Rapport de projet Analyse de données & Eléments de modélisation statistique

Xiaoya Wang, Mickael Song, Yessine Jmal, Florian Grivet

Contents

1 Introduction **2**

2 Analyse du jeu de données **2**

2.1 Statistiques descriptives et préparation du jeu de données 2

2.2 Analyse en composante principale 6

3 Modèle linéaire **8**

3.1 Etude de l'expression des gènes pour le traitement T3 à 6h 8

3.2 Etude sur tous les traitements et tous les temps 8

3.3 à faire 11

3.4 Etude de l'expression des gènes pour le traitement T1 à 6h : 11

3.5 Lasso 13

Tout ce qui se trouve dans le Rmarkdown mais pas dans le pdf est indiqué par ce symbole : (%%)

1 Introduction

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

$$Y_{gtsr} = log_2(Xgtsr + 1) - log_2(Xgt0 + 1)$$

avec
- $Xgtsr$ 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}
- $Xgt0$ l'expression du gène g pour un traitement de référence t0

Nous allons répartir l'étude de ce jeu de données en 4 parties :
- Analyse du jeu de données
- Clustering
- Etude de l'expression des gènes pour le traitement T3 à 6h
- 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 que notre jeu de données est complet. (%%)

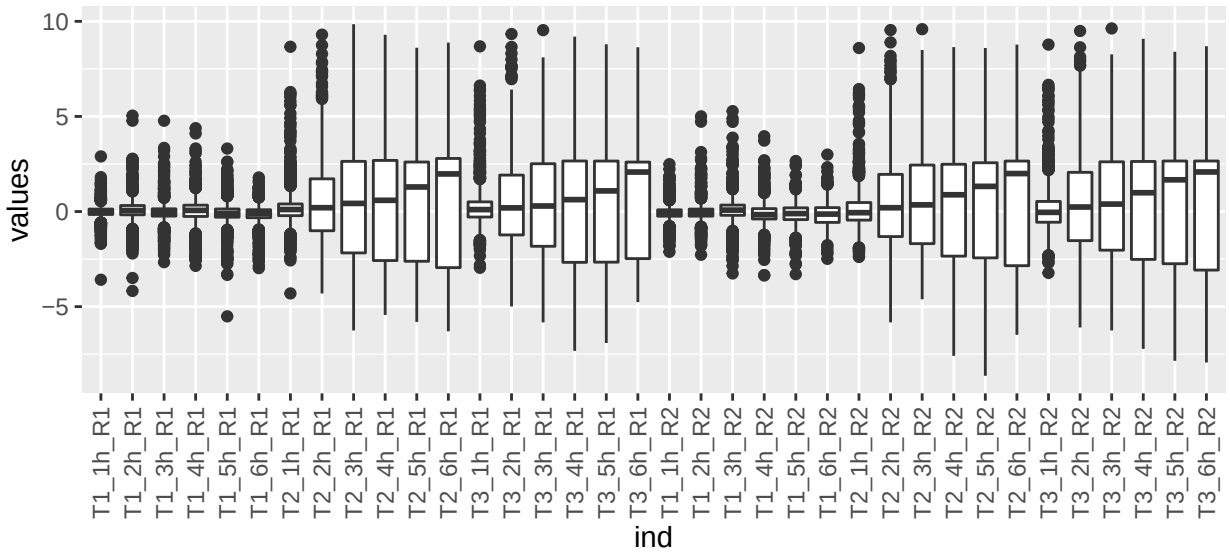


Figure 1: Boxplots des 36 variables

On remarque dans la figure 1 que les boxplots du traitement 2 et du traitement 3 ne sont pas centrés, ils ont donc un effet non nul sur les gènes. En étudiant la forme des boxplots, on remarque une dissymétrie pour ces deux traitements, des 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 beaucoup d'outliers.

