Projet d'étude

Castello Emeline, Seye Papa Samba et Aoun Ezzeddine

2025-01-30

Contents

Lorsque ce symbole (*) apparaît, la sortie est accessible dans le RMarkdown.

1 Introduction

Ce projet vise à analyser les données d'expression de 2144 gènes d'une plante modèle, mesurées sous trois traitements ($T1$, $T2$, $T3$), six temps ($1h$ à $6h$), et deux réplicats biologiques ($R1$ et $R2$).

L'objectif est de répondre à plusieurs problématiques clés : explorer la structure des données, étudier les effets des traitements et du temps, prédire l'expression des gènes à différents temps, et identifier les facteurs influençant l'expression. Ces analyses reposent sur des méthodes d'exploration de données, de clustering, et de modélisation statistique, implémentées en R.

Ce rapport présente notre démarche, les résultats obtenus et leur interprétation biologique, tout en mettant en avant la pertinence des outils statistiques utilisés.

2 Analyse descriptive du jeu de données

Nous commençons par étudier les corrélations entre les différentes variables.

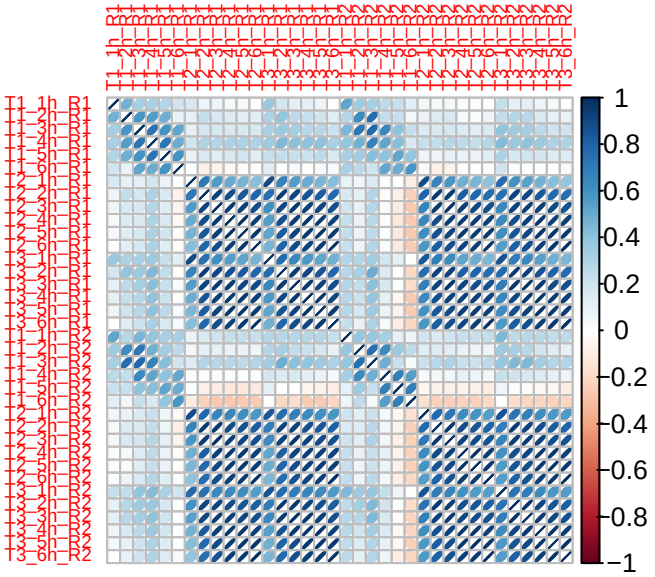


Figure 1: Corrélation entre les différentes variables

La Figure 1 met en évidence la répétition d'un même motif entre les réplicats 1 et 2. On observe une structure en blocs dans la matrice de corrélation. Le bloc situé en haut à gauche met en évidence de fortes corrélations du traitement 1 avec lui-même. De même, les traitements 2 et 3 présentent de fortes corrélations internes, mais aussi entre eux, formant un second bloc distinct. Pour une analyse plus détaillée, nous extrairons ces

éléments et les examinerons individuellement. Dans un premier temps, nous représenterons la corrélation propre à chaque traitement.

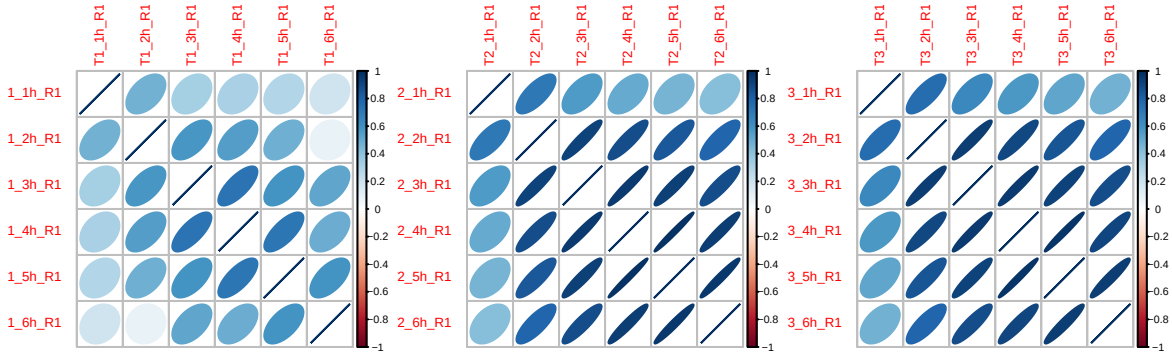


Figure 2: corrélation propre à chaque traitement

La figure 2 met en évidence l'influence du temps sur la corrélation. Plus deux mesures sont prises à des moments rapprochés, plus leur corrélation est élevée, illustrant ainsi l'évolution progressive de l'effet du traitement. Plus le traitement est appliqué longtemps, plus son impact sur l'expression du gène devient marqué. Nous allons ensuite extraire et analyser la corrélation entre chaque traitement.

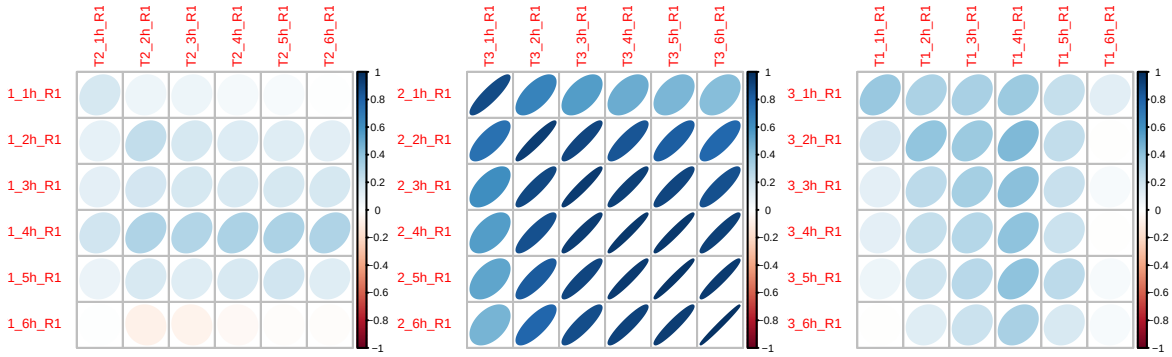


Figure 3: Corrélation entre les traitements

La Figure 3 montre une corrélation marquée entre les traitements 2 et 3, suggérant une expression similaire. En revanche, le traitement 1 présente une faible corrélation avec les traitements 2 et 3, indiquant une différence d'expression.

Nous traçons un viloin plot pour visualiser la distribution des données.

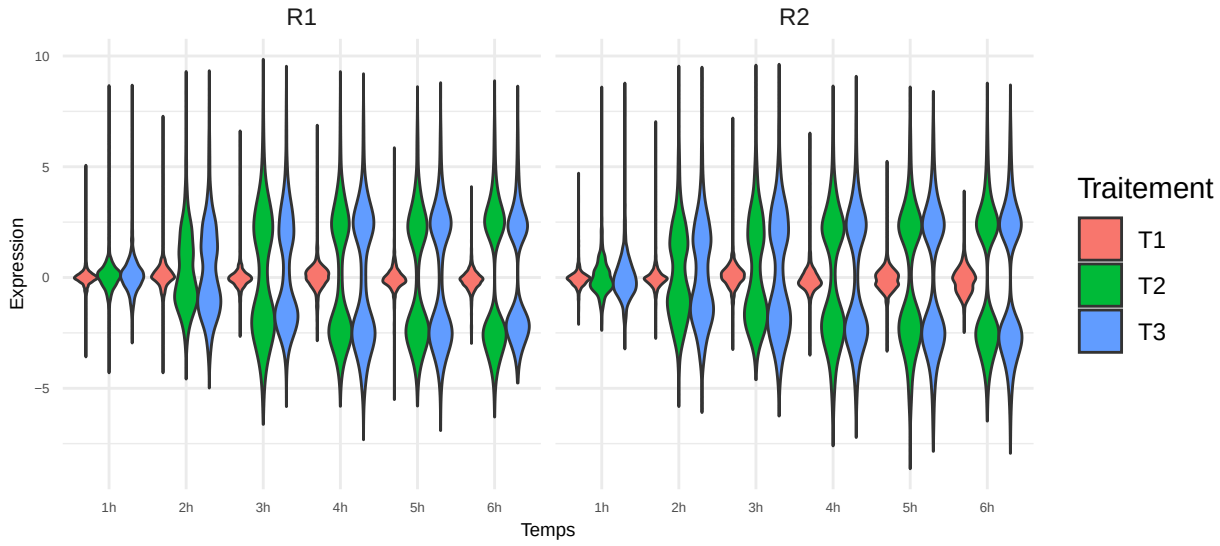


Figure 4: Distribution des Expressions par Réplicat, Temps et Traitement

Sur la Figure 4, une distinction claire entre les réplicats R1 et R2 est visible, suggérant une certaine périodicité dans les valeurs d'expression.

On cosntate trois groupes distincts, correspondant aux traitements T1, T2 et T3 :

- $T1$ présente des valeurs d'expression globalement proches de 0, suggérant une faible expression des gènes.
- $T2$ montre une distribution plus décalée, indiquant une réponse progressive des gènes au fil du temps.
- $T3$, combinaison de T1 et T2, affiche des profils similaires à T2, ce qui suggère que l'effet de T2 domine dans l'expression des gènes sous T3, bien que l'influence de T1 puisse moduler certaines réponses.

Ces tendances sont cohérentes entre les deux réplicats, témoignant de leur homogénéité globale.

3 Analyse des variables T_t s_H R_r

3.1 Analyse des composantes principales

Pour effectuer des analyses sur les variables, nous transposons notre jeu de données afin de considérer les variables initiales comme de nouveaux individus. Cette approche permet d'explorer les relations entre les variables et de visualiser leur structure dans un espace réduit.

Une Analyse en Composantes Principales (ACP) sera réalisée sur ce jeu de données transposé. L'objectif est d'identifier les vecteurs qui maximisent l'inertie de la projection des données sur ces vecteurs. Ces derniers correspondent aux vecteurs propres associés aux plus grandes valeurs propres de la matrice de covariance des données, représentant ainsi les axes principaux de l'ACP.

Dans notre cas, il n'est pas nécessaire de réduire les données, car toutes les variables sont exprimées sur la même échelle.

