Projet Hydrogen

Abdoulaye Diabakhaté

31 mai 2018

Table des matières

1 Analyse multivariée non paramétrique à partir de matrices de distances

1.1 Analyse multivariée classique

Soit un échantillon de taille n d'observations individuelles, indicées par i=1,...,n réalisations de variables aléatoires (y_i,x_i).

- y_i variable continue prenant ses valeurs dans $\mathbb{R}$.
- x_i variables en nombre K, de type quelconque.

1.2 Notations

- On confond les variables aléatoires et leurs réalisations
- On réserve les majuscules pour des vecteurs

1.2.1 Exemple :

$Y = (y_1, ..., y_n)'$

1.3 Définitions

La variable dépendante y_i s'écrit comme : $y_i = x_i\beta + u_i$

- β est un paramètre à estimer
- Le modèle est linéaire en β

2 Écriture matricielle du modèle : $Y = X\beta + U$

$$\text{Avec X=} \begin{pmatrix} \mathrm{x}_{11} & \cdots & \mathrm{x}_{1K} \\ \vdots & \ddots & \vdots \\ \mathrm{x}_{n1} & \cdots & \mathrm{x}_{nK} \end{pmatrix}$$

2.1 Propriétés

D'après le théorème de Gauss-Markov,qui énonce que dans un modèle linéaire dans lequel les erreurs ont une espérance nulle,sont non corrélées et dont les variances sont égales(homoscédasticité),le meilleur estimateur linéaire non biaisé des coefficients est son estimateur par les moindres carrés.

$\hat{\beta}= \underset{\beta}{\operatorname{argmin}}||Y - X\beta|| = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - x_i\beta)^2$

Donc le meilleur estimateur linéaire en Y et sans biais est :

$$\hat{\beta}=(X'X)^{-1}X'Y$$

En effet :

$||Y-X\beta||^2$=$(Y-X\beta)^{'}(Y-X\beta)$

$||Y-X\beta||^2$=$Y^{'}Y-Y^{'}X\beta-X^{'}\beta^{'}Y-X^{'}\beta^{'}X\beta$

Avec : $Y=X\beta$ alors $Y^{'}=X^{'}\beta^{'}$.

Par suite : $X^{'}\beta^{'}Y=Y^{'}X\beta$ et $X^{'}\beta^{'}X\beta$=$\beta^{'}X'X\beta$

$||Y-X\beta||^2$=$Y^{'}Y-2Y^{'}X\beta+\beta^{'}X'X\beta$

Aussi :

$$\frac{\partial}{\partial\beta}Y^{'}Y=0\,;$$
$$\frac{\partial}{\partial\beta}2Y'X\beta=2Y^{'}X$$

$$\frac{\partial}{\partial\beta}\beta^{'}X'X\beta=2\beta^{'}X'X$$

Ainsi :

$\frac{\partial}{\partial\beta}||Y-X\beta||^2=-2Y^{'}X+2\beta^{'}X^{'}X=0$

Donc on a :

$\beta=\hat{\beta}=(X^{'}X)^{-1}X'Y$

La matrice des valeurs prédites est :$\hat{Y}=X\hat{\beta}=HY$

Avec H=X(X$'X)^{-1}X'$

La matrice des résidus est : $R=Y-\hat{Y}=(I-H)Y$

La matrice totale SSCP est alors décomposée par les matrices SSCP prédites et résiduelles de la façon suivante :

$Y^{'}Y=\hat{Y}'\hat{Y}+R^{'}R$

où : $S_T=tr(Y^{'}Y)\,;\,S_H=tr(\hat{Y^{'}}\hat{Y})\,;\,S_R=tr(R'R)$

Une statistique appropriée pour tester l'hypothèse nulle de l'absence d'effet des paramètres du modèle est la pseudo statistique F :

$F=\frac{tr(\hat{Y^{'}}\hat{Y})/(m-1)}{tr(R'R)/(n-m)}$

où m est le nombre de paramètres.

Pour deux matrices $A_{n,p}$ et $B_{n,p}$: tr(AB)=tr(BA)

On a : $YY^{'} = \hat{Y}\hat{Y}' + RR^{'}$

Donc : $tr(YY^{'}) = tr(\hat{Y}\hat{Y}') + tr(RR^{'})$

La matrice H est symétrique :

En effet, on a : $H = X(X^{'}X)^{-1}X^{'}$

$H^{'} = [X(X^{'}X)^{-1}X^{'}]^{'} = H$

$$\hat{Y}\hat{Y}' = HY(HY)^{'} = H(YY^{'})H$$
$$RR^{'} = (I-H)Y.Y^{'}(I-H)^{'} = (I-H)(YY^{'})(I-H)$$

3 Analyse multivariée sur base de distances

Soit $D = (d_{ij})$, une matrice de distance de taille n*n.

Posons $A = (a_{ij}) = (\frac{-1}{2}d_{ij}^2)$.

Nous pouvons alors calculer la matrice centrée de Gower G en centrant les éléments de A :

$G = (\mathbb{I} - \frac{1}{n}11^{'})A(\mathbb{I} - \frac{1}{n}11^{'})$.

Avec 1 est une colonne de taille n, contenant uniquement des 1.

