes différents phénotypes (individus) ?

Objectifs de la PCA en biologie

- Analyse exploratoire des questions
- Comprendre la structure sous-jacente des données
- Identifier les biais, les erreurs expérimentales, les effets de lot
- Identifier les variables corrélées

Un exemple simple pour comprendre comment ça fonctionne

Notes de 11 élèves de 1 à 20 pour 5 disciplines

| | | Variables | | | | |
|-----------|----------|-----------|-------------------------|-------------|----------|--------------------------|
| Individus | | Maths | Histoire- Géographie | Philosophie | Physique | Biologie fondamentale |
| | eleve 1 | 10 | 19 | 9 | 4 | 7 |
| | eleve 2 | 12 | 12 | 13 | 6 | 9 |
| | eleve 3 | 16 | 18 | 14 | 10 | 13 |
| | eleve 4 | 7 | 12 | 16 | 1 | 4 |
| | eleve 5 | 18 | 9 | 11 | 12 | 15 |
| | eleve 6 | 16 | 12 | 17 | 10 | 13 |
| | eleve 7 | 13 | 14 | 10 | 7 | 10 |
| | eleve 8 | 12 | 14 | 7 | 6 | 9 |
| | eleve 9 | 11 | 12 | 15 | 5 | 8 |
| | eleve 10 | 9 | 16 | 11 | 3 | 6 |
| | eleve 11 | 9 | 16 | 11 | 3 | 6 |

Comment analyser simultanément ces 5 variables ?

Classiquement on aurait fait

Analyse univariée

→ Etudes des variables (colonnes)

| | Maths | Histoire Géographie | Philosophie | Physique | Biologie fondamentale |
|------------|-------|------------------------|-------------|----------|--------------------------|
| Moyenne | 10,9 | 13,8 | 12,3 | 6,7 | 8,8 |
| Ecart type | 4,6 | 2,9 | 3,1 | 3,2 | 3,6 |

→ Etude des individus (lignes)

| | eleve 1 | eleve 2 | eleve 3 | eleve 4 | eleve 5 | eleve 6 | eleve 7 | eleve 8 | eleve 9 | eleve 10 | eleve 11 |
|------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|----------|----------|
| Moyenne | 8,6 | 8,3 | 14,2 | 8,0 | 13,0 | 13,6 | 10,8 | 9,6 | 10,2 | 9,0 | 10,3 |
| Ecart type | 6,1 | 4,0 | 3,0 | 6,0 | 3,5 | 2,9 | 2,8 | 3,4 | 3,8 | 4,9 | 2,6 |

Classiquement on aurait fait

Analyse bivariée (matrice de corrélation)

→ Etudes des variables (colonnes)

| | Maths | Histoire-Géographie | Philosophie | Physique | Biologie fondamentale |
|-----------------------|-------|---------------------|-------------|----------|-----------------------|
| Maths | 1 | -0,28 | 0,11 | 0,71 | 0,95 |
| Histoire-Géographie | -0,28 | 1 | -0,33 | -0,23 | -0,28 |
| Philosophie | 0,11 | -0,33 | 1 | 0,02 | 0,08 |
| Physique | 0,71 | -0,23 | 0,02 | 1 | 0,89 |
| Biologie fondamentale | 0,95 | -0,28 | 0,08 | 0,89 | 1 |

→ Etude des individus (lignes)

| | eleve 1 | eleve 2 | eleve 3 | eleve 4 | eleve 5 | eleve 6 | eleve 7 | eleve 8 | eleve 9 | eleve 10 | eleve 11 |
|----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|----------|----------|
| eleve 1 | 1 | 0,73 | 0,71 | 0,6 | -0,75 | -0,13 | 0,62 | 0,63 | 0,48 | 0,88 | 0,93 |
| eleve 2 | 0,73 | 1 | 0,29 | 0,8 | -0,89 | 0,13 | 0,13 | 0 | 0,63 | 0,65 | 0,85 |
| eleve 3 | 0,71 | 0,29 | 1 | 0,63 | -0,09 | 0,38 | 0,99 | 0,92 | 0,68 | 0,92 | 0,74 |
| eleve 4 | 0,6 | 0,8 | 0,63 | 1 | -0,44 | 0,66 | 0,51 | 0,27 | 0,97 | 0,82 | 0,85 |
| eleve 5 | -0,75 | -0,89 | -0,09 | -0,44 | 1 | 0,34 | 0,05 | 0,04 | -0,22 | -0,46 | -0,71 |
| eleve 6 | -0,13 | 0,13 | 0,38 | 0,66 | 0,34 | 1 | 0,36 | 0,08 | 0,8 | 0,33 | 0,2 |
| eleve 7 | 0,62 | 0,13 | 0,99 | 0,51 | 0,05 | 0,36 | 1 | 0,95 | 0,59 | 0,84 | 0,62 |
| eleve 8 | 0,63 | 0 | 0,92 | 0,27 | 0,04 | 0,08 | 0,95 | 1 | 0,34 | 0,74 | 0,52 |
| eleve 9 | 0,48 | 0,63 | 0,68 | 0,97 | -0,22 | 0,8 | 0,59 | 0,34 | 1 | 0,79 | 0,75 |
| eleve 10 | 0,88 | 0,65 | 0,92 | 0,82 | -0,46 | 0,33 | 0,84 | 0,74 | 0,79 | 1 | 0,95 |
| eleve 11 | 0,93 | 0,85 | 0,74 | 0,85 | -0,71 | 0,2 | 0,62 | 0,52 | 0,75 | 0,95 | 1 |

Classiquement on aurait fait

Analyse bivariable (matrice de corrélation)

→ Etudes des variables (colonnes)

| | Maths | Histoire-Géographie | Philosophie | Physique | Biologie fondamentale |
|-----------------------|-------|---------------------|-------------|----------|-----------------------|
| Maths | 1 | -0,28 | 0,11 | 0,71 | 0,95 |
| Histoire-Géographie | -0,28 | 1 | 0,33 | 0,23 | 0,28 |
| Philosophie | 0,11 | -0,33 | 1 | 0,02 | 0,08 |
| Physique | 0,71 | -0,23 | 0,02 | 1 | 0,89 |
| Biologie fondamentale | 0,95 | -0,28 | 0,08 | 0,89 | 1 |

→ Etude des individus (lignes)

| | eleve 1 | eleve 2 | eleve 3 | eleve 4 | eleve 5 | eleve 6 | eleve 7 | eleve 8 | eleve 9 | eleve 10 | eleve 11 |
|----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|----------|----------|
| eleve 1 | 1 | 0,73 | 0,71 | 0,6 | -0,75 | 0,62 | 0,63 | 0,48 | 0,88 | 0,93 | |
| eleve 2 | 0,73 | 1 | 0,29 | 0,8 | -0,89 | 0,13 | 0 | 0,63 | 0,65 | 0,85 | |
| eleve 3 | 0,71 | 0,29 | 1 | 0,63 | -0,09 | 0,99 | 0,92 | 0,68 | 0,92 | 0,74 | |
| eleve 4 | 0,6 | 0,8 | 0,63 | 1 | -0,44 | 0,51 | 0,27 | 0,97 | 0,82 | 0,85 | |
| eleve 5 | -0,75 | -0,89 | -0,09 | -0,44 | 1 | 0,34 | 0,05 | 0,04 | -0,22 | -0,46 | -0,71 |
| eleve 6 | -0,13 | 0,13 | 0,38 | 0,66 | 0,34 | 1 | 0,36 | 0,08 | 0,8 | 0,33 | 0,2 |
| eleve 7 | 0,62 | 0,13 | 0,99 | 0,51 | 0,05 | 0,36 | 1 | 0,95 | 0,59 | 0,84 | 0,62 |
| eleve 8 | 0,63 | 0 | 0,92 | 0,27 | 0,04 | 0,08 | 0,95 | 1 | 0,34 | 0,74 | 0,52 |
| eleve 9 | 0,48 | 0,63 | 0,68 | 0,97 | -0,22 | 0,8 | 0,59 | 0,34 | 1 | 0,79 | 0,75 |
| eleve 10 | 0,88 | 0,65 | 0,92 | 0,82 | -0,46 | 0,33 | 0,84 | 0,74 | 0,79 | 1 | 0,95 |
| eleve 11 | 0,93 | 0,85 | 0,74 | 0,85 | -0,71 | 0,2 | 0,62 | 0,52 | 0,75 | 0,95 | 1 |

Et si on essayait la PCA ?

Un exemple simple pour comprendre comment ça fonctionne

Notes de 11 élèves de 1 à 20 pour 5 disciplines

| | | Variables | | | | |
|-----------|----------|-----------|-------------------------|-------------|----------|--------------------------|
| Individus | | Maths | Histoire- Géographie | Philosophie | Physique | Biologie fondamentale |
| | eleve 1 | 10 | 19 | 9 | 4 | 7 |
| | eleve 2 | 12 | 12 | 13 | 6 | 9 |
| | eleve 3 | 16 | 18 | 14 | 10 | 13 |
| | eleve 4 | 7 | 12 | 16 | 1 | 4 |
| | eleve 5 | 18 | 9 | 11 | 12 | 15 |
| | eleve 6 | 16 | 12 | 17 | 10 | 13 |
| | eleve 7 | 13 | 14 | 10 | 7 | 10 |
| | eleve 8 | 12 | 14 | 7 | 6 | 9 |
| | eleve 9 | 11 | 12 | 15 | 5 | 8 |
| | eleve 10 | 9 | 16 | 11 | 3 | 6 |
| | eleve 11 | 9 | 16 | 11 | 3 | 6 |

Comment analyser simultanément ces 5 variables ?

Etude des individus

- Un individu = 1 ligne du tableau \rightarrow 1 point dans un espace à p (n =variables) dimensions
 - si $p = 2 \rightarrow$ nuage de points
 - si $p = 3 \rightarrow$ espace 3D
 - si $p \geq 4 \rightarrow$ représentation impossible
- Notion de ressemblance entre les individus
 - deux individus se ressemblent \rightarrow valeurs proches sur l'ensemble des p variables
 - mesure de la distance entre les individus (somme des carrés des écarts pour chaque variable)
- Visualisation de la forme du nuage de points \rightarrow étude des individus

Etude des individus

- Visualisation d'un nuage de point en 2D (photo) à partir d'un espace 3D

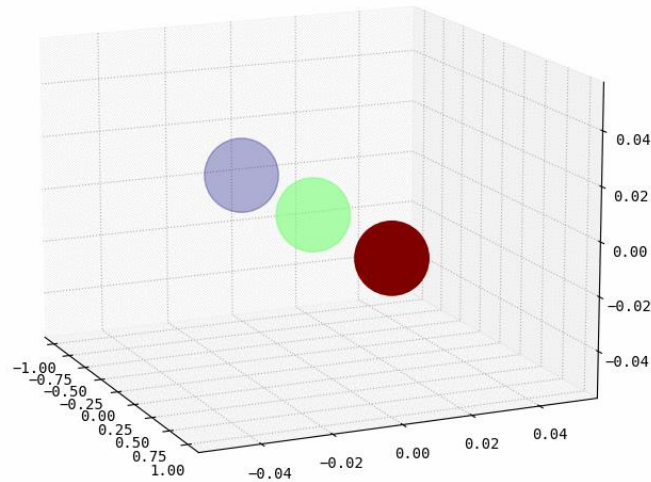


[Comment les oiseaux synchronisent-ils leur vol ? \(vidéo\) | Etrange et Insolite \(jack35.fr\)](#)

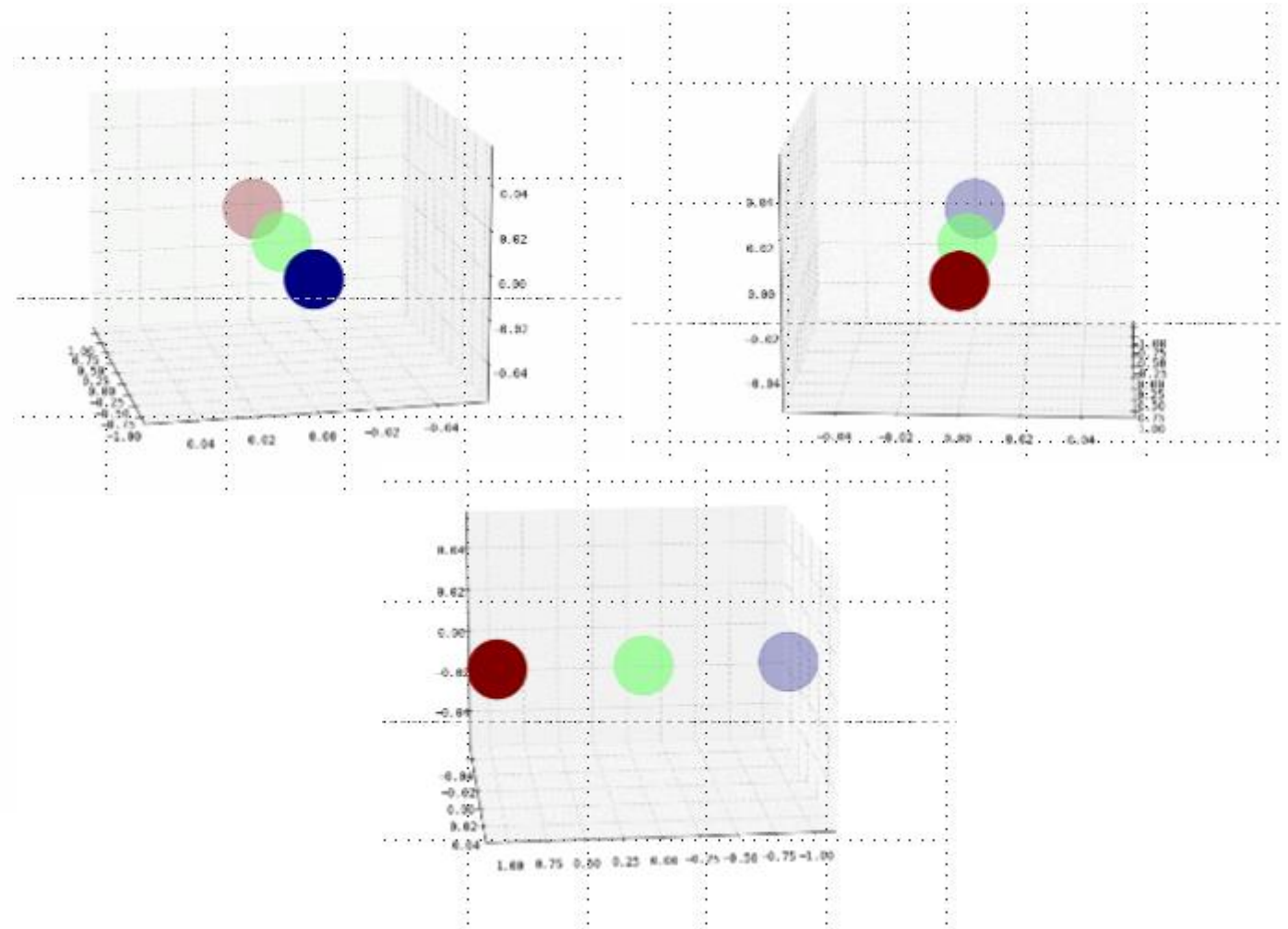
- L'ACP va fournir une image simplifiée
- Trouver le sous espace qui résume le plus fidèlement les données (restitution de l'image originale)

Etude des individus

- Quelle représentation choisir ?

