Rapport de projet Analyse de données & Eléments de modélisation statistique

Xiaoya Wang, Mickael Song, Yessine Jmal, Florian Grivet

Contents

1 Introduction

On observe pour G = 1615 gènes d'une plante modèle les valeurs suivantes :

$$Y_{gtsr} = log_2(X_{gtsr} + 1) - log_2(X_{gt_0} + 1)$$

avec
• X_{gtsr} la mesure d'expression du gène g ∈ {G1, . . . , G1615} pour le traitement t ∈ {T1, T2, T3} pour le réplicat r ∈ {R1, R2} et au temps s ∈ {1h, 2h, 3h, 4h, 5h, 6h}
• X_{gt_0} l'expression du gène g pour un traitement de référence t_0

Nous allons répartir l'étude de ce jeu de données en 4 parties :
1 Analyse du jeu de données
2 Clustering
3 Etude de l'expression des gènes pour le traitement T3 à 6h
4 Etude de l'expression des gènes pour le traitement T1 à 6h

2 Analyse du jeu de données

Nous allons dans cette partie effectuer une analyse des statistiques descriptives et préparer le jeu de données afin d'en sortir les variables redondantes, transformations, outliers et visualiser le jeu de données dans un espace de faible dimension (en particulier l'aspect réplicat biologique, l'effet traitement et l'effet temps)

2.1 Statistiques descriptives et préparation du jeu de données

Table 1: Les premières lignes du jeu de données.

	T1_1h_R1	T1_2h_R1	T1_3h_R1	T1_4h_R1	T1_5h_R1	T1_6h_R1	T2_1h_R1	T2_2h_R1
G1	0.17	0.68	-0.18	0.08	0.00	0.50	-0.59	0.19
G2	0.19	0.82	-0.05	0.18	0.47	-0.76	0.38	2.51
G3	-0.05	-0.03	0.26	-0.32	-0.39	0.42	0.21	-1.00
G4	-0.23	-0.75	-0.24	-0.70	-0.12	-0.38	0.41	0.23
G5	-0.21	-0.69	-0.18	-0.07	0.52	0.45	-0.45	-1.51
G6	-0.62	-0.86	-0.02	-0.14	0.49	0.45	-0.57	-1.48

Le jeu de données contient 1615 individus et 36 variables, toutes quantitatives.
Les attributs du jeu de données sont :
T1_1h_R1, T1_2h_R1, T1_3h_R1, T1_4h_R1, T1_5h_R1, T1_6h_R1, T2_1h_R1, T2_2h_R1, T2_3h_R1, T2_4h_R1, T2_5h_R1, T2_6h_R1, T3_1h_R1, T3_2h_R1, T3_3h_R1, T3_4h_R1, T3_5h_R1, T3_6h_R1, T1_1h_R2, T1_2h_R2, T1_3h_R2, T1_4h_R2, T1_5h_R2, T1_6h_R2, T2_1h_R2, T2_2h_R2, T2_3h_R2, T2_4h_R2, T2_5h_R2, T2_6h_R2, T3_1h_R2, T3_2h_R2, T3_3h_R2, T3_4h_R2, T3_5h_R2, T3_6h_R2

Avec le résultat de la commande python `datapy.isnull().sum()`, on voit bien sur le Rmd que notre jeu de données est complet.

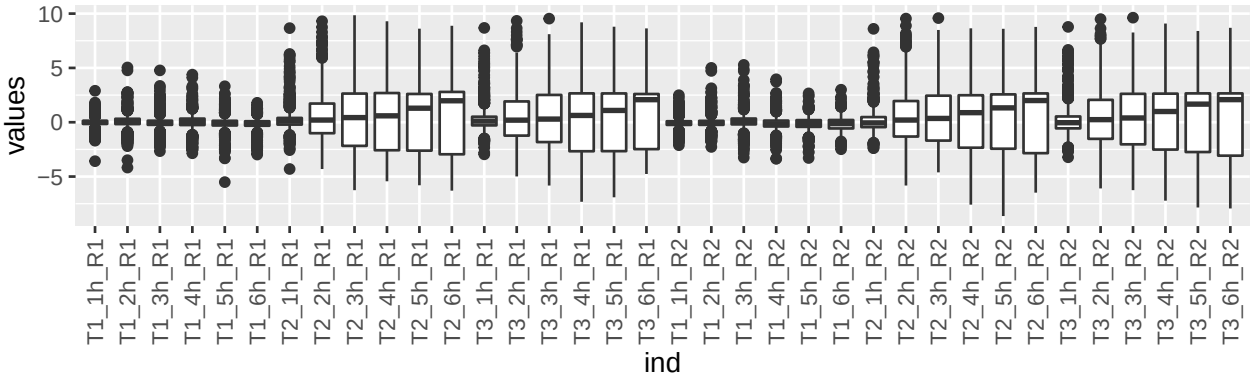


Figure 1: Boxplots des 36 variables

On remarque dans la figure 1 que les boxplots du traitement 2 et du traitement 3 ne sont pas centrés. On peut penser qu'ils ont donc un effet non nul sur les gènes. En étudiant la forme des boxplots, on remarque une dissymétrie pour ces deux traitements, de nombreux outliers ainsi qu'une forte variabilité entre les individus. Les boxplots du traitement 2 et du traitement 3 sont d'ailleurs similaires, quelque soit le réplicat. On peut donc faire l'hypothèse que ces deux traitements donnent des résultats similaires. Le traitement 1 est quant à lui beaucoup plus réduit et centré en 0. Ce traitement semble donc ne pas avoir d'effet sur les gènes. Les boxplots du traitement 1 sont symétriques mais possèdent également beaucoup d'outliers.

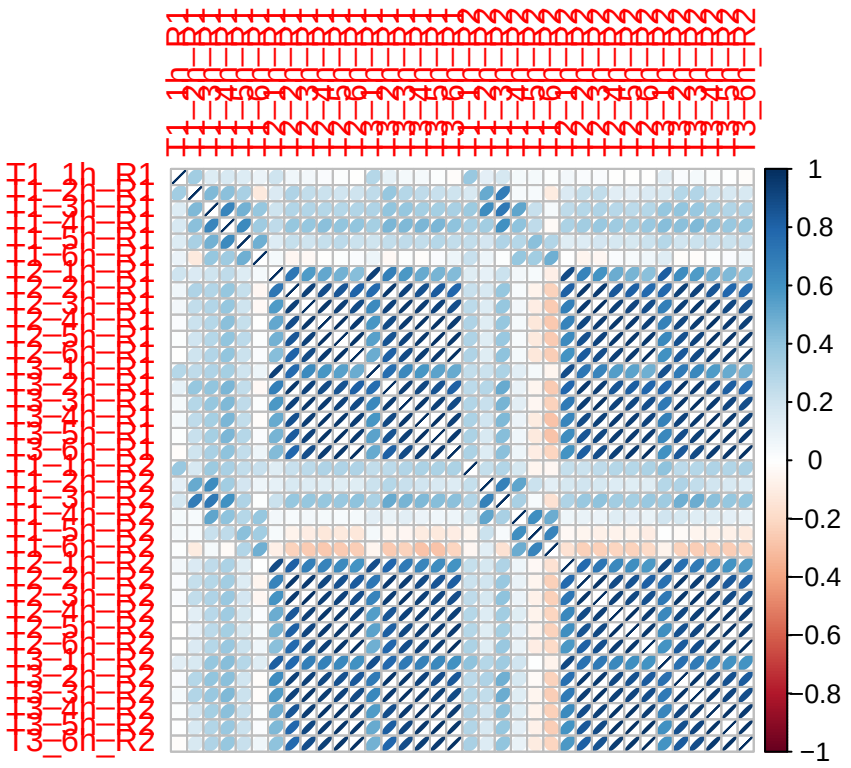


Figure 2: Graphique des corrélations des 36 variables

La figure 2 des corrélations nous confirme bien l'hypothèse précédente, les traitements 2 et 3 sont fortement corrélés alors que ces traitements semblent totalement décorrélés du traitements 1.
On peut également noter le fait que, pour un traitement donné, le réplicat 1 et le réplicat 2 sont fortement corrélés entre eux, ce qui est cohérent puisque ce sont des réplicats biologiques. On pourra donc par la suite faire notre étude uniquement sur un seul réplicat (réplicat 2) sans perdre trop d'informations.

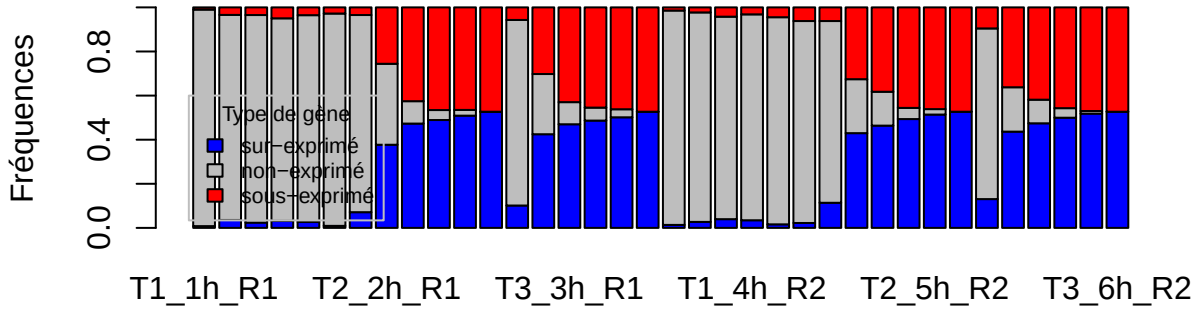


Figure 3: Fréquence de l'expression des gènes sous-exprimés, normaux et sur-exprimés en fonction des traitements

Le graphique 3 représente la fréquence des gènes "sous-exprimés", "non-exprimés" et "sur-exprimés" pour chaque traitement à toute heure sur les deux réplicats.
Il appuie notre hypothèse que les traitements 2 et 3 sont similaires et que le traitement 1 n'a pas beaucoup d'effet.
Avec l'évolution du temps, les traitements T2 et T3 ont tendance à regrouper les données en deux classes : sur-exprimé et sous-exprimé. Ces classes ont l'air d'avoir le même nombre d'individu.