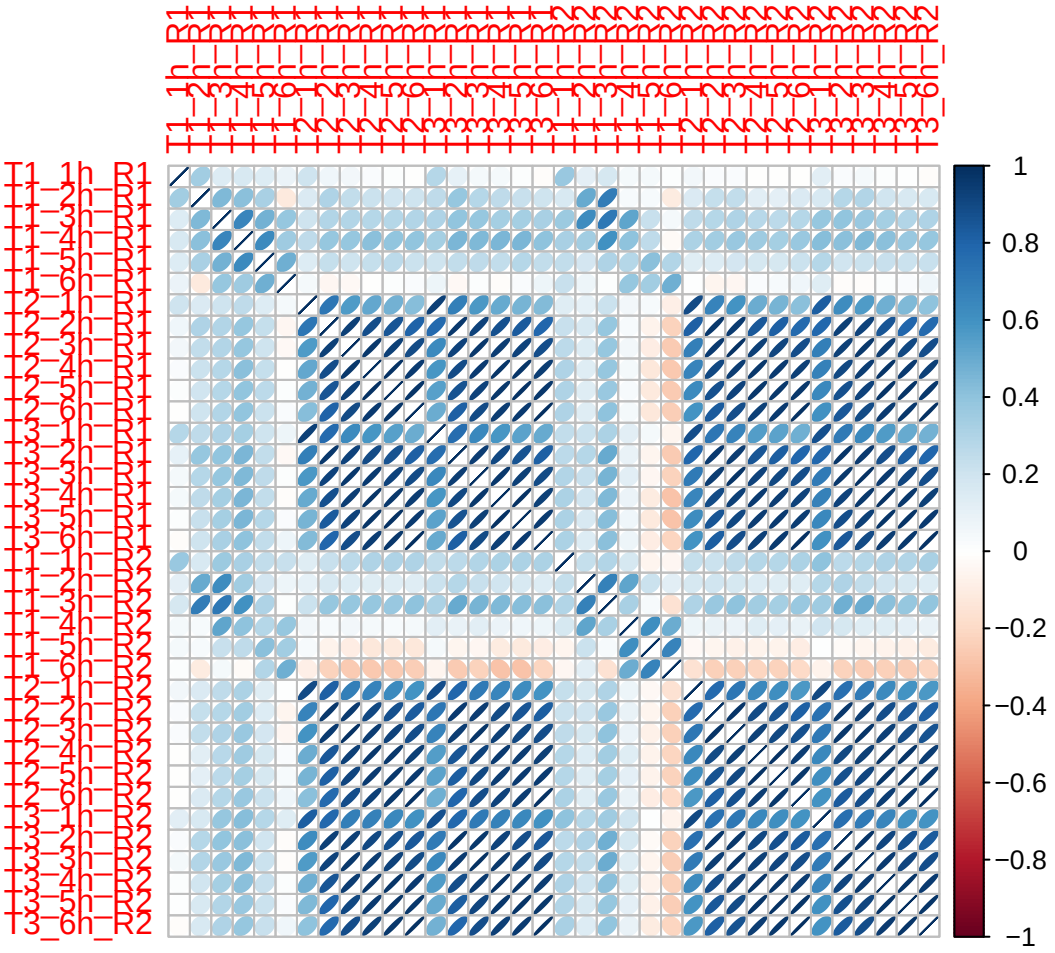


Figure 2: Graphique des corrélations des 36 variables

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 décorellés du traitements 1.
On peut également noter le fait que, pour un traitement donné, le réplicat 1 et le réplicat 2 sont fortement correlé entre eux, ce qui est cohérent puisque ce sont des réplicats biologiques. On pourra donc par la suite uniquement faire notre étude uniquement sur 1 seul réplicat, sans perdre trop d'informations.