Ainsi en remplaçant (YY$^{'}$)$parG, nousavonsS_T = tr(G)$ et la pseudo statistique F est :

$F = \frac{tr(HGH)/(m-1)}{tr[(\mathbb{I}-H)G(\mathbb{I}-H)]/(n-m)}$

4 Distances entre valeurs prédites

Soit Y le jeu de données centré contenant toutes nos observations.

Nous nous plaçon dans un cadre linéaire, c'est-à-dire : $Y = X\theta + \epsilon$, où X est la matrice de design contenant les covariables.

Nous disposons uniquement de la matrice de distances $D^2 = (||Y_{i,.} - Y_{j,.}||^2)_{i,j}$, de métrique variable(jaccard,braycurtis,...).

Supposons que D est une matrice de distances euclidiennes. Notre objectif est de calculer une matrice de distances entre valeurs prédites (ou ajustées) à partir de la prédication $\hat{Y}$ de Y :

Supposons que D est une matrice de distances euclidiennes.

Notre objectif est de calculer une matrice de distances entre valeurs prédites(ou ajustées)à partir de la prédication $\hat{Y}$ de Y : $\hat{Y} = X\hat{\theta} = HY$.

Notons par : $\hat{D^2} = (||\hat{Y_{i,.}} - \hat{Y_{j,.}}||^2)_{i,j}$,une telle matrice.

Cette dernière est calculable en fonction de :
- X et $\tilde{Y}$, où $\tilde{Y}$ correspond aux observations placées dans un cadre euclidien et obtenues à partir d'une matrice de distance quelconque à l'aide du MDS,
- X et D

Avec MDS

Nous avons ici la matrice de distances euclidiennes :

$D^2 = (||\tilde{Y_{i,.}} - \tilde{Y_{j,.}}||^2)_{i,j}$

$D^2 = (||(H\tilde{Y})_{i,.} - (H\tilde{Y})_{j,.}||^2)_{i,j}$.

Sans MDS

Calculons $(\hat{D}^2)_{i,j}$ en fonction de X et D.

$\hat{D}^2_{i,j} = ||(HY)_{i,.} - (HY)_{j,.}||^2$

$\hat{D}^2_{i,j} = ||(HY)_{i,.}||^2 + ||(HY)_{j,.}||^2 - 2 < (HY)_{i,.}, (HY)_{j,.} >$

$< (HY)_{i,.}, (HY)_{j,.} >=< H_{i,.}Y, H_{j,.}Y >$

$< (HY)_{i,.}, (HY)_{j,.} >= (H_{i,.}Y)(H_{j,.}Y)^{'}$

$< (HY)_{i,.}, (HY)_{j,.} >= H_{i,.}(YY^{'})H^{'}_{j,.}$

Avec $G = YY^{'}$ et $H^{'}_{j,.} = H_{.,j}$

Donc : $< (HY)_{i,.}, (HY)_{j,.} >= H_{i,.}GH_{.,j}$

Par ailleurs on a :

$(HY)_{i,.} = (H_{i,.}Y_{.,1}, ..., H_{i,.}Y_{.,p}) = (Y_{1,.}H_{.,i}, ..., Y_{p,.}H_{.,i})$

5 Distances entre résidus

Cette fois-ci nous cherchons à calculer la matrices de distances entre résidus,c'est-à dire :

$$D_R^2 = (||R_{i,.} - R_{j,.}||^2)_{i,j}$$

Avec : $R = Y - \hat{Y} = (I - H)Y$ est la matrice des résidus.

En nous plaçant dans un cadre euclidien, cette dernière est calculable en fonction de :
- X et $\tilde{Y}$, où $\tilde{Y}$ correspond aux observations placées dans un cadre euclidien et obtenues à partir d'une matrice de distance quelconque à l'aide du MDS.
- X et D

Avec MDS

Nous avons ici la matrice de distances euclidiennes :

$$D^2 = (||\tilde{Y}_{i,.} - \tilde{Y}_{j,.}||^2)_{i,j}$$

$D_R^2 = (||R_{i,.} - R_{j,.}||^2)_{i,j}$

$D_R^2 = (||((I-H)\tilde{Y})\{i,. - ((I-H)\tilde{Y})_{j,.}||^2)_{i,j}$

Sans MDS

Calculons $D_R^2(i,j)$ en fonction de X et D :

$D_R^2(i,j) = ||((I-H)Y)_{i,.} - ((I-H)Y)_{j,.}||^2$

$D_R^2(i,j) = ||((I-H)Y)_{i,.}||^2 + ||((I-H)Y)_{j,.}||^2 - 2 < ((I-H)Y)_{i,.}, ((I-H)Y)_{j,.} >$

Or : $((I-H)Y)_{i,.} = ((I-H)_{i,.}Y_{.,1}, ..., (I-H)_{i,.}Y_{.,p}) = (Y_{1,.}(I-H)_{.,i}, ..., Y_{p,.}(I-H)_{.,i})$

Donc : <((I-H)Y)$_{i,.}$, $((I-H)Y)_{j,.} >= \sum_{k=1}^{p}((I-H)_{i,.}Y.,k)((I-H)_{j,.}Y.,k)$

<((I-H)Y)$_{i,.}$, $((I-H)Y)_{j,.} >= \sum_{k=1}^{p}(I-H)_{i,.}Y_{.,k}Y_{k,.}(I-H)_{.,j}$

<((I-H)Y)$_{i,.}$, $((I-H)Y)_{j,.} >= \sum_{k=1}^{p}(I-H)_{i,.}Y_{.,k}Y'_{.,k}(I-H)_{.,j}$

<((I-H)Y)$_{i,.}$, $((I-H)Y)_{j,.} >= (I-H)_{i,.}(\sum_{k=1}^{p}Y_{.,k}Y'_{.,k})(I-H)_{.,j}$

<((I-H)Y)$_{i,.}$, $((I-H)Y)_{j,.} >= (I-H)_{i,.}G(I-H)_{.,j}$

Car : $G = YY'$ quand D est euclidienne.