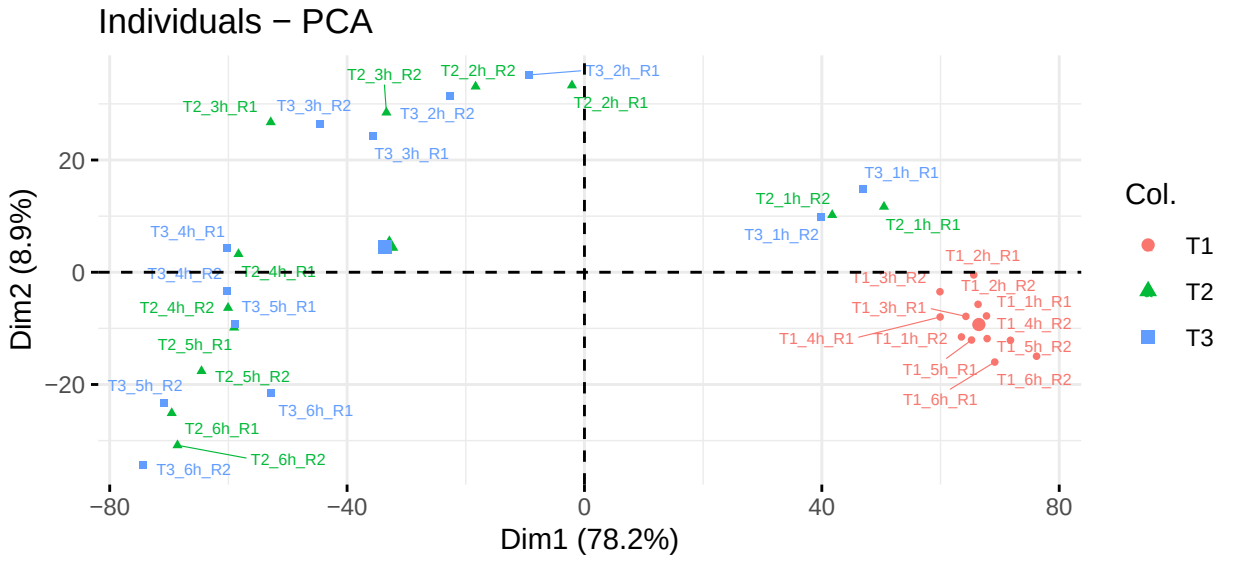


Figure 5: Projection des individus

D'après la figure 5, on observe :

- Les réplicats (R1 et R2) montrent une cohérence, car ils se regroupent dans l'espace factoriel pour un même traitement et une même heure. Cela confirme que les variations observées dans les données sont principalement dues aux effets des traitements et des temps, et non à des différences techniques entre les réplicats.
- Le traitement 1 se distingue des traitements 2 et 3 en formant un cluster isolé. L'axe 1 est tiré d'un coté par T1 à toutes les heures et dans les deux réplicats, et de l'autre coté par T2 et T3, surtout après 4h. On observe également des petits groupes isolés en haut, correspondants aux premières heures (1h, 2h, 3h) des traitements T2 et T3. Cela peut être interprété comme le passage d'un état où les gènes réagissent faiblement (à droite) à un état où ils commencent à répondre plus fortement au traitement. On peut interpéter Cette dimension comme le niveau d'expression des gènes.
- La dimension 2 semble être liée à la temporalité, avec une décroissance des heures du haut vers le bas.

3.2 Clustering

3.2.1 K-means

Pour cet algorithme, on cherche à minimiser l'`inertie intraclasse`, qui est définie par :

$$I_{\text{inter}} = \sum_{k=1}^{K} |C_k| \times d(m_k, c)^2 \quad \text{où} \quad m_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i \text{ est le centre de gravité de la classe } C_k.$$

Pour faire cette étude, nous devons avant tout déterminer le nombre de clusters K. Pour cela, nous optons pour le choix du critère `silhouette` permettant d'évaluer la qualité du clustering en mesurant la cohésion intra-classe et la séparation entre les classes. Le nombre optimal de clusters $\hat{K}$ est celui qui maximise $S(K)$. Nous traçons aussi l'inertie intraclasse.

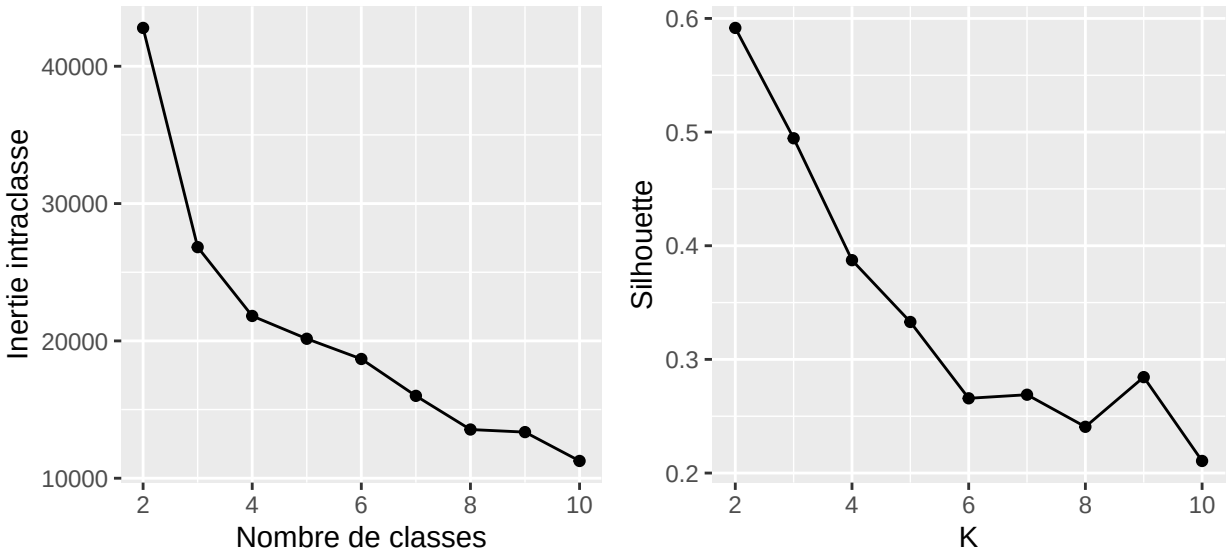


Figure 6: Sélection du nombre de classes optimal

D'après la figure 6, on observe un coude marqué à 3 ou 4 classes pour l'inertie intraclasse. Dilhouette indique un nombre optimal de classes égal à 2. Cependant, nous optons pour 4 classes, car la projection des individus obtenue par l'ACP révèle une distinction claire entre 4 groupes, ce qui facilite une interprétation pertinente de chaque classe.

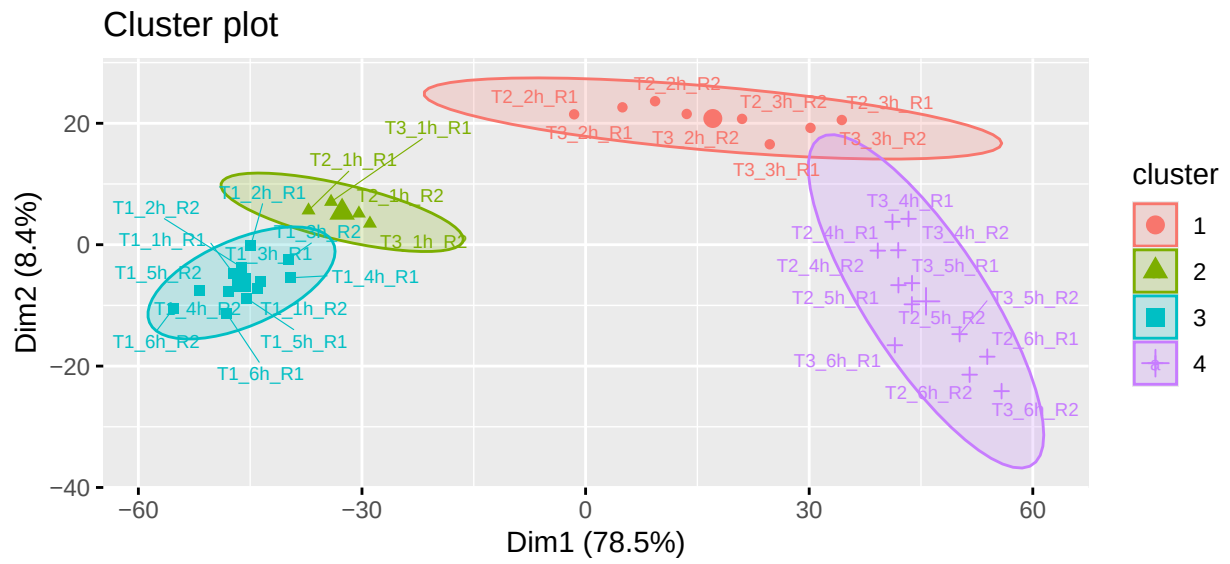


Figure 7: Clustering à 4 classes avec Kmeans

Sur la figure 7, on observe :

- Une première classe regroupant les traitements T1 appliqués à toutes les heures pour les deux réplicats.
- Une deuxième classe contenant T2 et T3 à 1h pour les deux réplicats, située très proche de la classe T1. Cela suggère qu'à 1h, les gènes n'ont pas eu suffisamment de temps pour réagir aux traitements, expliquant l'absence d'effet visible et leur proximité avec la classe T1.

- Une troisième classe, composée de T2 et T3 à 2h et 3h pour les deux réplicats. À ce stade, les gènes commencent à réagir progressivement aux traitements. Cette classe se positionne à mi-chemin entre les classes avec une faible réaction et celle présentant une réaction significative.
- Une quatrième classe, contenant T2 et T3 à des heures supérieures à 4h, où les gènes réagissent aux traitements, montrant un effet significatif.

3.2.2 Modèles de mélange

Nous nous intéressons ici à la méthode des mélanges gaussiens, basée sur l'hypothèse que les gènes suivent une distribution de probabilité composée de plusieurs distributions gaussiennes :

$$f(.|\theta_K) = \sum_{k=1}^{K} \pi_k f_k(.|\alpha_k)$$

Pour sélectionner le bon modèle, nous optons pour le critère de sélection `ICL`, défini par :

$$\text{crit}(K) = L(x|\hat{\theta}_K) - \frac{\nu_K}{2}\ln(n) - \text{Ent}(K)$$

Nous utilisons les coordonnées issues de l'ACP comme matrice de données pour classer les variables $T_t s_H R_r$. Ce choix est motivé par la taille de la matrice de variance-covariance du jeu de données initial, qui est $p*p = 2144*2144$. Une matrice aussi grande serait coûteuse en temps de calcul et ralentirait considérablement la convergence vers le mélange optimal. En utilisant les coordonnées principales, nous réduisons la dimension tout en préservant l'essentiel de l'information.

Nous gardons les 5 premières composantes principales ce qui représente plus de 95% de l'inertie.

```
Best ICL values:
             VEV,8        EEV,9      EEV,12
ICL      -1279.788 -1283.690725 -1319.28804
ICL diff     0.000    -3.902672   -39.49998
```

A partir des resultats observés, on va garder 8 classes de volume variant (V), de même étendu (E) et d'orientation libre (V) puisque l'ICL correspondant est le maximum.