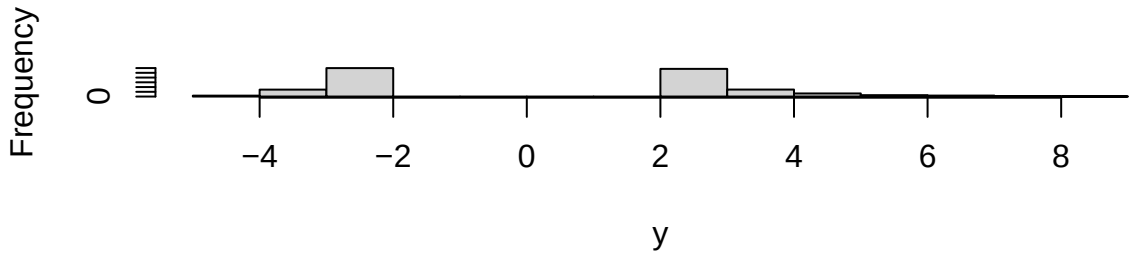


Figure 4: Fréquence de la valeur des gènes du traitement 3 à l'heure 6

Sur la figure 4, on remarque qu'à la dernière heure (6h) du traitement 3, tous les gènes sont soit très sur-exprimé (valeurs ≥ 2), soit très sous-exprimé (valeurs ≤ 2). Les gènes du jeu de données ont donc été choisis en fonction de T3_6H.

2.2 Analyse en composante principale

2.2.1 Analyse en composante principale sur les individus

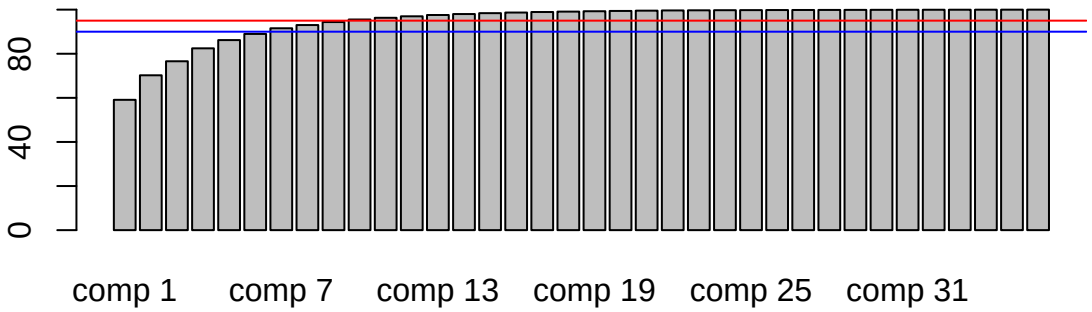


Figure 5: Variance expliquée cumulée (en %) des différentes composantes principales

D'après le graphique 5, on note que :
- Pour avoir 90% de l'information, on peut réduire nos données à 7 dimensions. On utilisera ces 7 composantes pour la partie clustering.
- Pour avoir 95% de l'information il suffit de se placer en dimension 10.

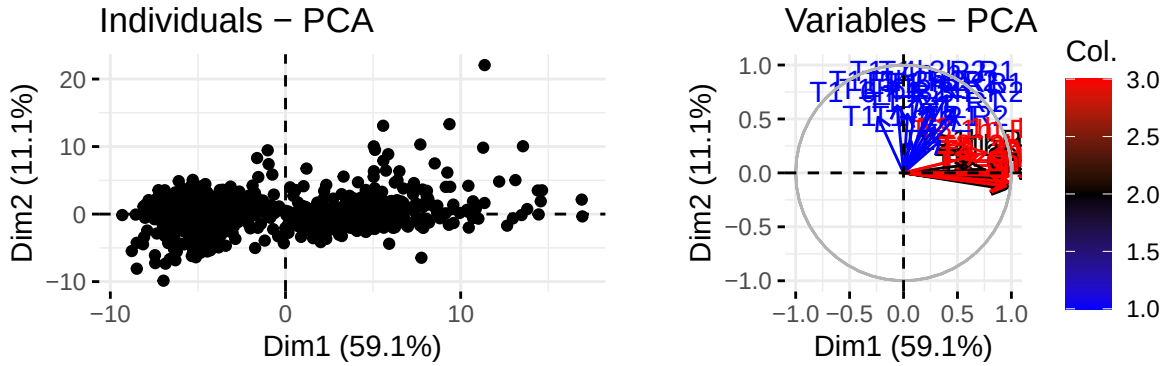


Figure 6: Visualitaion de l'ACP sur les deux premières composantes principales, pour les individus (à gauche) et pour les variables (à droite)

La première composante principale de la figure 6 nous dit si un gène réagit au traitement 2 ou 3. Si le gène réagit fortement à l'un de ces traitements, c'est-à-dire s'il est sur-exprimé (resp. sous-exprimé), il se trouvera à droite (resp. à gauche) du graphique et s'il ne réagit pas beaucoup à l'un de ces traitement il se trouvera au centre.

La deuxième composante principale nous dit si un gène réagit au traitement 1. Si le gène réagit fortement à ce traitement, c'est-à-dire s'il est sur-exprimé (resp. sous-exprimé), il se retrouve en haut (resp. en bas) du graphique. Par contre, si le gène ne réagit pas beaucoup au traitement 1, il se trouve au milieu.

2.2.2 Analyse en composante principale sur les données transposées (variables)

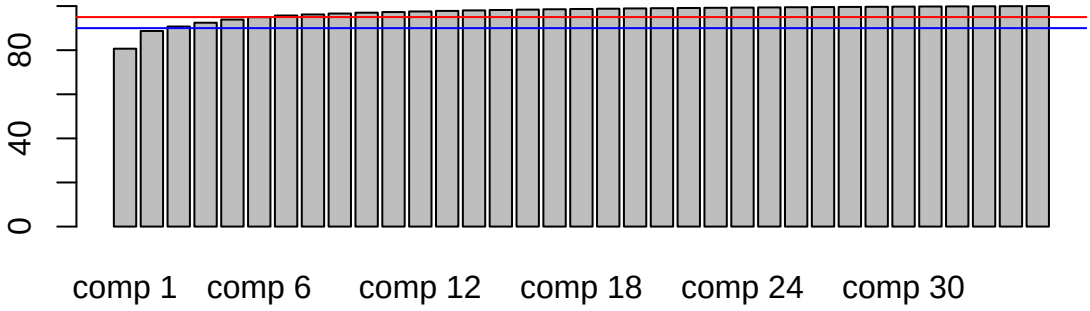


Figure 7: Variance expliquée cumulée (en %) des différentes composantes principales

D'après le graphique 7, on note que :
Pour avoir 90% de l'information, on peut réduire nos données à 3 dimensions. On utilisera ces composantes dans la partie clustering.
Pour avoir 95% de l'information il suffit de se placer en dimension 7.

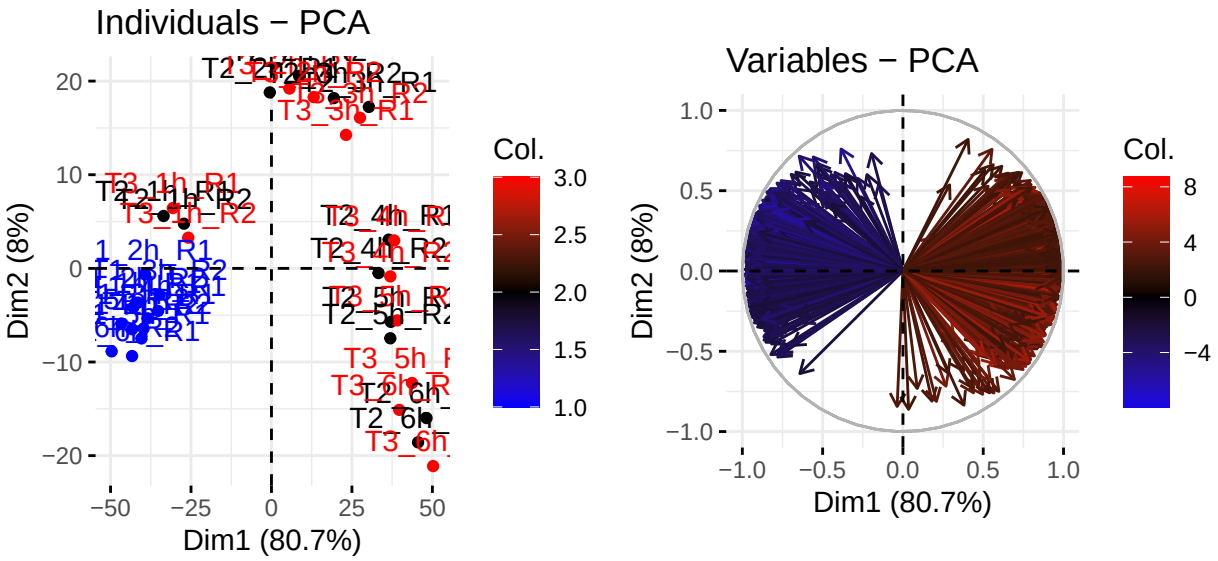


Figure 8: Visualitaion de l'ACP sur les deux premières composantes principales, pour les individus (à gauche) et pour les variables (à droite)

La première composante principale de la figure 8 nous dit si le traitement correspond à T1 (à gauche) ou à T2/T3 (à droite).