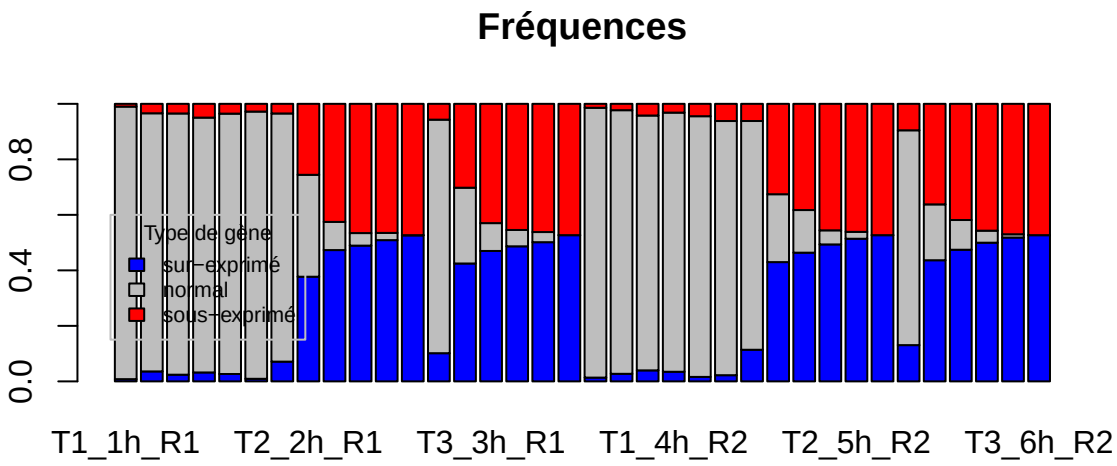


Figure 3: Fréquence de l'expression des gènes sous-exprimés, normaux et sur-exprimés en fonction des traitements

La graphique 3 représente la fréquence des gènes "sous-exprimés", "normaux" 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.

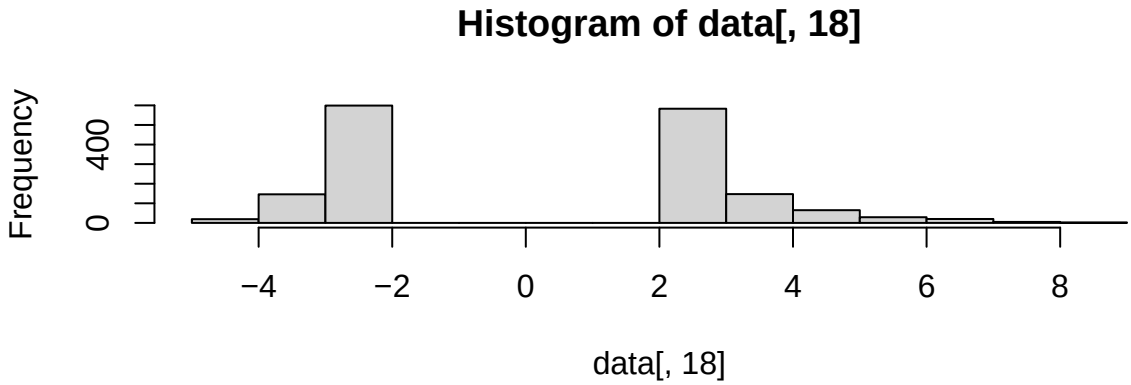


Figure 4: Fréquence de la valeur des gènes du traitement 3 à l'heure 6

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

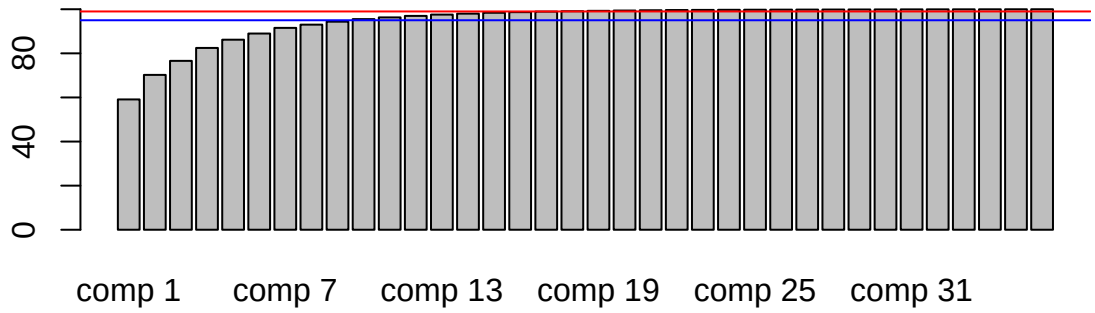


Figure 5: Variance expliquée cumulée (en %) des différentes composantes principales