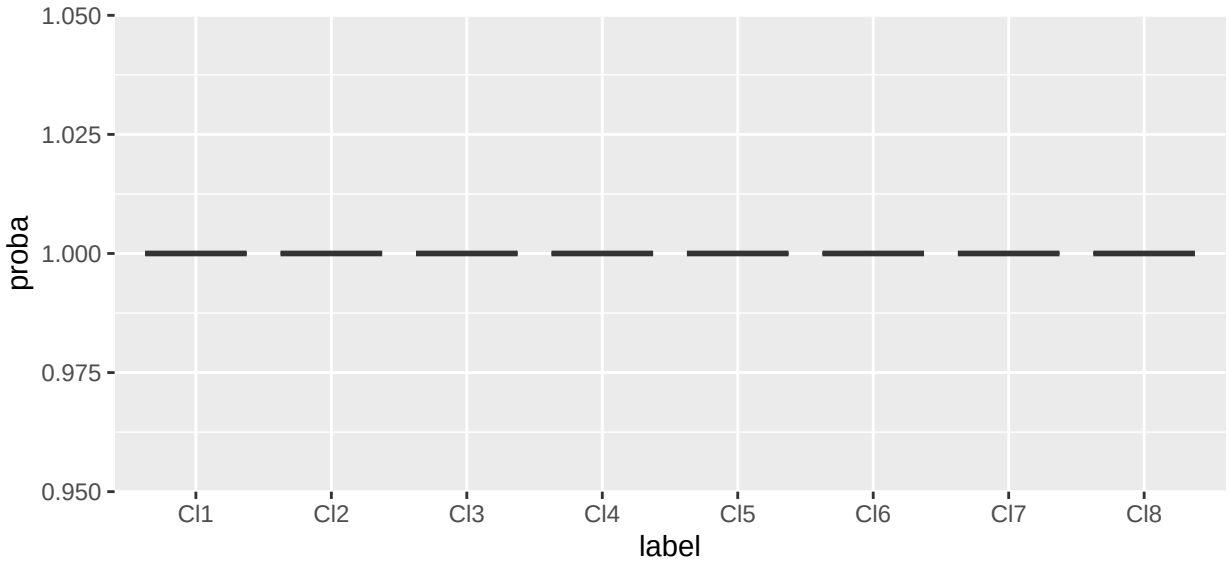


Figure 8: Les probabilités d'appartenance

```
            classes_kmeans
classes_MEL 1 2 3 4
          1 0 0 6 0
          2 0 0 6 0
          3 0 4 0 0
          4 5 0 0 0
          5 0 0 0 4
          6 0 0 0 4
          7 3 0 0 0
          8 0 0 0 4
```

Les points sont parfaitement classés, comme en témoigne la figure 8. Toutes les probabilités d'appartenance sont à 1.

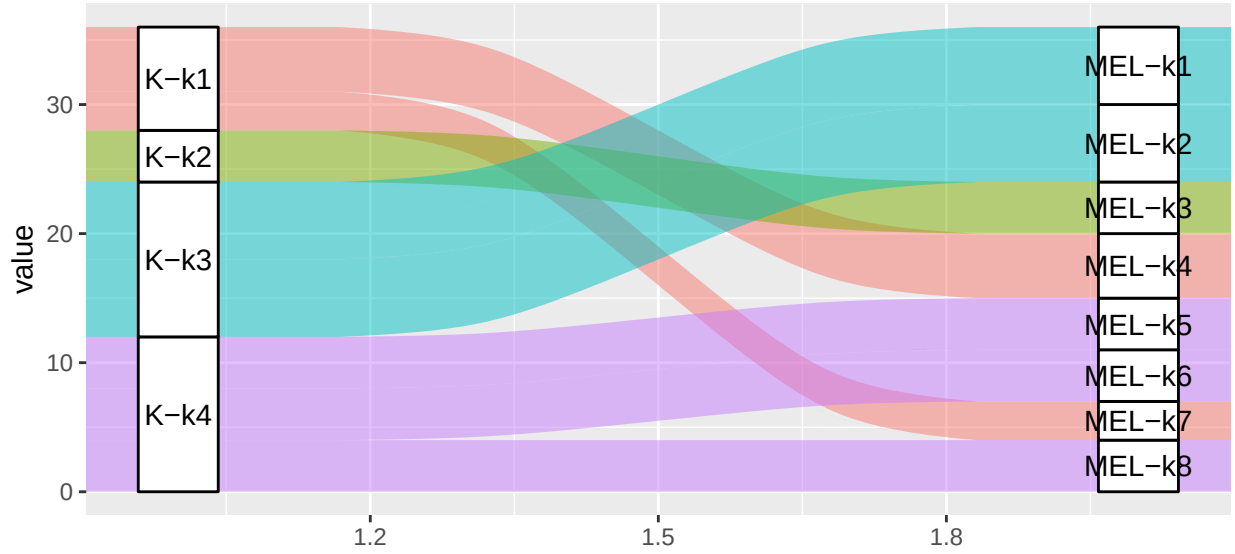


Figure 9: Comparaison entre les classifications Kmeans et MEL

La figure 9 montre que les classes générées par le modèle de mélanges sont des sous-classes de celles définies par K-means, sans aucun chevauchement.

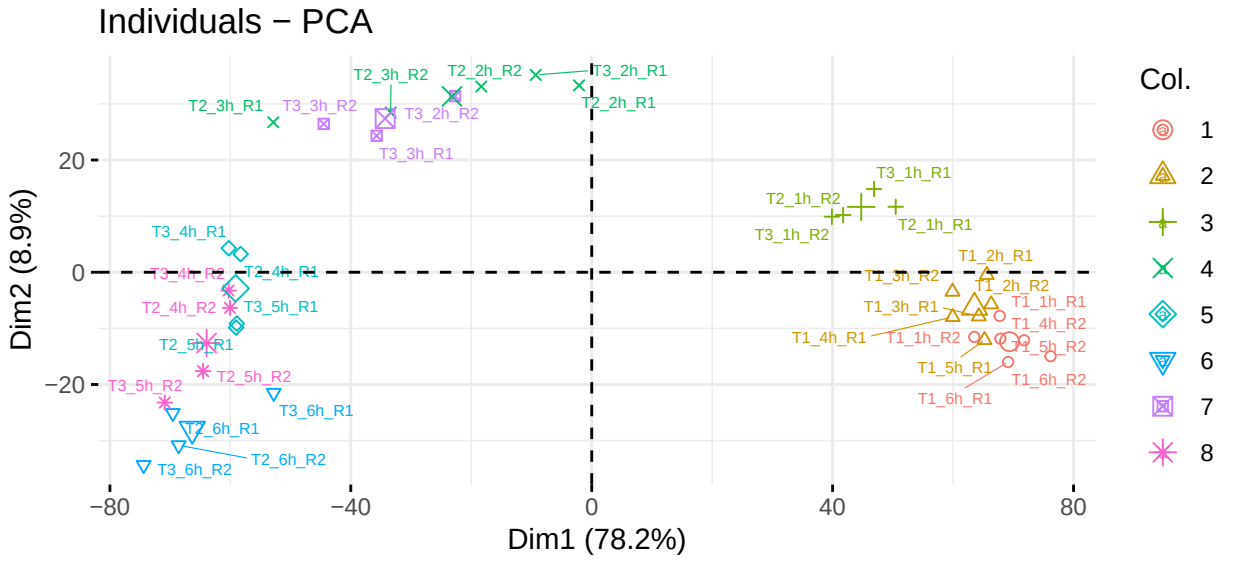


Figure 10: Projection des individus classés par modèles de mélange

D'après la figure 10, on voit que le modèle de mélanges affine davantage la distinction entre les variables en segmentant les classes initiales en sous-classes plus spécifiques. Cette classification permet d'extraire davantage d'informations sur l'expression des gènes en fonction du traitement. Elle révèle que, dans une classe obtenue par k-means, il est possible d'identifier deux ou trois sous-classes présentant une expression similaire.

4 Analyse des individus Gènes

4.1 Partie ACP des individus

Nous nous intéressons ici aux individus. La figure 11 représente ces individus en fonction des variables dans un espace réduit qui résume le mieux les données.

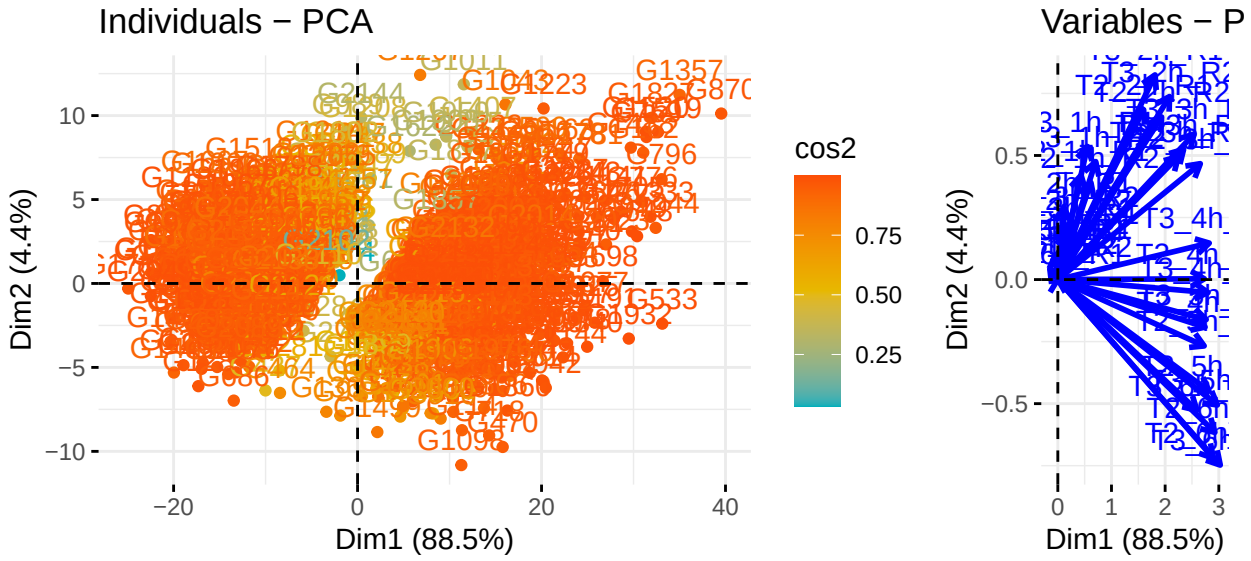


Figure 11: Représentation des individus sur les axes principaux de l'ACP

Ce graphique met en évidence la formation de deux groupes distincts de gènes, répartis de part et d'autre du premier axe principal. Ce premier axe est celui qui sépare le mieux les gènes, indiquant que les deux groupes contiennent des gènes présentant des profils d'expression très différents. Les gènes proches sur le graphique ont des profils d'expression similaires, tandis que ceux éloignés montrent des différences significatives. La coloration des points, basée sur le `cos2`, reflète la qualité de la projection des gènes : les couleurs chaudes indiquent une meilleure qualité de projection, ce qui signifie que ces gènes contribuent davantage à la variance expliquée dans cet espace réduit. Le cercle de corrélation nous montre que les traitements $t \in \{\mathrm{T}_2, \mathrm{T}_3\}$, pour le réplicat $r \in \{\mathrm{R}_1, \mathrm{R}_2\}$, et au temps $s \in \{3\mathrm{h}, 4\mathrm{h}, 5\mathrm{h}, 6\mathrm{h}\}$ sont fortement corrélés à l'axe 1. On en déduit que cet axe correspond à l'expression des gènes, avec les gènes surexprimés à droite et les gènes sous-exprimés à gauche.