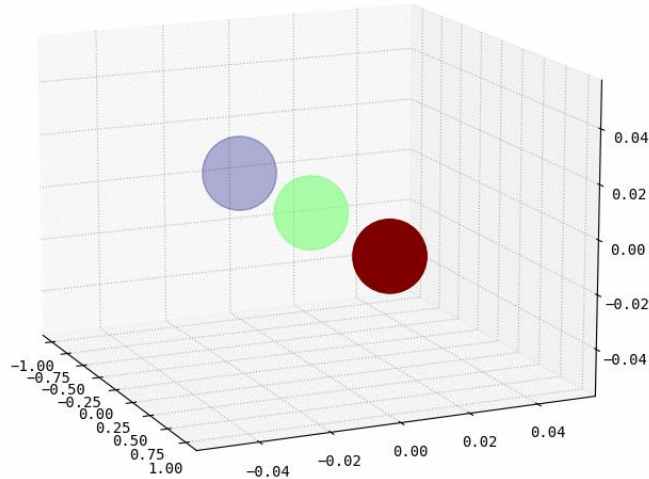


<https://github.com/matplotlib/matplotlib/issues/5830/>

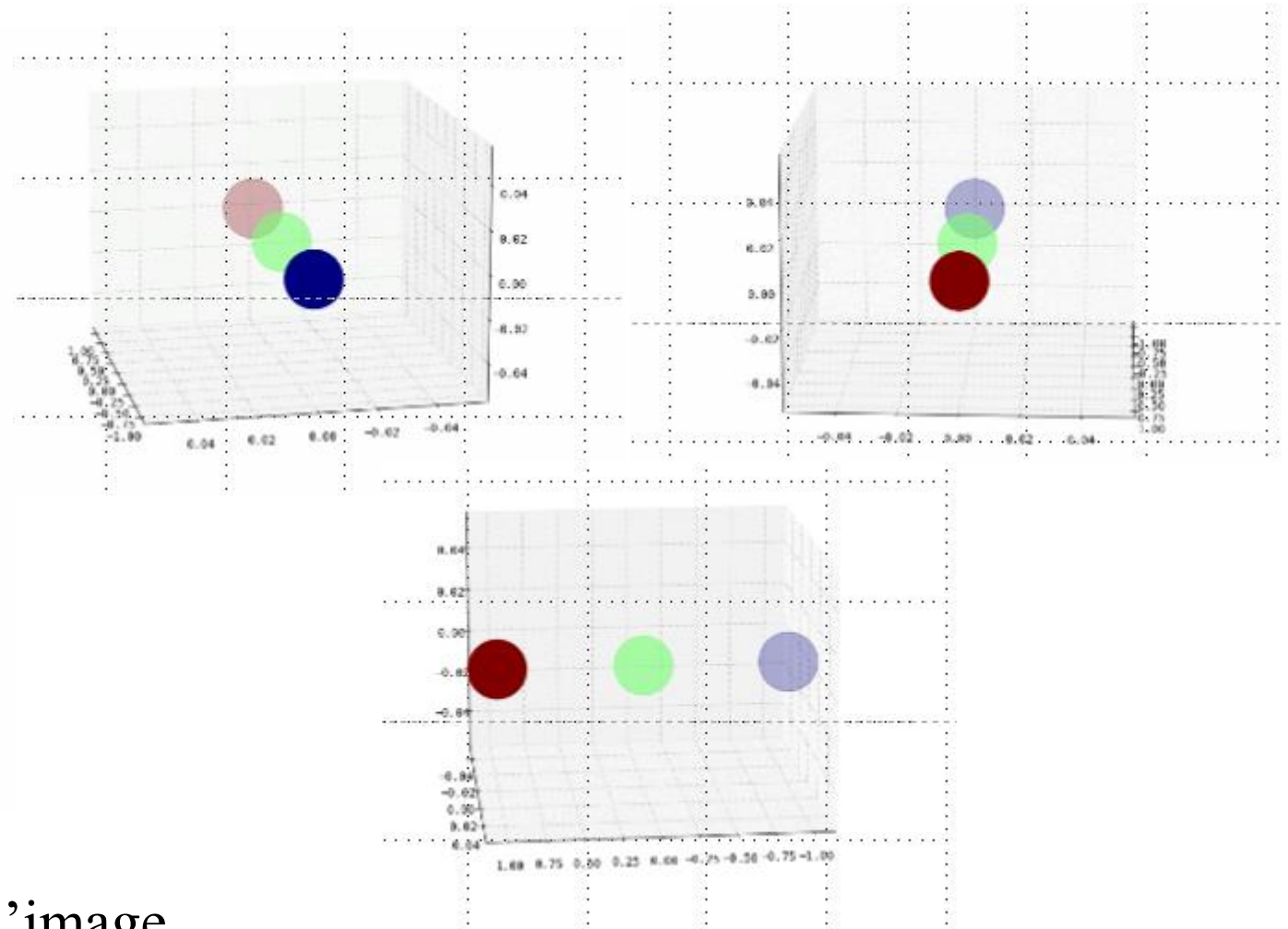


Etude des individus

- Quelle représentation choisir ?



<https://github.com/matplotlib/matplotlib/issues/5830/>



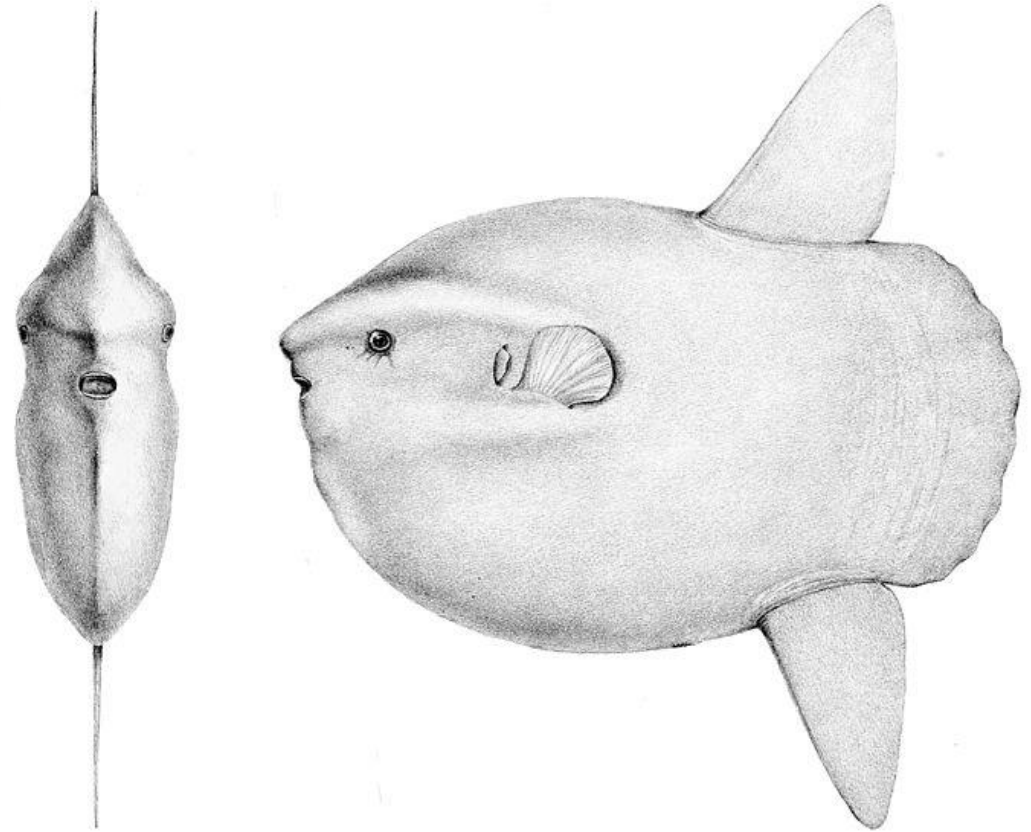
- Une image est bonne :
 - si elle restitue la forme initiale de l'image
 - si elle ne déforme pas les distances entre les individus
 - si elle représente au mieux la diversité et la variabilité des données

Etude des individus

- Comment dire qu'une image est de bonne qualité ?
 - Notion de variabilité ou de dispersion sur plusieurs dimensions = inertie
 - Inertie = variance généralisée sur plusieurs dimensions

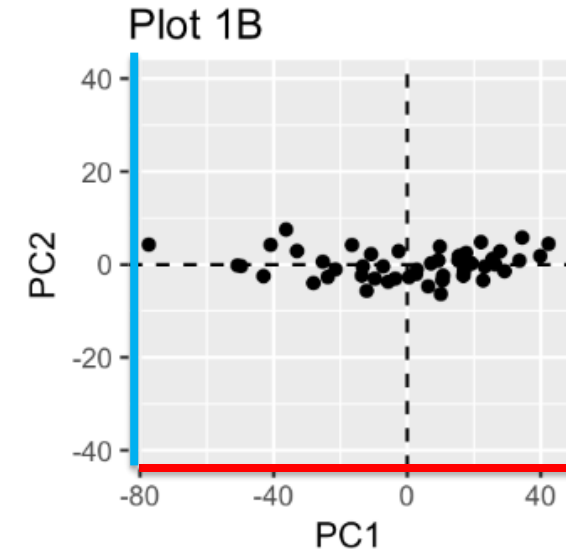
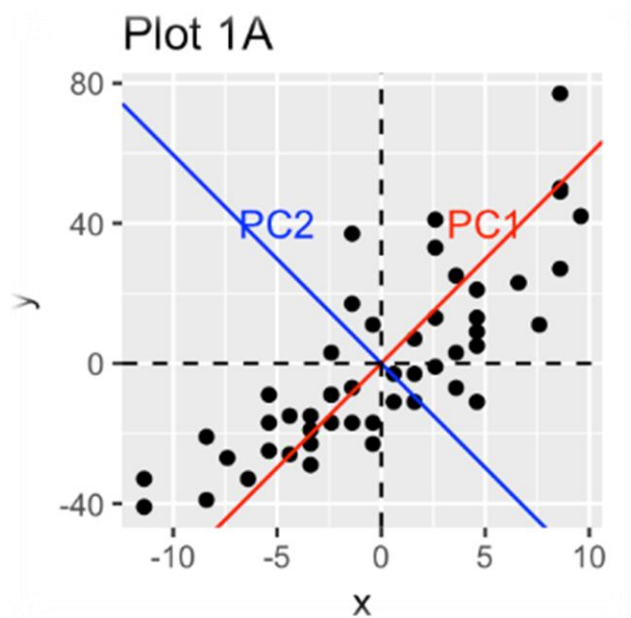


[https://fr.wikipedia.org/wiki/M%C3%B4le_\(poisson\)#/media/Fichier:Sunfish2.jpg](https://fr.wikipedia.org/wiki/M%C3%B4le_(poisson)#/media/Fichier:Sunfish2.jpg)



https://fr.m.wikipedia.org/wiki/Fichier:Mola_mola_face_profile.jpg

Etude des individus



<https://www.analyticsvidhya.com/blog/2020/12/an-end-to-end-comprehensive-guide-for-pca/>

→ Composantes principales :

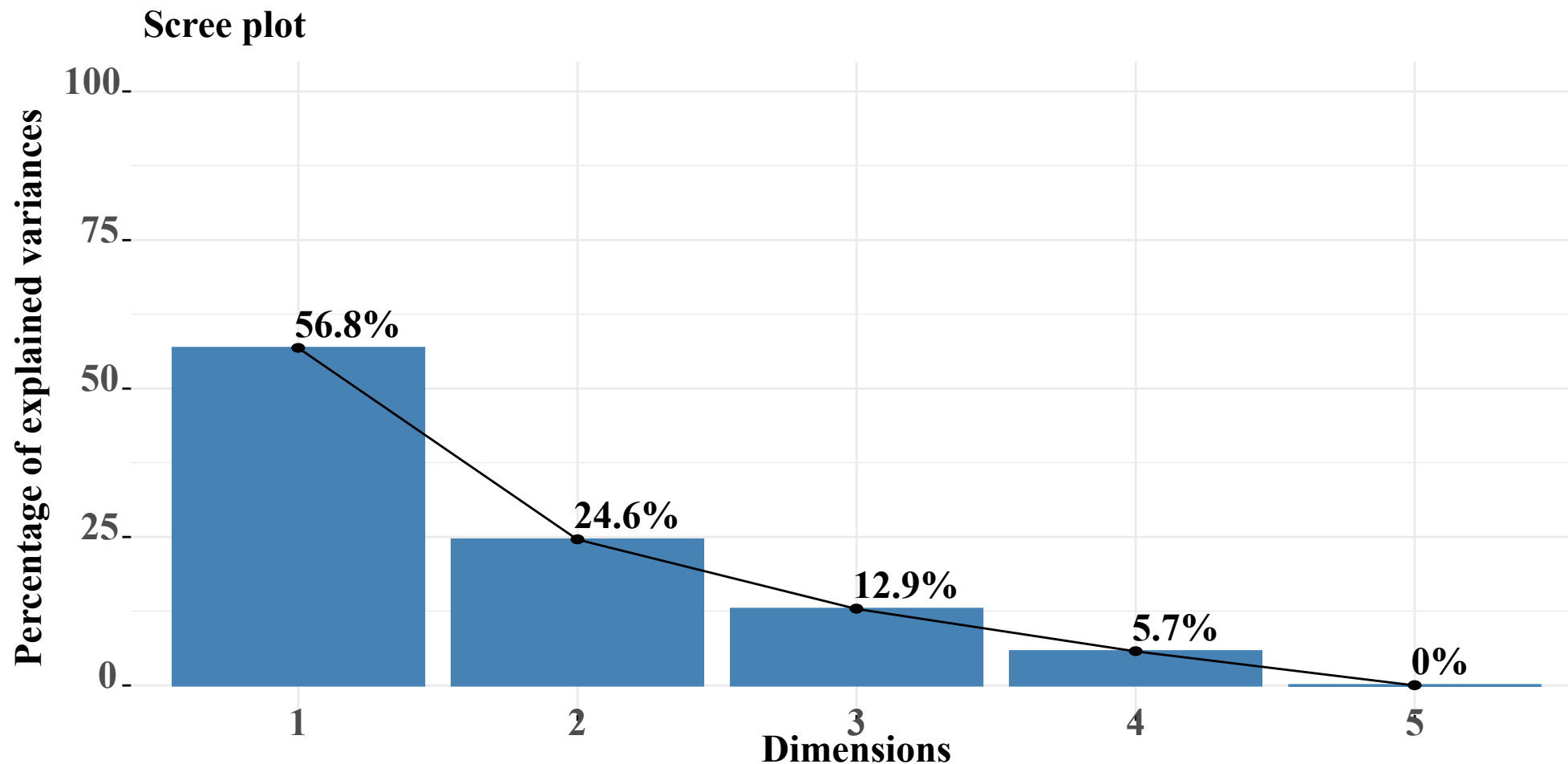
- Les variables originales = variable artificielles pour expliquer l'information (ici la variance)

→ Séparation avec les composantes principales (PC1 et PC2) :

- La direction des axes = maximisation de la variance
- PC1 : premier axe principal → direction selon le maximum de variance entre les individus
- PC2 : deuxième axe principal → seconde direction la plus importante, orthogonale à PC1