La deuxième composante principale nous dit si l'heure du traitement est supérieure ou égale à 4h (en haut) ou inférieur à 4h (en bas).

3 Clustering

Pour mieux comprendre les relations entre les variables et les gènes dans les différentes conditions, nous allons utiliser, dans cette partie, différentes méthodes de clustering pour obtenir une classification des variables "Tx_xh_Rx" et des gènes ayant des profils d'expression similaires.

3.1 Obtention d'une classification des variables "Tx_xh_Rx"

On reprend les 3 premières composantes principales de l'ACP effectué précédemment sur les variables. Les premières 3 composantes principales résument 90% de l'information.

3.1.1 K means

Avant d'appliquer l'algorithme K-means sur notre jeu de données, nous allons déterminer le nombre optimal de classes. Pour cela, nous allons tracer l'évolution de l'inertie intraclasse en fonction du nombre de classes.

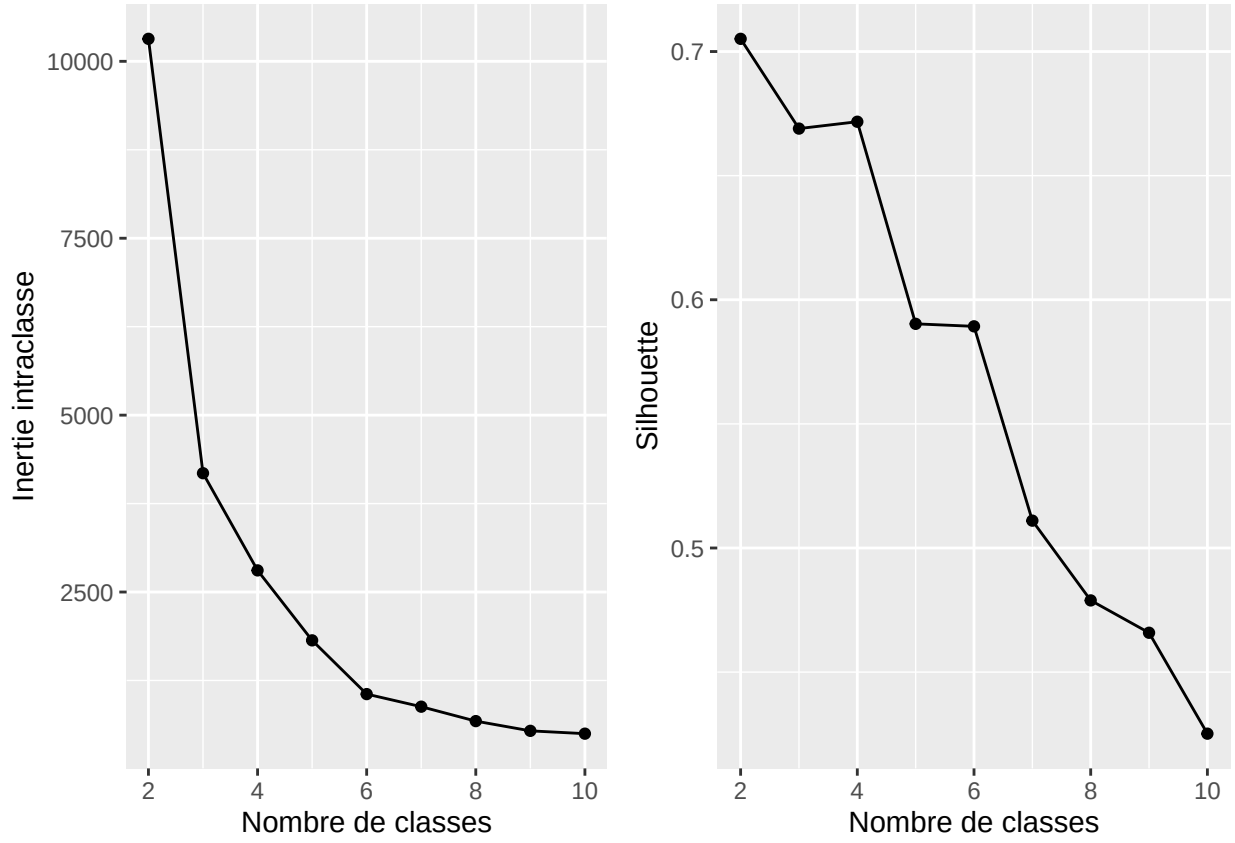


Figure 9: Evolution de l'inertie intraclasse (gauche) et du critère silhouette (droite) en fonction du nombre de classes

On voit sur la figure 9, pour l'inertie intraclasse, qu'il y a un coude pour K = 3 ou K = 6. On retient donc 3 classes par ce critère. Le critère silhouette montre un pic à K = 2, donc on retient 2 classes avec ce critère.

On a représenté sur la figure 10 les résultats de K-means à trois classes (gauche) et à deux classes (droite) sur les 2 premières composantes principales de l'ACP.

Pour un K-means à trois classes, on remarque que le cluster 1 regroupe les traitements T2 et T3 du réplicat R1/R2 à 1h et le traitement T1 du réplicat R1/R2 pour toutes les heures. Le cluster 2 regoupe les traitements T2 et T3 de 2h et 3h. Le cluster 3 regroupe les traitements T2 et T3 de 4h à 6h.

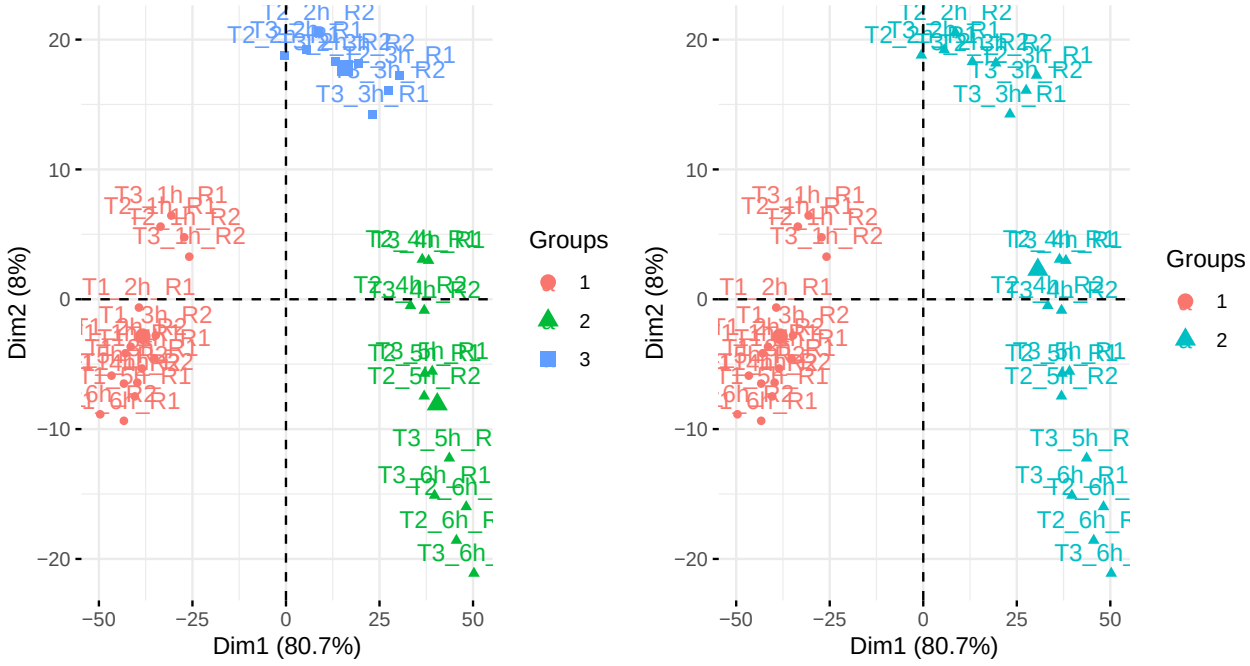


Figure 10: Résultat du clustering Kmeans à deux et trois classes sur les 2 premiers axes de l'ACP

A l'heure de début (1h), les traitements T2/T3 ont le même effet que T1 sur les gènes, on peut supposer que l'effet du traitement T2/T3 se déroule graduellement donc ne s'est pas encore manifesté à 1h.

Le K-means à deux classes, regroupe seulement les classes deux et trois du K-means à trois classes.

Effectuons maintenant un diagramme de silhouette afin de déterminer l'homogénéité de nos 3 clusters.
On voit sur le diagramme de silhouette figure 11 que les clusters 1 et 2 ont des scores de silhouette inférieurs au cluster 3. La cohésion des points du cluster 3 est donc plus grande. C'est-à-dire que le cluster 3 a moins d'outliers et est plus proche de son centre de gravité que les deux autres clusters. On peut cependant noter que le score de silhouette de chaque cluster est supérieur à 0.5, les partitions de chaque cluster sont globalement homogènes.