D'après le graphique 5, on note que :
- Pour avoir 95% de l'information, on peut réduire nos données à 10 dimensions.
- Pour avoir 99% de l'information il suffit de se placer en dimension 18.

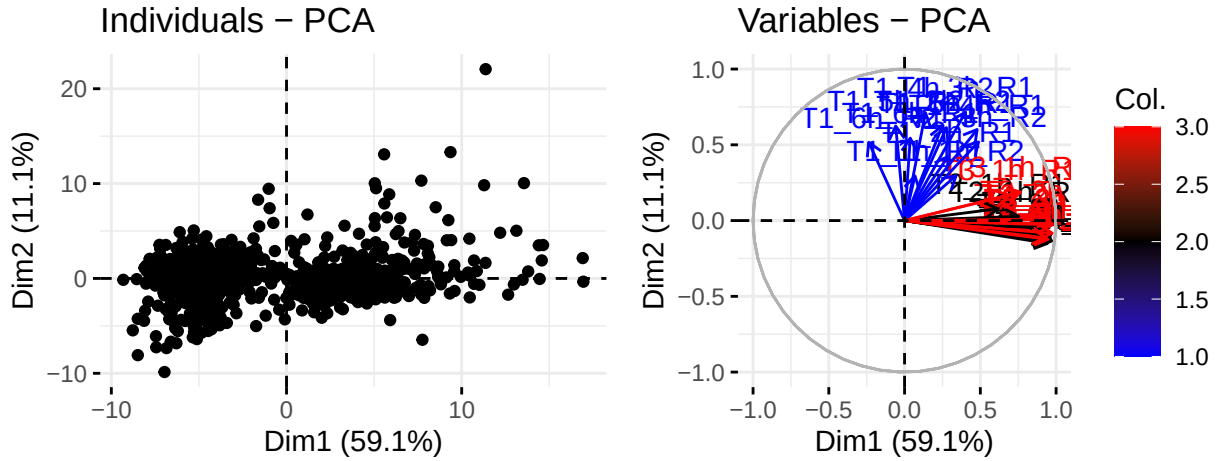


Figure 6: Visualitaion de l'ACP sur les deux premières composantes principales, pour les individus (à gauche) et pour les variables (à droite)

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, il se trouve aux extrémités du cercle (si le gène devient très sous-exprimé ou très sur-exprimé) et s'il ne réagit pas beaucoup à l'un de ces traitement il se trouve au milieu du cercle.

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 cercle. Par contre, si le gène ne réagit pas beaucoup au traitement 1, il se trouve au milieu.