Valeurs propres/Variances

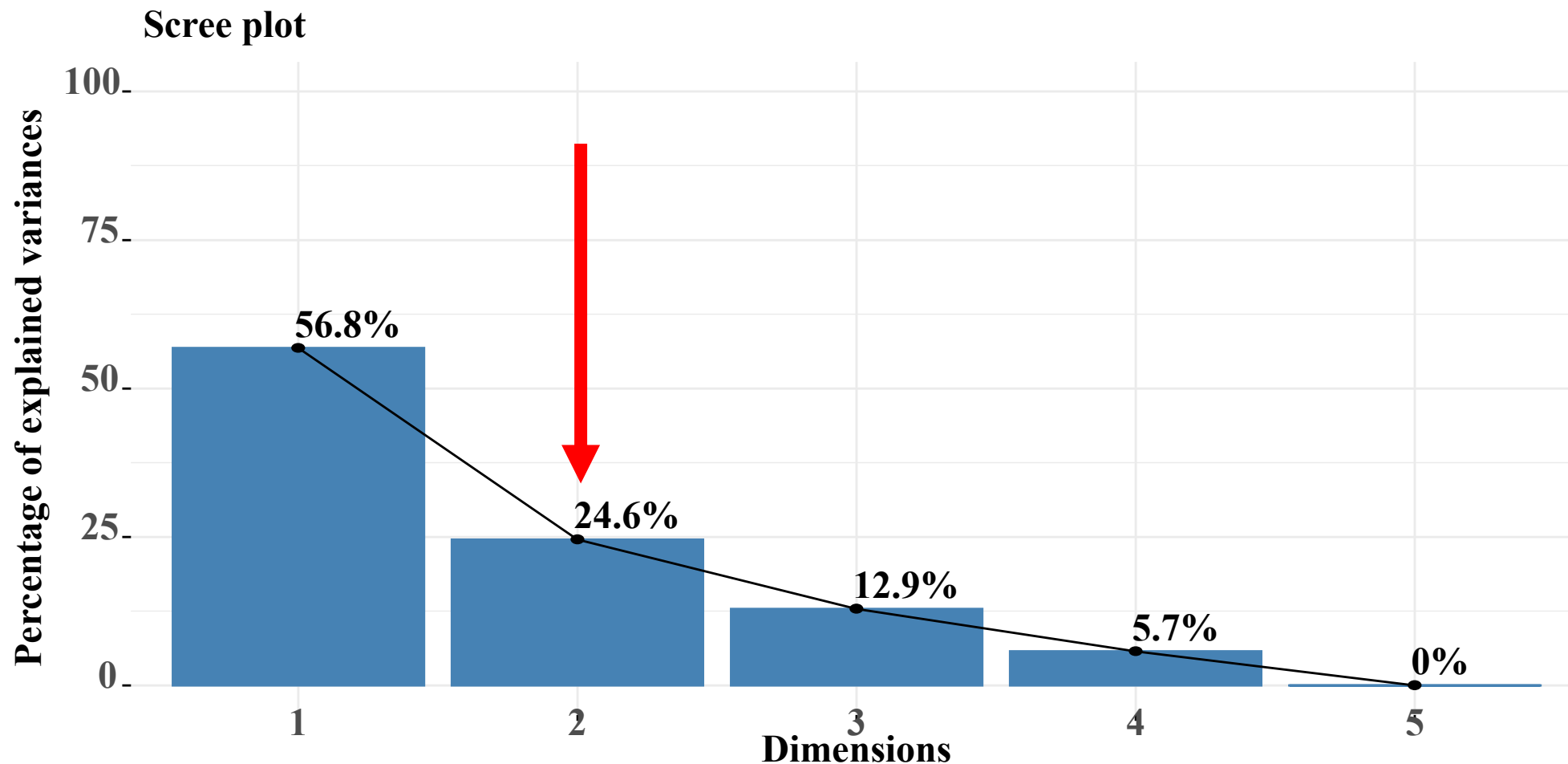
Valeurs propres (eigenvalues en anglais) : mesure de la quantité de variance par composante



Quantité d'information (variance) par composante

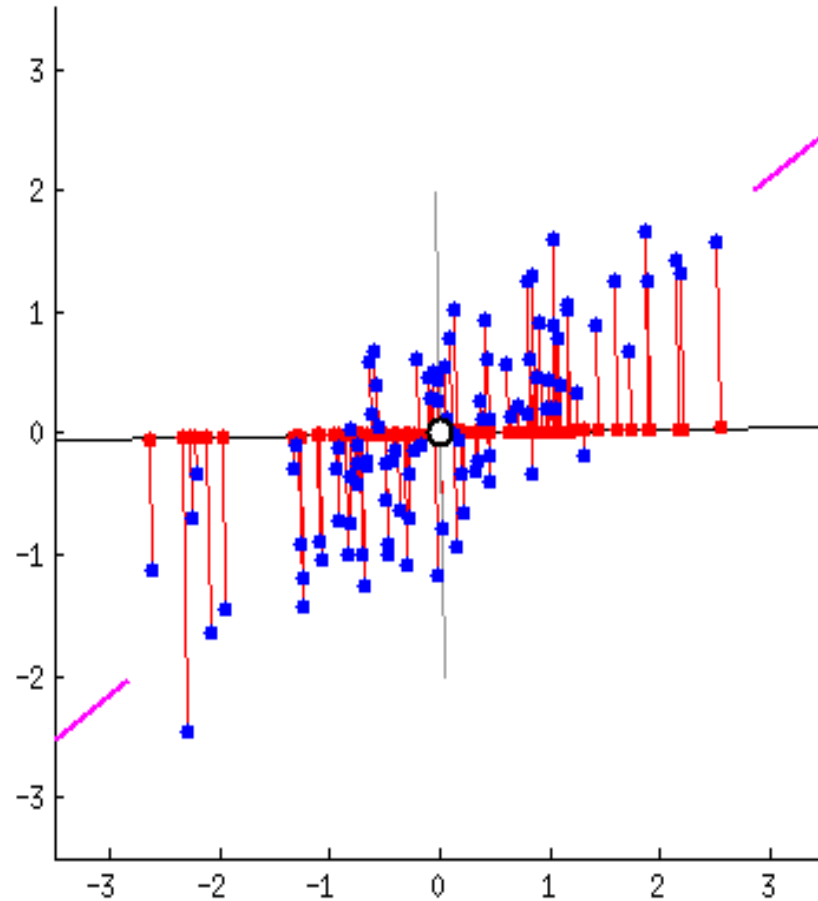
Valeurs propres/Variances

Valeurs propres (eigenvalues en anglais) : mesure de la quantité de variance par composante



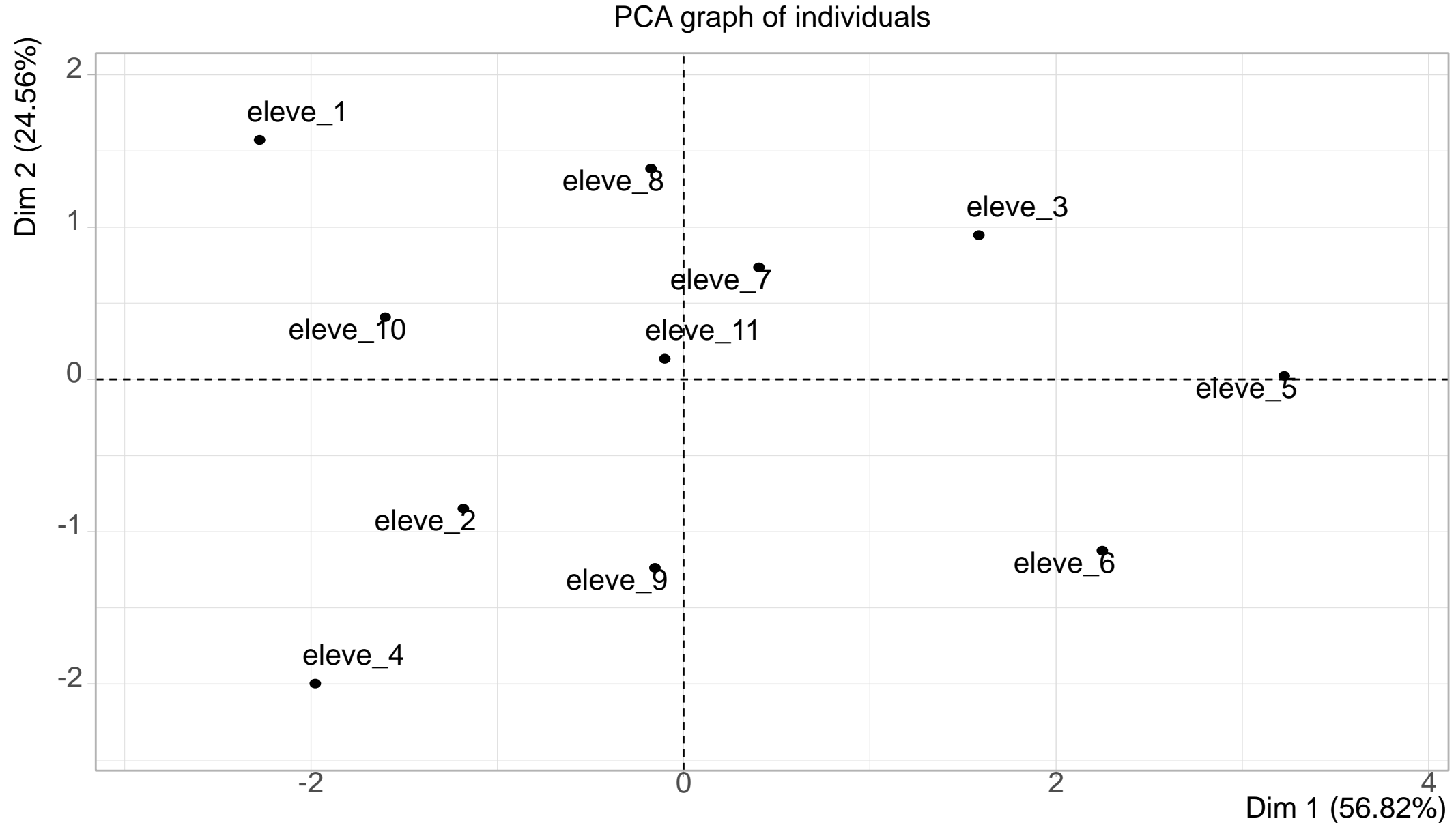
Quantité d'information (variance) par composante

Etudes des individus

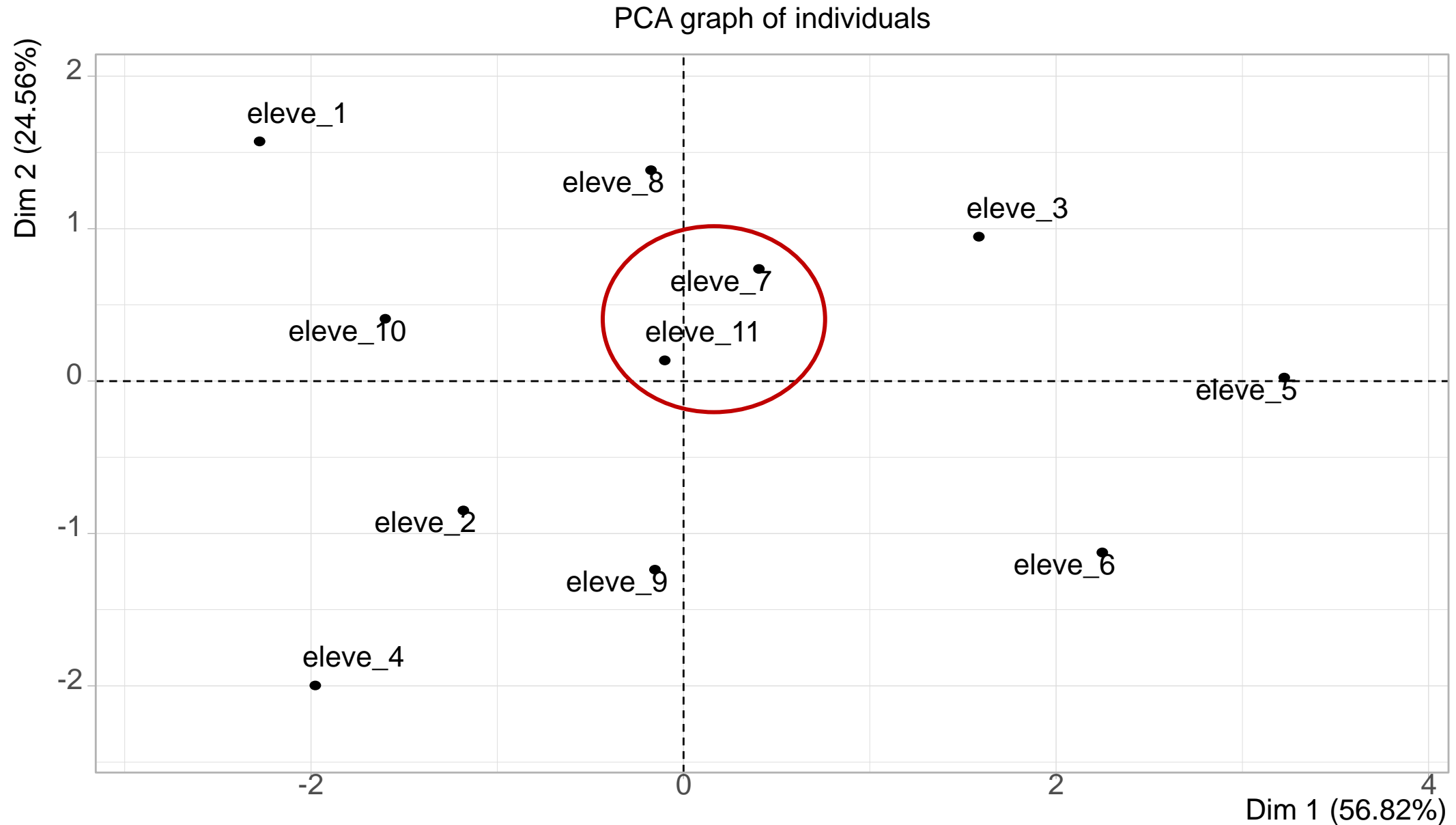


Composantes = combinaisons linéaires des variables initiales (les valeurs propres)

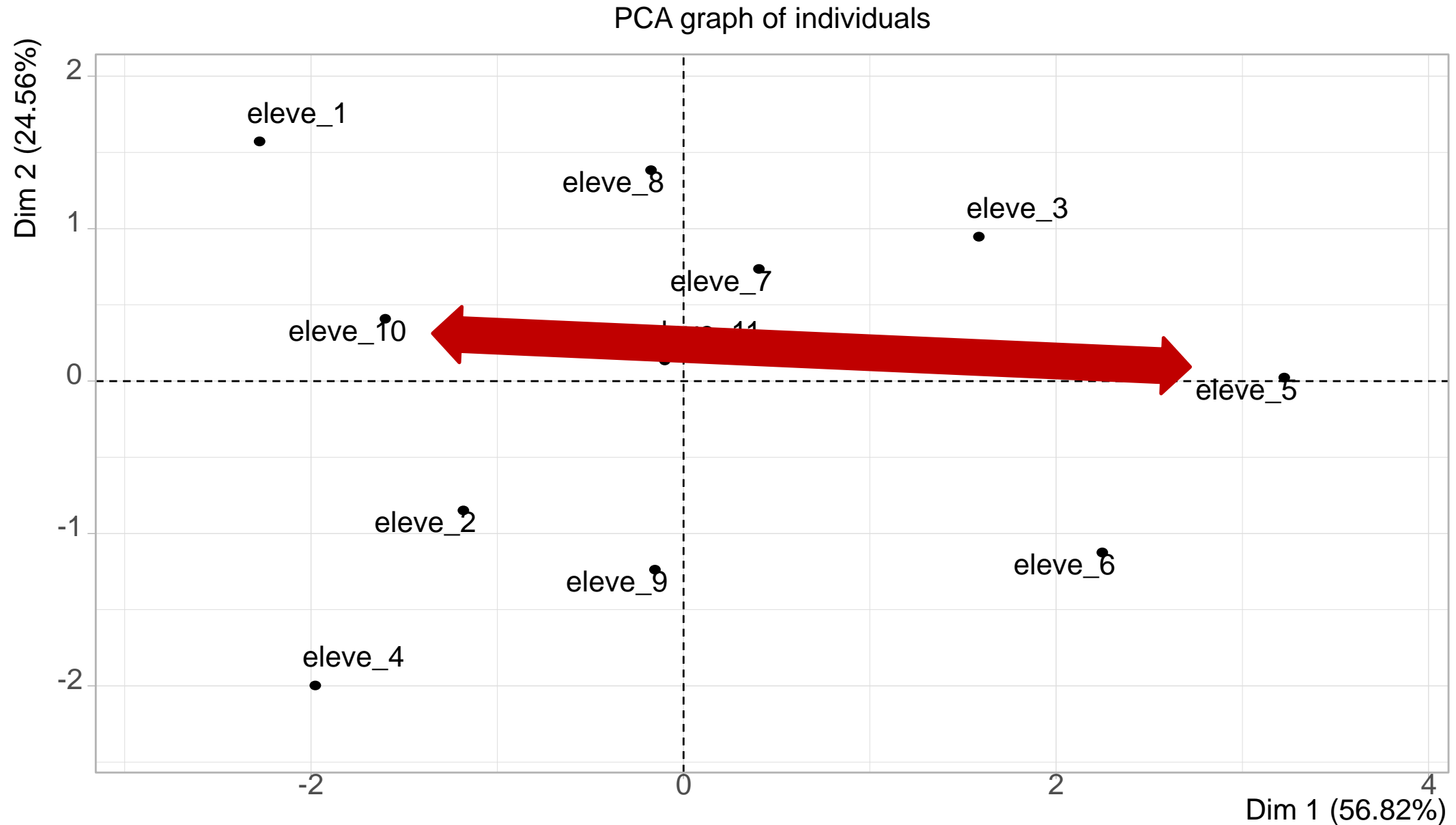
Etudes des individus/Graphe des individus