4.2 Clustering K-means

Tout d'abord, nous utiliserons la méthode `k-means` pour étudier le clustering des individus.

La figure 12 permet de déterminer le nombre optimal de clusters en analysant l'évolution de l'inertie intra-classe et en utilisant le critère silhouette pour évaluer la qualité du clustering.

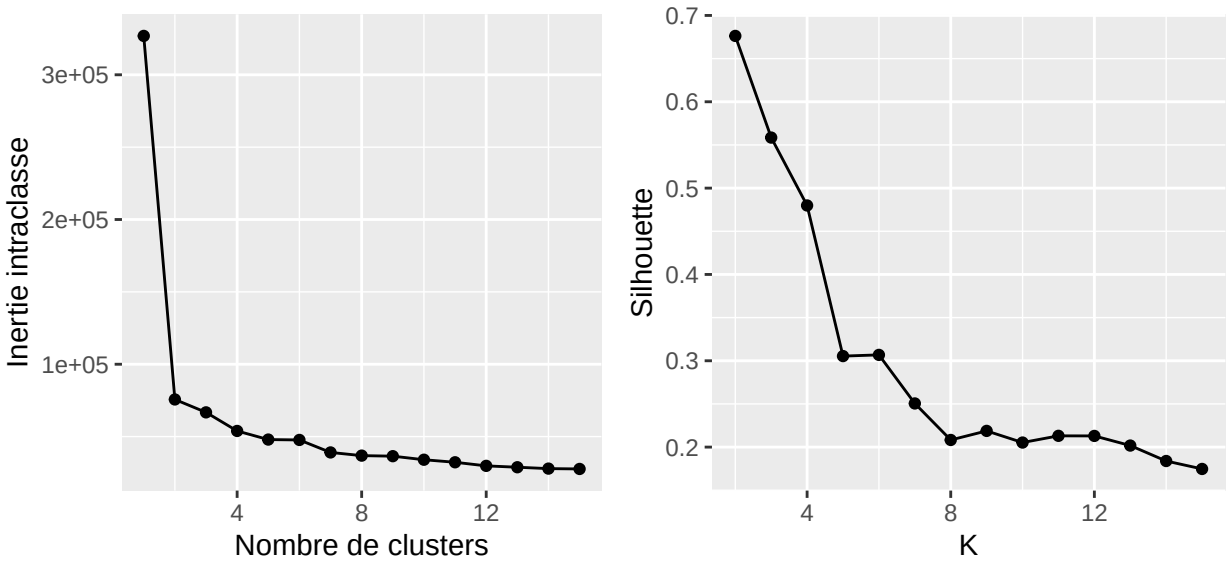


Figure 12: Évolution de l'inertie intra-classe et du critère silhouette

La figure de gauche représente l'évolution de l'inertie intra-classe, qui diminue à mesure que le nombre de clusters augmente. Cependant, lorsque le nombre de clusters atteint un certain seuil, l'inertie devient stable et ne diminue plus de manière significative. Le point de coude dans le graphique nous aide à identifier le nombre optimal de clusters. Dans ce cas, le coude se forme lorsque le nombre de clusters est égal à 4, ce qui suggère que 4 clusters représentent le meilleur compromis entre la compacité des groupes et la simplicité du modèle.

La courbe de l'évolution du critère silhouette montre que le nombre optimal de clusters est égal à 2. Cependant, le critère silhouette est basé sur une moyenne, ce qui signifie qu'il peut être sensible à la présence d'outliers, et dans ce cas, il peut attribuer des clusters incorrects.

Nous pouvons maintenant représenter les individus dans l'espace réduit de l'ACP et les afficher selon leur cluster. La figure 13 représente les individus selon leur cluster dans l'espace réduit de l'ACP.

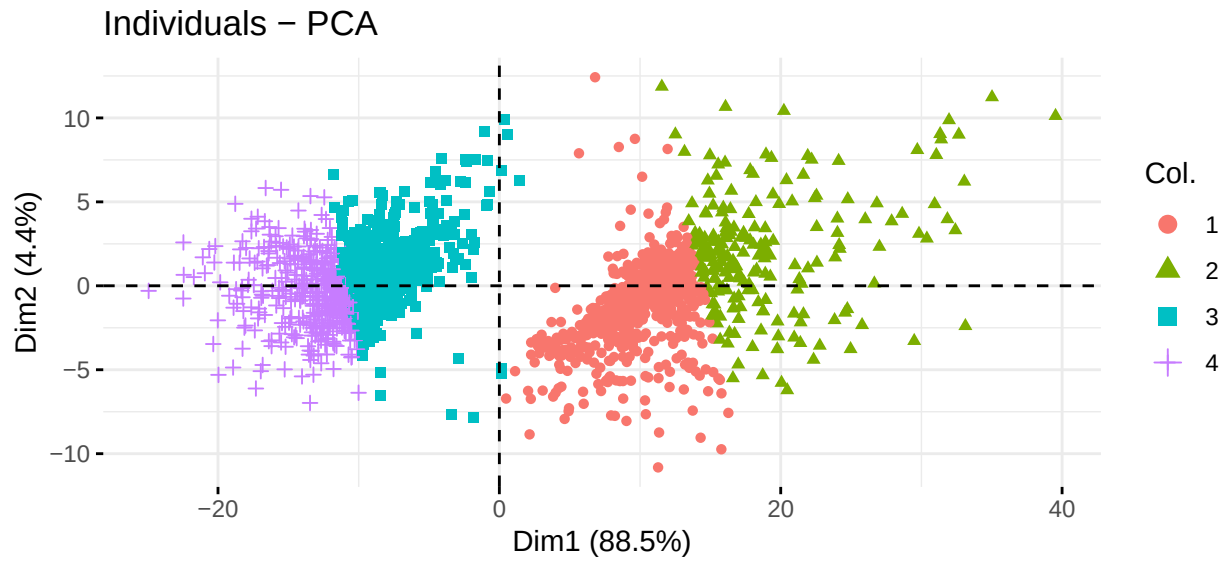


Figure 13: Clusters représentés sur les axes principaux

4.3 Clustering Hiérarchique (HAC)

Dans cette partie, nous allons mettre en place une classification hiérarchique ascendante. La méthode `Euclidienne` utilisée pour calculer la distance entre les gènes est la distance `k-means` définie par :

$$d(x_i, x_\ell) = \|x_i - x_\ell\|_2 = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{\ell j})^2}$$

Pour l'agrégation des clusters, nous optons pour la `mesure d'agrégation de Ward`, donnée par :

$$D(C_k, C_{k'}) = \frac{|C_k|\,|C_{k'}|}{|C_k| + |C_{k'}|}\, d(m_k, m_{k'})^2 \quad \text{où} \quad m_k \,(\text{resp. } m_{k'}) \text{ est le centre de gravité de } C_k \,(\text{resp. } C_{k'})$$

Pour déterminer le nombre de clusters, nous introduisons le critère `Calinski-Harabasz` donné par :

$$\text{PseudoF}(K) = \frac{I_{\text{inter}}(P_K)}{K-1} \Big/ \frac{I_{\text{intra}}(P_K)}{n-K}$$

La figure 14 ci-dessous représente le clustering hiérarchique, permettant de créer un dendrogramme et de déterminer les clusters en coupant l'arbre à un niveau spécifique.

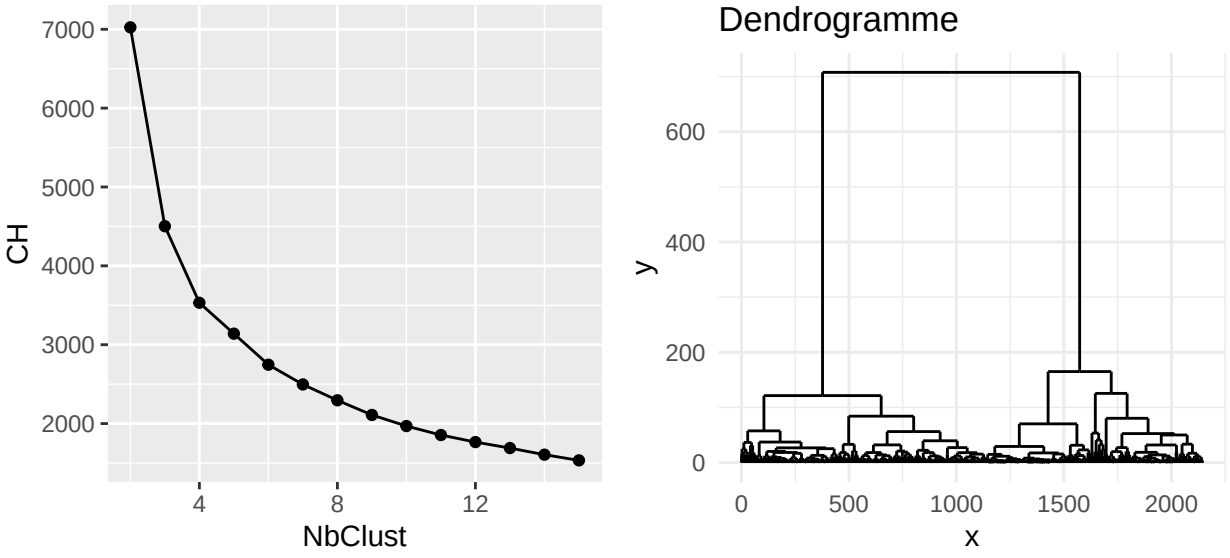


Figure 14: Evolution du critère CH et Dendrogramme

L'analyse de l'indice `Calinski-Harabasz` montre que 2 clusters maximisent ce critère. Le dendrogramme montre que les groupes de gènes qui se fusionnent tôt ont des profils d'expression très similaires, tandis que ceux qui se rejoignent plus tard présentent des différences plus marquées. La figure 15 représente les individus classés par la méthode hiérarchique ascendante selon leur cluster dans l'espace réduit de l'ACP.

