Projet

2024-12-04

Contenu du jeu de données :

- 3 variables qualitatives nominales représentant l'expression du gêne

  $$g$$

  dont les modalités sont

  $$\{"sur", "sous", "non"\}$$

  . chaque variable correspond respectivement à la différence d'expression du gêne mesurée à la 6ème heure lors du traitement

  $$T \in \{T1, T2, T3\}$$

  en moyenne, sur les réplicats

  $$\{R1, R2\}$$

- $$3 * 6 + 3 * 6 = 36$$

  variables quantitatives continues représentant les effets des traitements sur l'expression des gênes T1 T2 et T3 à 1h,2h,3h,4h,5h,6h après l'administration pour les replicats R1 et R2, par rapport à leur expression à T=0 (sans traitement).

- Ce jeu de données contient des relevés sur 542 individus, ici des gênes.

Analyse unidimentionnelle :

Expression des gênes lors des traitements T1,T2 et T3

Nous avons réalisé 3 histogrammes pour visualiser les effectifs des gènes en fonction de leur expression relative moyenne à 6h suite aux traitements T1,T2 et T3, cette moyenne est représenté par les variables qualitatives nominales ExpT1,ExpT2 et ExpT3

Analyse des résultats

On remarque que les traitements T2 et T3 semblent avoir un effet assez similaire sur l'expression des gênes relevée à la 6ème heure : Une polarisation entre la sous expression et la sur expression qui se partagent presque la totalité des relevés, avec un poids légèrement superieur à 55% pour la sur-expression au détriment de la sous-expression. Cela a peut être un rapport avec le fait que T3 est une combinaison des traitement T1 et T2.

T1 quant à lui se démarque grandement par une large majorité (Un peu plus de 80%), de gêne n'ayant pas changé d'expression après 6h de traitement.

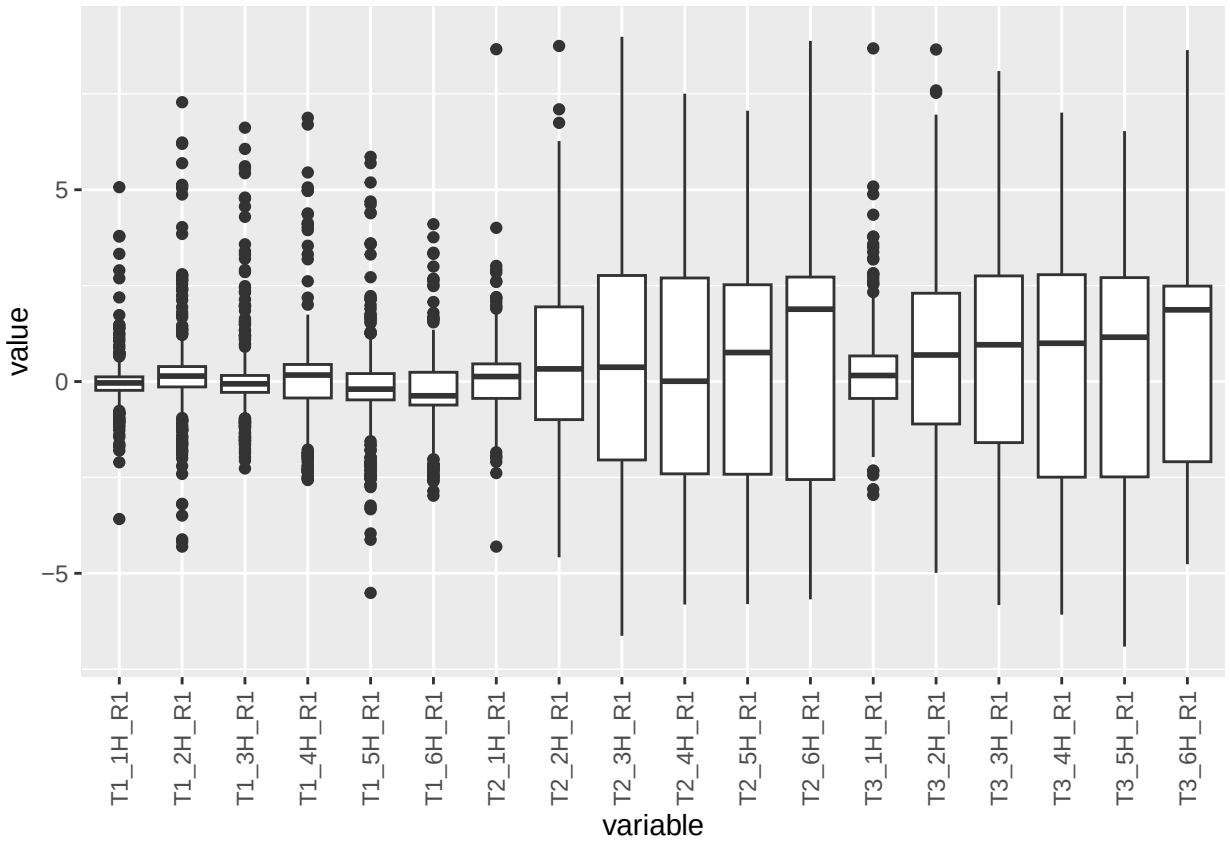
Expression relative des gênes mesurées à intervalle régulier

