Rapport de projet Analyse de données & Eléments de modélisation statistique

Xiaoya Wang, Mickael Song, Yessine Jmal, Florian Grivet

2022-11-27

Contents

1 Analyse du jeu de données

1.1 Statistiques descriptives et préparation du jeu de données

```
# Chargement des données
data = read.delim("Data_Etudiants_2023.txt",header=TRUE, sep=";")
```

```
datapy = pd.read_csv("Data_Etudiants_2023.txt", sep=";")
```

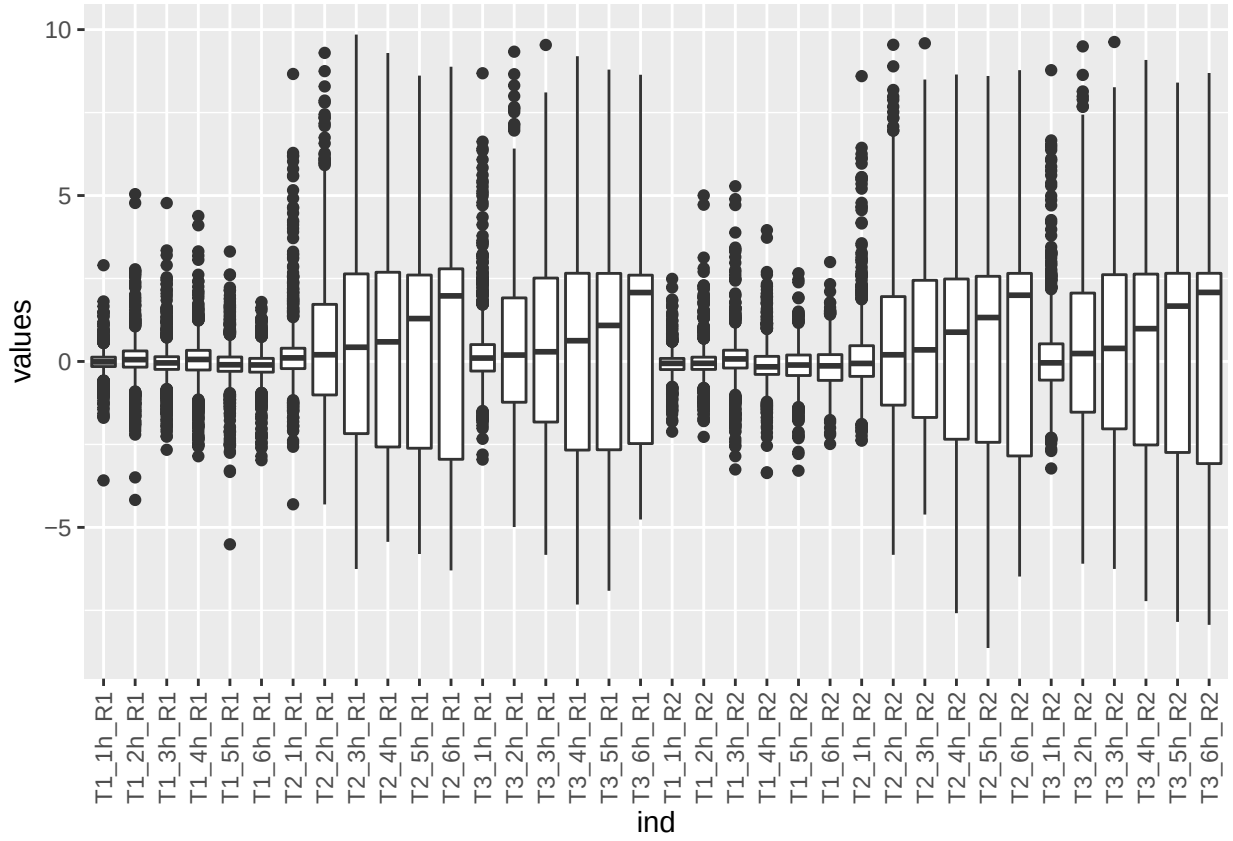
Table 1: Les premières lignes du jeu de données.