Par suite : $D_R^2(i,j) = (I-H)_{i,.}G(I-H)_{.,i} + (I-H)_{j,.}G(I-H)_{.,j} - 2(I-H)_{i,.}G(I-H)_{.,j}$

6 Multidimensional scaling (MDS)

Le MDS («positionnement multidimensionnel») est un ensemble de techniques statistiques utilisées dans le domaine de la visualisation d'information pour explorer les similarités dans les données.
Considérons n individus. Contrairement aux chapitres précédents,on ne connait pas les observations de p variables sur ces n individus. Ces informations sont contenues dans une matrice $(n * n)D$. L'objectif du MDS (ou ACP d'un tableau de distances) est de construire, à partir de cette matrice, une représentation euclidienne des individus dans un espace de dimension réduite q, qui approche au "mieux" les indices observées.

Autrement dit, visuellement le graphique obtenu représente en dimension (en général) 2, la meilleure approximation des distances observées entre les individus pouvant être des gènes ou des échantillons biologiques.

6.1 Cas de la matrice Jaccard abundance et de la fraction $0_0.2$

MDS en dimension 1 et 2 :

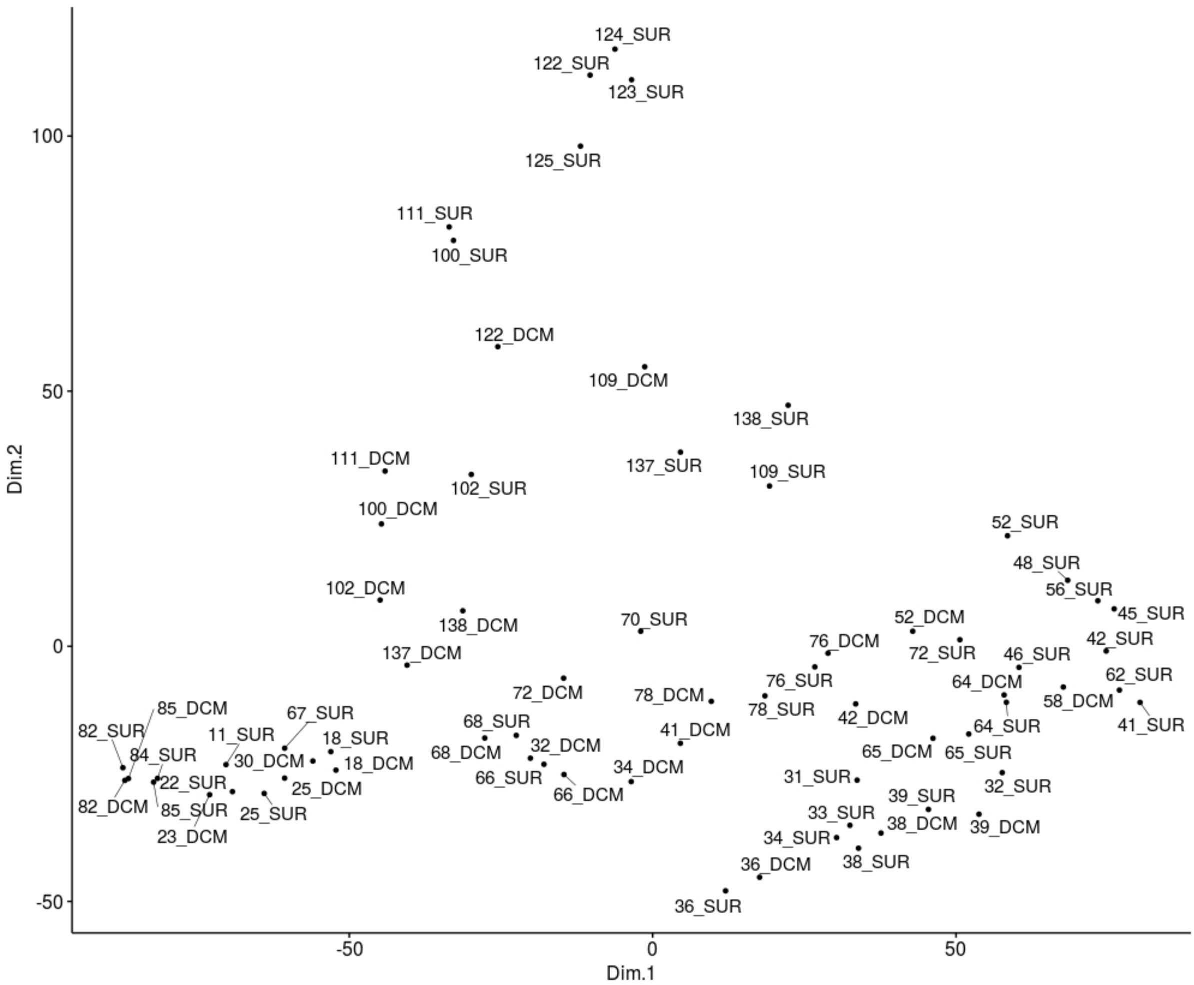


Figure 1 – MDS dim1-2

MDS en dimension 1 et 3 :