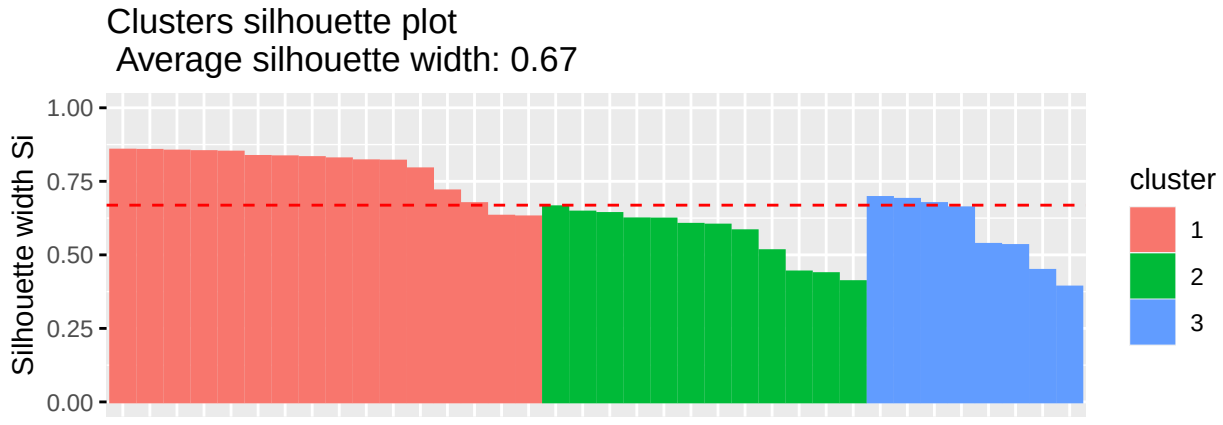


Figure 11: Graphique des silhouettes scores pour nos 3 clusters

3.1.2 PAM

On visualise le nombre de classes optimal par le critère Silhouette avec l'algorithme PAM :

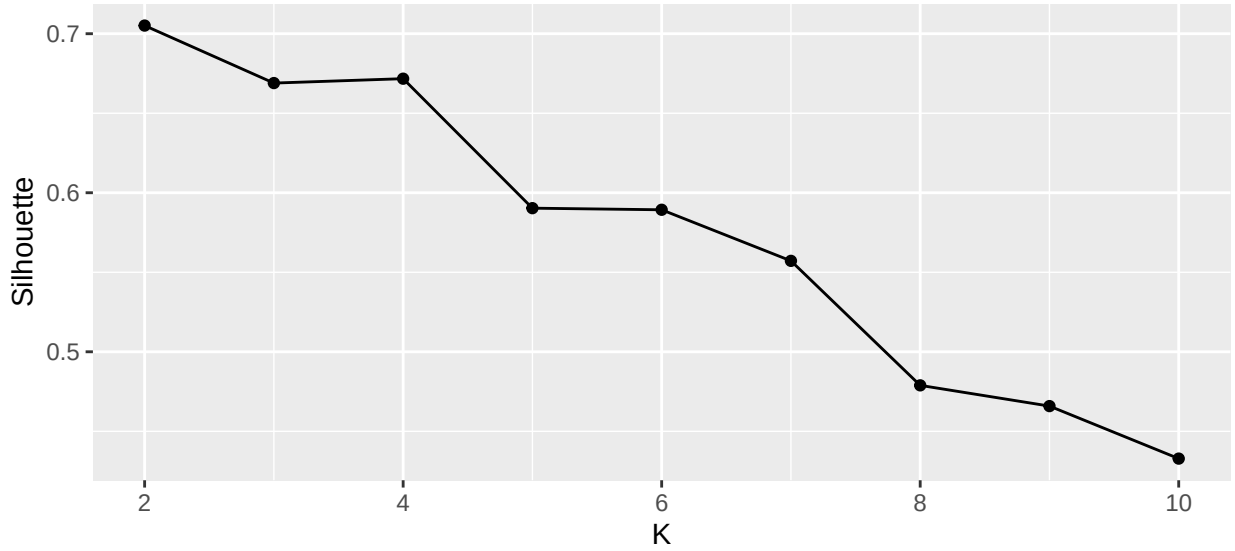


Figure 12: Critère silhouette en fonction du nombre de classe

On obtient le même résultat : 2 classes, et ils sont identiques qu'avec k-means.

```
##
##       1  2
##   1 16  0
##   2  0 20
```

3.1.3 Classification hiérarchique

Effectuons maintenant une classification hiérarchique avec la mesure d'agrégation de Ward.

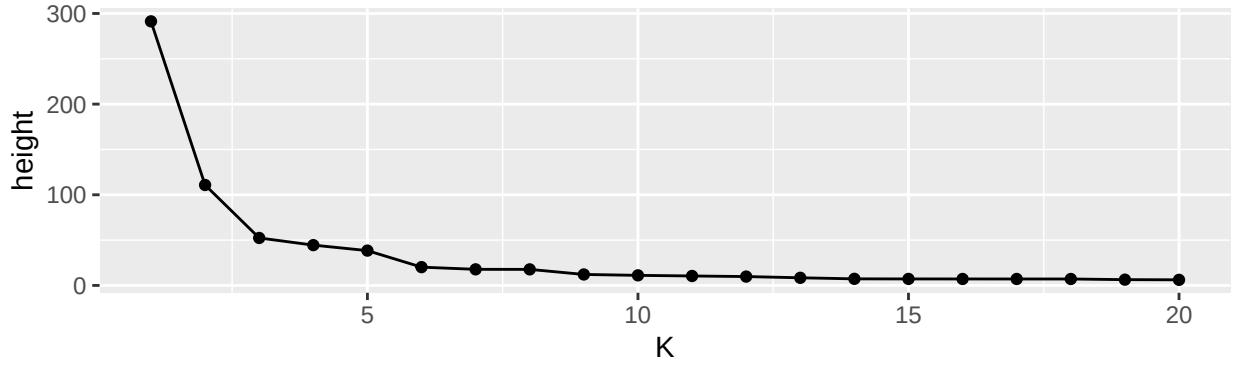


Figure 13: Dendogramme de nos données avec classification hiérarchique avec la mesure d'agrégation de Ward

D'après la figure 13, on voit qu'il y a un saut en k = 3 ou 6 puis la courbe s'aplanit, on retient donc 3 ou 6 classes.

On a également essayé dans le Rmd de déterminer le nombre de classes à retenir avec l'indice de Calinski-Harabasz mais cela n'a pas marché. On voit que plus k est grand, plus l'indice Calinski-Harabasz augmente, il n'y a pas de pic sur la courbe, donc on peut pas déterminer une classification avec le critère Calinski-Harabasz.

3.1.3.1 Modèle de mélange

BIC : figure 14

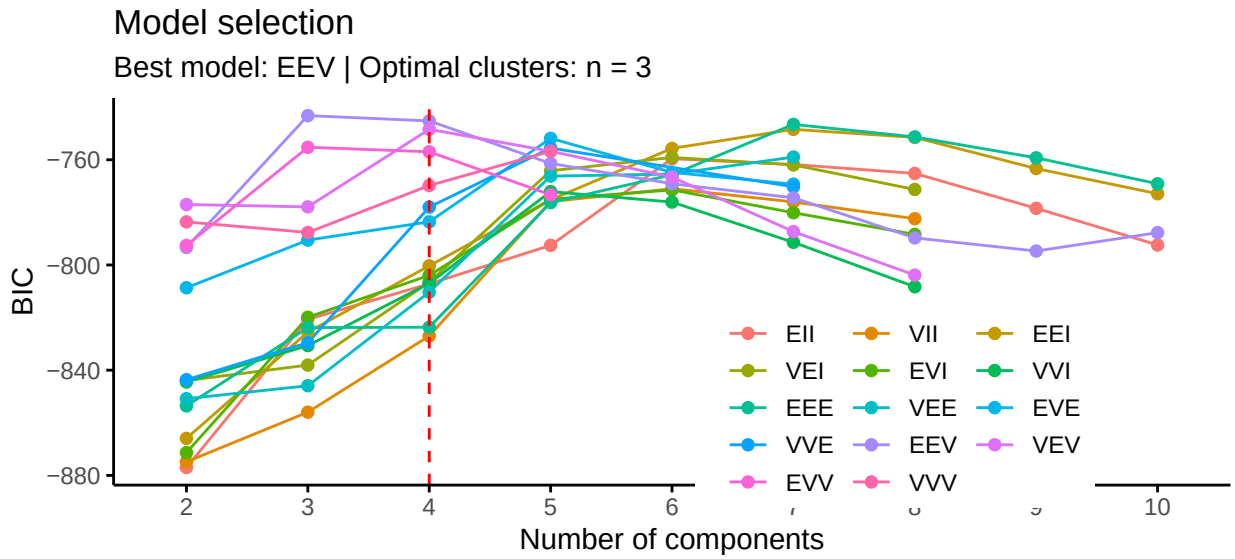


Figure 14: Modèle de mélange BIC

ICL :

```
## Best ICL values:
##              EEV,3       EEV,4       EEE,7
## ICL      -743.2232 -745.239478 -746.582550
## ICL diff    0.0000   -2.016256   -3.359329
```

Si on retient 3 classes : CAH, modèle de mélange BIC et ICL donnent le même résultat de classification. Les trois combinaisons donnent un ARI = 1 (voir Rmd).