	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

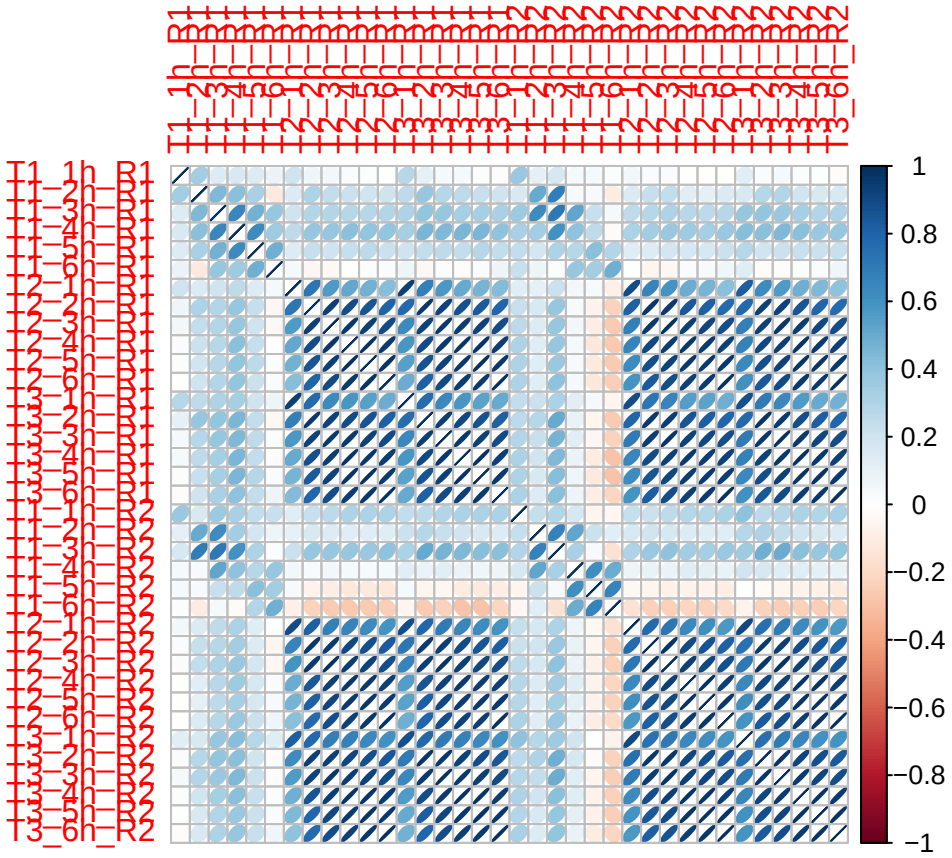
Avec le résultat de la commande `datapy.isnull().sum()`, on voit bien que notre jeu de données est complet.

```r
# Boxplot
ggplot(stack(data), aes(x = ind, y = values))+
  geom_boxplot()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



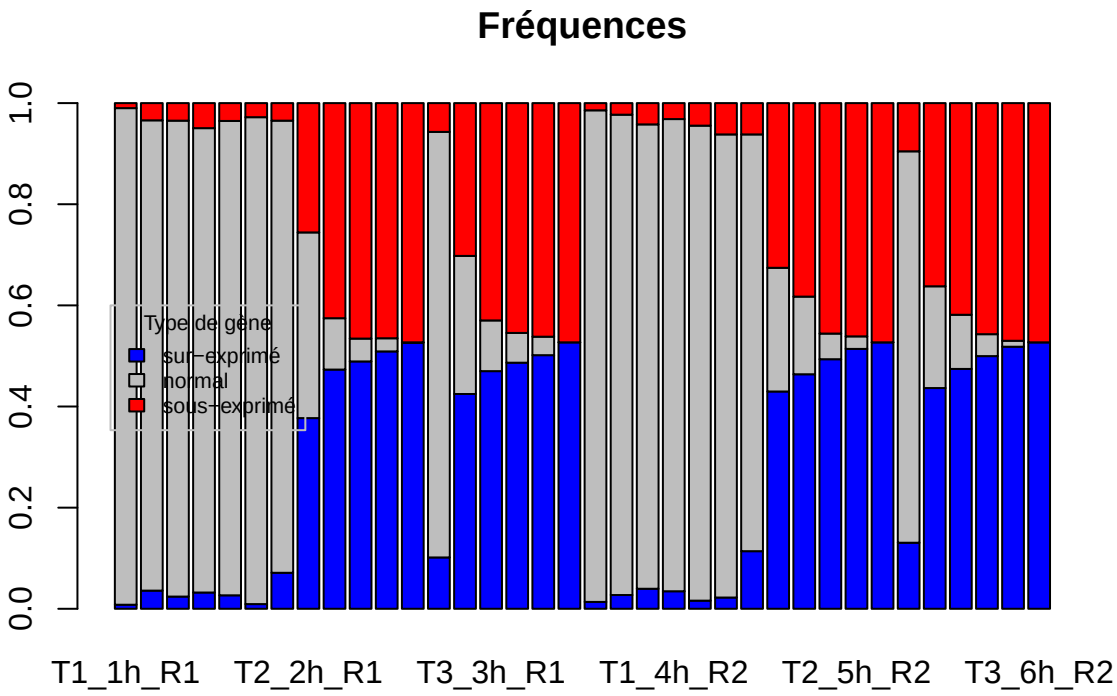
On remarque que les traitements 2 et 3 ont des boxplots similaires que ça soit pour le réplicat 1 ou pour le réplicat 2. On peut faire l'hypothèse que ces deux traitements ont des résultats similaires. On remarque que ces deux traitements ne sont pas centrés donc qu'ils ont un effet non nul sur les gènes. On note également une dissymétrie pour ces traitements ainsi que de nombreux outliers et une forte variabilité entre les individus. Le traitement 1 est quant à lui beaucoup plus réduit et centré en 0. Le traitement semble donc ne pas avoir d'effet sur les gènes. On remarque que le traitement 1 est symétrique mais possède également beaucoup d'outliers.