Lors d'une analyse supplémentaire des fréquences d'expression des gènes à chaque heure et chaque traitement, nous observons bien une concordance avec l'analyse des expressions des gênes figure . En effet, les histogrammes en rapport avec le traitement 1 sont très nettement regroupés vers 0, soit une expression relative des gênes qui ne change peu. Les histogrammes pour les relevés des variables en lien avec T2 et T3 sont tout aussi similaires aux résultats précédents : La variance de l'expression relative des gênes est plus élevée et on observe bien une polarisation "sous-exprimé-"sur-exprimé" sur les relevés à 6h. Attention, ici on observe aussi que T2 et T3 n'ont pas leur effet caractéristique directement : à 2h, la distribution de l'expression des genes semble presque Gaussienne, et à 1h elle ne se distingue pas beaucoup du traitement 1 avec un regroupement sur 0.

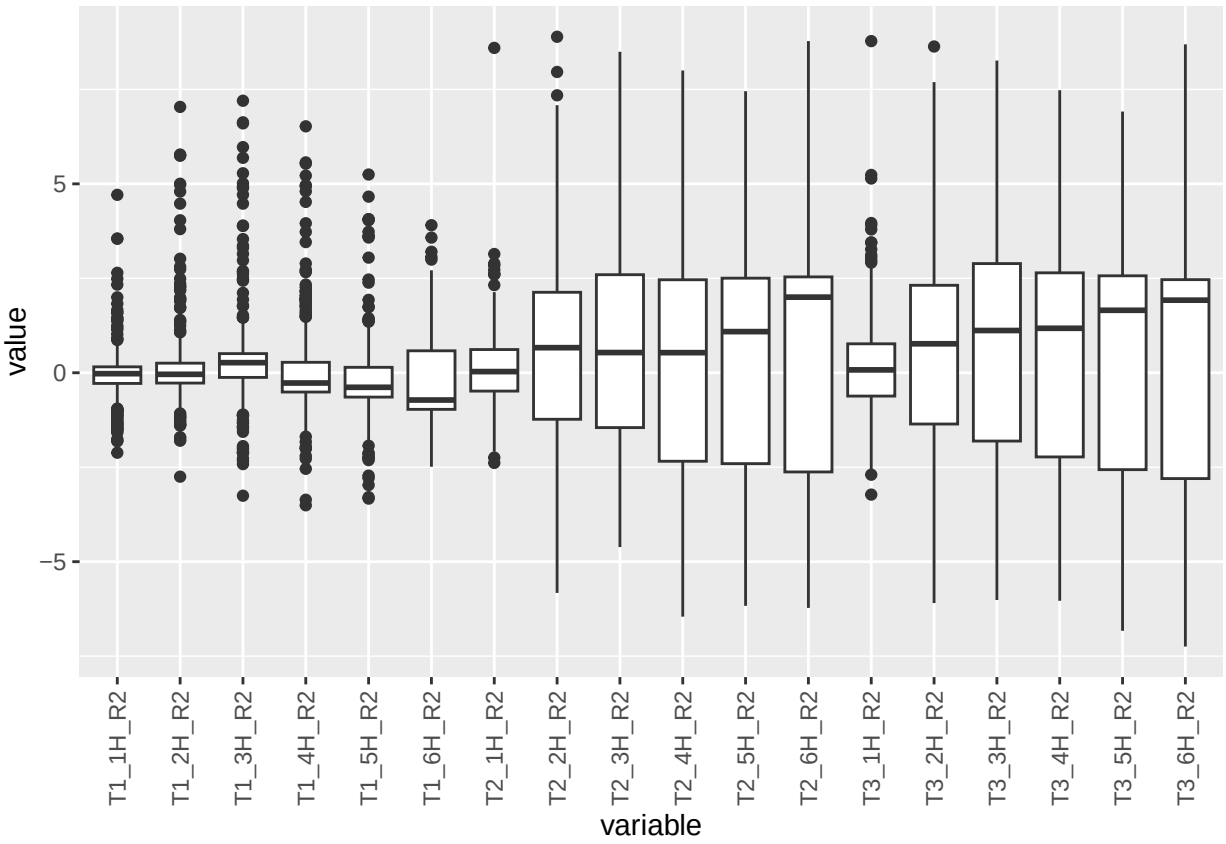
c.f le document RMarkdown pour les histogrammes.

boxplots pour faire joli

```
## No id variables; using all as measure variables
```




```
## No id variables; using all as measure variables
```