2.2.1 Analyse en composante principale sur les données transposées

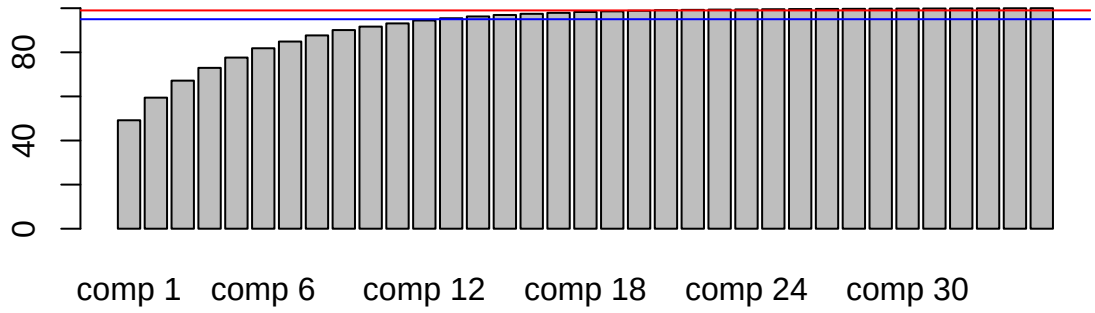
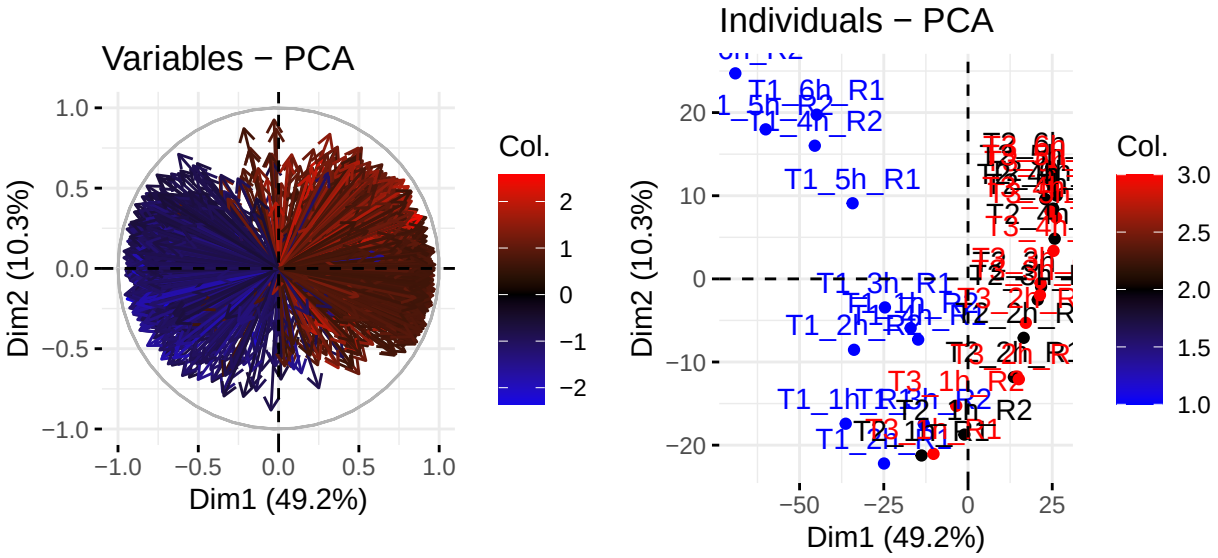


Figure 7: Variance expliquée cumulée (en %) des différentes composantes principales

D'après le graphique 7, on note que :
Pour avoir 95% de l'information, on peut réduire nos données à 13 dimensions.
Pour avoir 99% de l'information il suffit de se placer en dimension 21.



Commentaires à faire

3 Modèle linéaire

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

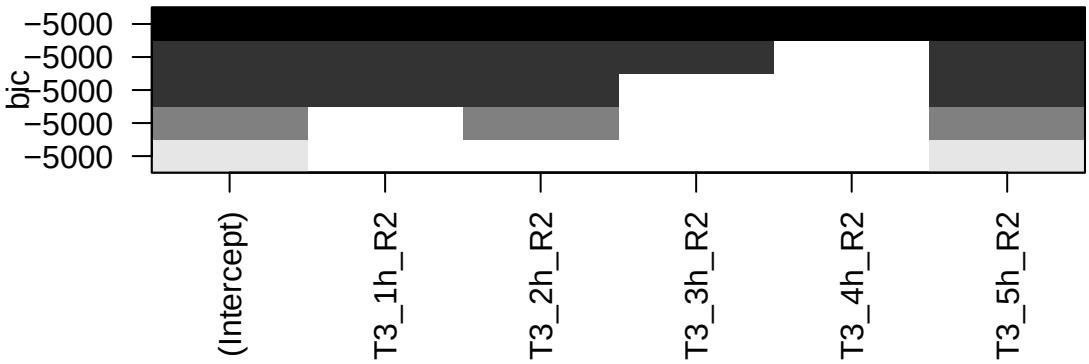


Figure 8: Selection de variable du traitement 3 selon le critère BIC et la méthode backward

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 variable sur plus d'heures afin valider ou non cette hypothèse.