```
corrplot(cor(data), method="ellipse")
```



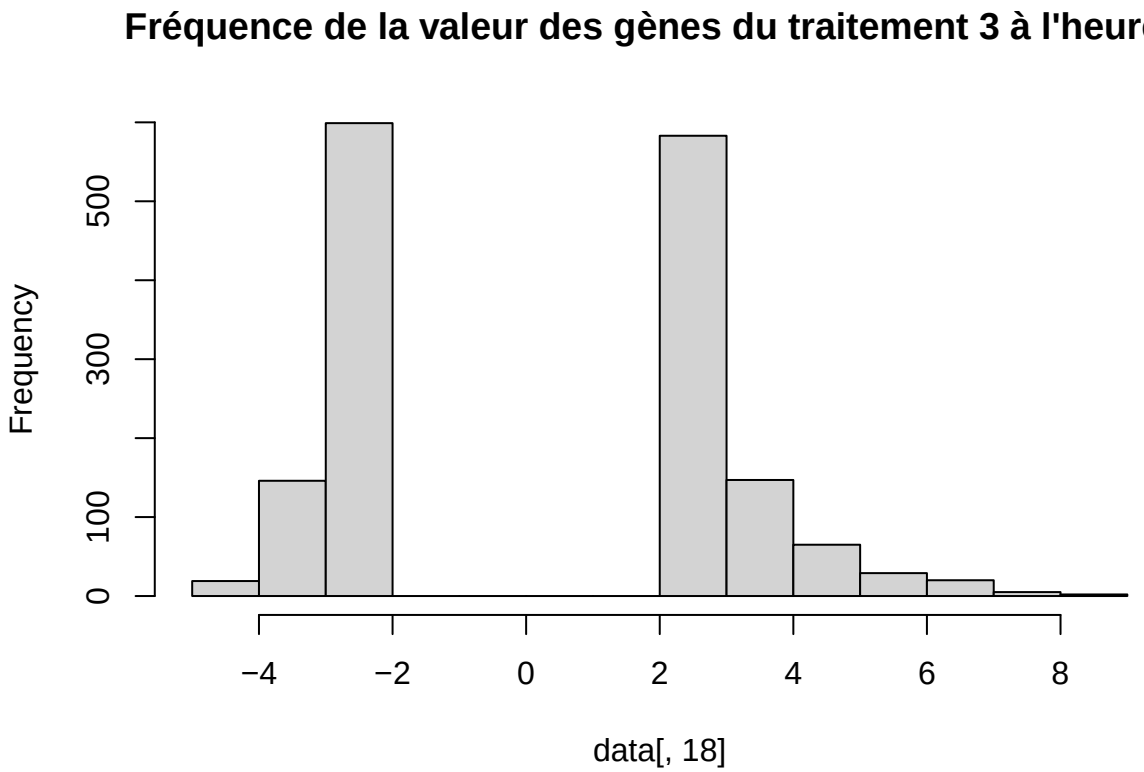
Sur ce graphique des corrélations on remarque que le traitement 1 réplicat 1 est fortement correlé au traitement 1 réplicat 2 mais ces derniers semblent totalement décorellés des traitements 2 et 3. On remarque également que les traitements 2 et 3 que ce soit le réplicat 1 ou 2 sont fortement corrélés.