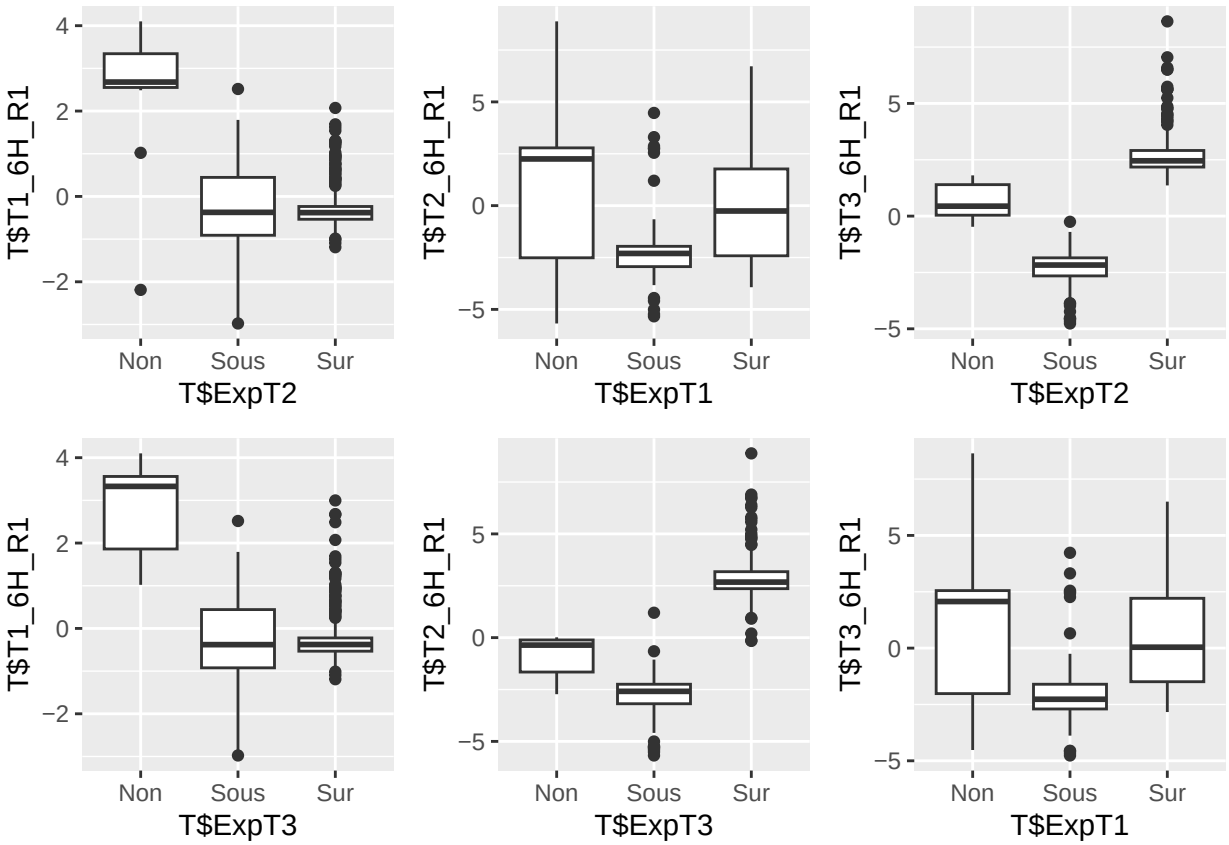


L'analyse des boxplots montre que, même sans effectuer de réduction des données, chaque variable présente une dispersion similaire, ce qui suggère qu'il n'est pas nécessaire de procéder à un centrage et à une réduction avant de réaliser l'Analyse en Composantes Principales.

Analyse bi-dimensionnelle

Boxplots par paire de variables (qualitative,quantitative)

Nous avons ensuite générés des graphes contenants des boxplots d'expression des gènes moyennes d'un traitement par rapport à l'expression des gènes à 6h sur un autre traitement. Cela nous a permis d'identifier les changements d'expressions de chaque groupes de gènes sur, sous ou non-exprimés entre un traitement et les deux autres. Les deux réplicats ayant des résultats très similaires, pour ne pas surcharger le rapport nous avons décidé d'afficher seulement le réplicat R1.