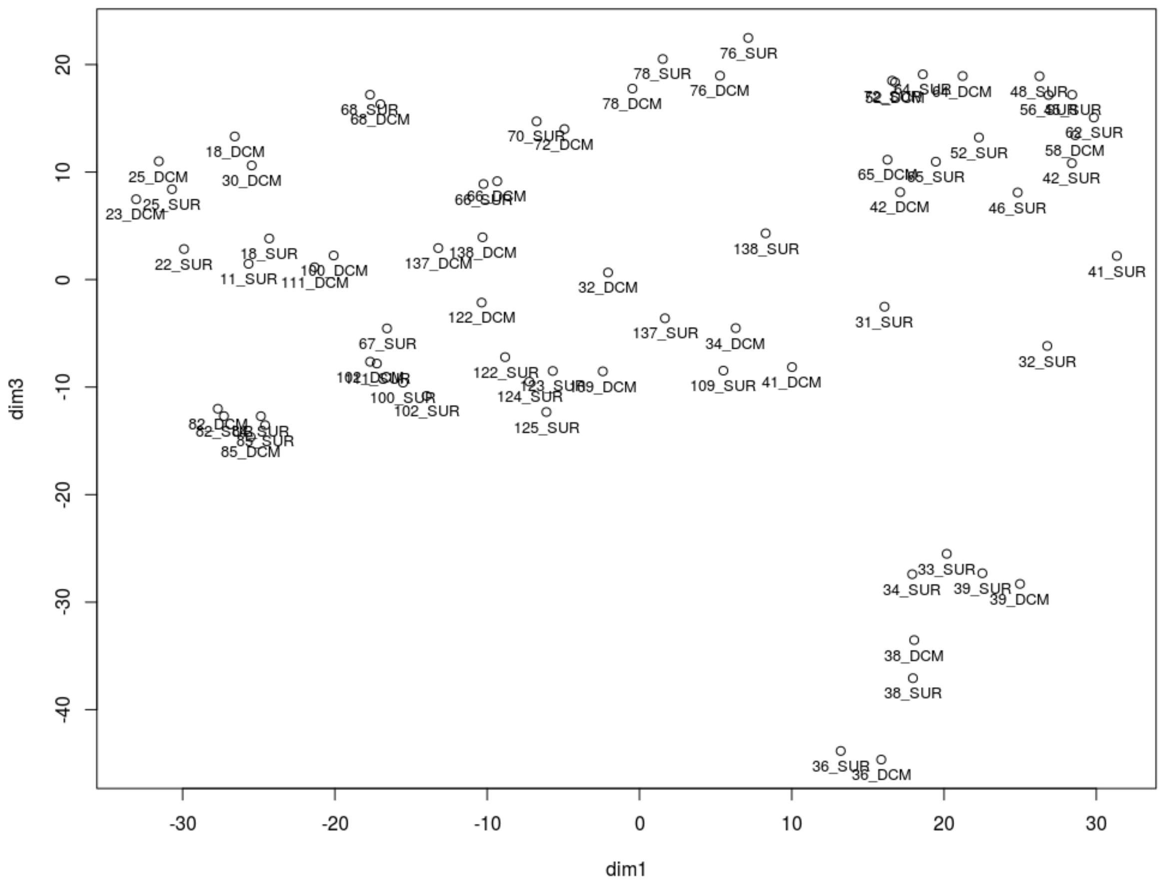


Figure 2 – MDS dim1-3

MDS en dimension 2 et 3 :