3.2 Etude sur tous les traitements et tous les temps

D'après la figure 9 (et les autres figures qui ont été réalisé sur le R-markdown), 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 traitemet T3 ont des comportements similaires.
- On retrouve, par ailleurs, les résultats de l'analyse de la figure 8 puisque les heures les plus influentes sont les heures les plus proches de 6h.

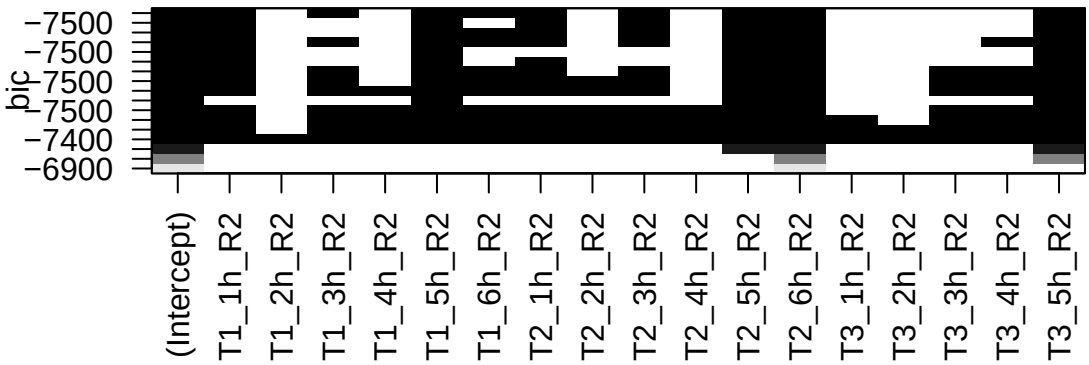


Figure 9: Selection de variable sur tous les traitement selon le critère BIC et la méthode backward

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 14.937
## 2   1597 14.846  8  0.091151 1.2256 0.2798
```

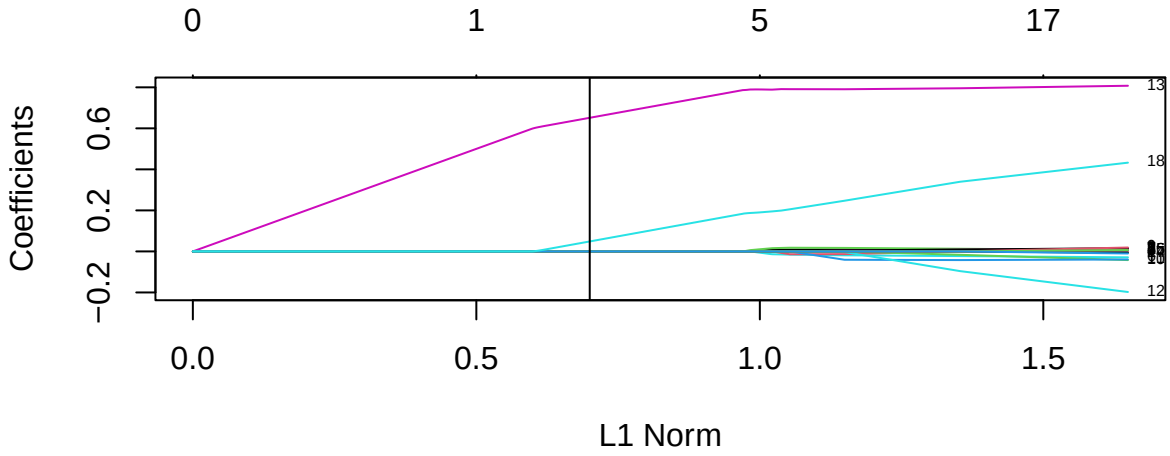
La p-valeur est égale 0.2798 et est supérieure à 0.05, on ne rejette don pas H0 au risque de 5%, on accepte donc le sous_modèle.

3.2.1 Lasso