On visualise les 3 clusters en boxplot des probabilités d'appartenance et dans le plan de l'ACP :

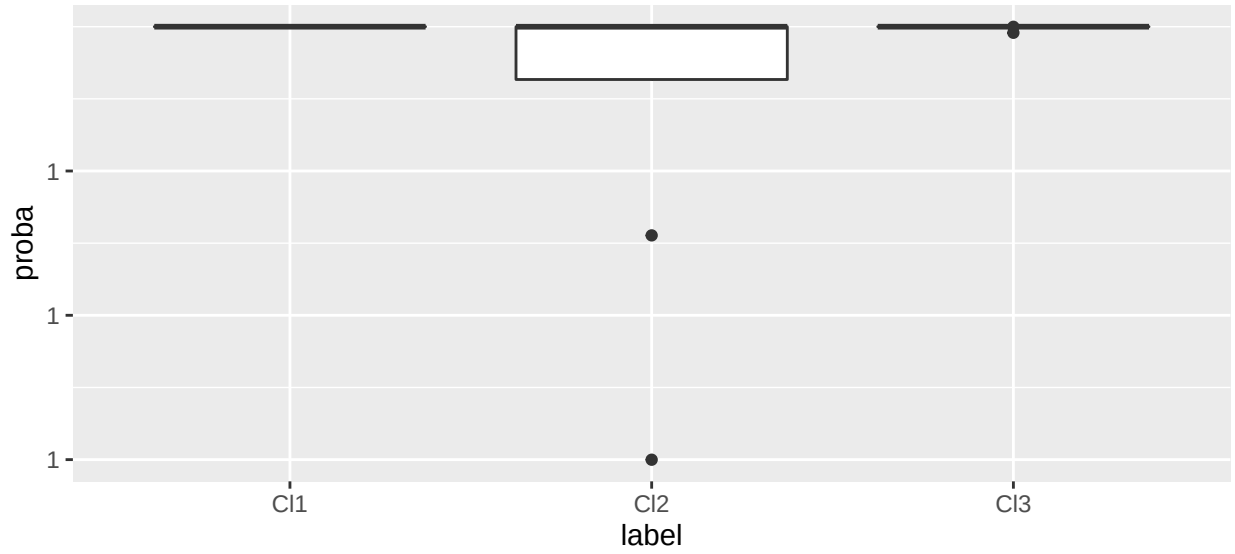


Figure 15: Boxplots des probabilités d'appartenance

Tous les individus ont bien été classé dans chacune des classes car toutes les probas sont très proches de 1.

On compare les deux classifications : k = 3 avec K-means et k = 3 avec CAH Ward :

```
##           clust2
## clust1     CAH3-1 CAH3-2 CAH3-3
##   Kmeans-1     16      0      0
##   Kmeans-2      0      0     12
##   Kmeans-3      0      8      0
```

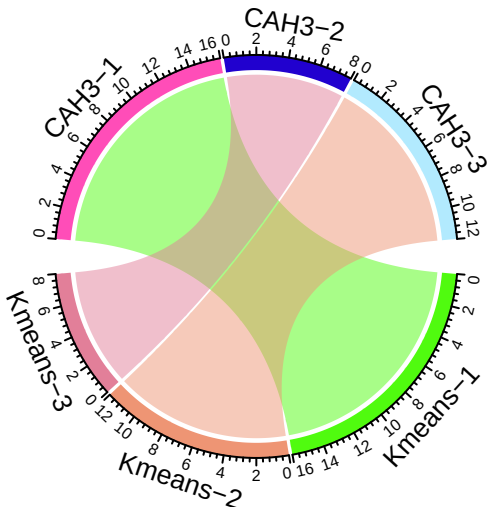


Figure 16: Diagramme de Chord entre les clusters de Kmeans et les clusters de CAH Ward

Ce sont exactement les mêmes cluster.“‘

3.2 Obtention d'une classification des gènes ayant des profils d'expression similaires (co-exprimés) dans les différentes conditions

Nous allons utiliser dans cette partie, différentes méthodes de clustering pour obtenir une classification des gènes ayant des profils d'expression similaires. Pour cela nous utilisons les 7 premières coordonnées de l'ACP réalisée sur les données (non transposées).

3.2.1 K-means

Implémentons l'algorithme K-means.
Commençons par déterminer le nombre de classe optimal. Nous avons implémenté sur la figure 17 le gap statistique en fonction du nombre de classe. On trouve un pic pour K = 2. On choisi donc 2 classes. La méthode silhouette et la méthode de l'inertie intraclasse ont également été implémentées sur le Rmd et nous donnent le même résultat.

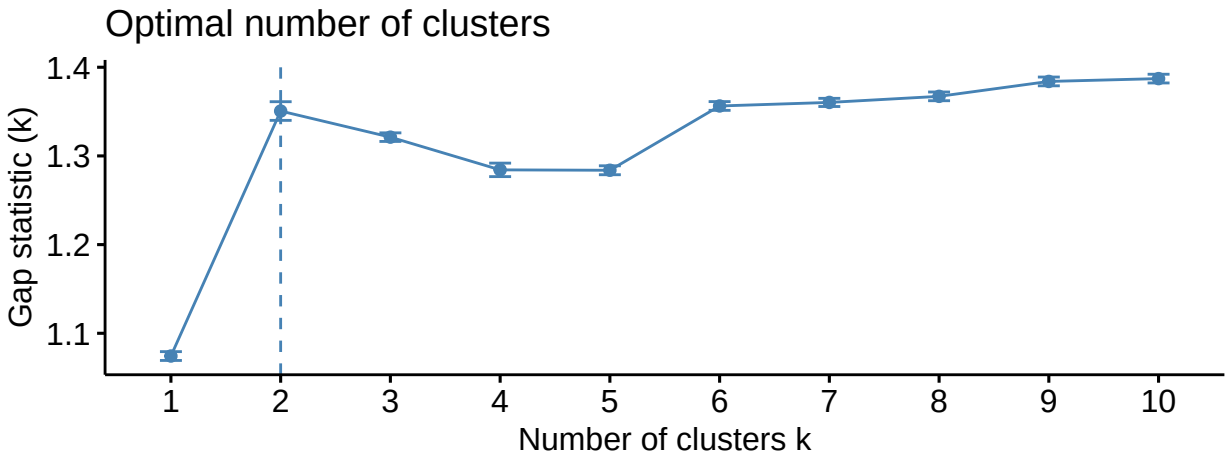


Figure 17: Gap statistique en fonction du nombre de cluster

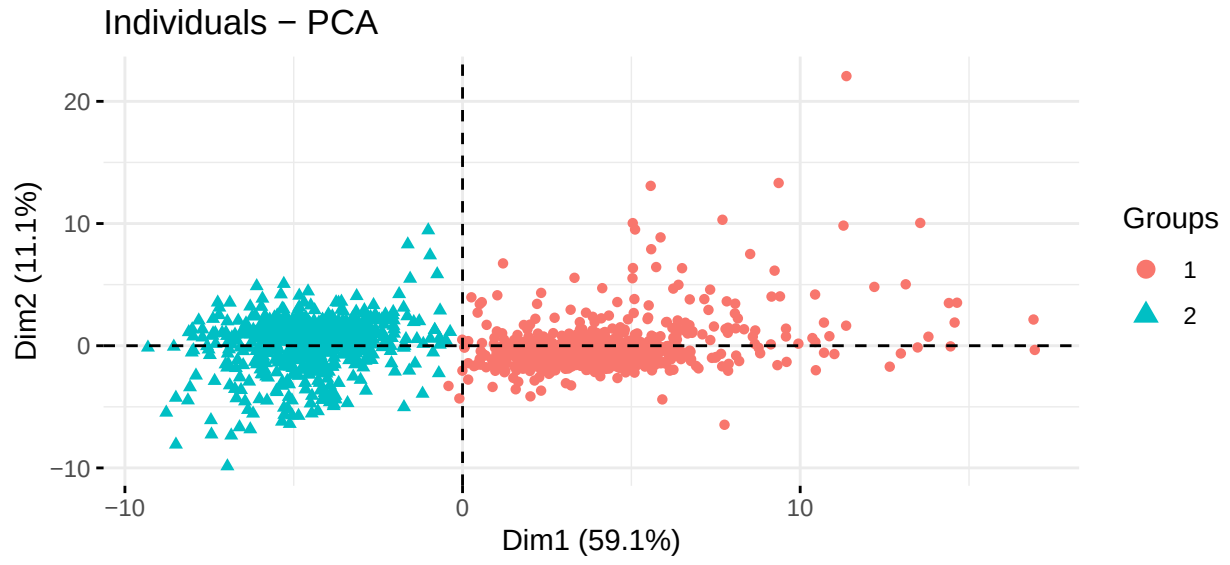


Figure 18: Classification KMeans obtenue sur les gènes sur les axes de l'ACP

On visualise la classification obtenue sur le plan de l'ACP figure 18. L'axe 1 de l'ACP sépare bien les deux clusters.

3.2.2 Classification hiérarchique avec mesure de Ward

Effectuons maintenant une classification hiérarchique des gènes avec la mesure d'agrégation de Ward.

Afin de déterminer le nombre de classe optimal pour la classification hiérarchique, on trace la hauteur du dendrogramme (fig 19) en fonction du nombre de classe K. Cette figure nous donne 2 ou 3 classes. On retrouve le même résultat avec l'indice de Calinski-Harabasz et l'indice Silhouette implémentés sur le Rmd.