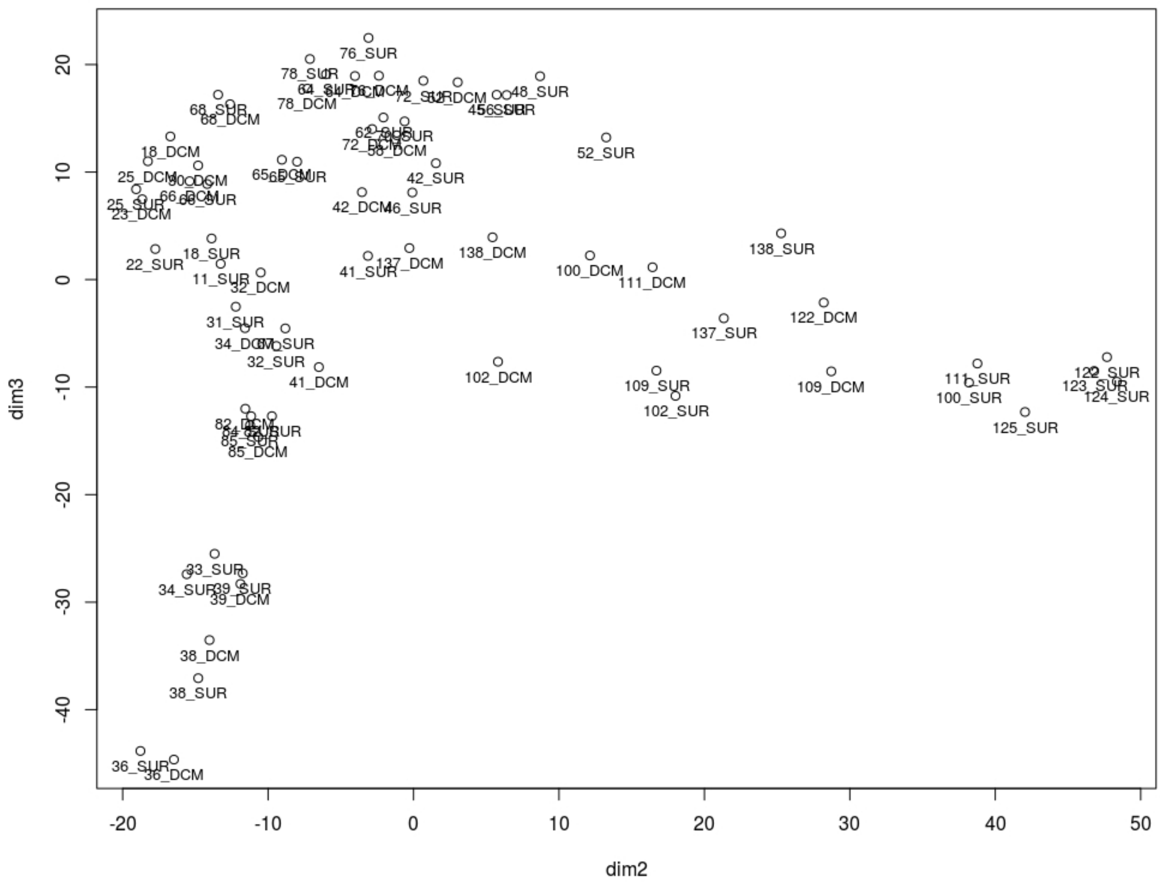


Figure 3 – MDS dim2-3

6.2 Stress plot VS dimensions

Le MDS étant encore une technique factorielle, comme en ACP il est nécessaire de déterminer le nombre de dimensions fixant la taille de l'espace de représentation.

Le graphique représentant la décroissance des valeurs propres aide à ce choix.

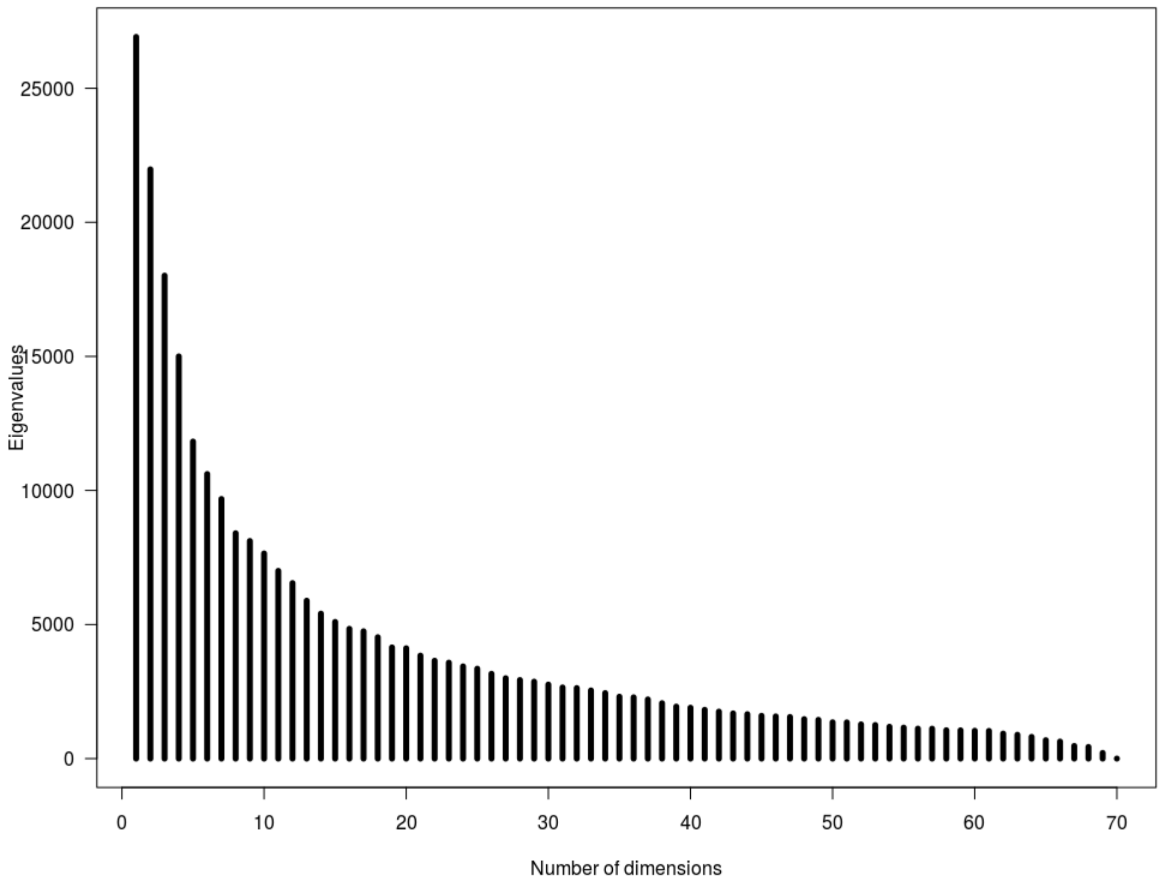


Figure 4 – "Scree Plot"