```
lambda_seq=seq(0,1,0.001)
x = model.matrix(T3_6h_R2~.,data=R2)
y = data$T3_6h_R2
fitlasso <- glmnet(x, y , alpha = 1, lambda = lambda_seq, family=c("gaussian"), intercept=F)
plot(fitlasso,label= TRUE)
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```

```
abline(v = 0.7)
```



On voit que les variables les plus affectantes sont (13, 18): T3_1h_R2, T3_6h_R2 ??

3.3 à faire

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.

```
sur_exp = T3R2[T3R2>1]
```

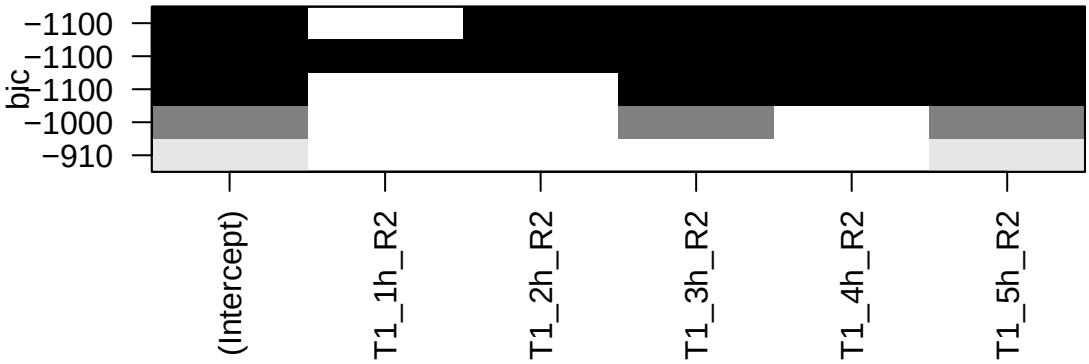
```
apply(T3R2["T3_1h_R2">1], 2, sum)
```

```
##      T3_1h_R2      T3_2h_R2      T3_3h_R2      T3_4h_R2      T3_5h_R2
## -1.170375e-14  2.327478e-14 -5.571932e-15  2.195119e-14  1.385697e-14
##      T3_6h_R2
##  5.062617e-14
```

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

```
T1 = data[grep("T1", names(data), value=TRUE)]
T1R2 = T1[grep("R2", names(T1), value=TRUE)]
```

```
choix=regsubsets(T1_6h_R2~., data = T1R2, nbest = 1, nvmax = 5, method = "backward")
plot(choix,scale = "bic")
```




```
mod.debut = lm(T1_6h_R2 ~ ., data = T1R2)
mod.fin = lm(T1_6h_R2 ~ T1_2h_R2+T1_3h_R2+T1_4h_R2+T1_5h_R2, data = T1R2)
anova(mod.fin,mod.debut)
```