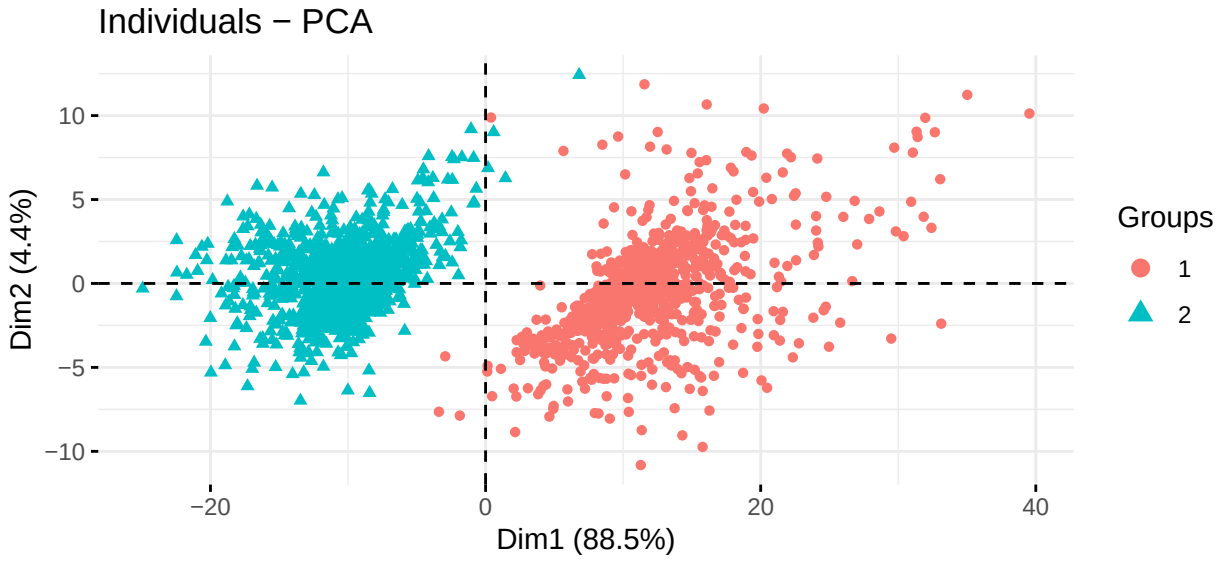


Figure 15: Clusters représentés sur les axes principaux

4.4 Mélange gaussien

Comme on sait que l'expression des gènes est dépendante, on peut d'ores et déjà retenir que les modèles permettant des degrés variés de liberté sont les plus adaptés. On en déduit que seuls les modèles suivants sont retenus (`modelNames`) :

```
mg_res <- mclustICL(Data, G = 2:20, modelNames = c("EVE", "EEV", "EVV", "VEV", "VVV"))
```

```
  Best.ICL.values       ICL  ICL.diff
1           VEV,4 -61554.37     0.000
2           VVV,4 -61722.51  -168.143
3           VEV,5 -63474.42 -1920.053
```

Le modèle maximisant le critère `ICL` est "VEV", avec un nombre de clusters égal à 4. La figure 17 suivante représente les clusters obtenus des individus.

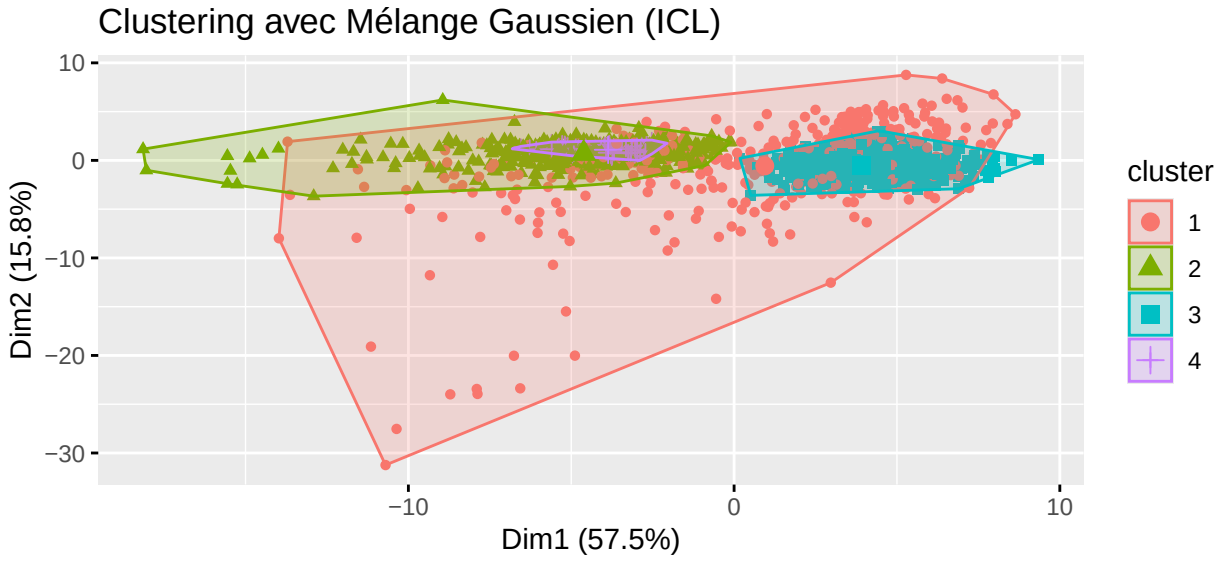


Figure 16: Clusters représentés avec un Mélange Gaussien

4.5 Comparaison des différents algorithmes

4.5.1 Diagramme d'alluvion :

Nous allons ici comparer les trois méthodes de clustering réalisées précédemment en utilisant un diagramme d'alluvions (figure 17).

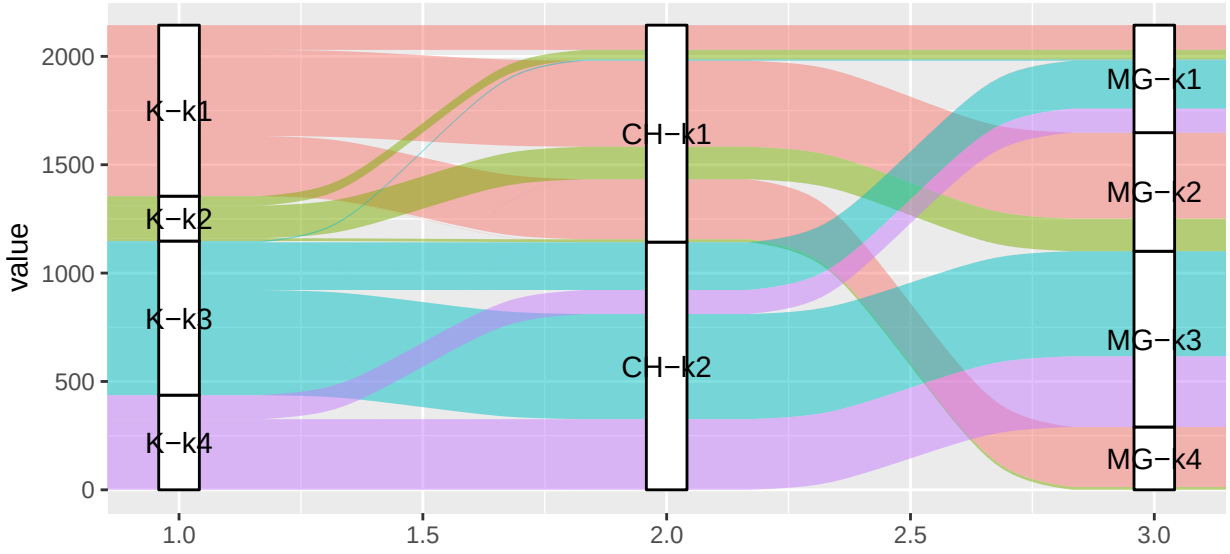


Figure 17: Diagramme d'alluvions

La première observation marquante est le flux important partant du cluster 1 des `k-means` vers le cluster 1 de la classification hiérarchique, et se terminant dans les clusters 1, 2 et 4 du mélange gaussien. Les classes obtenues via la classification hiérarchique sont une agrégation des classes de `k-means`. On en déduit qu'elles sont concordantes.

4.5.2 MCA sur les différentes méthodes :

Une autre méthode de comparaison consiste à créer un tableau disjonctif complet qui présente, pour chaque gène, les clusters obtenus selon les trois méthodes de clustering : `k-means`, classification hiérarchique et mélanges gaussiens.

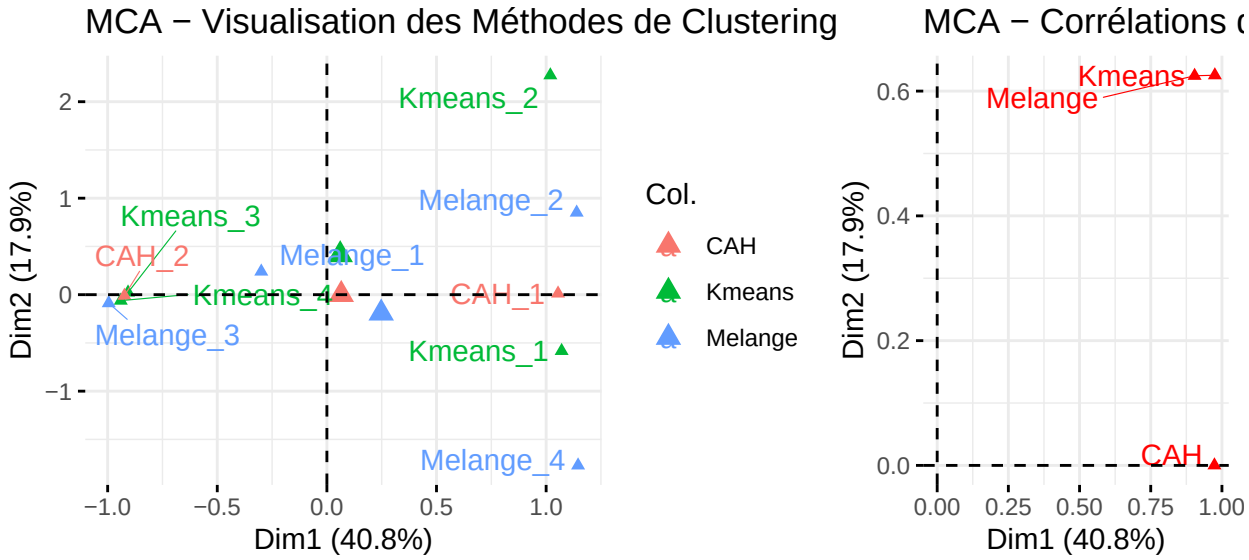


Figure 18: Représentation des modalités sur les axes principaux de la MCA

La première dimension de la figure 18 est associée aux trois méthodes, tandis que la deuxième dimension semble être principalement influencée par `k-means` et le mélange gaussien. Les proximités observées entre les modalités sur le plan révèlent une forte association entre elles. Par exemple, on constate que le cluster 1 obtenu par `k-means` et le cluster 1 de la méthode hiérarchique sont très proches, ce qui suggère que plusieurs gènes classés dans le cluster 1 de `k-means` se retrouvent également dans le cluster 1 de la méthode hiérarchique. Cette observation confirme la comparaison réalisée à l'aide du diagramme d'alluvions.