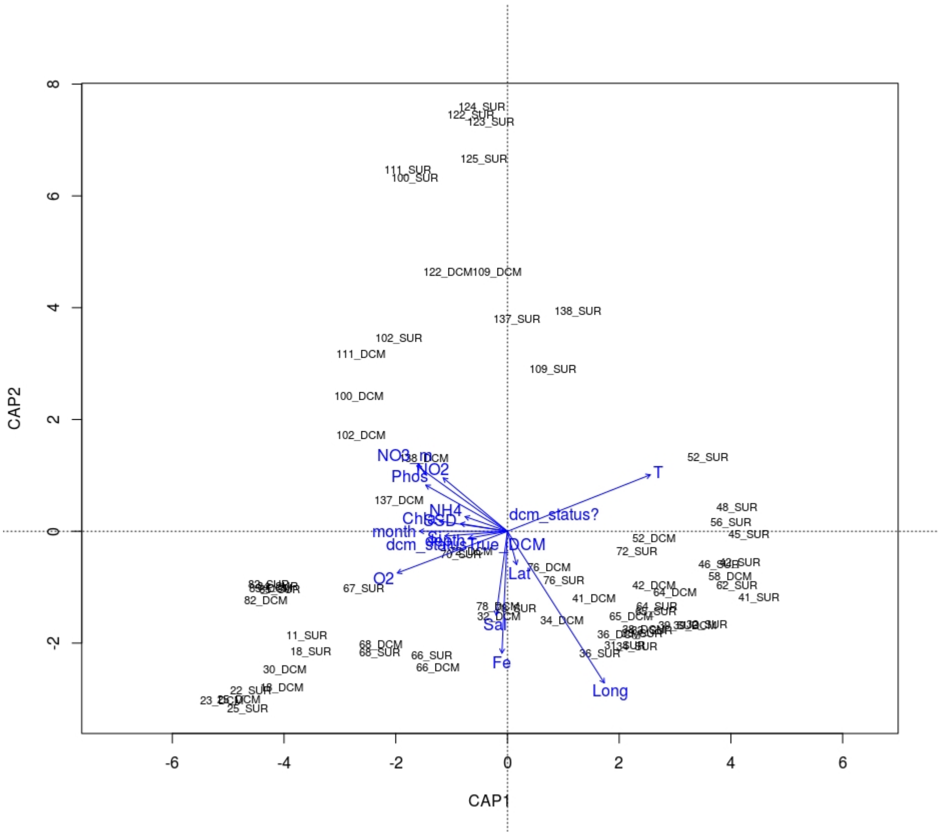
Dendogramme : CAH

La classification ascendante hiérarchique (CAH) organise les observations définies par un certain nombre de variables, elles-mêmes divisées en modalités, en les regroupant de façon hiérarchique.

Elle commence par agréger celles qui sont les plus semblables entre elles, puis les observations ou groupes d'observations un peu moins semblables et ainsi de suite jusqu'au regroupement trivial de l'ensemble de l'échantillon.
Ces agrégations se font deux à deux.

Le dendogramme ou arbre hiérarchique montre non seulement les liaisons entre les classes mais la hauteur des branches nous indique leur niveau de proximité.

7 Distance-based Redundancy Analysis (db-RDA)




```
> anova(dbRDA)## overall test of the significance of the analysis

#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Permutation test for capscale under reduced model
Permutation: free
Number of permutations: 999

Model: capscale(formula = jaccard_abundance ~ Lat + Long + T + Sal +
Chla + O2 + NO3_m + NO2 + NH4 + Fe + SSD
+ Phos + Si +depth + month + dcm_status,
data = design, distance = "bray")

         Df Variance      F Pr(>F)
Model    16   2049.9 3.3899  0.001 ***
Residual 53   2003.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
> anova(dbRDA, by="axis", perm.max=500) ## test axes for significance

Permutation test for capscale under reduced model
Forward tests for axes
Permutation: free
Number of permutations: 999

Model: capscale(formula = jaccard_abundance ~ Lat + Long + T +
Sal + Chla + O2 + NO3_m + NO2 + NH4 + Fe + SSD + Phos +
Si + depth + month + dcm_status,
data = design, distance = "bray")

         Df Variance      F Pr(>F)
CAP1      1   356.91 9.4432  0.001 ***
CAP2      1   301.69 7.9820  0.001 ***
CAP3      1   230.62 6.1018  0.001 ***
CAP4      1   197.95 5.2375  0.001 ***
CAP5      1   138.79 3.6720  0.001 ***
CAP6      1   132.09 3.4948  0.001 ***
CAP7      1   107.83 2.8529  0.001 ***
CAP8      1   107.36 2.8406  0.001 ***
CAP9      1    94.97 2.5128  0.001 ***
CAP10     1    83.69 2.2142  0.001 ***
CAP11     1    72.95 1.9301  0.006 **
CAP12     1    53.17 1.4068  0.427
CAP13     1    51.26 1.3562  0.391
CAP14     1    47.42 1.2545  0.440
CAP15     1    41.04 1.0859  0.626
CAP16     1    32.20 0.8520  0.800
Residual 53  2003.17
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

> anova(dbRDA, by="terms", permu=200)## test for sig.environ. variables
Permutation test for capscale under reduced model
Terms added sequentially (first to last)
Permutation: free
```

```
Number of permutations: 999

Model: capscale(formula = jaccard_abundance ~ Lat + Long + T +
Sal + Chla + O2 + NO3_m + NO2 + NH4 + Fe + SSD +
Phos + Si + depth + month + dcm_status,
data = design, distance = "bray")

           Df Variance      F Pr(>F)    
Lat         1   187.01 4.9478  0.001 ***
Long        1   288.41 7.6307  0.001 ***
T           1   322.71 8.5382  0.001 ***
Sal         1   126.68 3.3518  0.001 ***
Chla        1    95.06 2.5152  0.001 ***
O2          1   122.03 3.2286  0.001 ***
NO3_m       1   143.74 3.8030  0.001 ***
NO2         1    98.73 2.6121  0.001 ***
NH4         1   125.87 3.3303  0.001 ***
Fe          1    56.12 1.4848  0.028 *  
SSD         1   108.48 2.8702  0.001 ***
Phos        1    55.08 1.4574  0.030 *  
Si          1    80.47 2.1291  0.001 ***
depth       1    67.95 1.7978  0.002 ** 
month       1    94.75 2.5068  0.001 ***
dcm_status  1    76.87 2.0339  0.001 ***
Residual    53  2003.17                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

8 Adonis sur la variable month