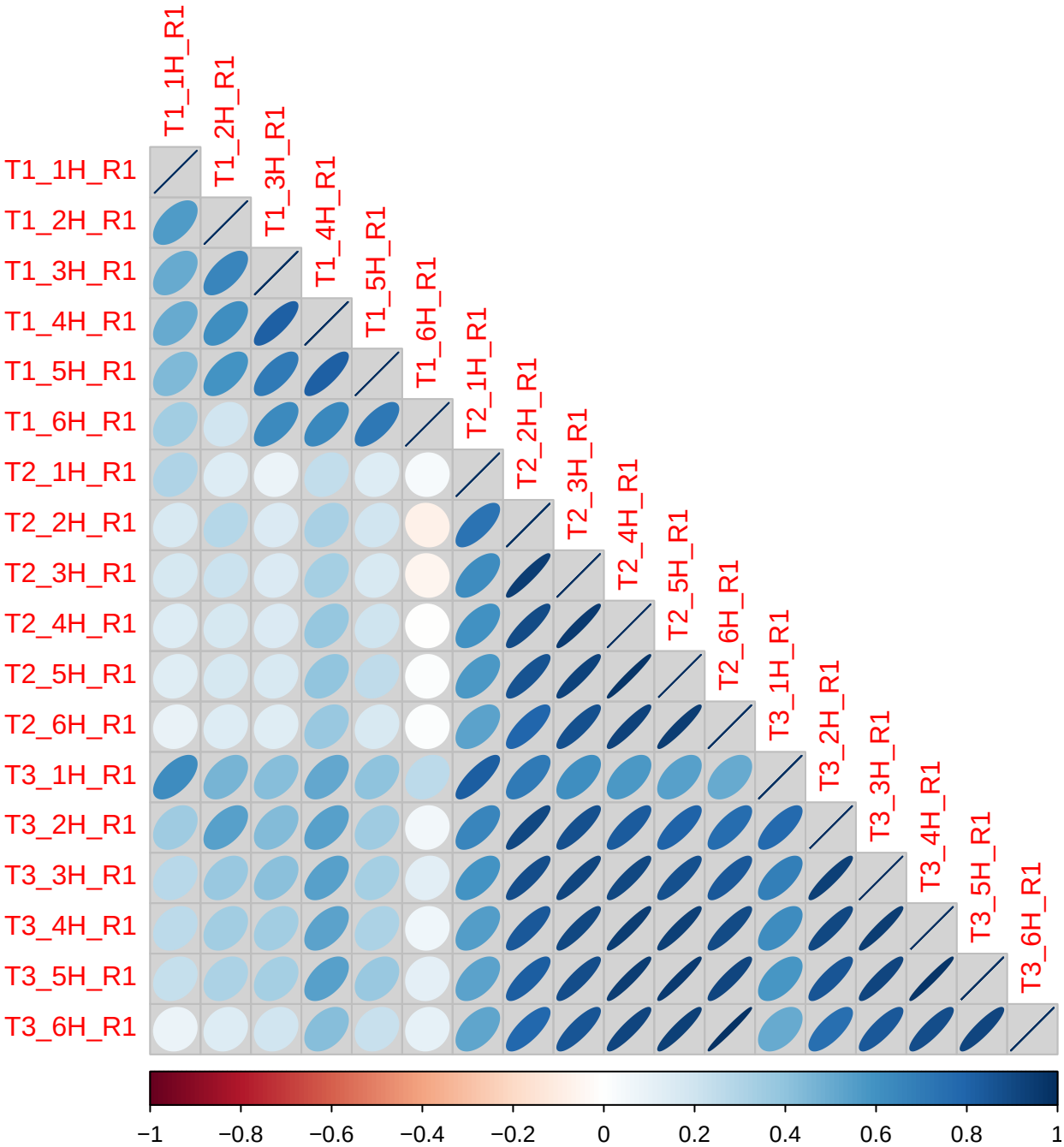


Analyse des boxplots : > Traitement 1 Les genes sur-exprimés au T1 sont non-exprimé durant le T2 et T3 (Boxplots T1/ExpT2 et T1/ExpT3) . Il est difficile d'observer une catégorie de genes de T1 qui se soient sous exprimés ou sur exprimés dans T2 et T3, ceux qui n'avaient pas changés d'expression relative durant T1, se sont soit sous exprimé soit sur exprimé (Boxplots T1/ExpT2 et T1/ExpT3).

traitements 2 et 3 On observe une légère tendance des genes s'étant sous-exprimé avec T1 à se sous exprimer avec T2 et T3 (Boxplots T2/ExpT1 et T3/expT1). En revanche, il est très clair que T2 et T3 ont les même effet sur les mêmes genes à 6h, toutes les expressions relevées par T2 concordent aux modalités qualitatives moyennes calculées sur T3 (Boxplots T2/ExpT3 et T3/ExpT2).

matrice de covariance des variables quantitatives

Nous avons généré la matrice de covariance de nos données sans les variables quantitatives. Suite à cela nous avons remarqué qu'il était inutile d'afficher le réplicat R2 dans la matrice et que cela rendait le graphe moins lisible. Nous avons donc décidé de seulement afficher le graphe de la matrice de covariance avec le réplicat R1.



analyse de la matrice de covariance On observe clairement de grandes zones de de corrélation entre T2 et T3 délimitées par une correlation plus moyenne à la première heure des deux traitements sûrement dues au fait que T3 est influencé par T1 aux premières heures. On remarque que T3 est plus fortement corrélé à T1 à 1H que T2 (qui ne l'est que légèrement), semblant indiquer que T1 s'exprime avant T2; T3 étant la

combinaison des deux traitements.