```
# Fréquence de l'expression des gènes ("sous-exprimés", "normaux" et "sur-exprimés") en fonction des tr
BP = array(rep(rep(0,3),36), dim=c(3,36))
BP[1,] = apply(data>1, 2, sum)
BP[2,] = apply(data<=1&data>=(-1), 2, sum)
BP[3,] = apply(data<(-1), 2, sum)
BP = BP/nrow(data)
barplot(BP, main="Fréquences", col=c("blue", "grey", "red"), names.arg=c(attributes(data)$names))
legend(0,0.6, legend=c("sur-exprimé", "normal", "sous-exprimé"), title="Type de gène", box.col="grey",
```



Ce graphique représente la fréquence des gènes “sous-exprimés”, “normaux” et “sur-exprimés” pour chaque traitement à toute heure sur les deux réplicats. Ce graphique appuie notre hypothèse que les traitements 2 et 3 sont similaires et que le traitement 1 n’a pas beaucoup d’effet.

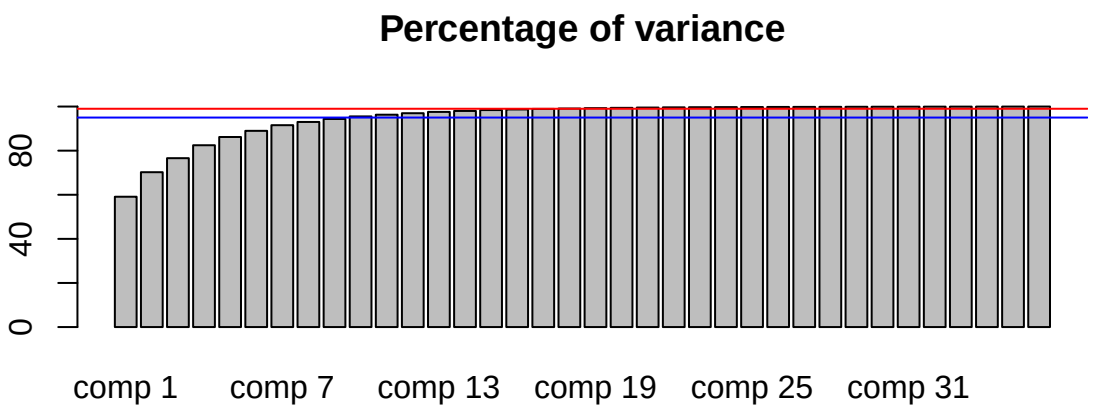
```
hist(data[,18], main="Fréquence de la valeur des gènes du traitement 3 à l'heure 6")
```



On remarque qu’à l’heure 6 (la dernière) du traitement 3, tous les gènes sont soit très sur-exprimé (valeurs ≥ 2), soit très sous-exprimé (valeurs ≤ 2). C’est comme ça que les gènes du jeu de données ont été choisi.

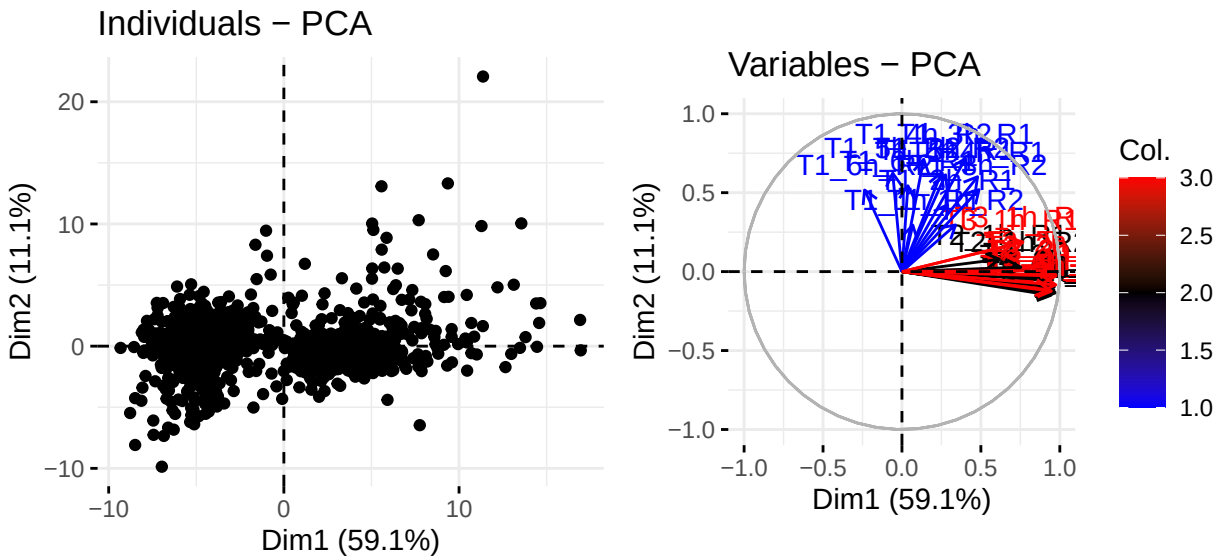
1.2 Analyse en composante principale