```
adonis(as.dist(jaccard_abundance)~month,data=design,method="euclidian")
Call:

adonis(formula = as.dist(jaccard_abundance) ~ month, data = design,method = "euclidian")

Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)

          Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
month      1     13741 13740.5  3.5136 0.04913  0.001 ***
Residuals 68    265925  3910.7         0.95087
Total     69    279665                 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


anova.cca(capscale(as.dist(jaccard_abundance) ~ month, data = design),by="terms")

Permutation test for capscale under reduced model
Terms added sequentially (first to last)
Permutation: free
Number of permutations: 999

Model: capscale(formula = as.dist(jaccard_abundance) ~ month, data = design)
         Df Variance      F Pr(>F)
month     1    199.1 3.5136  0.001 ***
Residual 68   3854.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = distances ~ group, data = df)

$group
            diff         lwr         upr     p adj
3-1   -11.6505423 -26.2443102   2.9432255 0.2372726
```

4-1	-1.4095848	-15.5676193	12.7484496	0.9999998
5-1	-4.2798887	-20.1090524	11.5492750	0.9977583
6-1	-28.4398398	-51.9182838	-4.9613958	0.0064239
7-1	-2.8490227	-16.6485677	10.9505223	0.9997820
8-1	-39.1230398	-62.6014838	-15.6445958	0.0000327
9-1	-8.6347975	-26.6828210	9.4132260	0.8752277
10-1	-5.7778346	-22.5298469	10.9741777	0.9849144
11-1	-5.5900719	-20.1838397	9.0036960	0.9683040
12-1	1.6652619	-13.4703136	16.8008374	0.9999994
4-3	10.2409575	-4.3528103	24.8347254	0.4158266
5-3	7.3706537	-8.8494125	23.5907198	0.9066337
6-3	-16.7892975	-40.5330431	6.9544481	0.4045150
7-3	8.8015197	-5.4447277	23.0477670	0.6034725
8-3	-27.4724975	-51.2162431	-3.7287519	0.0112431
9-3	3.0157449	-15.3760814	21.4075711	0.9999735
10-3	5.8727077	-11.2491507	22.9945662	0.9855172
11-3	6.0604705	-8.9563927	21.0773337	0.9550991
12-3	13.3158042	-2.2281259	28.8597344	0.1588711
5-4	-2.8703039	-18.6994676	12.9588598	0.9999329
6-4	-27.0302550	-50.5086990	-3.5518110	0.0119170
7-4	-1.4394378	-15.2389828	12.3601072	0.9999996
8-4	-37.7134550	-61.1918990	-14.2350110	0.0000686
9-4	-7.2252127	-25.2732361	10.8228108	0.9574108
10-4	-4.3682498	-21.1202620	12.3837625	0.9983421
11-4	-4.1804870	-18.7742549	10.4132808	0.9964050
12-4	3.0748467	-12.0607287	18.2104222	0.9998114
6-5	-24.1599511	-48.6823861	0.3624838	0.0568994
7-5	1.4308660	-14.0784836	16.9402157	0.9999999
8-5	-34.8431511	-59.3655861	-10.3207162	0.0006293
9-5	-4.3549088	-23.7415959	15.0317783	0.9995393
10-5	-1.4979459	-19.6842704	16.6883786	1.0000000
11-5	-1.3101832	-17.5302493	14.9098830	1.0000000
12-5	5.9451506	-10.7640723	22.6543735	0.9810926
7-6	25.5908172	2.3267927	48.8548417	0.0197871
8-6	-10.6832000	-40.7169265	19.3505265	0.9811284
9-6	19.8050423	-6.2049278	45.8150124	0.2986621
10-6	22.6620052	-2.4660132	47.7900236	0.1130845
11-6	22.8497680	-0.8939776	46.5935135	0.0692000
12-6	30.1051017	6.0245486	54.1856548	0.0042521
8-7	-36.2740172	-59.5380417	-13.0099927	0.0001218
9-7	-5.7857748	-23.5539670	11.9824174	0.9901787
10-7	-2.9288119	-19.3789614	13.5213375	0.9999433
11-7	-2.7410492	-16.9872965	11.5051982	0.9998846
12-7	4.5142846	-10.2864974	19.3150666	0.9940738
9-8	30.4882423	4.4782722	56.4982124	0.0096380
10-8	33.3452052	8.2171868	58.4732236	0.0018327