Rapport de correlation Eta2

```
## [1] "T1 vs T2"
## Replicat 1 :  0.1874682
## Replicat 2 :  0.1987298
## [1] "T1 vs T3"
## Replicat 1 :  0.1516016
## Replicat 2 :  0.1423772
## [1] "T2 vs T3"
## Replicat 1 :  0.9022439
## Replicat 2 :  0.8877959
```

Le calcul du rapport de correlation eta^2 bien notre observation de la grande similarité d'expression des genes traités avec T2 et T3 et la dissimilarité des expression des genes lorsque la plante est traitée avec T1 comparée à T2 et T3, chose normale au vu du peu de genes affectés par T1.

table de contingence pour les variables quali 2 à 2, mosaic plot ?

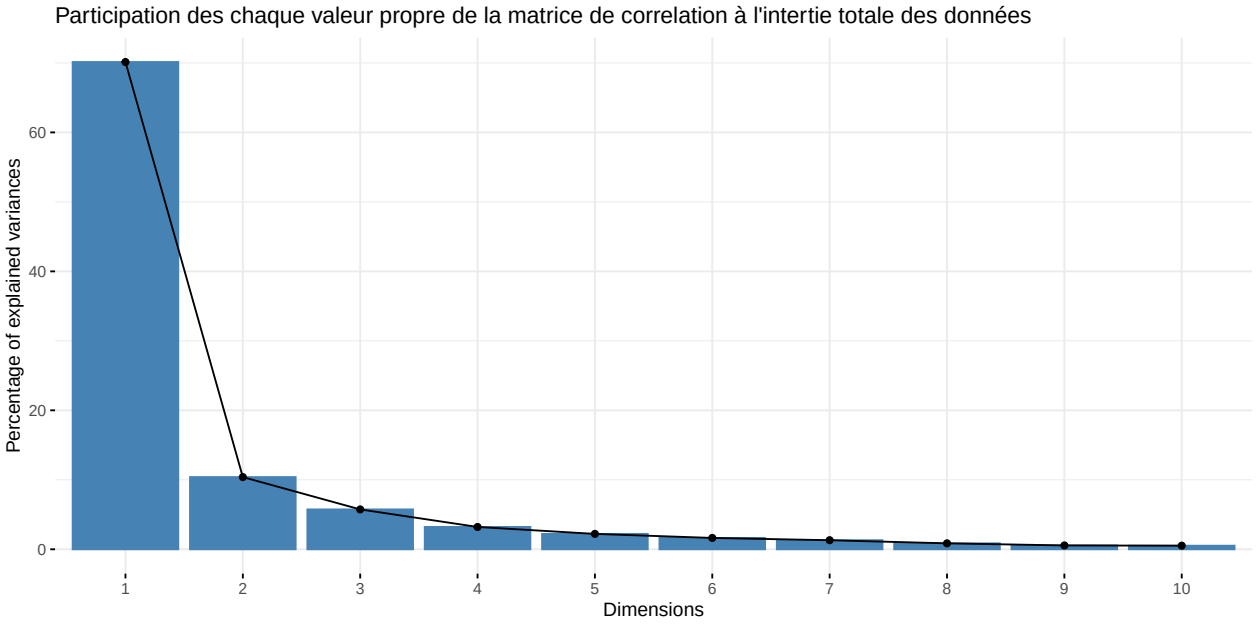
```
## [1] "table de contingence entre T1 et T2"
##
##        Non Sous Sur
##   Non    0  178 263
##   Sous   1   50   6
##   Sur   10   19  15
## [1] "table de contingence entre T1 et T3"
##
##        Non Sous Sur
##   Non    0  178 263
##   Sous   0   51   6
##   Sur    7   18  19
## [1] "table de contingence entre T2 et T3"
##
##        Non Sous Sur
##   Non    6    1   4
##   Sous   1  246   0
##   Sur    0    0 284
```

Nouvelle confirmation de nos résultats de manière encore plus précise, on observe que T1 ne change pas l'expression de la très grande majorité des genes. Plus finement, on peut confirmer l'observation faite sur les boxplots tendant à dire que le peu de genes s'étant sous exprimés avec T1 se sont aussi sous-exprimés avec T2 et T3.