```
barplot(cumsum(res.pca$eig[, "percentage of variance"]), main="Percentage of variance")
abline(h=95, col="blue")
abline(h=99, col="red")
```



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.

```
col_trait = rep(rep(c(1,2,3), each=6),2)
g1 = fviz_pca_ind(res.pca, label="none")
g2 = fviz_pca_var(res.pca, col.var=col_trait) + scale_color_gradient2(low="blue", mid="black", high="red
grid.arrange(g1,g2,ncol=2)
```



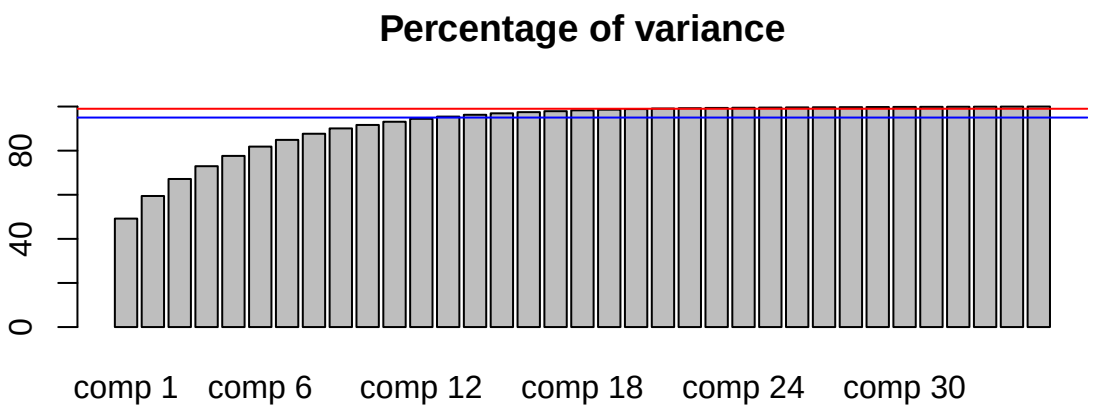
Les traitements 2 et 3 semblent avoir le même comportement et son selon l'axe 1.
Le traitement 1 suit l'axe 2.

La commande `res.pca$var$cor` nous donne la composition des axes.
L'axe 1 nous dit si un gène réagit au traitement 2 ou 3. Si le gène réagit fortement à l'un de ces traitements, il se retrouve sur un côté du graphique (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.

L'axe 2 nous dit si un gène réagit au traitement 1. Si le gène réagit fortement à ce traitement, il se retrouve en haut s'il devient sur-exprimé, en bas s'il devient sous-exprimé et s'il ne réagit pas beaucoup, il se trouve au milieu.

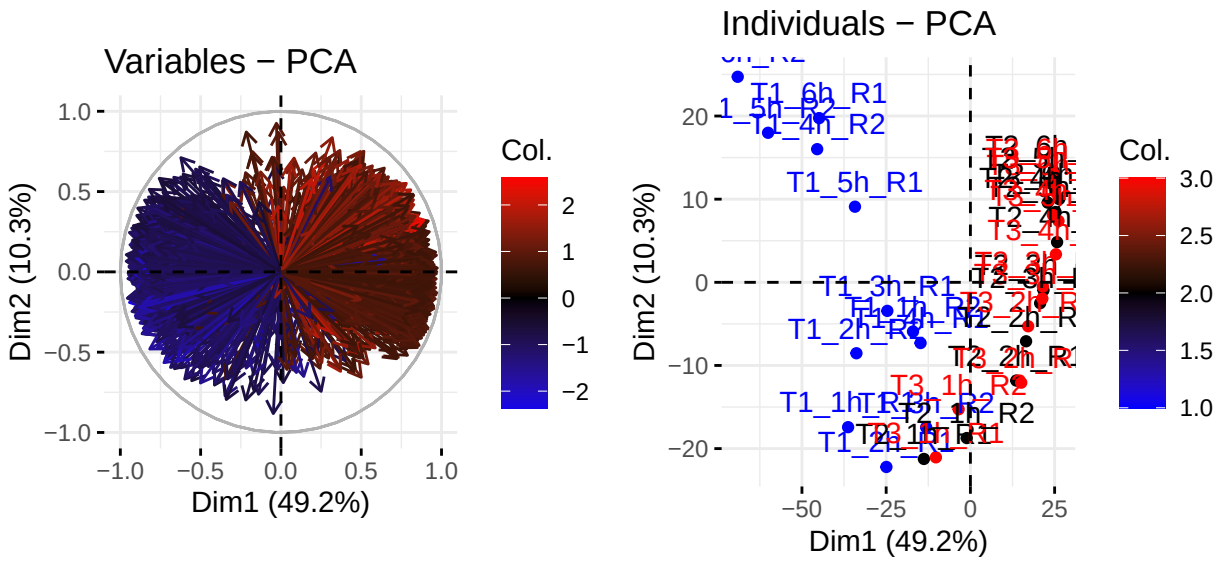
1.2.1 Analyse en composante principale sur les données transposées