5 Etude des différences entre les deux réplicats

5.1 Comparaison Statistique des Lois de Probabilité des Deux Réplicats

Dans cette partie, nous allons appliquer le `test de Kolmogorov-Smirnov (KS)`, un test statistique non paramétrique qui permet de comparer deux distributions de probabilité empiriques ou une distribution empirique avec une distribution théorique. La statistique de test est définie par :

$$D_{n,m} = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - \hat{G}_m(t)|$$

où $\hat{F}_n(t)$ et $\hat{G}_m(t)$ représentent les fonctions de répartition empiriques respectives des deux échantillons. Sous l'hypothèse nulle $H_0 : F = G$, où F et G désignent les fonctions de répartition des deux populations, la distribution de $D_{n,m}$ est indépendante de F si F est continue.

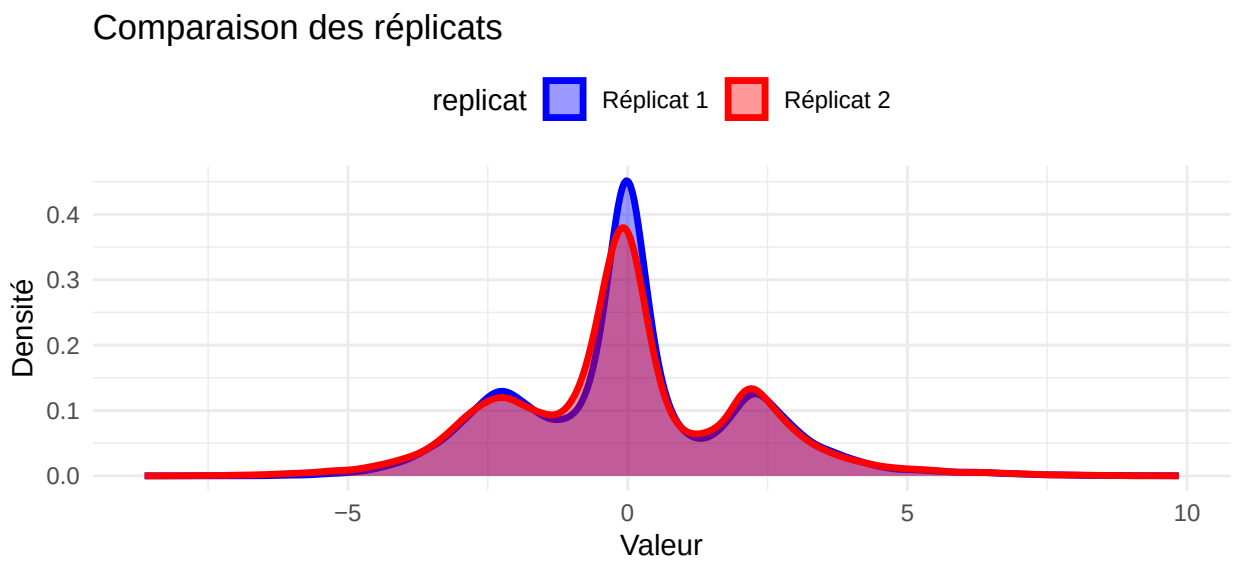


Figure 19: Comparaison des deux réplicats

La figure 19 represente le graphique de densité comparatif qui permet de visualiser les distributions des deux réplicats. Les hypothèses du test sont les suivantes :

- H_0 : Les deux réplicats comparés suivent la même loi de probabilité.
- H_1 : Les lois de probabilité des deux réplicats comparés sont différentes.

Rejet de H0 : Les lois de probabilité des deux réplicats sont significativement différentes (p-value = 1.111068e-32).(*)

5.2 Analyse de Significativité des Lois de Probabilité pour chaque Traitement pris séparément

Ici, nous étudions si les sous-groupes de traitements (`T1, T2, T3`) sont significativement différents. Nous interpréterons la statistique de test obtenue et sa région critique en fonction du niveau de signification choisi.

Rejet de H0 : Les lois de probabilité des deux réplicats pour T1 sont significativement différentes (p-value = 1.143769e-14).(*)

Rejet de H0 : Les lois de probabilité des deux réplicats pour T2 sont significativement différentes (p-value = 7.358161e-12).(*)

Rejet de H0 : Les lois de probabilité des deux réplicats pour T3 sont significativement différentes (p-value = 7.112548e-17).(*)

En somme, l'ensemble des tests de Kolmogorov-Smirnov a conduit au rejet systématique de l'hypothèse nulle (H_0) pour tous les traitements ($T1, T2, T3$). Cela signifie que les distributions des deux réplicats ($R1$ et $R2$) sont significativement différentes pour chaque traitement.

5.3 Effet combiné du temps et du traitement

Nous cherchons à étudier l'effet combiné du temps et du traitement sur la différence des deux réplicats. Pour ce faire, nous définissons une nouvelle variable Y_{tsg} :

$$Y_{tsg} = Y_{tsgR1} - Y_{tsgR2}, t \in \{1, \dots, 3\},\ s \in \{1, \dots, 6\},\ g \in \{1, \dots, 2144\}.$$

L'objectif est de modéliser cette variable en fonction de deux facteurs principaux : le temps (variable quantitative) et le traitement (variable qualitative).

Pour ce faire, nous utilisons un **modèle ANCOVA avec interaction**, qui permet de modéliser la relation entre la variable dépendante Y_{tsg} et les deux facteurs (temps et traitement), ainsi que leur interaction.

$$Y_{tsg} = \beta_0 + \beta_1 \cdot Temps + \beta_2 \cdot Traitement + \beta_3 \cdot (Temps \times Traitement) + \epsilon_{tsg}$$

- Temps : Variable quantitative représentant le temps.

- Traitement : est une variable indicatrice (0 ou 1) indiquant si le gène a été soumis au traitement.
- Les coefficients β représentent les effets respectifs de l'ordonnée à l'origine, du temps, du traitement et de l'interaction entre le temps et le traitement.
- ϵ_{tsg} est un terme d'erreur aléatoire.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0487461260	0.012641807	3.8559461	1.154698e-04
traitementT2	-0.0072778929	0.017878215	-0.4070816	6.839503e-01
traitementT3	-0.0196789728	0.017878215	-1.1007236	2.710238e-01
temps	-0.0049365532	0.003246116	-1.5207567	1.283291e-01
traitementT2:temps	0.0008565131	0.004590702	0.1865756	8.519944e-01
traitementT3:temps	0.0333072705	0.004590702	7.2553767	4.080708e-13

On observe :

- L'intercept est significatif.
- T2 et T3 seuls ne sont pas significativement différents du traitement de référence.
- Le temps seul n'est pas significatif.
- L'interaction T3:temps est hautement significative, ce qui indique que pour T3, l'effet du temps est très important et influence fortement la réponse.

Pour vérifier l'optimalité de notre modèle, nous établissons le modèle additif sans intéraction.

$$Y_{tsg} = \beta_0 + \beta_1 \cdot Temps + \beta_2 \cdot Traitement + \epsilon_{tsg}$$

Avant de procéder au test de sous-modèle pour déterminer si le modèle avec interaction doit être conservé, nous comparons les valeurs prédites des deux modèles en les visualisant graphiquement.

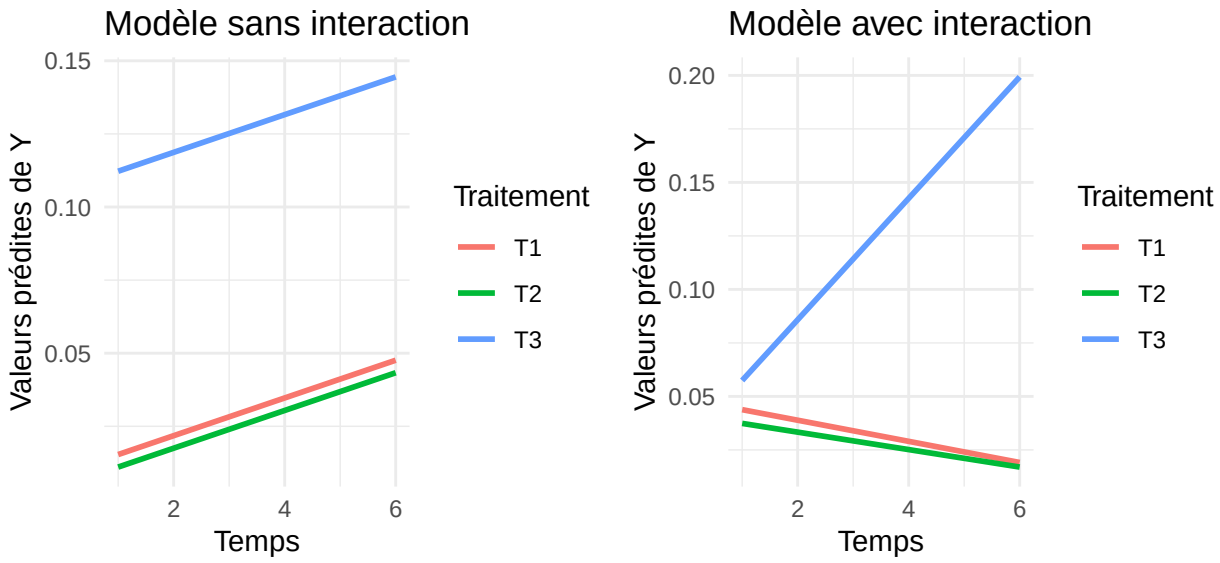


Figure 20: Interaction plot

D'après la figure 20, on observe que dans le modèle sans interaction, les courbes sont parallèles, reflétant l'absence d'interdépendance entre le traitement et le temps. En revanche, dans le modèle avec interaction, les courbes montrent des variations de pente ou des croisements, indiquant une présence probable d'interaction entre ces deux variables. À partir de ces deux représentations, un test de sous-modèle est effectué pour déterminer si l'interaction doit être conservée dans le modèle final.