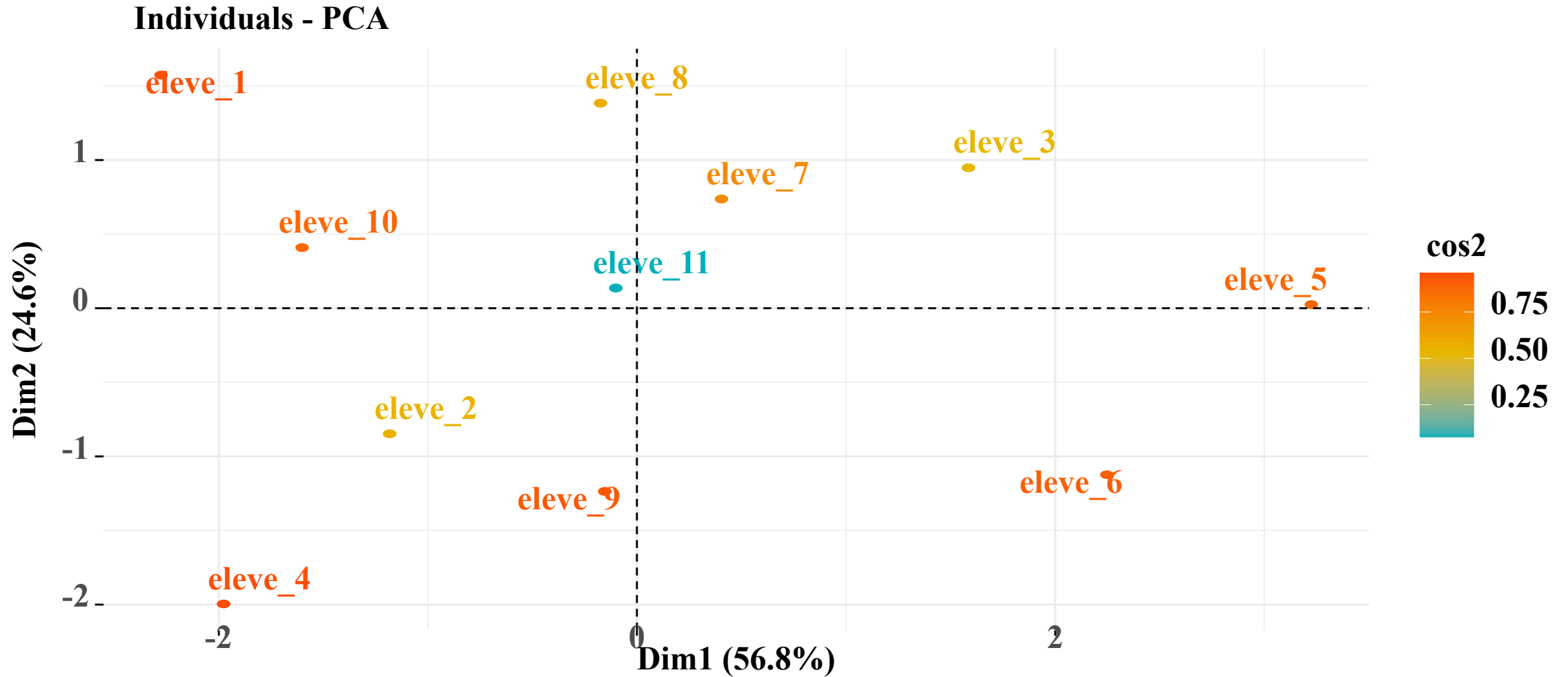
Etudes des individus/Graphe des individus



Etudes des individus/Graphe des individus



Etudes des individus/Graphe des individus



Cosinus carré → qualité de la représentation de chaque individu sur chaque axe

Etude des variables

- Une variable = 1 colonne du tableau → 1 point dans un espace à N (n =individus) dimensions
- Variables = représentées par des flèches

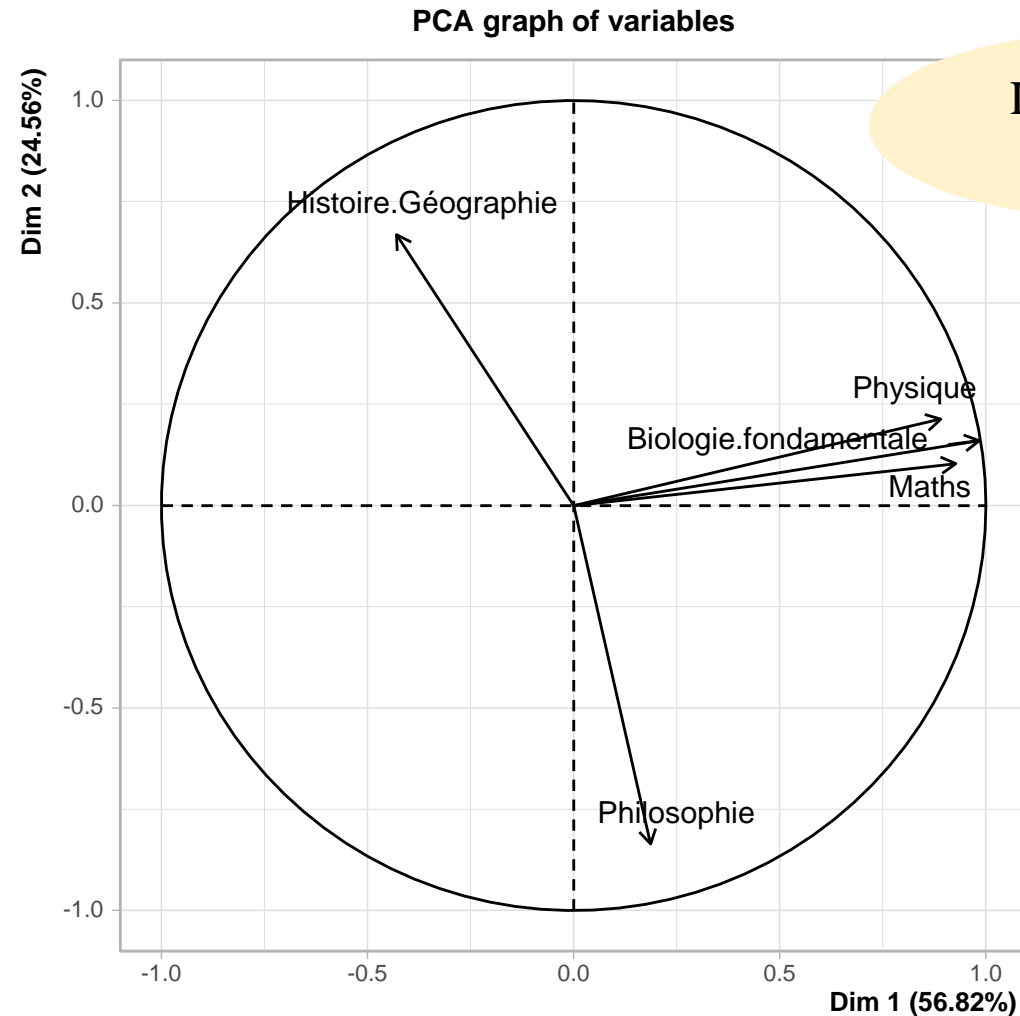
```
> res.pca.raw$var$coord
```

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 | Dim.5 |
|-----------------------|------------|------------|------------|---------------|---------------|
| Maths | 0.9271567 | 0.1032435 | 0.08920569 | -0.3489464200 | 3.778953e-32 |
| Histoire.Géographie | -0.4298643 | 0.6691947 | 0.60613132 | -0.0000828758 | 1.672270e-47 |
| Philosophie | 0.1866839 | -0.8357073 | 0.51396389 | 0.0508290836 | 7.838766e-48 |
| Physique | 0.8907839 | 0.2135242 | 0.01972722 | 0.4006524254 | 2.643445e-32 |
| Biologie.fondamentale | 0.9839317 | 0.1603253 | 0.06537553 | -0.0435917477 | -5.954092e-32 |

Etude de la corrélation des variables avec les composantes principales

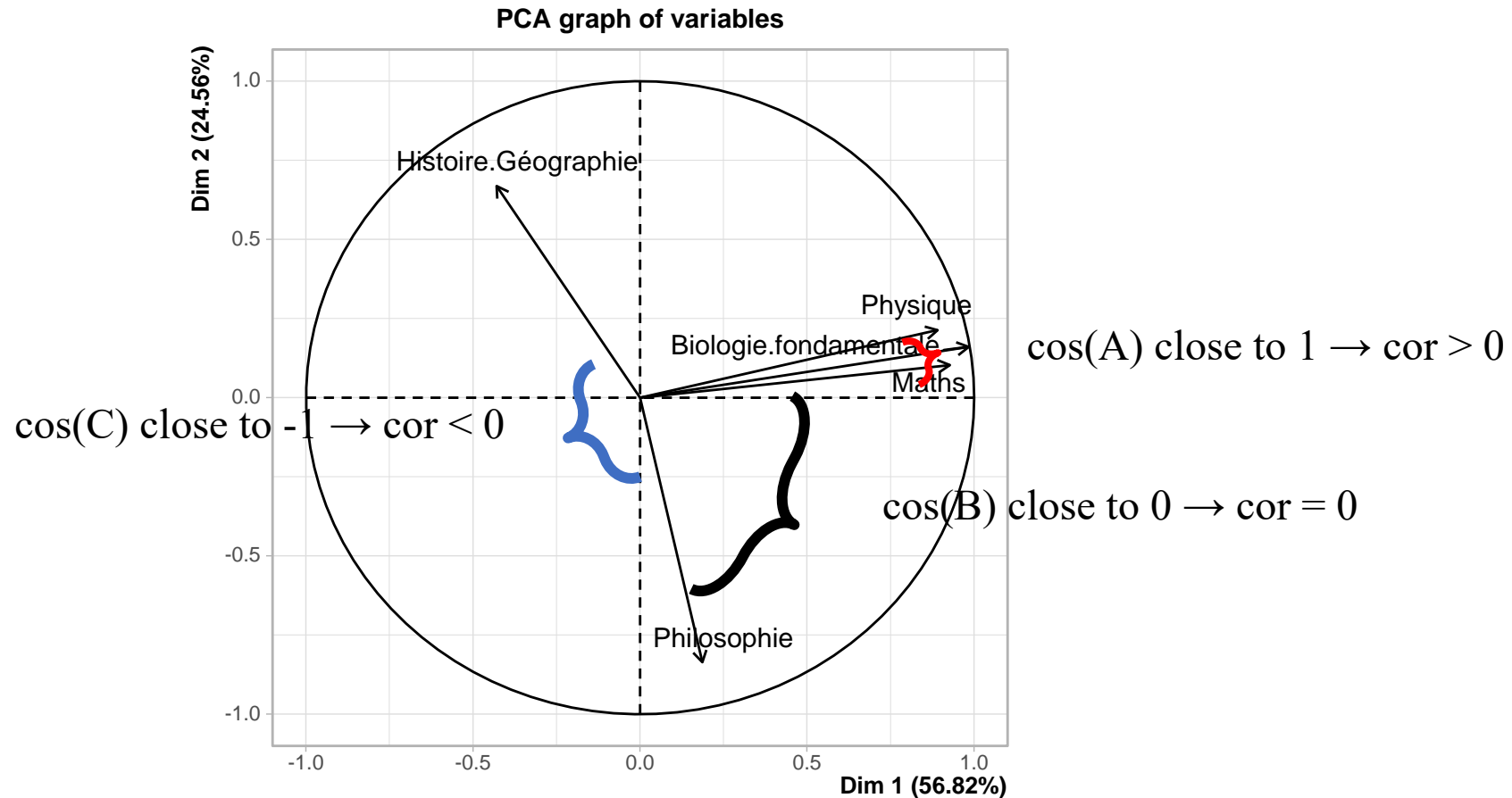
Etude des variables

- Représentation des variables dans espaces déterminés par les composantes
- Coordonnées de la variable = corrélation entre la variable et chaque composante



Etude des variables

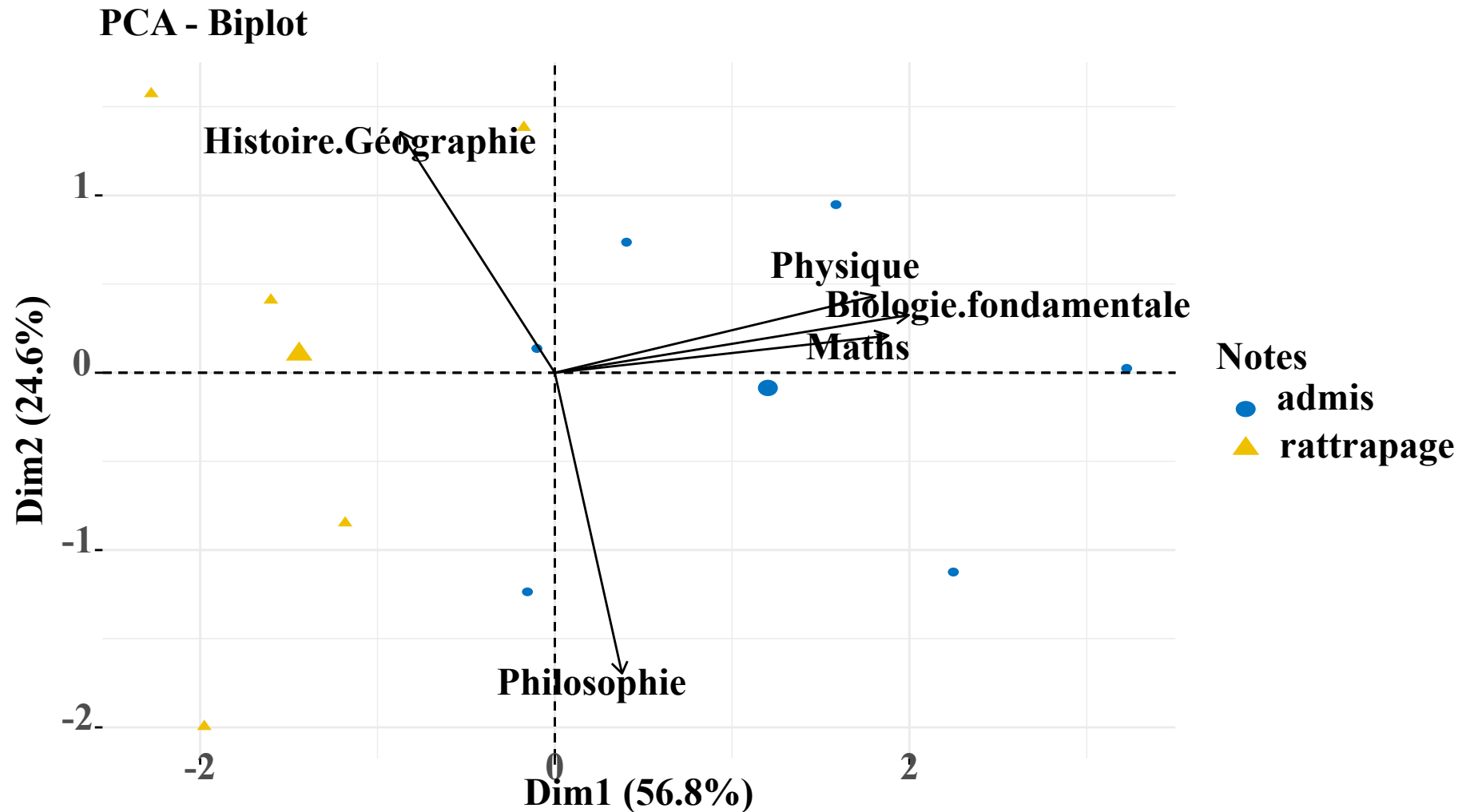
- Représentation des variables dans espaces déterminés par les composantes
- Coordonnées de la variable = corrélation entre la variable et chaque composante



Visualisation de la corrélation entre les variables

Identification des groupes de variables corrélées entres elles

Biplot/Analyse simultanée



Un individu qui se trouve du même côté d'une variable donnée a une valeur élevée pour cette variable ;
Un individu qui se trouve sur le côté opposé d'une variable donnée a une faible valeur pour cette variable.