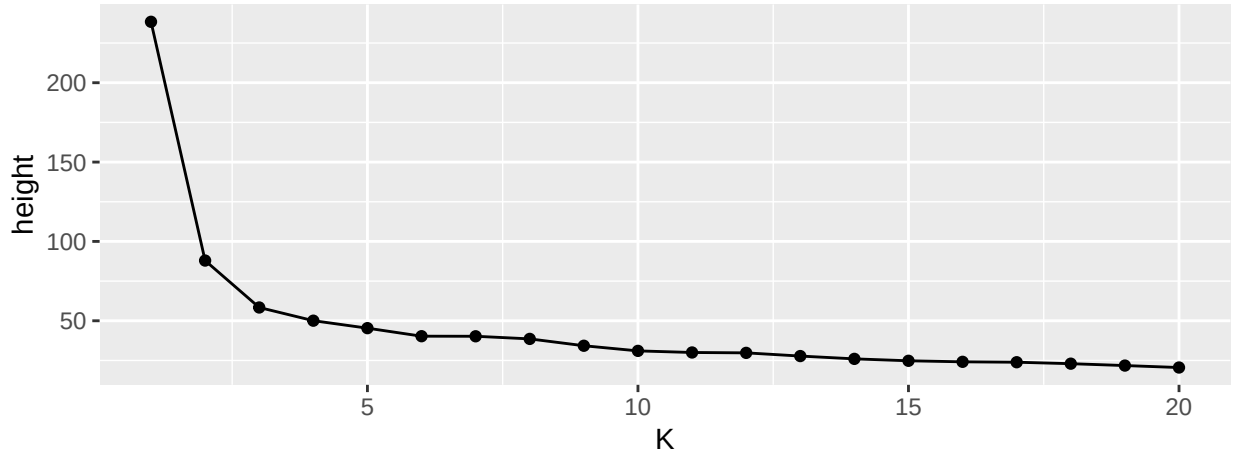


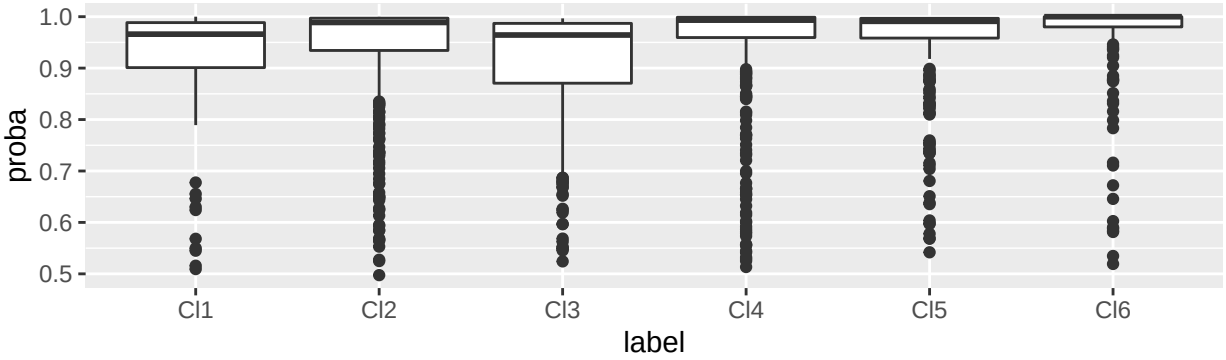
Figure 19: Hauteur du dendogramme en fonction de K

3.2.3 Modèle de mélange

Le modèle de mélange avec BIC ne fonctionne pas : il a tendance à prendre le plus grand nombre de classe. (Voir Rmd)

Le modèle mélange avec ICL retient 6 classes avec la forme VVV.

```
## Best ICL values:
##              VVV,6        VEV,5       VVV,5
## ICL      -35216.24 -35240.78587 -35244.1317
## ICL diff      0.00    -24.54394    -27.8898
```



La classification n'a pas été bien faite, il y a beaucoup d'outliers avec une probabilité d'appartenance descendant jusqu'à 50%.

Transformons les données en données qualitative avec les modalités -1, 0 et 1 :

D'après analyse descriptive précédente, on sait que T3_6h_R2 possède que des gènes sur-exprimé et sous-exprimé, donc en comparant avec T3_6h_R2, on en déduit que les deux clusters qu'on a obtenu sépare les profils d'expression non-similaires (sur et sous-exprimé), qui veut dire regroupe les profils co-exprimés :

```
##
##      -1   1
##   1   3 839
##   2 761  12
```

4 Etude de l'expression des gènes pour le traitement T3 à 6h

Nous allons dans cette partie étudier l'expression des gènes pour le traitement T3 à 6h. Nous allons notamment évaluer les temps clés qui influencent l'expression des gènes et étendre cette analyse à tous les traitements et temps. Nous allons également découvrir les facteurs prédictifs qui permettent de distinguer les gènes sur-exprimés et les gènes sous-exprimés pour le traitement T3 à 6 heures.

4.1 Modèle linéaire

Nous allons étudier l'expression des gènes pour le traitement T3 à 6 heures par un modèle linéaire par rapport aux autres heures.

```
##
## Call:
## lm(formula = T3_6h_R2 ~ ., data = T3R2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6068 -0.3814  0.0046  0.3445  3.7964
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06441    0.01815  -3.549 0.000397 ***
## T3_1h_R2     0.13580    0.02491   5.451 5.79e-08 ***
## T3_2h_R2    -0.24370    0.03783  -6.441 1.56e-10 ***
## T3_3h_R2     0.18888    0.04027   4.691 2.95e-06 ***
## T3_4h_R2    -0.18577    0.04335  -4.285 1.93e-05 ***
## T3_5h_R2     1.17862    0.02970  39.686  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7061 on 1609 degrees of freedom
## Multiple R-squared:  0.9563, Adjusted R-squared:  0.9561
## F-statistic:  7035 on 5 and 1609 DF,  p-value: < 2.2e-16
```

Pour identifier les temps qui ont une réelle influence sur l'expression des gènes à ce stade nous effectuons une sélection de variables avec différents critères.

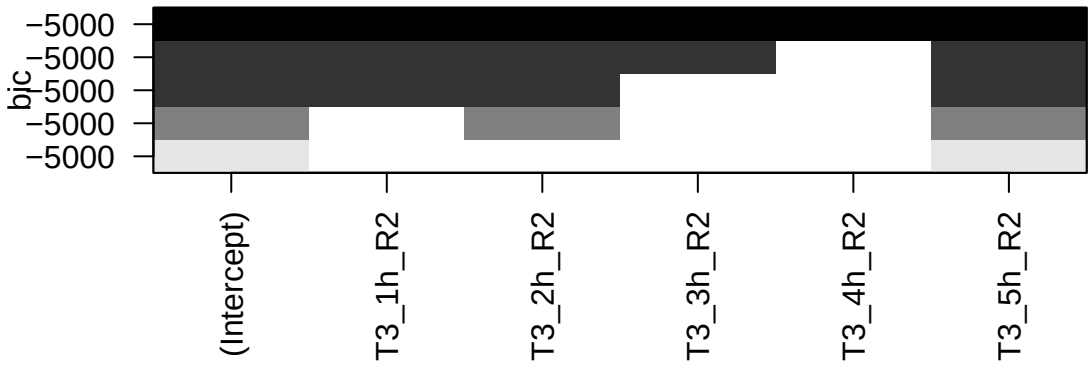


Figure 20: Selection de variable du traitement 3 selon le critère BIC et la méthode backward

On a réalisé notre sélection de variables avec tous les critères (BIC, adjr2, Cp) et avec les méthodes forward et backward. Nous avons eu les mêmes résultats.

On garde toutes les variables mais on observe quand même une gradation. Le temps précédent (5h) est le plus influent suivi du temps de démarrage (1h, 2h). On peut faire l'hypothèse d'une périodicité de temps sur l'influence des traitements sur les gènes. Il faudrait tester cette sélection de variables sur plus d'heures afin de valider ou non cette hypothèse.

4.1.1 Etude sur tous les traitements et tous les temps

Réalisons maintenant la même étude mais cette fois ci sur tous les traitements et tous les temps.

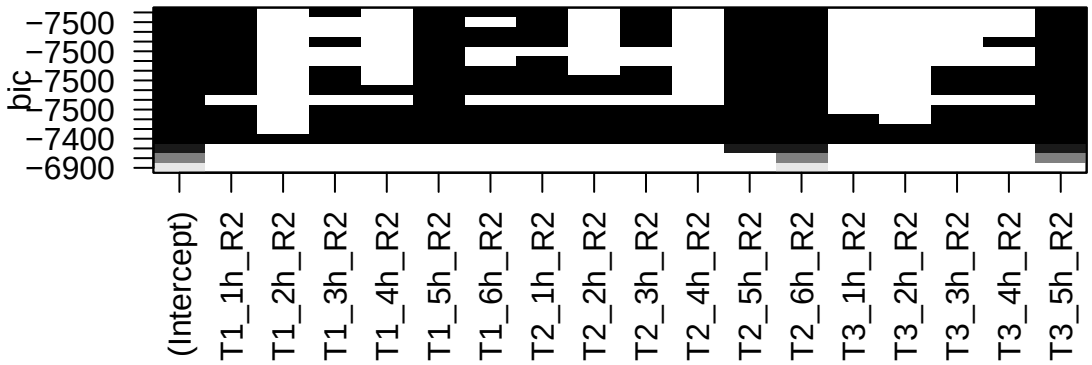


Figure 21: Selection de variable sur tous les traitement selon le critère BIC et la méthode backward

D'après la figure 21 (et les autres figures qui ont été réalisé sur le Rmd), on trouve que :
- On sélectionne les variables suivantes pour T1 : 1h, 3h, 5h, 6h, pour T2 : 1h, 3h, 5h, 6h et pour T3 : 5h. Cela rejoint l'analyse descriptive précedente : les gènes qui ont eu le traitement T2 ou le traitement T3 ont des comportements similaires.
- On retrouve, par ailleurs, les résultats de l'analyse de la figure 20 puisque les heures les plus influentes sont les heures les plus proches de 6h.