```
11-8    33.5329680   9.7892224  57.2767135 0.0006950
12-8    40.7883017  16.7077486  64.8688548 0.0000232
10-9     2.8569629 -17.2902733  23.0041991 0.9999933
11-9     3.0447256 -15.3471006  21.4365519 0.9999711
12-9    10.3000594  -8.5245749  29.1246937 0.7565803
11-10    0.1877628 -16.9340957  17.3096212 1.0000000
12-10    7.4430965 -10.1428530  25.0290460 0.9391584
12-11    7.2553338  -8.2885964  22.7992640 0.8911830
```

8.1 Visualisation graphique du résultat précédent

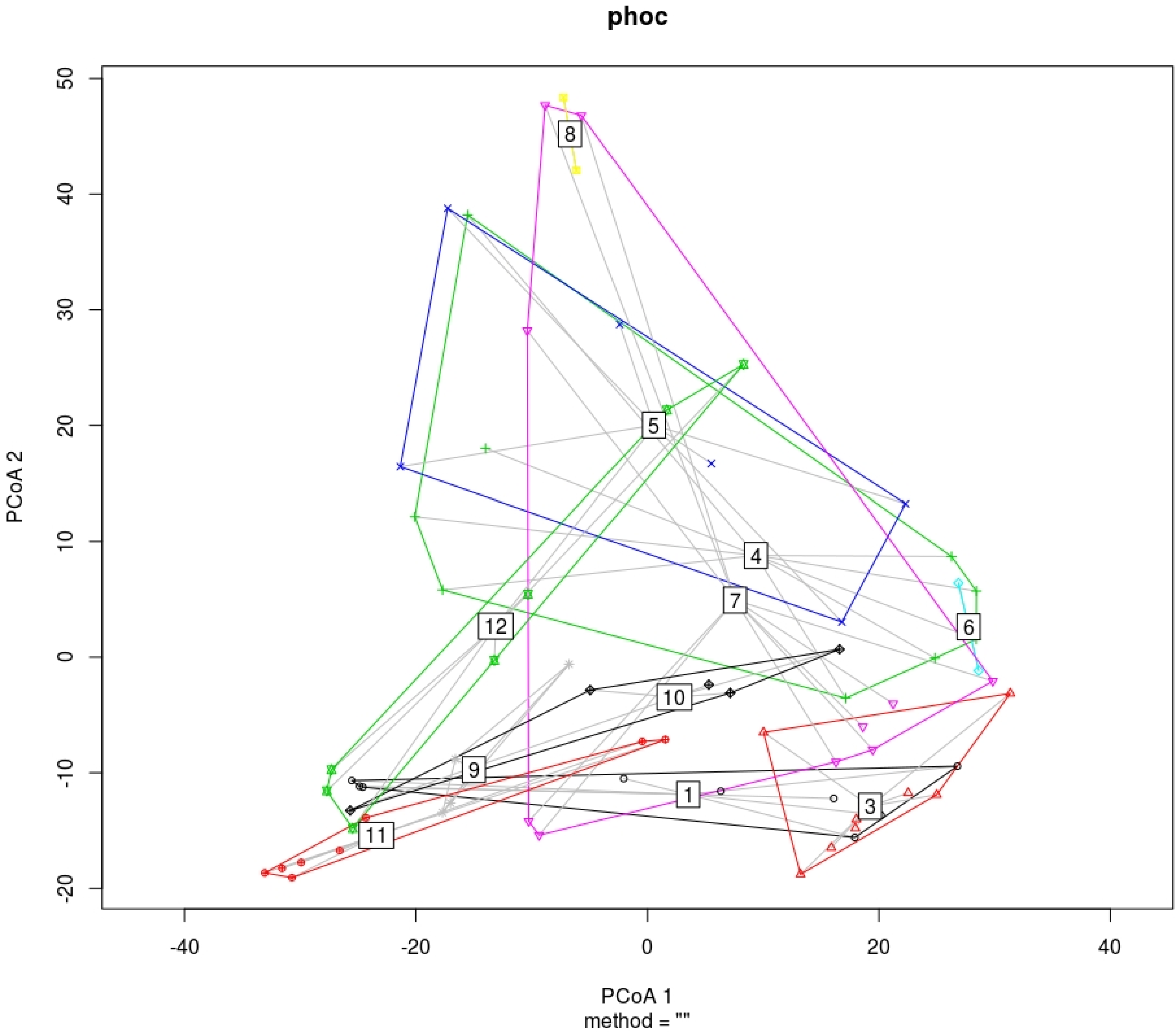


Figure 5 – Pairwise comparisons