Exemple biologique simple

| | penicillin | streptomycin | neomycin | gramstain |
|-----------------------------------|------------|--------------|----------|-----------|
| Aerobacter aerogenes | 870.000 | 1.00 | 1.600 | neg |
| Brucella abortus | 1.000 | 2.00 | 0.020 | neg |
| Escherichia coli | 100.000 | 0.40 | 0.100 | neg |
| Klebsiella pneumoniae | 850.000 | 1.20 | 1.000 | neg |
| Mycobacterium tuberculosis | 800.000 | 5.00 | 2.000 | neg |
| Proteus vulgaris | 3.000 | 0.10 | 0.100 | neg |
| Pseudomonas aeruginosa | 850.000 | 2.00 | 0.400 | neg |
| Salmonella typhosa | 1.000 | 0.40 | 0.008 | neg |
| Salmonella schottmuelleri | 10.000 | 0.80 | 0.090 | neg |
| Bacillus anthracis | 0.001 | 0.01 | 0.007 | pos |
| Diplococcus pneumoniae | 0.005 | 11.00 | 10.000 | pos |
| Staphylococcus albus | 0.007 | 0.10 | 0.001 | pos |
| Staphylococcus aureus | 0.030 | 0.03 | 0.001 | pos |
| Streptococcus fecalis | 1.000 | 1.00 | 0.100 | pos |
| Streptococcus hemolyticus | 0.001 | 14.00 | 10.000 | pos |
| Streptococcus viridans | 0.005 | 10.00 | 40.000 | pos |

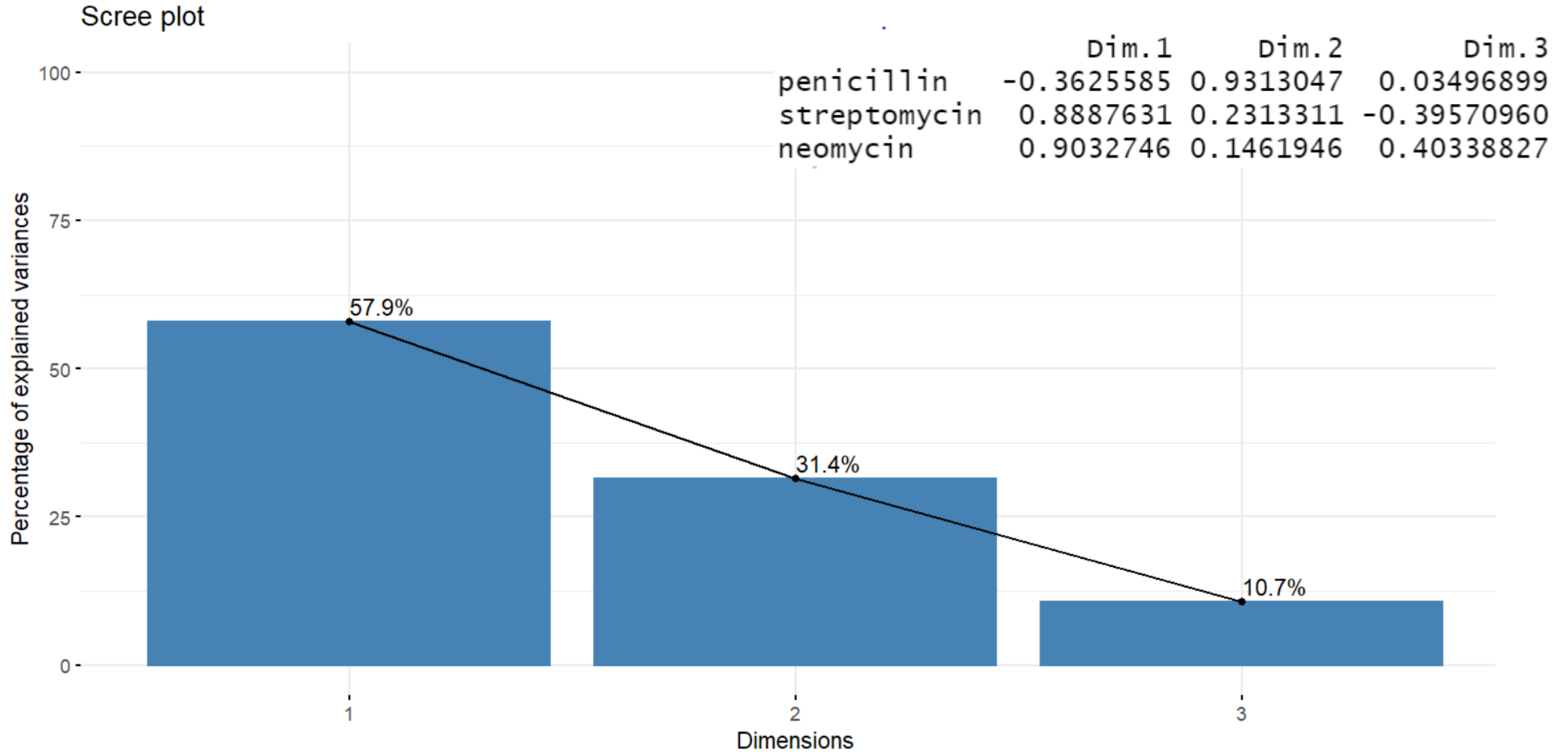
- Données provenant du package R « Lucid »
- “Effectiveness of 3 antibiotics against 16 bacterial species”
- “16 observations on the following 5 variables”

Exemple biologique simple

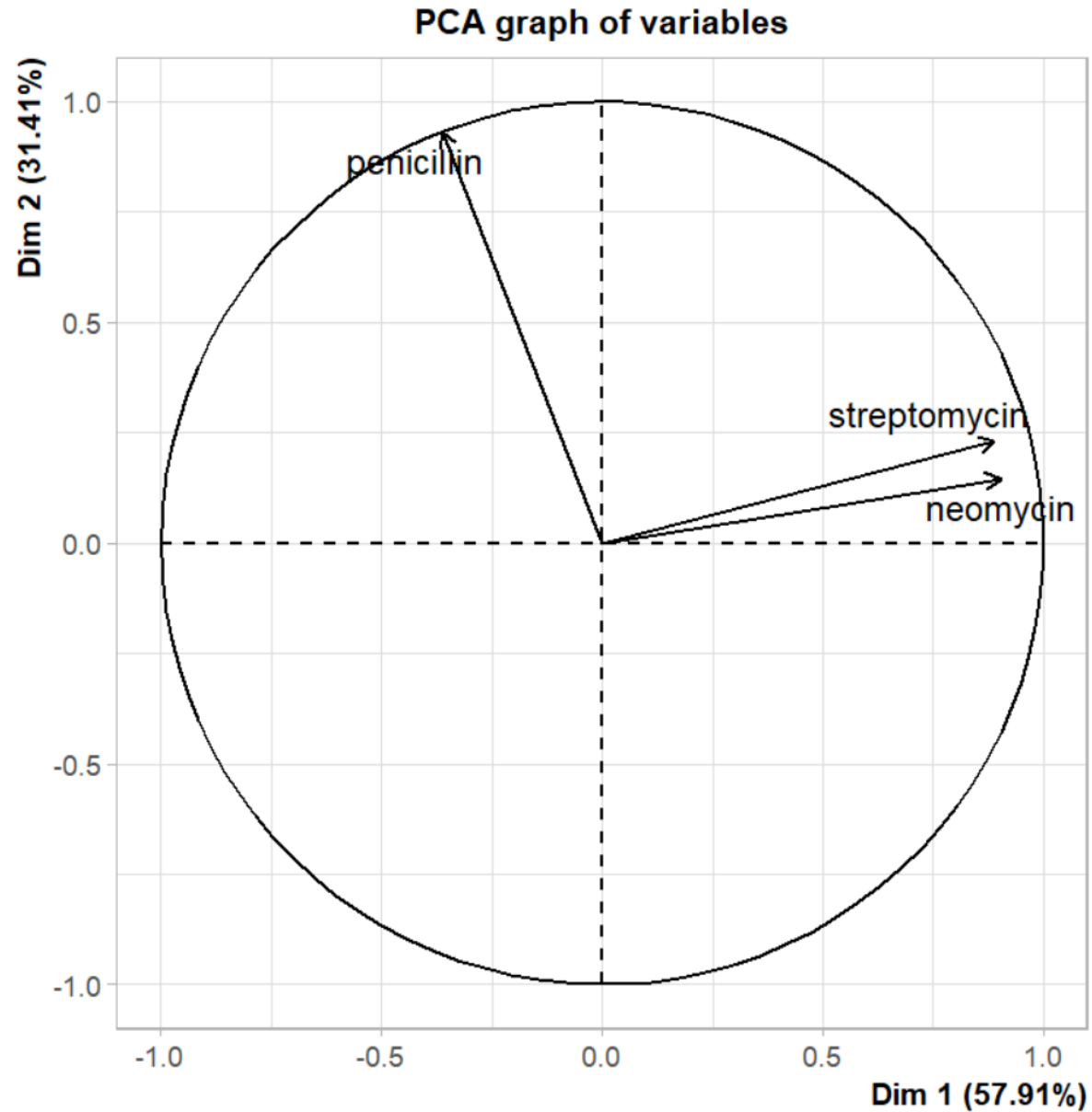
- Définir les variables et les individus ?
- Quelles sont les variables à garder pour la PCA ?

| | penicillin | streptomycin | neomycin | gramstain |
|-----------------------------------|------------|--------------|----------|-----------|
| Aerobacter aerogenes | 870.000 | 1.00 | 1.600 | neg |
| Brucella abortus | 1.000 | 2.00 | 0.020 | neg |
| Escherichia coli | 100.000 | 0.40 | 0.100 | neg |
| Klebsiella pneumoniae | 850.000 | 1.20 | 1.000 | neg |
| Mycobacterium tuberculosis | 800.000 | 5.00 | 2.000 | neg |
| Proteus vulgaris | 3.000 | 0.10 | 0.100 | neg |
| Pseudomonas aeruginosa | 850.000 | 2.00 | 0.400 | neg |
| Salmonella typhosa | 1.000 | 0.40 | 0.008 | neg |
| Salmonella schottmuelleri | 10.000 | 0.80 | 0.090 | neg |
| Bacillus anthracis | 0.001 | 0.01 | 0.007 | pos |
| Diplococcus pneumoniae | 0.005 | 11.00 | 10.000 | pos |
| Staphylococcus albus | 0.007 | 0.10 | 0.001 | pos |
| Staphylococcus aureus | 0.030 | 0.03 | 0.001 | pos |
| Streptococcus fecalis | 1.000 | 1.00 | 0.100 | pos |
| Streptococcus hemolyticus | 0.001 | 14.00 | 10.000 | pos |
| Streptococcus viridans | 0.005 | 10.00 | 40.000 | pos |