```
barplot(cumsum(res.pca.transpose$eig[, "percentage of variance"]), main="Percentage of variance")
abline(h=95, col="blue")
abline(h=99, col="red")
```



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.

```
g1 = fviz_pca_var(res.pca.transpose, col.var=data$T3_6h_R2, label="none") + scale_color_gradient2(low="
g2 = fviz_pca_ind(res.pca.transpose, axes=c(1, 2), autoLab="yes", col.ind=col_trait) + scale_color_grad
grid.arrange(g1,g2,ncol=2)
```



Commentaires à faire

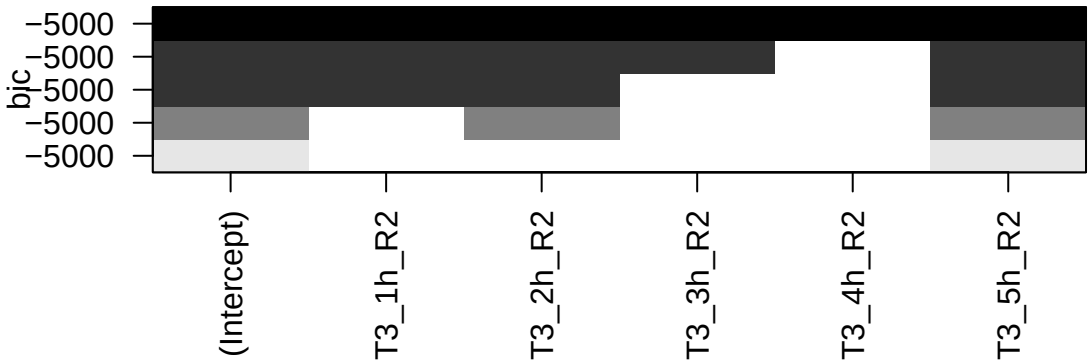
2 Modèle linéaire

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

On récupère les valeurs du traitement T3 à 6h grâce aux commandes suivantes :

```
T3 = data[grep("T3", names(data), value=TRUE)]
T3R2 = T3[grep("R2", names(T3), value=TRUE)]
```

```
choix=regsubsets(T3_6h_R2~., data = T3R2, nbest = 1, nvmax = 5, method = "backward")
plot(choix,scale = "bic")
```



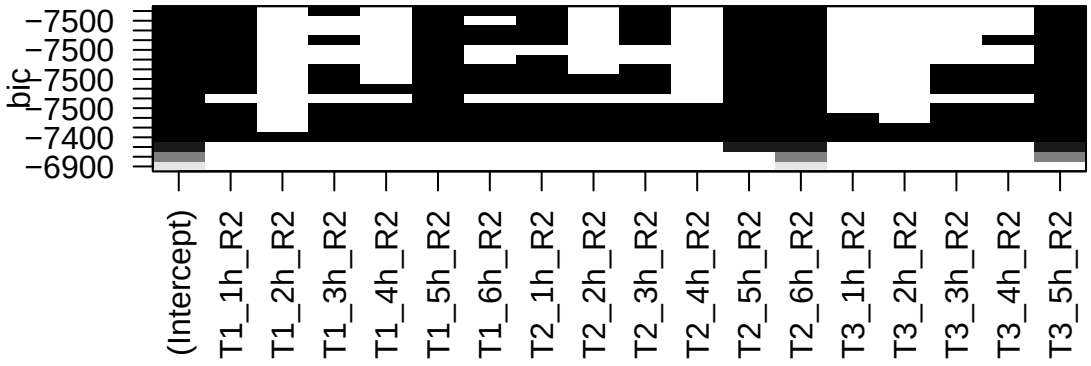
On a réalisé notre sélection de variables avec tous les critères (BIC, adjr2, Cp) et avec les méthodes forward et backward. Nous avons eu les mêmes résultats :

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 avoir tester sur plus d'heures pour valider ou non cette hypothèse.

2.1.1 Etude sur tous les traitements et tous les temps

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