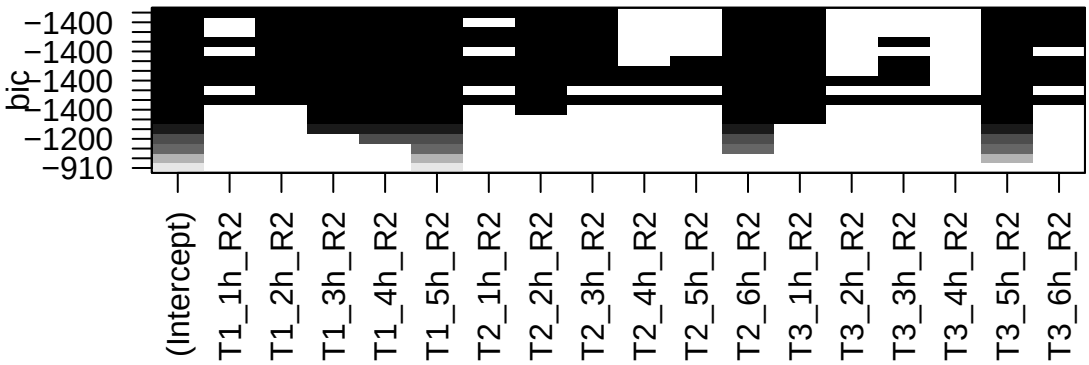
```
## Analysis of Variance Table
```

```
##
## 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 787.94
## 2   1609 787.78  1   0.15913 0.325 0.5687
```

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

On étudie sur tous les traitements et tous les temps:

```
choix=regsubsets(T1_6h_R2~., data = R2, nbest = 1, nvmax = 18, method = "backward")
plot(choix,scale = "bic")
```



On choisit le sous-modèle avec la criètre bic et méthode backward, on cherche à le valider:

```
mod.debut = lm(T1_6h_R2 ~ ., data = R2)
mod.fin = lm(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, data = R2)
anova(mod.fin,mod.debut)
```

```
## 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 636.58
```

```
## 2   1597 632.81  5    3.7733 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.

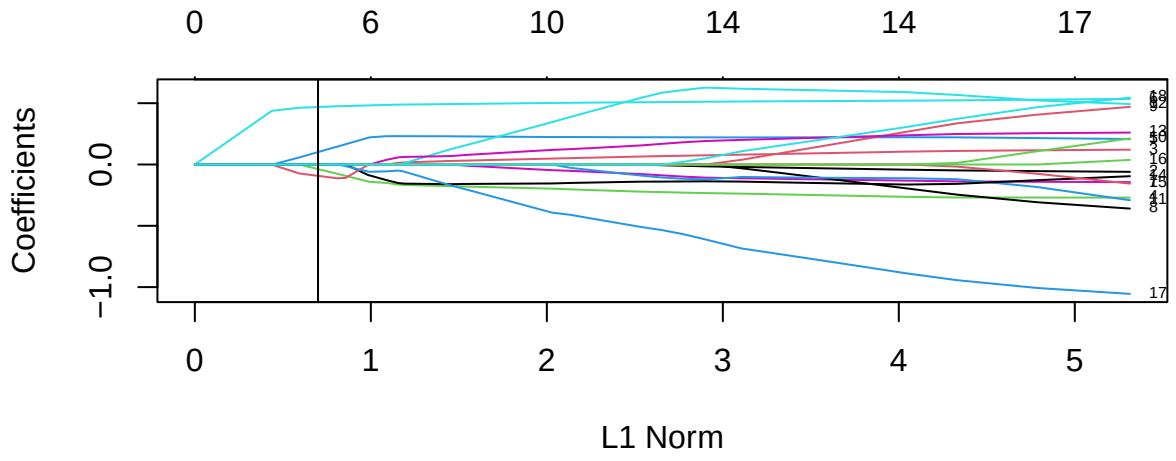
On voit que l'expression des gènes à 6h pour le traitement T1 est affecté par - les heures finales (3h, 4h, 5h) du traitement T1 - les heures débutantes (1h, 2h, 3h) et finale(6h) du traitements T2 - les heures débutantes (1h) et finales (5h, 6h) du traitement T3

3.5 Lasso

```
lambda_seq=seq(0,1,0.001)
x = model.matrix(T1_6h_R2~.,data=R2)
y = data$T1_6h_R2
fitlasso <- glmnet(x, y , alpha = 1, lambda = lambda_seq, family=c("gaussian"), intercept=F)
plot(fitlasso,label= TRUE)
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```

```
abline(v = 0.7)
```



On voit que les variables les plus affectantes sont ??