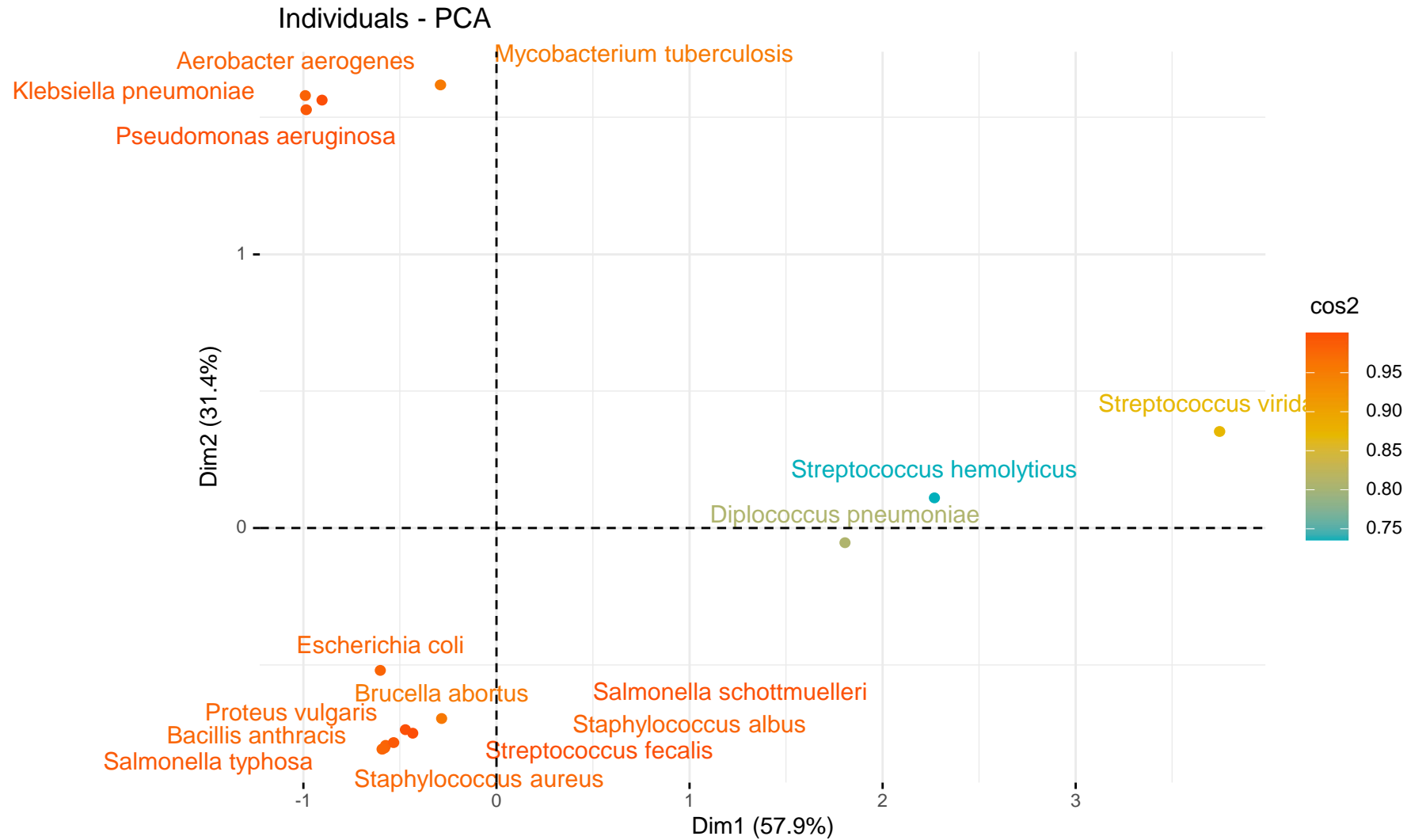
Exemple biologique simple



Exemple biologique simple

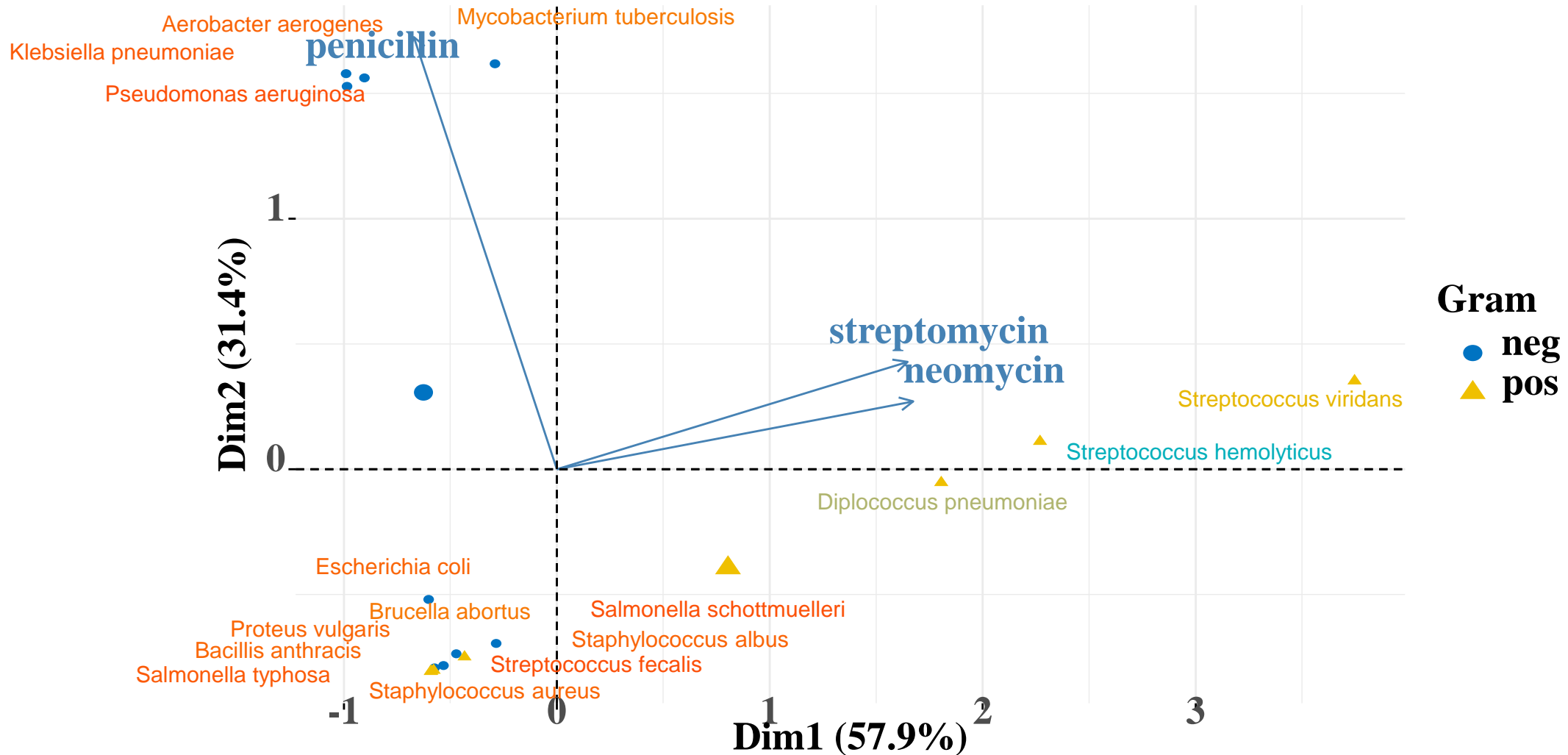


Exemple biologique simple



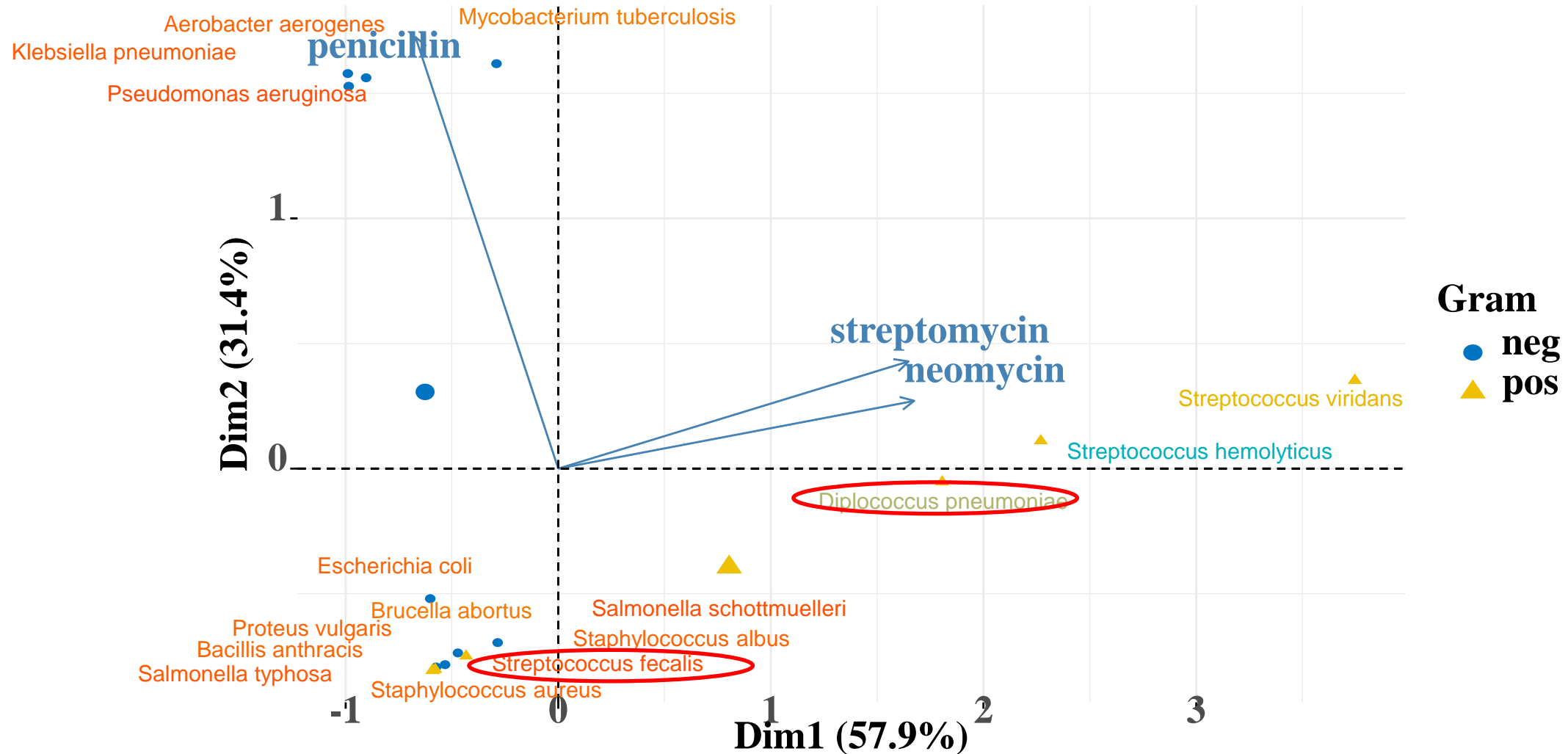
Exemple biologique simple

PCA - Biplot



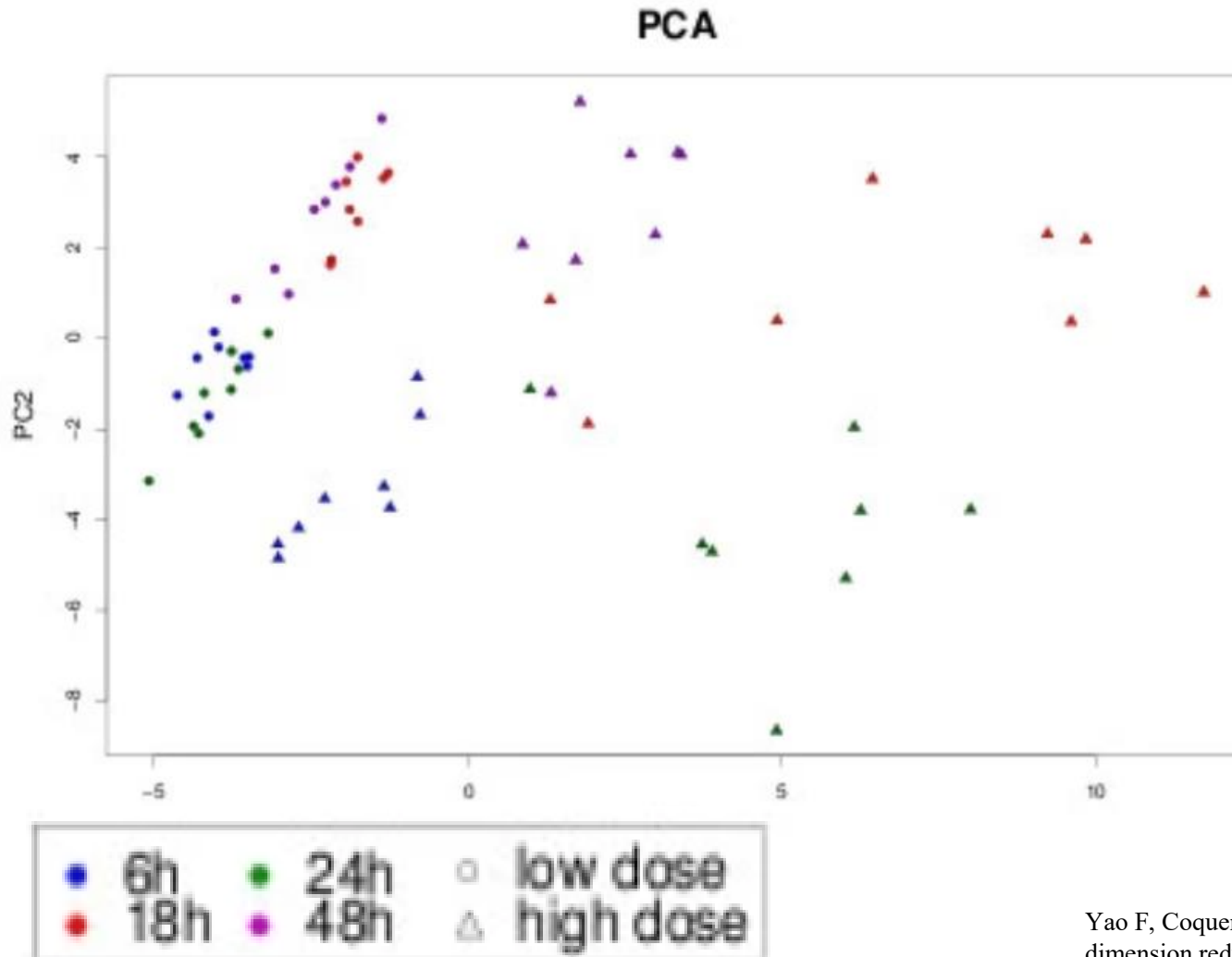
Exemple biologique simple

PCA - Biplot



Exemples biologique (1)

Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets

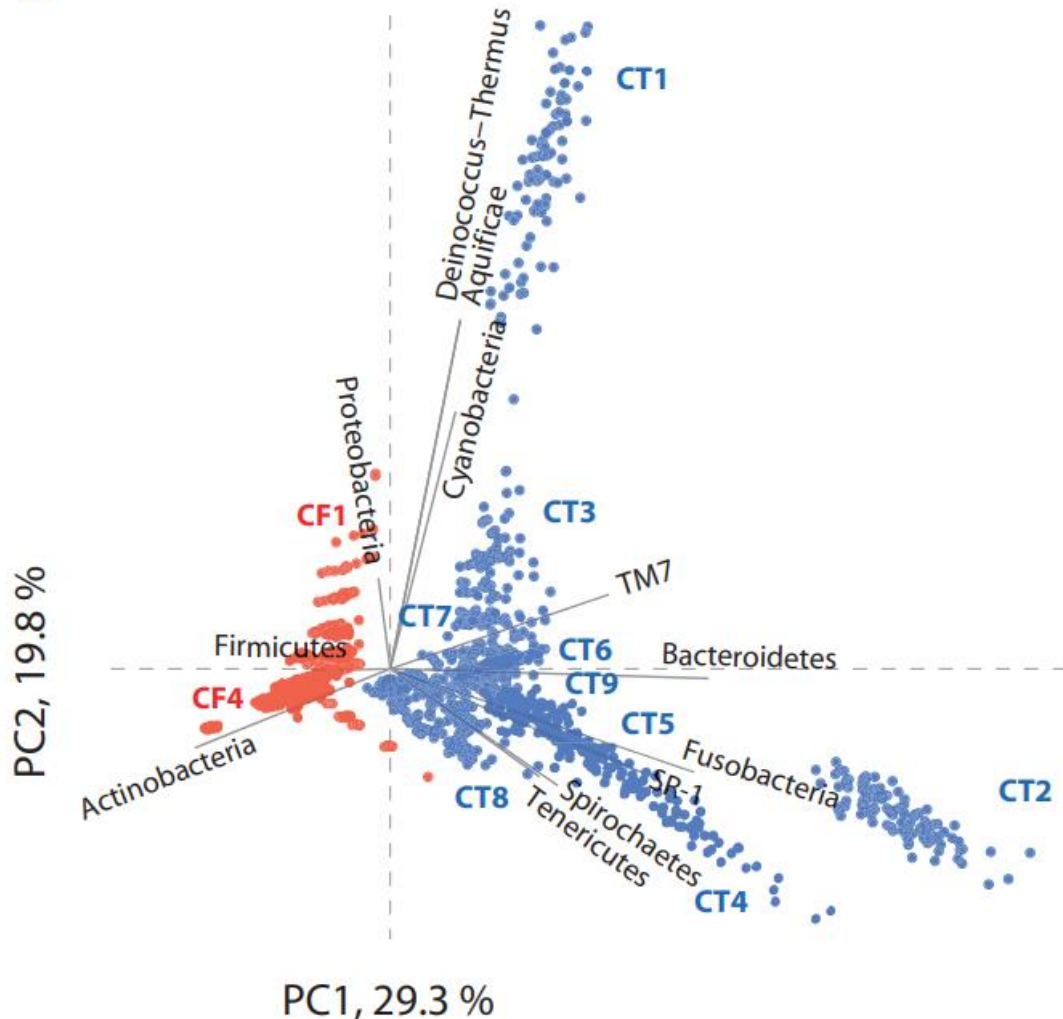


- Echantillons : rats (n = 64)
- Acétamonophène à différentes doses
- Analyse des données : étude transcriptomique

Exemples biologique (2)

Quantitative Analysis of the Human Airway Microbial Ecology Reveals a Pervasive Signature for Cystic Fibrosis