On cherche maintenant à valider ce sous-modèle en comparant avec le modèle de départ :

```
## Analysis of Variance Table
##
## Model 1: T3_6h_R2 ~ T1_1h_R2 + T1_3h_R2 + T1_5h_R2 + T1_6h_R2 + T2_1h_R2 +
##     T2_3h_R2 + T2_5h_R2 + T2_6h_R2 + T3_5h_R2
## Model 2: T3_6h_R2 ~ T1_1h_R2 + T1_2h_R2 + T1_3h_R2 + T1_4h_R2 + T1_5h_R2 +
##     T1_6h_R2 + T2_1h_R2 + T2_2h_R2 + T2_3h_R2 + T2_4h_R2 + T2_5h_R2 +
##     T2_6h_R2 + T3_1h_R2 + T3_2h_R2 + T3_3h_R2 + T3_4h_R2 + T3_5h_R2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1605 169.72
## 2   1597 168.68  8    1.0357 1.2256 0.2798
```

La p-valeur est égale 0.2798 et est supérieure à 0.05, on ne rejette donc pas H0 au risque de 5%, on accepte donc le sous modèle.

4.1.2 Lasso

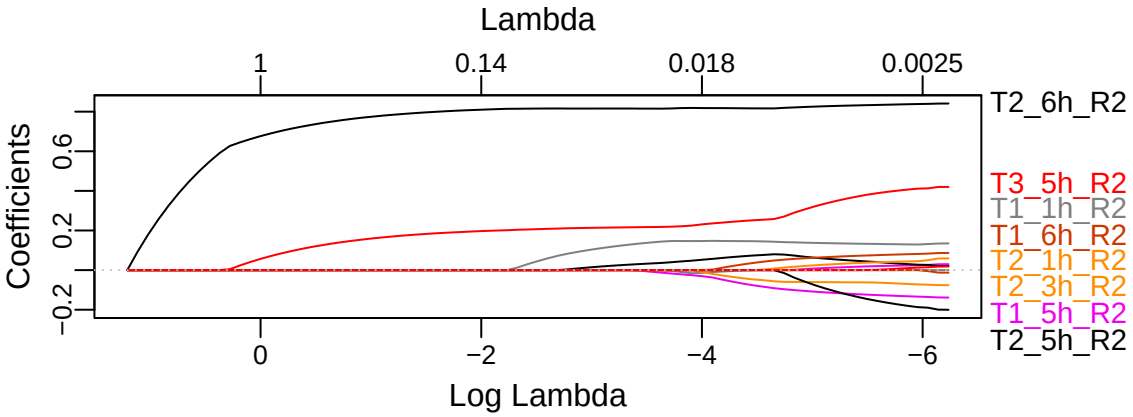


Figure 22: Sélection de variables avec le critère lasso

On voit que les variables les plus affectantes sont : -T1: 1h, 5h, 6h -T2 : 1h, 3h, 5h, 6h -T3 : 5h
Ce modèle est très proche du sous-modèle que l'on vient de valider, il manque seulement T1_3h.

4.2 Modèle linéaire généralisé

On veut chercher les variables prédictives qui permettent de discriminer les gènes sur-exprimés (Y>1) des gènes sous-exprimés (Y<-1) à 6h pour le traitement T3.
La sortie est binaire, nous allons donc chercher les variables prédictives par une régression logistique sur le réplicat 2 uniquement (puisque nous avions montré précédemment que le réplicat 1 était similaire en comportement au réplicat 2).

```
##
## Call:
## glm(formula = T36HR2_binomial$T3_6h_R2 ~ ., family = binomial(link = "logit"),
##     data = T36HR2_binomial, control = glm.control(maxit = 100))
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -5.840e-06  -2.110e-08   2.110e-08   2.110e-08   5.176e-06
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -1.334e+00  6.111e+04       0        1
## T1_1h_R2    -4.964e-01  1.913e+05       0        1
## T1_2h_R2    -2.698e+00  1.760e+05       0        1
## T1_3h_R2     5.340e+00  1.298e+05       0        1
## T1_4h_R2     1.060e+00  1.984e+05       0        1
## T1_5h_R2    -1.204e+00  2.134e+05       0        1
## T1_6h_R2    -3.859e-01  1.714e+05       0        1
## T2_1h_R2    -4.752e-01  1.925e+05       0        1
## T2_2h_R2     1.326e+00  1.541e+05       0        1
## T2_3h_R2     2.075e+00  1.655e+05       0        1
## T2_4h_R2    -4.261e+00  1.722e+05       0        1
## T2_5h_R2    -1.334e+00  1.624e+05       0        1
## T2_6h_R2     1.402e+01  9.531e+04       0        1
## T3_1h_R2    -2.748e-01  1.805e+05       0        1
## T3_2h_R2     4.596e-01  2.126e+05       0        1
## T3_3h_R2    -1.538e+00  1.525e+05       0        1
## T3_4h_R2     9.621e-01  1.485e+05       0        1
## T3_5h_R2     3.683e+00  1.932e+05       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.2342e+03  on 1614  degrees of freedom
## Residual deviance: 4.5579e-10  on 1597  degrees of freedom
## AIC: 36
##
## Number of Fisher Scoring iterations: 29
```

En prenant en compte toutes les variables, le modèle linéaire généralisé n'arrive pas à bien ajuster le modèle. On remarque que toutes les p-valeurs sont égales à 1.
Ceci est probablement dû au fait que les variables sont très liées les unes aux autres.

Cependant, d'après la table lorsqu'on fait une sélection de variables "backward" sur notre modèle, on obtient que T3_6h_R2 peut s'expliquer par les variables : T1_4h_R2, T1_6h_R2, T2_5h_R2, T3_3h_R2.

Peu importe les combinaisons de traitement qu'on prend en pour expliquer T3_6h_R2, on obtient la même erreur (des p-valeurs toutes égales à 1) sauf lorsqu'on prend seulement le traitement 1.
Dans ce cas, on obtient que T3_6h_R2 s'explique par T1_1h_R2, T1_2h_R2, T1_3h_R2, T1_4h_R2, T1_6h_R2 avec une sélection de variables "backward". Autrement dit, on enlève seulement l'heure 5 ce qui ne nous aide pas vraiment.
Toutes les sélections de variable et les combinaisons citées ici sont dans le Rmd.

5 Etude de l'expression des gènes pour le traitement T1 à 6h

Nous allons dans cette partie étudier l'expression des gènes pour le traitement T1 à 6h. Nous allons notamment repérer les temps influent l'expression de ces gènes ainsi que les variables prédictives qui permettent de discriminer les gènes sur-exprimés des gènes sous-exprimés, à 6h pour le traitement T1.

5.1 Modèle linéaire

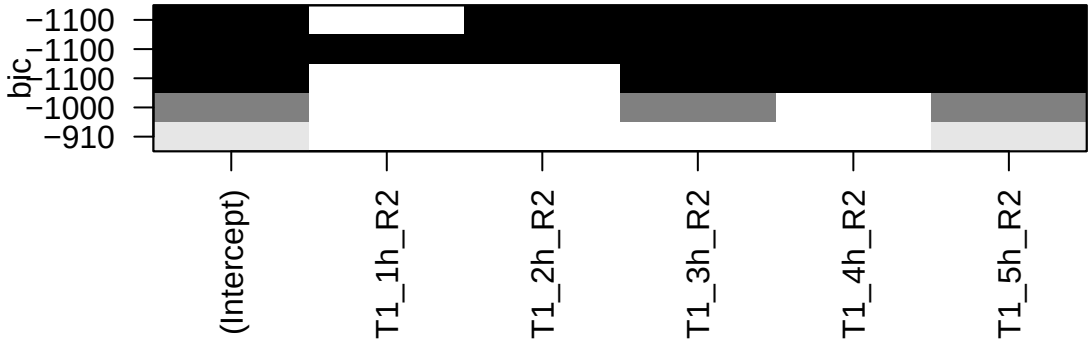


Figure 23: Selection de variable du traitement 1 selon le critère BIC et la méthode backward

On a réalisé notre sélection de variables avec tous les critères (BIC, adjr2, Cp) et avec les méthodes forward et backward. Nous avons eu les mêmes résultats :

On garde toutes les variables sauf T1_1h_R2.

On cherche à valider ce sous-modèle :

```
## Analysis of Variance Table
##
## Model 1: T1_6h_R2 ~ T1_2h_R2 + T1_3h_R2 + T1_4h_R2 + T1_5h_R2
## Model 2: T1_6h_R2 ~ T1_1h_R2 + T1_2h_R2 + T1_3h_R2 + T1_4h_R2 + T1_5h_R2
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1   1610 257.34
## 2   1609 257.29  1  0.051971 0.325 0.5687
```

p-valeur = 0.5687 > 0.05, on ne rejette pas H0 au risque de 5%, donc on valide le sous-modèle.