La p-valeur obtenue est de $1.426e-15 < 0.05$. Cela conduit à rejeter largement l'hypothèse nulle et à privilégier le modèle avec interaction.(*)

6 Etude de la dynamique de l'expression des gènes

6.1 Pédiction de l'expression des gènes à 6h à partir de l'expression à 1h

Nous avons calculé la moyenne des réplicats pour chaque gène et chaque temps, notée Y_{tsgMoy}. Un modèle ANCOVA avec interaction a été ajusté pour prédire l'expression finale (à t=6h) en fonction de l'expression initiale (à t=1h) et du traitement :

$$Y_{t6hgMoy} = \beta_0 + \beta_1 \cdot Y_{t1hgMoy} + \beta_2 \cdot Traitement + \beta_3 \cdot (Y_{t1hgMoy} \times Traitement) + \epsilon_{tsg}$$

où :

- $Y_{t6hgMoy}$ est l'expression moyenne du gène g au temps t=6h.
- $Y_{t1hgMoy}$ est l'expression moyenne du gène g au temps t=1h (expression initiale).
- Traitement : est une variable indicatrice (0 ou 1) indiquant si le gène a été soumis au traitement.
- Les coefficients β représentent les effets respectifs de l'ordonnée à l'origine, de l'expression initiale, du traitement et de l'interaction entre l'expression initiale et le traitement.
- ϵ_{tsg} est un terme d'erreur aléatoire.

```
Call:
lm(formula = Y_moy_6h ~ Y_moy_1h * traitement, data = matrixMoy)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8229  -1.4374  -0.0992   1.3805   8.8671

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.10335    0.04521  -2.286   0.0223 *
Y_moy_1h               0.31983    0.12556   2.547   0.0109 *
traitementT2          -0.20072    0.06393  -3.140   0.0017 **
traitementT3          -0.16336    0.06393  -2.555   0.0106 *
Y_moy_1h:traitementT2  1.48931    0.13620  10.935   <2e-16 ***
Y_moy_1h:traitementT3  1.28212    0.13379   9.583   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.084 on 6426 degrees of freedom
Multiple R-squared:  0.2706,    Adjusted R-squared:   0.27
F-statistic: 476.7 on 5 and 6426 DF,  p-value: < 2.2e-16
```

Nous traçons deux graphes qui nous aideront à déterminer la qualité de l'ajustement du modèle :

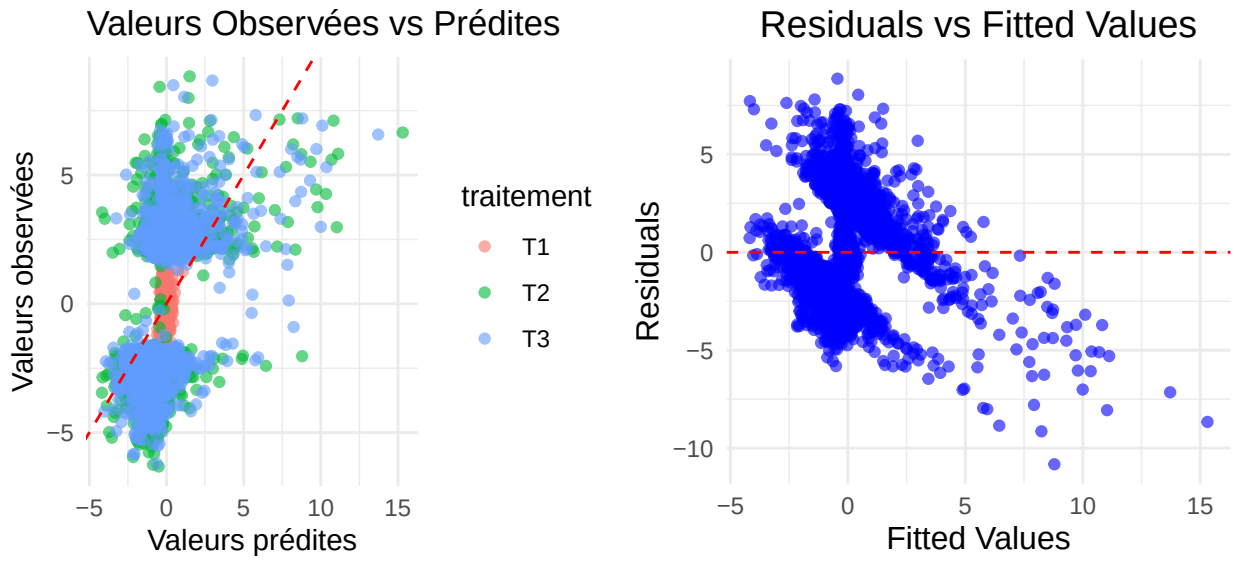


Figure 21: Repésentation de la qualité du modèle

D'après la figure 21, on observe :

- les valeurs prédites sont principalement proches de 0, tandis que les valeurs observées, en particulier pour les traitements T2 et T3, diffèrent significativement de 0. Cela suggère que l'expression des gènes à 1h ne constitue pas un bon indicateur pour prédire celle à 6h. En effet, 1 heure représente un stade relativement précoce, et les gènes n'ont probablement pas eu suffisamment de temps pour réagir aux traitements.

- le graphique des résidus en fonction des valeurs ajustées, on constate que La dispersion non uniforme des résidus suggère une hétéroscédasticité: la variabilité des erreurs n'est pas constante. De plus, la légère tendance non-linéaire des résidus indique que la relation entre les variables pourrait être plus complexe que ce que notre modèle linéaire simple ne capture. En d'autres termes, le modèle que nous utilisons actuellement ne semble pas être le plus adapté à nos données.

Conclusion: Le modèle ajusté montre une qualité d'ajustement modérée, avec un R^2 de 27.06%. Les termes d'interaction entre Y_{1h} et le traitement sont significatifs (p<2e−16), ce qui montre que l'effet de Y_{1h} sur Y_{6h} varie selon le traitement. En se basant sur les graphiques obtenus, le modèle que nous utilisons actuellement ne semble pas être le plus adapté à nos données. Des ajustements sont nécessaires pour obtenir des résultats plus fiables.

6.2 A partir de l'expressions des gènes à 3h

Nous souhaitons prédire Y_{6hMoy} en utilisant Y_{3hMoy}. Pour ce faire, nous allons reprendre le modèle précédent et l'adapter en remplaçant toutes les variables correspondant aux données au temps $1h$ par leurs équivalents au temps $3h$.

```
Call:
lm(formula = Y_moy_6h ~ Y_moy_3h * traitement, data = matrixMoy2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3774 -0.5527 -0.0160  0.5124  6.4759

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.12411    0.02529  -4.908 9.44e-07 ***
Y_moy_3h              0.21078    0.03880   5.432 5.78e-08 ***
traitementT2         -0.16173    0.03574  -4.525 6.14e-06 ***
traitementT3         -0.15031    0.03575  -4.204 2.65e-05 ***
Y_moy_3h:traitementT2  0.83582    0.04004  20.873  < 2e-16 ***
Y_moy_3h:traitementT3  0.80412    0.04006  20.071  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.168 on 6426 degrees of freedom
Multiple R-squared:  0.7708,    Adjusted R-squared:  0.7706
F-statistic:  4322 on 5 and 6426 DF,  p-value: < 2.2e-16
```

Le modèle basé sur les données de 3h semble être plus performant que celui de 1h, avec un R^2 ajusté presque trois fois supérieur et un RSE significativement plus faible. Cela pourrait signifier que les valeurs à 3h prédisent mieux les valeurs à 6h.

On trace la courbe des valeurs observées par rapport aux valeurs prédites, ainsi que les résidus par rapport aux valeurs prédites.

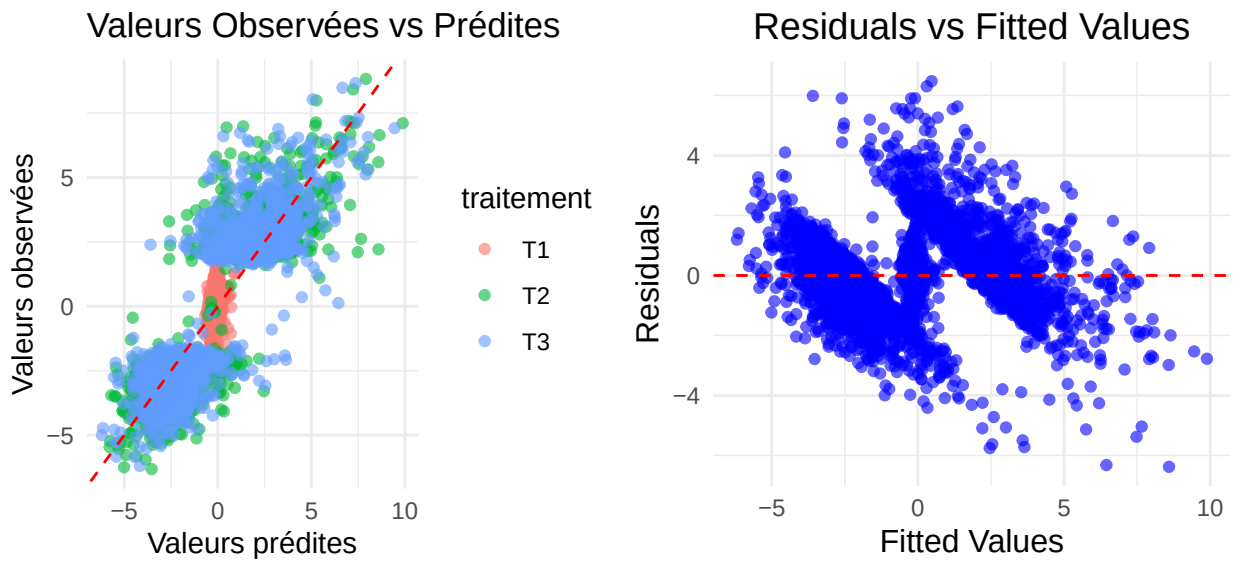


Figure 22: Repésentation de la qualité du modèle

A partir de la figure 22 on observe que les valeurs prédites s'alignent bien avec les valeurs observées, et que les résidus sont plus uniformément dispersés autour de 0. Cela indique une amélioration notable de la qualité des prédictions par rapport au modèle précédent.

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

7.1 Les variables prédictives pour le traitement T3 à 6h parmi les différents temps observés pour les traitements T1 et T2

Nous allons réaliser une regression linéaire pour prédire les gènes du traitement 3 à 6h en fonction des gènes des autres traitements à toutes les heures.