A



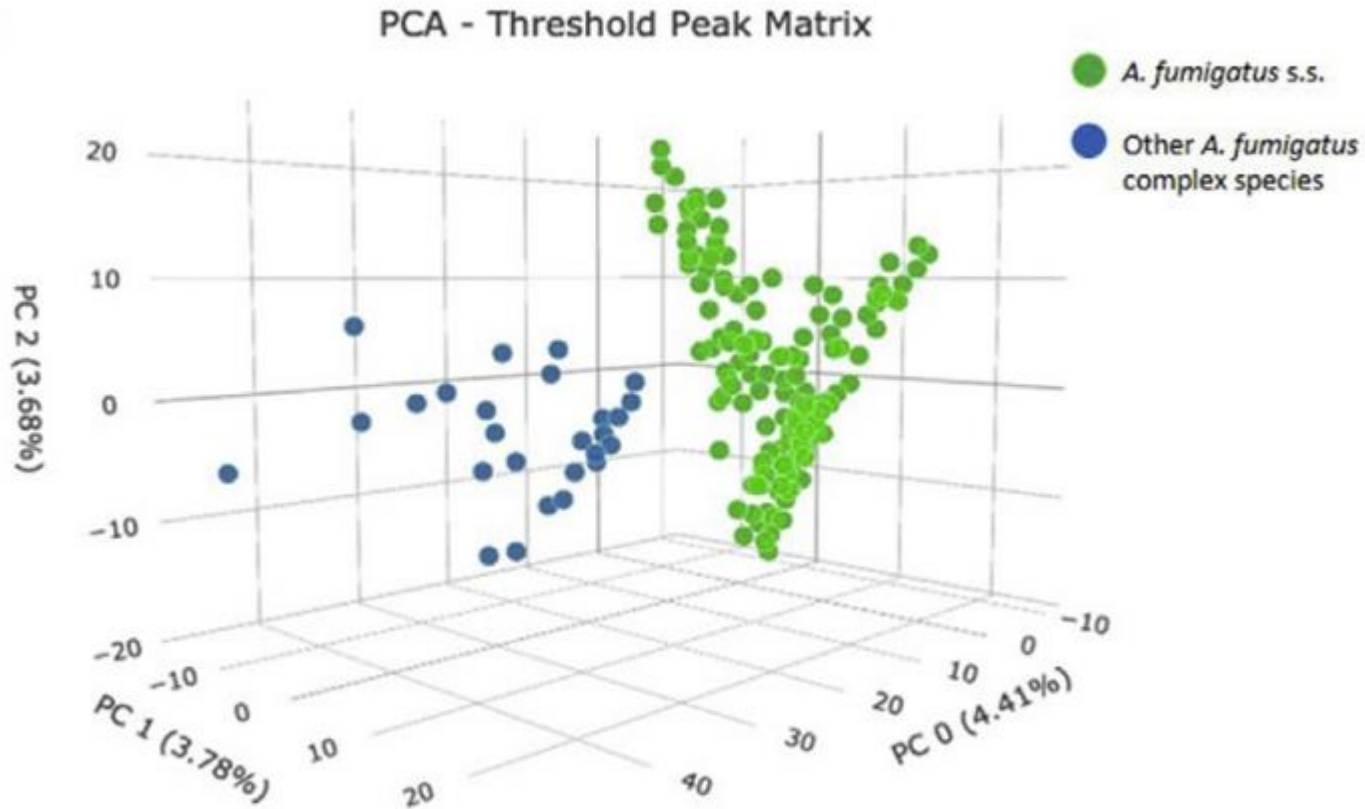
- Echantillons : crachats (n = 25)
 - patients sains (n = 9)
 - Patients mucoviscidose (n = 16)
- Analyse des données : NGS

Blainey PC, Milla CE, Cornfield DN, Quake SR. Quantitative analysis of the human airway microbial ecology reveals a pervasive signature for cystic fibrosis. *Sci Transl Med*. 2012;4(153):153ra130. doi:10.1126/scitranslmed.3004458

Exemples biologique (3)

Detection of azole resistance in *Aspergillus fumigatus* complex isolates using MALDI-TOF mass spectrometry

A



- Echantillons : souches de *A. fumigatus*
- Analyse des données : MALDI-TOF SM

Discrimination of *Aspergillus fumigatus* sensu stricto from the cryptic species of the *Aspergillus fumigatus* complex

Zvezdanova ME, Arroyo MJ, Méndez G, Candela A, Mancera L, Rodríguez JG, Serra JL, Jiménez R, Loza I, Castro C, López C, Muñoz P, Guinea J, Escribano P, Rodríguez-Sánchez B; ASPEIN group. Detection of azole resistance in *Aspergillus fumigatus* complex isolates using MALDI-TOF mass spectrometry. Clin Microbiol Infect. 2022 Feb;28(2):260-266. doi: 10.1016/j.cmi.2021.06.005. Epub 2021 Jun 17. PMID: 34147673.

Plan

Partie I : L'Analyse en Composantes Principales (ACP)

Partie II : Clustering



Clustering

Rassembler les objets en groupes ou clusters :

- (i) **homogènes** : notion de similarité au sein d'un groupe ou cluster
- (ii) **séparés** : notion de différence ou dissimilarités entre les groupes ou clusters







Un cluster est une collection d'objets similaires au sein du même cluster mais dissimilaires aux objets appartenant aux autres clusters

Un exemple biologique

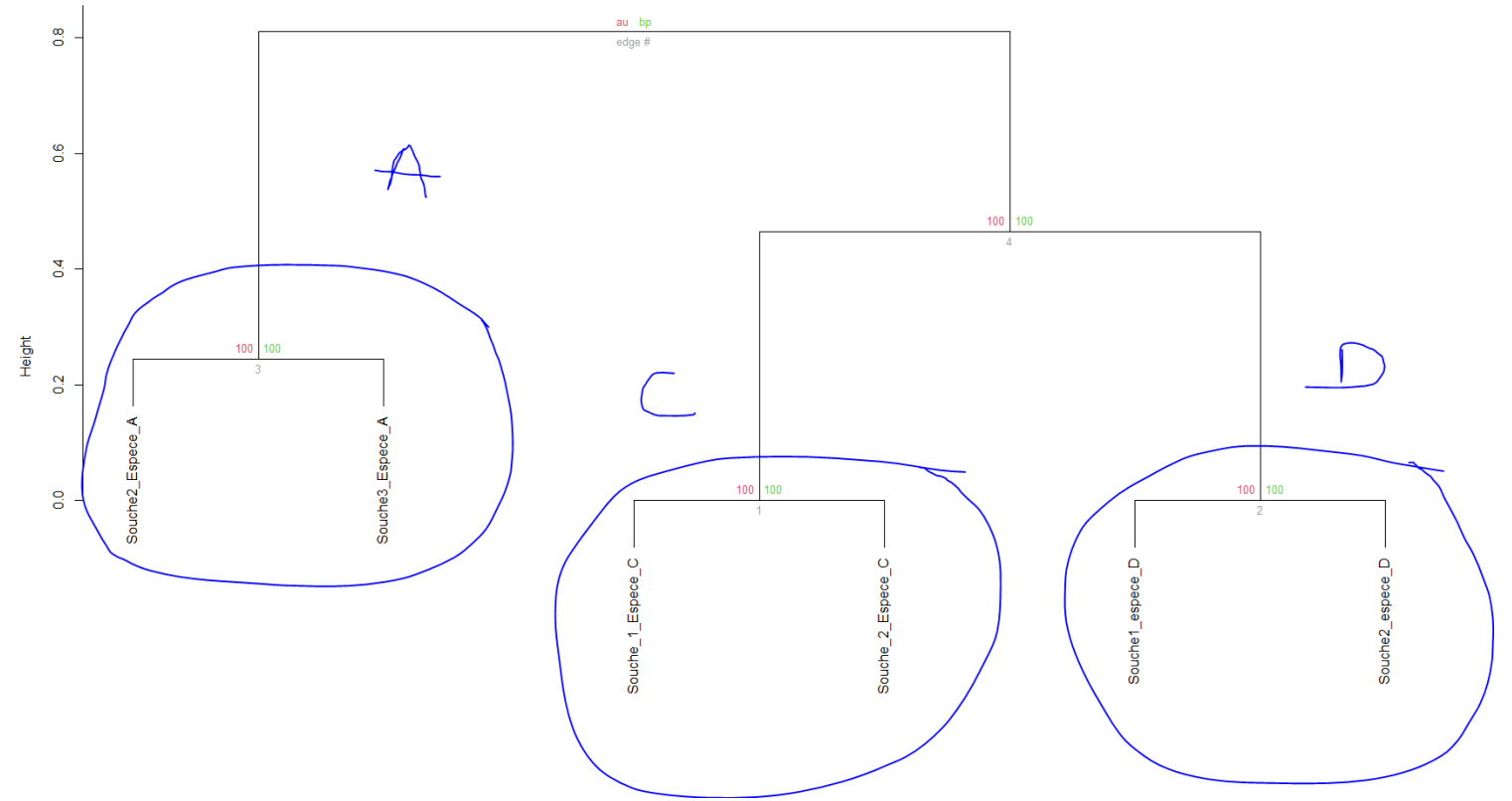
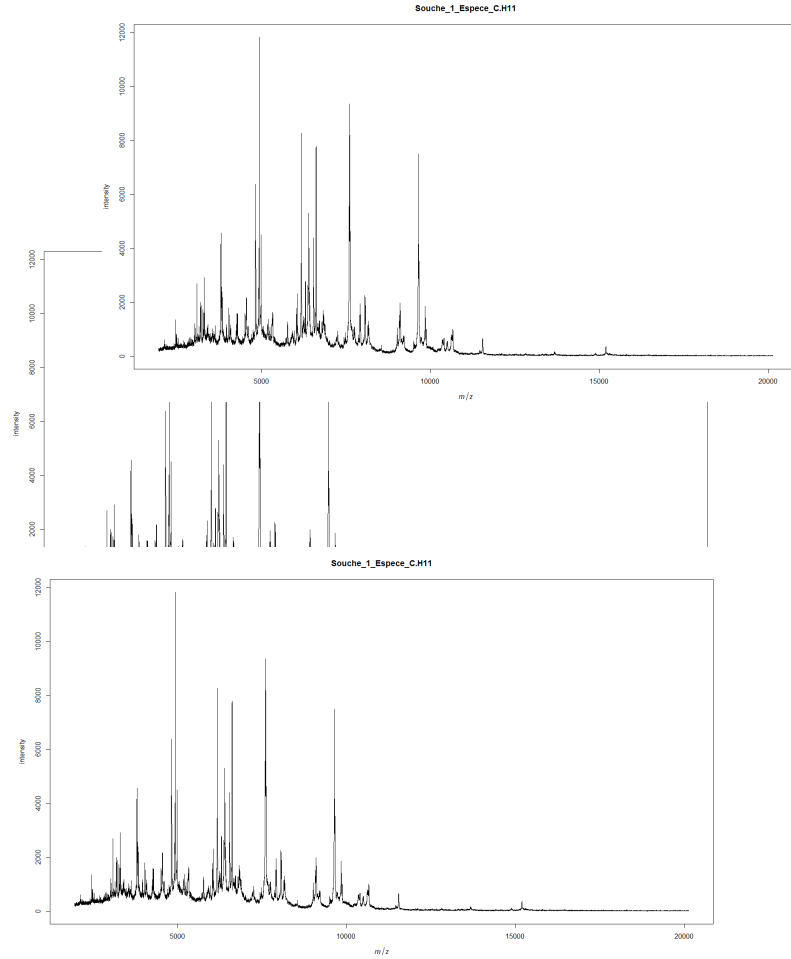
On dispose de souches bactériennes analysées en spectrométrie de masse de type MALDI-TOF

Les 6 souches ont été caractérisée génétiquement en trois espèces distinctes (espèces : A, C et D)

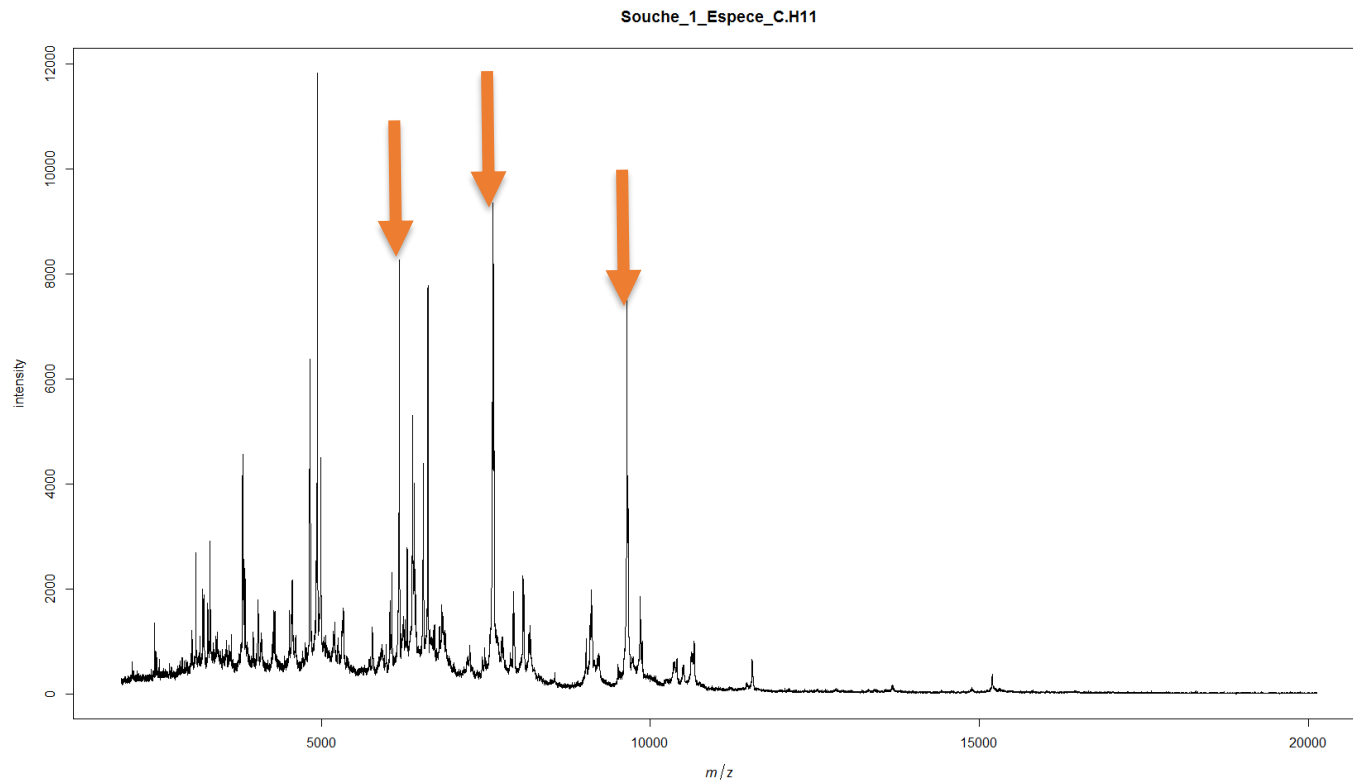
Problématique : existe-t-il une relation entre la spectrométrie de masse et la génétique pour l'identification de ces souches ?

| | | |
|---|------------------|---------------------|
|  Souche_1_Espece_C | 27/06/2020 12:09 | Dossier de fichiers |
|  Souche_2_Espece_C | 27/06/2020 12:02 | Dossier de fichiers |
|  Souche1_espece_D | 27/06/2020 12:02 | Dossier de fichiers |
|  Souche2_Espece_A | 27/06/2020 11:59 | Dossier de fichiers |
|  Souche2_espece_D | 27/06/2020 12:08 | Dossier de fichiers |
|  Souche3_Espece_A | 27/06/2020 11:59 | Dossier de fichiers |