5.1.1 Etude sur tous les traitements et tous les temps

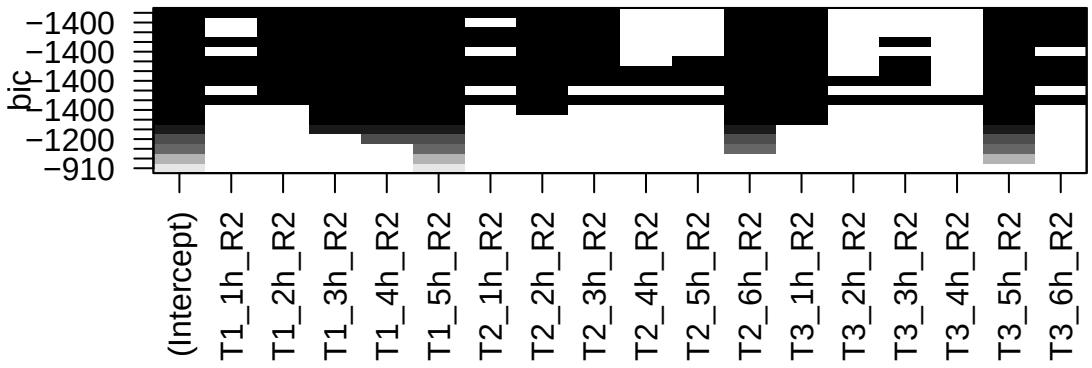


Figure 24: Selection de variable sur tous les traitement selon le critère BIC et la méthode backward

D'après la figure 24 (et les autres figures qui ont été réalisé sur le Rmd), on sélectionne les variables suivantes pour T1 : 1h, 2h, 3h, 4h, 5h, 6h, pour T2 : 1h, 2h, 3h, 6h et pour T3 : 1h, 5h, 6h.

On cherche maintenant à valider ce sous-modèle en comparant avec le modèle de départ :

```
## Analysis of Variance Table
##
## Model 1: T1_6h_R2 ~ T1_1h_R2 + T1_2h_R2 + T1_3h_R2 + T1_4h_R2 + T1_5h_R2 +
##     T2_1h_R2 + T2_2h_R2 + T2_3h_R2 + T2_6h_R2 + T3_1h_R2 + T3_5h_R2 +
##     T3_6h_R2
## Model 2: T1_6h_R2 ~ T1_1h_R2 + T1_2h_R2 + T1_3h_R2 + T1_4h_R2 + T1_5h_R2 +
##     T2_1h_R2 + T2_2h_R2 + T2_3h_R2 + T2_4h_R2 + T2_5h_R2 + T2_6h_R2 +
##     T3_1h_R2 + T3_2h_R2 + T3_3h_R2 + T3_4h_R2 + T3_5h_R2 + T3_6h_R2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1602 207.91
## 2   1597 206.68  5    1.2324 1.9045 0.0906 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-valeur $= 0.09 > 0.05$, on ne rejette pas H0 au risque de 5%, on accepte le sous-modèle.

On voit que l'expression des gènes à 6h pour le traitement T1 est affecté par - les heures finales (3h, 4h, 5h) du traitement T1 - les heures débutantes (1h, 2h, 3h) et finale(6h) du traitements T2 - l'heure de début (1h) et les heures finales (5h, 6h) du traitement T3.

5.1.2 Lasso

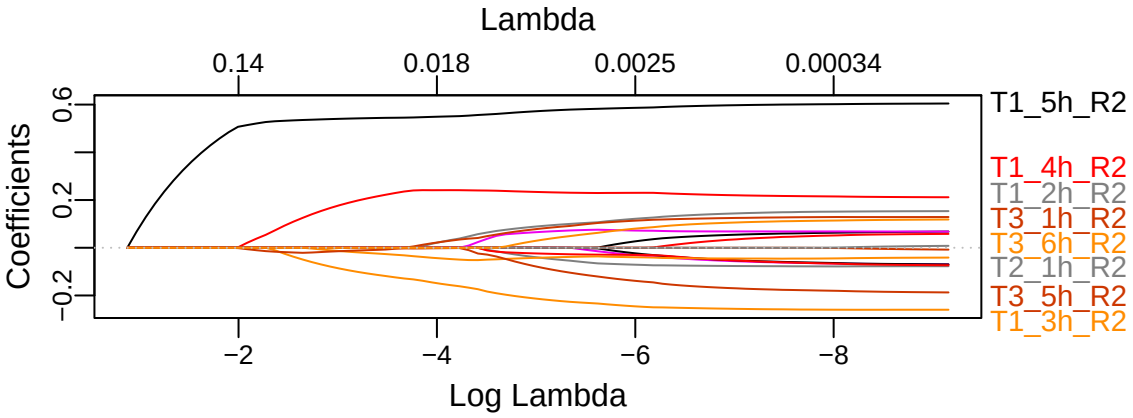


Figure 25: Sélection de variables avec le critère lasso

On voit que les variables les plus affectantes sont : -T1 : 2h, 3h, 4h, 5h -T2 : 1h -T3 : 1h, 5h, 6h

5.2 Modèle linéaire généralisé

On veut chercher les variables prédictives qui permettent de discriminer les gènes sur-exprimés (Y>1), les gènes sous-exprimés (Y<-1), et les gènes non-exprimés à 6h pour le traitement T1.
La sortie n'est pas binaire, nous allons donc chercher les variables prédictives par une régression logistique multinomiale sur le réplicat 2.

```
## Call:
## multinom(formula = Y ~ ., data = dfmodel, trace = F)
##
## Coefficients:
##              (Intercept)    T1_1h_R2   T1_2h_R2   T1_3h_R2   T1_4h_R2  T1_5h_R2
## sous-exprime   -4.905461  0.03987164 -0.6813844  0.6556245 -0.5017774 -3.256918
## sur-exprime    -5.280181 -0.81065181  0.4547041 -1.4371988  2.4496719  1.417759
##                T2_1h_R2  T2_2h_R2   T2_3h_R2    T2_4h_R2  T2_5h_R2  T2_6h_R2
## sous-exprime  0.3941555 0.5988196 -0.6684853 -0.09437566 0.4183293 0.2568884
## sur-exprime  -0.3300347 0.4580245 -0.5060857 -0.51678735 1.1800846 0.3912118
##                T3_1h_R2   T3_2h_R2  T3_3h_R2    T3_4h_R2   T3_5h_R2   T3_6h_R2
## sous-exprime -0.7000935 -0.2273308 0.7588703 -0.08713926  0.4014076 -0.9723514
## sur-exprime   0.2728899 -1.0812748 1.0944825  0.15461699 -2.1851137  0.8626062
##
## Std. Errors:
##              (Intercept)  T1_1h_R2  T1_2h_R2  T1_3h_R2  T1_4h_R2  T1_5h_R2
## sous-exprime   0.3156944 0.3813541 0.4058223 0.3367225 0.3741317 0.3577547
## sur-exprime    0.4140214 0.6247254 0.5856648 0.4807367 0.4765345 0.5190984
##                T2_1h_R2  T2_2h_R2  T2_3h_R2  T2_4h_R2  T2_5h_R2  T2_6h_R2
## sous-exprime 0.3631824 0.4340794 0.4363998 0.3669008 0.4648246 0.4446681
## sur-exprime  0.5459498 0.4953546 0.5387247 0.5351794 0.6314579 0.4908601
##                T3_1h_R2  T3_2h_R2  T3_3h_R2  T3_4h_R2  T3_5h_R2  T3_6h_R2
## sous-exprime 0.3039686 0.4338753 0.4426995 0.4133474 0.4388571 0.4447722
## sur-exprime  0.5180650 0.5888178 0.6086898 0.4801942 0.5948142 0.5228239
##
## Residual Deviance: 699.067
## AIC: 771.067
```

En faisant une sélection de variables en mode "backward" on peut exprimer T1_6h_R2 par : T1_3h_R2, T1_4h_R2, T1_5h_R2, T2_2h_R2, T2_5h_R2, T3_1h_R2, T3_5h_R2, T3_6h_R2

6 Conclusion

Dans une première partie, nous avons vu que les traitements 2 et 3 étaient fortement corrélés et que les réplicats 1 et 2 ont des résultats similaires. Nous avons également vu que le jeu de données a été fait de sorte que les gènes soient soit très sur-exprimés (valeurs ≥ 2), soit très sous-exprimés (valeurs $\leq$ -2) à 6h pour le traitement 3 (et donc pour le traitement 2 aussi).

Dans un second temps, nous avons essayé de regrouper les traitements et les gènes. Pour les traitements, nous sommes arrivés à les classer en trois groupes :
- T1 + la première heure de T2 et T3
- T2 et T3 aux heures 2 et 3
- T2 et T3 aux heures 4, 5 et 6

On a classé les gènes en deux groupes qui peuvent se distinguer par leur valeur à T3_6h : ceux sous-exprimés d'un côté et ceux sur-exprimés de l'autre.

Nous avons ensuite étudié l'expression des gènes pour le traitement T3 à 6h et remarqué qu'il pouvait s'exprimer en fonction d'une valeur au début, une au milieu et une à la fin des autres traitements ainsi que de T3 à 5h. Cependant, le modèle linéaire et le modèle linéaire généralisé n'ont pas été très efficaces. Nous pensons que cela est dû au fait que les variables et les gènes sont très liés les uns aux autres.

Pour finir, nous avons fait la même chose mais cette fois-ci pour T1 à 6h. L'expression des gènes à ce temps et pour ce traitement est un peu plus complexe à prévoir que celle de T3 à 6h puisqu'elle nécessite les heures finales de T1, les heures de début et l'heure de fin de T2 et l'heure de début et celles de fin de T3.