```
choix=regsubsets(T3_6h_R2~., data = R2, nbest = 1, nvmax = 18, method = "backward")
plot(choix,scale = "bic")
```



Critère bic, méthode backward/forward: T1: 1h, 3h, 5h, 6h T2: 1h, 3h, 5h, 6h T3: 5h

ça correspond à l'analyse descriptive précedente, les gènes qui ont eu le traitement T2 ou le traitemet T3 ont des comportements similaires. Cela est cohérent avec le fait que T2 est plus influent sur T3 que T1 sur T3.

On retrouve par ailleurs les résultats de l'analyse précédente puisque les heures les plus influentes sont les heures les plus proches.

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

```
mod.debut = lm(T3_6h_R2 ~ ., data = R2)
mod.fin = lm(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_R
anova(mod.fin,mod.debut)
```

```
## 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
```

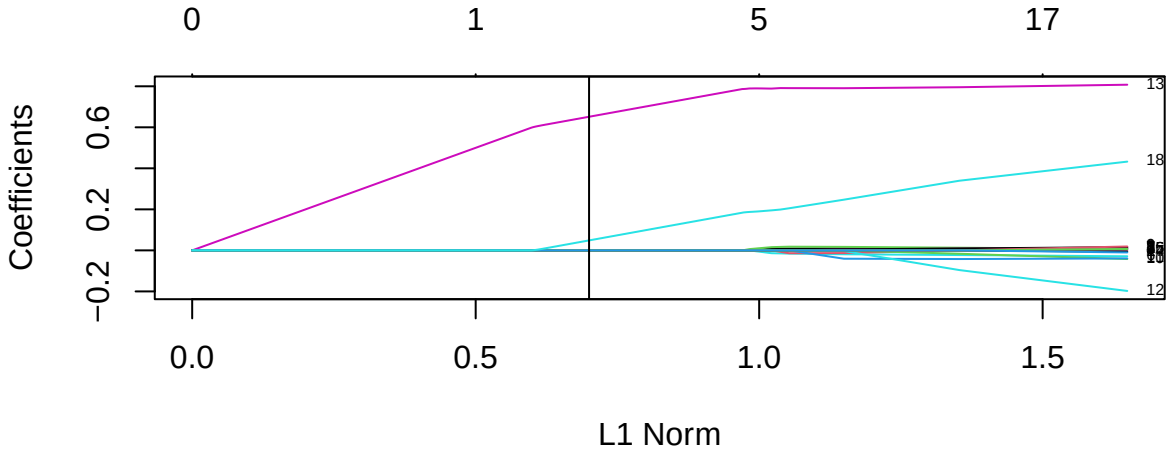
p-valeur $= 0.2798 > 0.05$, on ne rejette pas H0 au risque de 5%, on accepte le sous_modèle.

2.2 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):
## suppression des ex-aequos de 'x'
```

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



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

2.3 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]
```

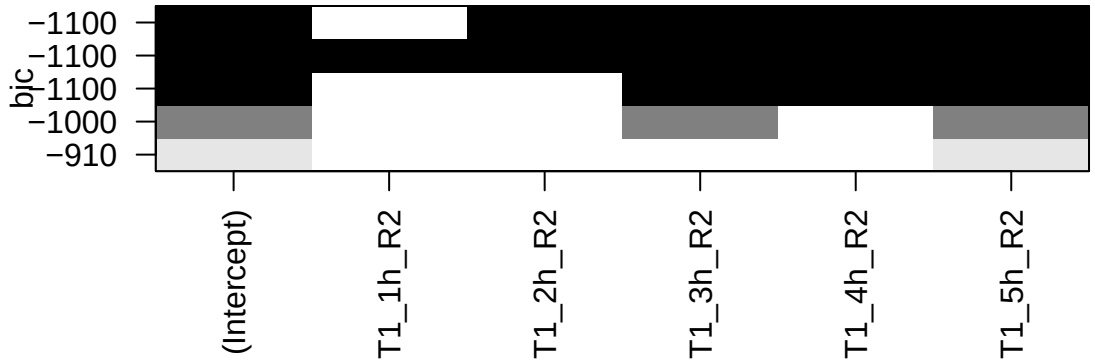
```
apply(T3R2["T3_1h_R2">1], 2, sum)
```