Comment réaliser un clustering

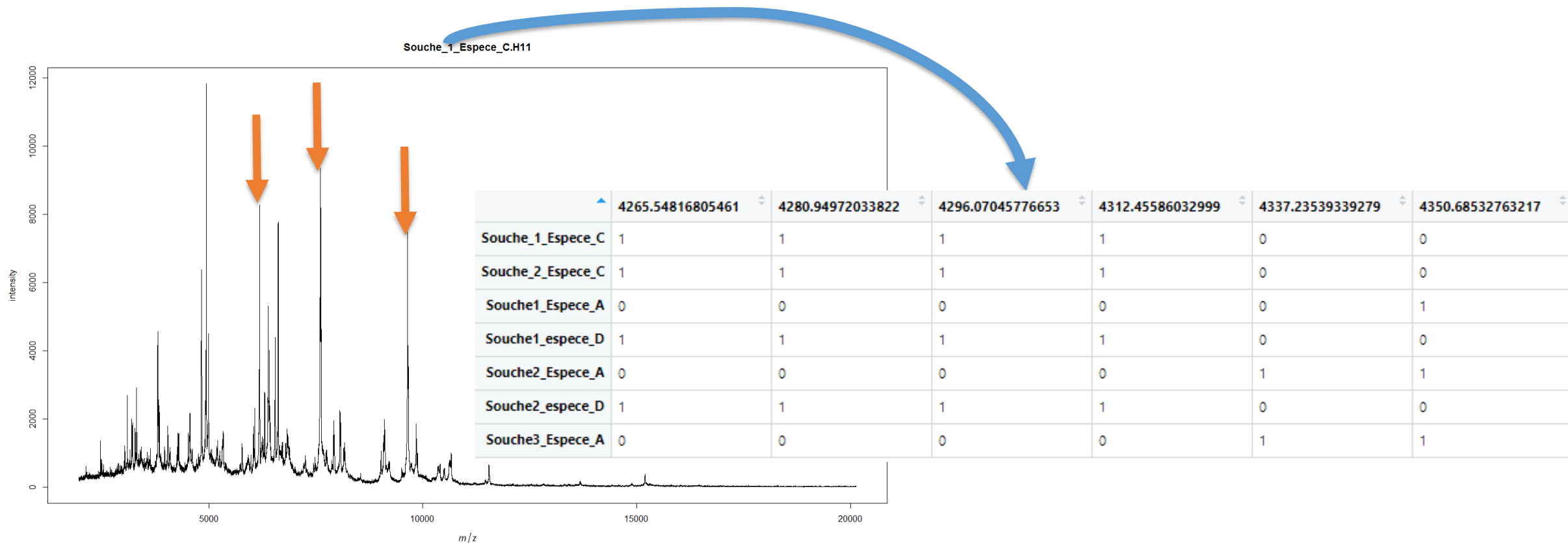


Comment réaliser un clustering



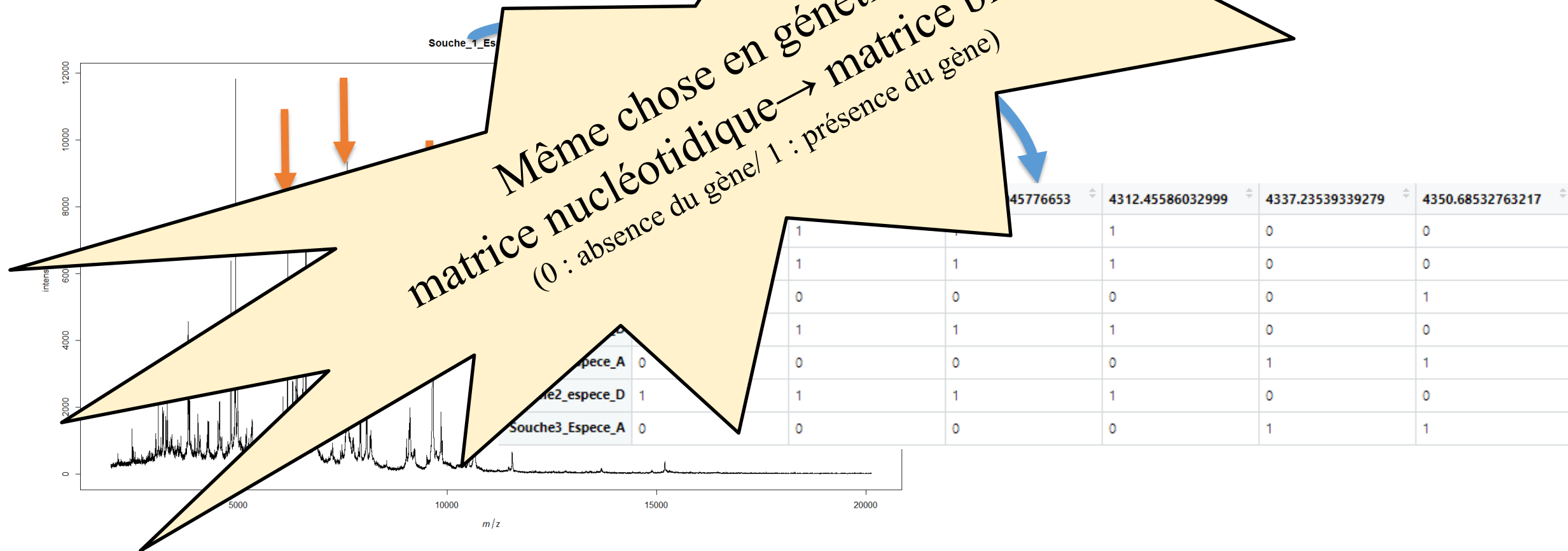
Transformation du spectre en matrice binaire (0 : absence du pic/ 1: présence du pic)

Comment réaliser un clustering



Transformation du spectre en matrice binaire (0 : absence du pic/ 1: présence du pic)

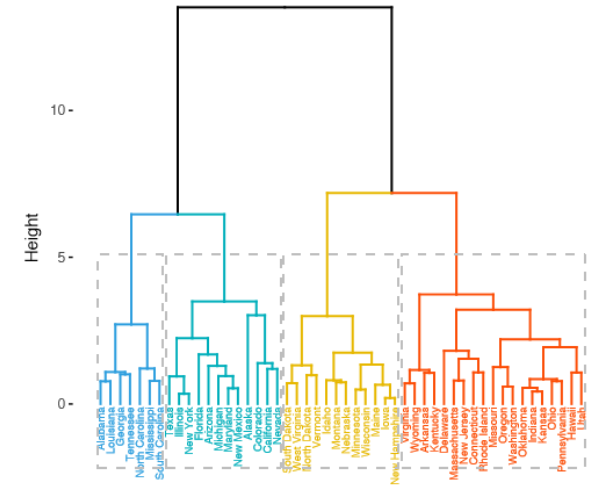
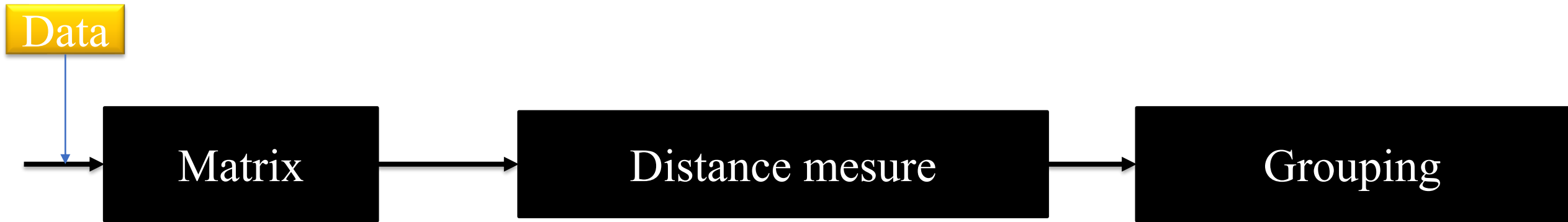
Comment réaliser un clustering



Transformation du spectre en matrice binaire (0 : absence du pic/ 1 : présence du pic)

Mode opératoire

- Pipeline



Mode opératoire

Matrix

Distance mesure

Grouping (linkage)

- File .csv (comma delimited)

| name | bike | natation | long running | short_running |
|----------|------|----------|--------------|---------------|
| PAUL | 11.8 | 18.9 | 20.9 | |
| JULIETTE | 13.4 | 67.0 | 28.1 | 30.7 |
| AGATHE | 6.6 | 23.2 | 27.3 | 32.8 |
| PIERRE | 8.9 | | 24.9 | 41.7 |
| MICHEL | 10.3 | 10.8 | 20.0 | 18.9 |
| FLORE | 30.3 | 23.7 | 24.2 | 13.9 |
| JEAN | 11.2 | 20.7 | 27.9 | |

Wrong data

Bad name

Mode opératoire

Matrix

Distance mesure

Grouping (linkage)

- File .csv (comma delimited)

| name | bike | natation | long_running | short_running |
|----------|------|----------|--------------|---------------|
| PAUL | 11.8 | 18.9 | 20.9 | NA |
| JULIETTE | 13.4 | 67.0 | 28.1 | 30.7 |
| AGATHE | 6.6 | 23.2 | 27.3 | 32.8 |
| PIERRE | 8.9 | NA | 24.9 | 41.7 |
| MICHEL | 10.3 | 10.8 | 20.0 | 18.9 |
| FLORE | 30.3 | 23.7 | 24.2 | 13.9 |
| JEAN | 11.2 | 20.7 | 27.9 | NA |

Observation

Variables

Mode opératoire

→ **Matrix** →

Distance mesure →

Grouping (linkage)

Matrice de distance

- Euclidean distance :
$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- Manhattan distance :
$$d_{man}(x, y) = \sum_{i=1}^n |(x_i - y_i)|$$

- Pearson correlation distance
$$d_{cor}(x, y) = 1 - \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Degré de relation linéaire entre
deux profils

- Kendall correlation distance
$$d_{kend}(x, y) = 1 - \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

- n_c : total number of concordant pairs
- n_d : total number of discordant pairs
- n : size of x and y

Mode opératoire



- Quelle distance choisir ?

gene expression data analysis ► correlation based distance
gene presence or peak presence ► binary method

Matrix

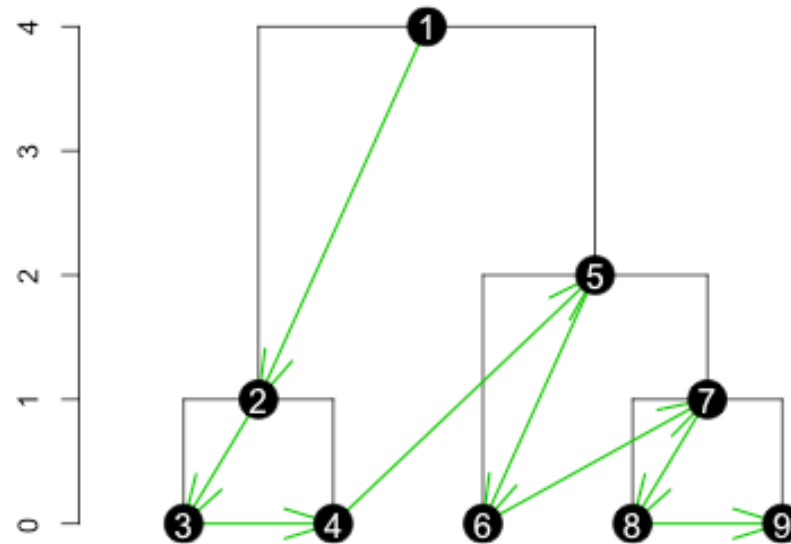
Distance measure

Grouping (linkage)

Clustering

- Single Link
- Complete Link
- ...

Nodes order when using
Depth-first search in a dendrogram





| | A | B | C | D |
|----------|-----|-----|-----|-----|
| Gene_X | 0 | 0 | 1 | 0 |
| Gene_Y | 0 | 0 | 1 | 1 |
| Gene_Z | 0 | 0 | 0 | 1 |
| Gene_T | 0 | 0 | 1 | 1 |
| Gene_... | ... | ... | ... | ... |



| | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 5 | 2 | 8 |
| B | | 0 | 5 | 7 |
| C | | | 0 | 8 |
| D | | | | 0 |



| | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 5 | 2 | 8 |
| B | | 0 | 5 | 7 |
| C | | | 0 | 8 |
| D | | | | 0 |

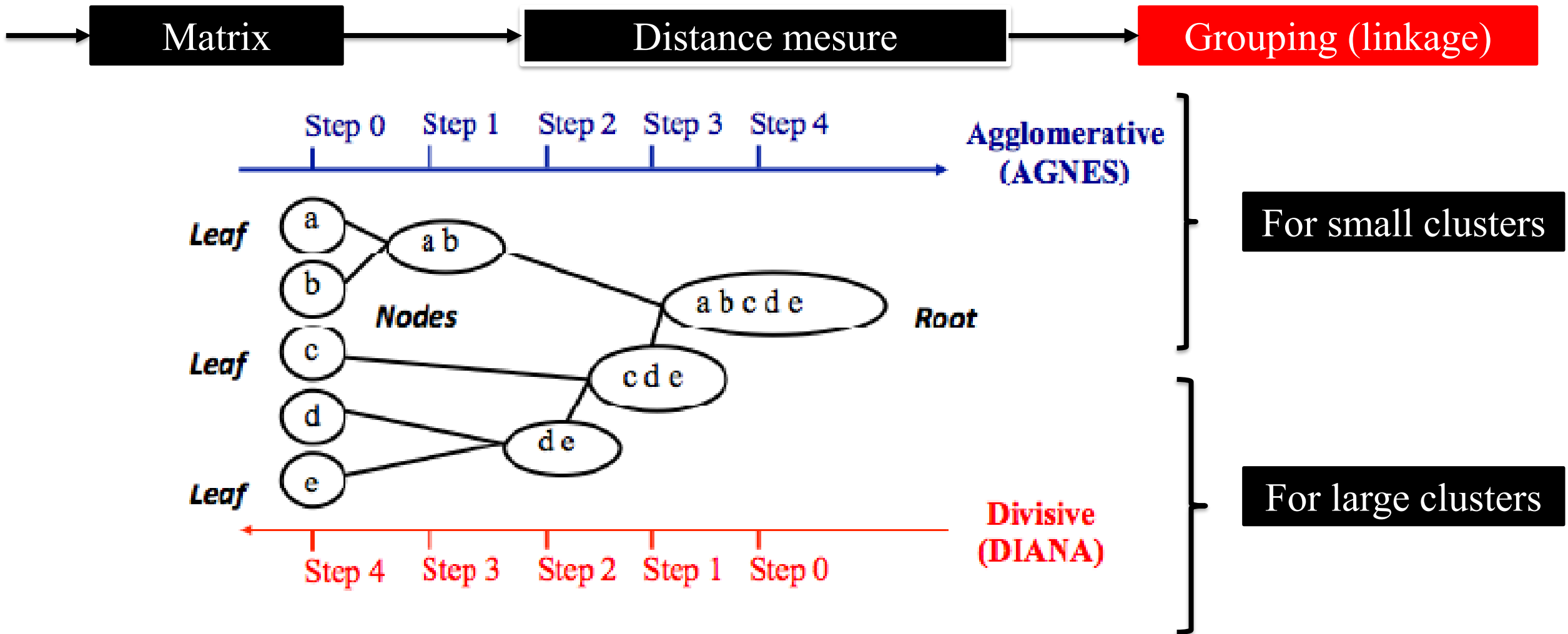


| | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 5 | 2 | 8 |
| B | | 0 | 4 | 1 |
| C | | | 0 | 2 |
| D | | | | 0 |



| | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 5 | 2 | 8 |
| B | | 0 | 4 | 1 |
| C | | | 0 | 2 |
| D | | | | 0 |

Hierarchical clustering (HCA)



```
graph LR; A[ ] --> B[Matrix]; B --> C[Distance measure]; C --> D[Grouping (linkage)];
```

Matrix

Distance measure

Grouping (linkage)

- **Maximum or *complete linkage*:** The distance between two clusters is defined as the maximum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce more compact clusters.
- **Minimum or *single linkage*:** The distance between two clusters is defined as the minimum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce long, “loose” clusters.
- **Mean or *average linkage*:** The distance between two clusters is defined as the average distance between the elements in cluster 1 and the elements in cluster 2 = **UPGMA**
- ***Centroid linkage*:** The distance between two clusters is defined as the distance between the centroid for cluster 1 (a mean vector of length p variables) and the centroid for cluster 2 = **UPGMC**
- ***Ward's minimum variance method*:** It minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged.

→ Matrix

→ Distance measure

→ Grouping (linkage)

- **Maximum or *complete linkage*:** The distance between two clusters is defined as the maximum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce more compact clusters.
- **Minimum or *single linkage*:** The distance between two clusters is defined as the minimum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce long, “loose” clusters.
- **Mean or *average linkage*:** The distance between two clusters is defined as the average distance between the elements in cluster 1 and the elements in cluster 2 = **UPGMA**
- ***Centroid linkage*:** The distance between two clusters is defined as the distance between the centroid for cluster 1 (a mean vector of length p variables) and the centroid for cluster 2 = **UPGMC**
- ***Ward's minimum variance method*:** It minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged.

Evaluer la solidité de l'arbre

Cophenetic
distance

- The linking of objects in the cluster tree should have a strong correlation with the distances between objects in the original distance matrix
- **>0,75: acceptable correlation**