On remarque dans cette sortie, que la p-valeur pour le modèle constant est $< 2.2e\text{-}16$, on rejette le modèle constant. On se demande maintenant quelle est la meilleur modèle qu'on puisse établir. Pour selectionner le modèle parcimonieux, on décide d'utiliser le critère de Mallows avec une méthode backward. On cherche à minimiser ce critère.(*)

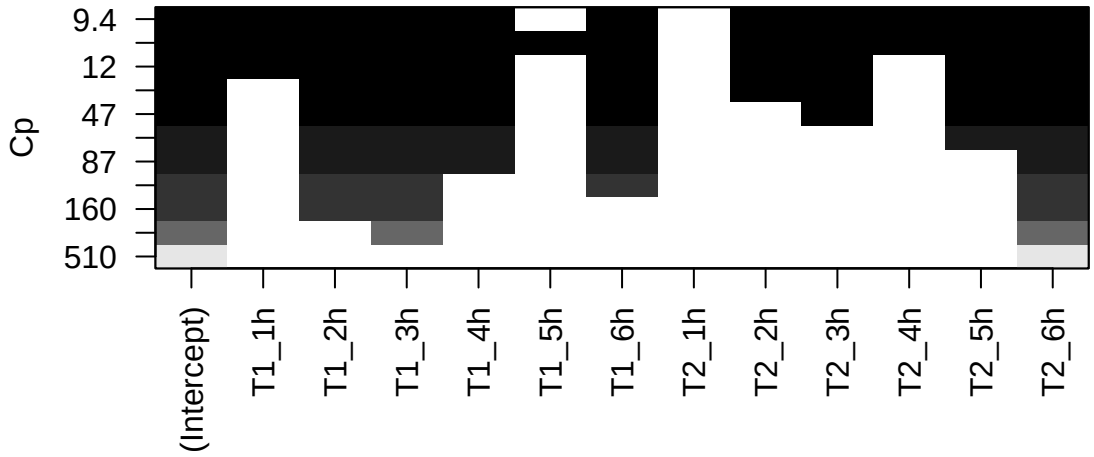
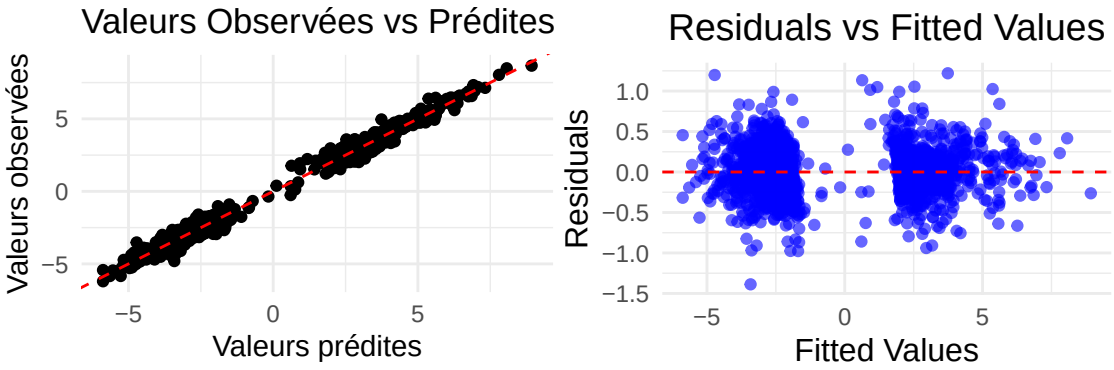


Figure 23: Critère de Mallows

Sur la Figure 23, on remarque le critère Cp selectionne toutes les variables sauf T1_5h et T2_1h. On décide de tester ce sous-modèle en le comparant au modèle de regression linéaire complété implémenté précédemment.

Pour l'analyse de variable entre les 2 modèles, on obtient une p-valeur $= 0.8284 > 0.05$, on accepte donc le sous-modèle.(*)

On souhaîte maintenant vérifier si notre nouveau modèle permet de bien prédire les expression des gènes à 6h.



Sur la figure **??**, on observe une répartition linéaire des points suivant la droite $y = x$, indiquant une forte concordance entre les valeurs prédites et observées. Cela suggère que notre modèle prédit efficacement l'expression des gènes à 6h.

7.2 Prédiction des gènes sur-exprimés et des gènes sous-exprimés à 6h pour le traitement 3 à partir des traitements T1 et T2 et les heures 1 à 3 pour ces mêmes gènes

Nous allons utiliser un modèle linéaire généralisé pour prédire l'expression des gènes, car la variable cible est binaire : un gène peut être soit sous-exprimé ($Y = 0$), soit sur-exprimé ($Y = 1$). Pour cela, nous appliquons une régression logistique en choisissant la fonction de lien logit, qui est définie comme suit :

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

où p représente la probabilité qu'un gène soit sur-exprimé.

L'équation de notre modèle est donc la suivante :

$$\log\left(\frac{P(Y_{6hMoy}=1)}{1-P(Y_{6hMoy}=1)}\right) = \beta_0 + \beta_1 Y_{T1,1h} + \beta_2 Y_{T1,2h} + \beta_3 Y_{T1,3h} + \beta_4 Y_{T2,1h} + \beta_5 Y_{T2,2h} + \beta_6 Y_{T2,3h}$$

où $Y_{Tt,sh}$ représente l'expression du gène sous le traitement Tt à l'instant sh.(*)

On essaie ensuite de réduire les paramètres du modèle grace au critère AIC que l'on souhaîte minimiser. Pour cela, on utilise la commande suivant :

```
step.backward = step(model, direction = "backward", k = log(nrow(filtered_data)))
```

Les resultats nous donne un AIC de 422.59 pour un modèle prédit en fonction de T2_2h, T1_2h, T1_3h et T2_3h. On génère ensuite et compare ce nouveau modèle au modèle complet.

On obtient ici une p-valeur = 0.785 > 0.05, on accepte donc sous modèle.(*)

Nous évaluons la performance de notre modèle de régression logistique en utilisant une courbe ROC (Receiver Operating Characteristic). Cette courbe permet d'analyser la capacité du modèle à distinguer correctement les classes en traçant le taux de vrais positifs contre le taux de faux positifs pour différents seuils de classification.

L'aire sous la courbe ROC (AUC) permet de quantifier la performance globale du modèle : plus l'AUC est proche de 1, meilleure est la capacité de discrimination du modèle.

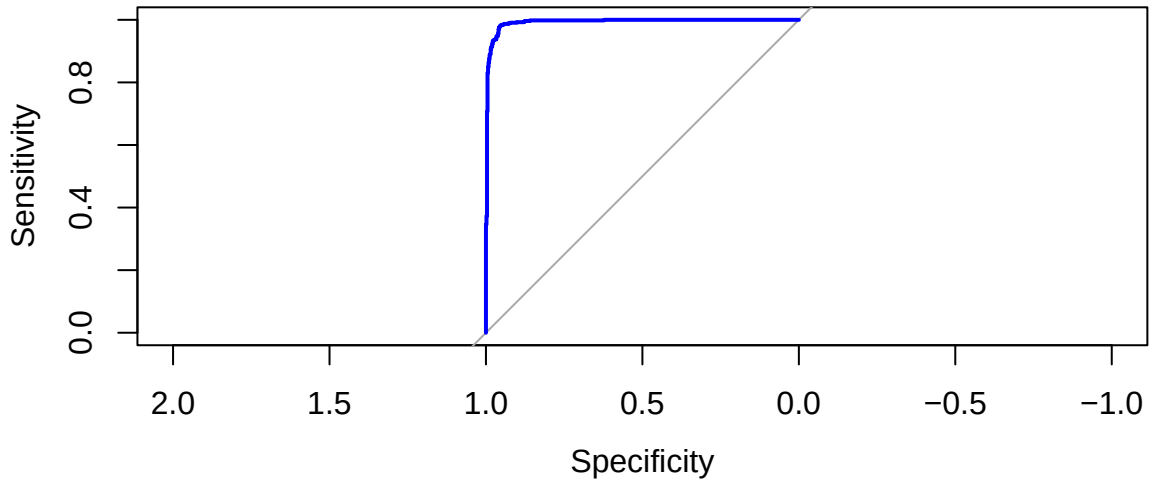


Figure 24: Courbe ROC

```
[1] "AUC: 0.992903855478609"
```

D'après la figure 24, la est très proche du coin supérieur gauche du graphique, ce qui indique une excellente capacité de discrimination du modèle. le critère AUC est égal à 0.9947, ce qui est très proche de 1. On conlut que notre modèle est performant pour prédire les gènes sur-exprimés et sous-exprimés à 6h.

8 Test d'indépendance

8.1 Pour tous les traitements

Nous souhaitons déterminer si le caractère des gènes à 6h est influencé par le traitement. Pour ce faire, nous effectuerons un test non paramétrique d'indépendance de Khi-deux. H_0: "le caractère est indépendant du

traitement" contre H_1: "il existe une relation entre ces deux variables". Pour mener ce test, nous construirons un tableau de contingence.

	T1	T2	T3
Non exprimé	2043	11	8
Sous-exprimé	57	1141	1140
Sur-exprimé	44	992	996

En effectuant le test, on obtient les resultats suivants :

```
    Pearson's Chi-squared test

data:  table_contingence
X-squared = 5903.3, df = 4, p-value < 2.2e-16
```

Compte tenu d'une p-valeur inférieure à 2.2e-16, nous rejetons l'hypothèse nulle selon laquelle le traitement n'influence pas l'expression génique à 6h. Cette conclusion est cohérente avec l'observation d'une interaction significative entre le traitement et le niveau d'expression initial. En effet, nous avons constaté que certains gènes sur-exprimés sous les traitements T2 et T3 ne le sont pas sous le traitement T1, et vice versa.

8.2 Pour les traitements T2 et T3

On se limite aux 2 traitements T2 et T3 et on refait le même test du Khi-deux. On obtient la table de contingence ainsi que le test suivants:

	T2	T3
Non exprimé	11	8
Sous-exprimé	1141	1140
Sur-exprimé	992	996

```
    Pearson's Chi-squared test

data:  table_T2_T3
X-squared = 0.48217, df = 2, p-value = 0.7858
```

La p-valeur obtenue est de 0.7858, ce qui est bien supérieur au seuil de significativité de 0.05. Par conséquent, nous ne pouvons pas rejeter l'hypothèse nulle H_0. L'hypothèse d'indépendance entre les traitements T2 et T3 et l'expression des gènes est donc confirmée. Cela signifie qu'il n'y a pas de différence significative dans l'expression des gènes à 6 heures en fonction du traitement (T2 ou T3).

En conclusion, les traitements T2 et T3 n'ont pas d'effet distinct sur l'expression des gènes dans le contexte étudié.

9 Conclusion

Ce projet nous a permis d'explorer les données d'expression de 2144 gènes d'une plante modèle à travers différentes méthodes d'analyse statistique et de modélisation. Grâce aux techniques de réduction de dimension, de clustering et de modélisation prédictive, nous avons pu identifier des tendances et des relations significatives entre les traitements, le temps et l'expression des gènes.

L'analyse des clusters a révélé des structures cohérentes entre les différentes méthodes utilisées (`k-means`, classification hiérarchique et modèles de mélanges gaussiens), tout en mettant en évidence les spécificités de chaque approche.

L'étude des réplicats a montré des différences significatives dans l'expression des gènes, confirmées par les tests statistiques.

Les modèles de prédiction de l'expression des gènes à 6h ont mis en évidence l'importance du temps d'observation, le modèle basé sur l'expression à 3h offrant de meilleures performances que celui utilisant les données à 1h. Enfin, les tests d'indépendance ont permis d'évaluer l'impact des traitements sur l'expression génique et de mieux comprendre les interactions entre les différentes variables étudiées.

Ce travail met ainsi en lumière la complexité des dynamiques d'expression des gènes et la pertinence des outils statistiques pour leur analyse. Des perspectives d'amélioration pourraient inclure l'intégration de modèles non linéaires ou d'approches de machine learning plus avancées afin d'affiner les prédictions et d'explorer d'autres aspects de la régulation génique.