```
##      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
```

2.4 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")
```



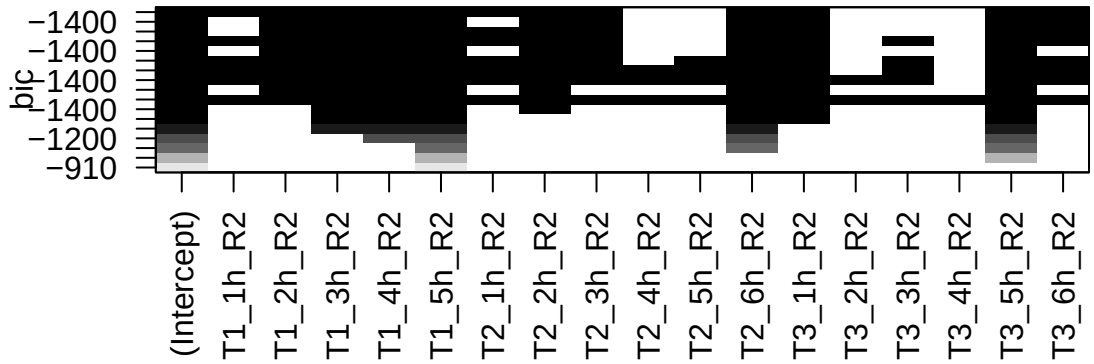

```
mod.debut = lm(T1_6h_R2 ~ ., data = T1R2)
mod.fin = lm(T1_6h_R2 ~ T1_2h_R2+T1_3h_R2+T1_4h_R2+T1_5h_R2, data = T1R2)
anova(mod.fin,mod.debut)
```

```
## 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.

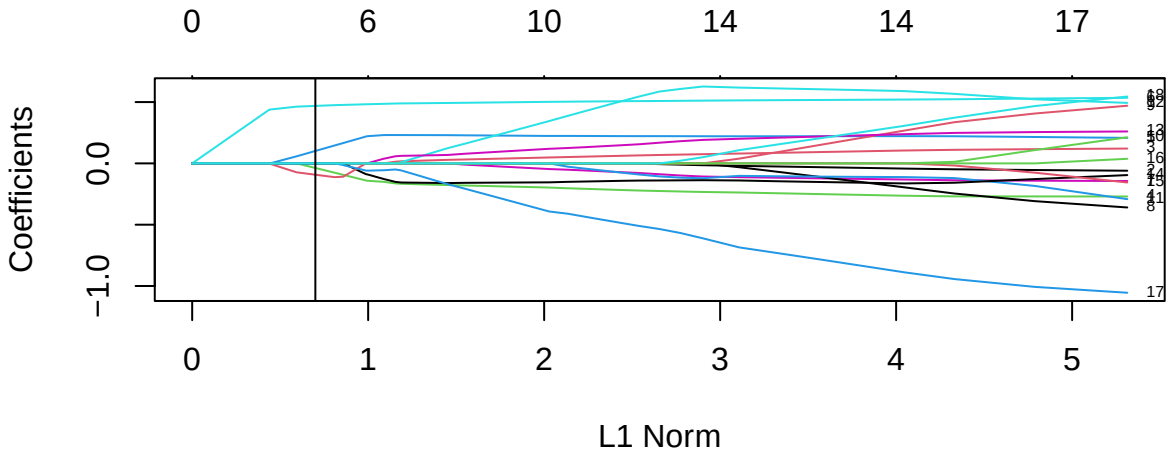
On voit que l'expression des gènes à 6h pour le traitement T1 est affecté par - les heures finales (3h, 4h, 5h) du traitement T1 - les heures débutantes (1h, 2h, 3h) et finale(6h) du traitements T2 - les heures débutantes (1h) et finales (5h, 6h) du traitement T3

2.5 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):
## suppression des ex-aequos de 'x'
```

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



On voit que les variables les plus affectantes sont ??