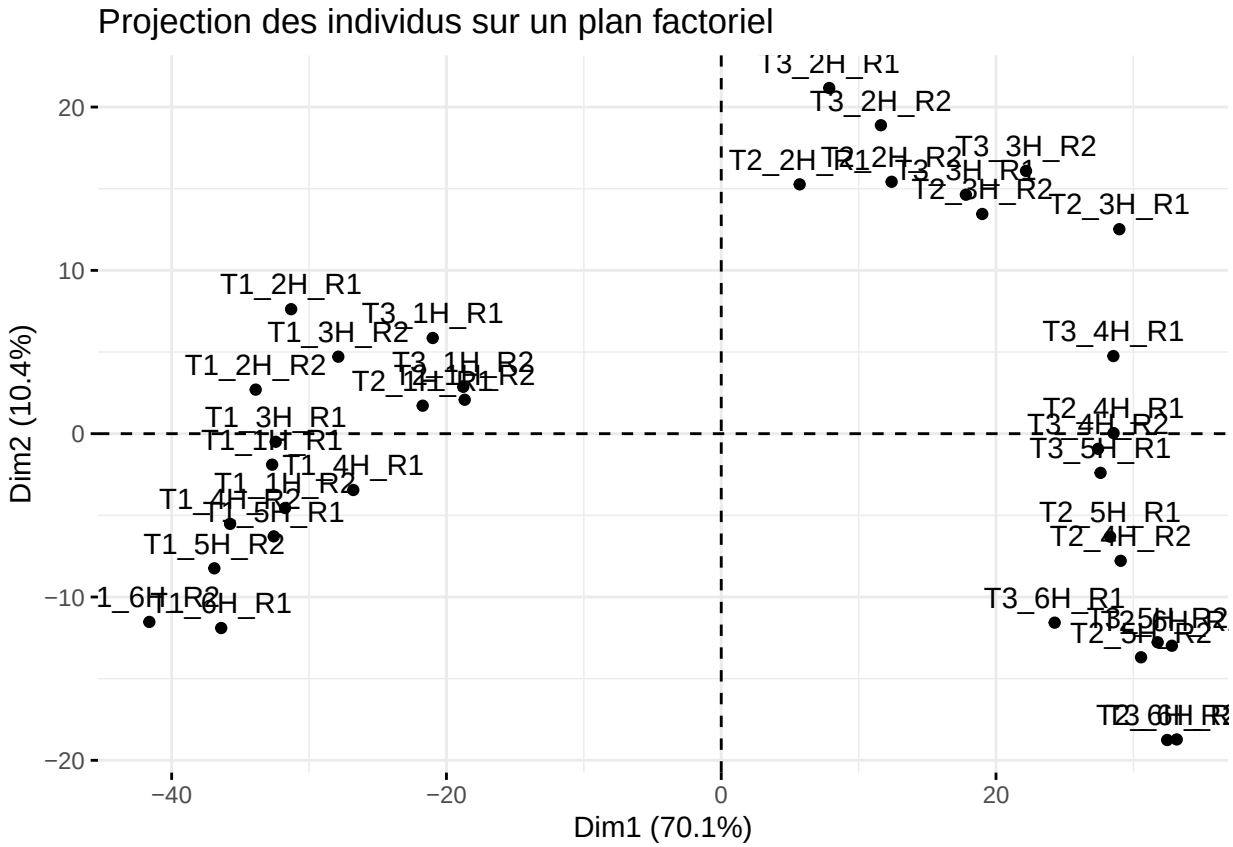
La grande valeur des effectifs partiels sur la diagonale de la table de contingence entre T2 et T3 montre bien la similarité de l'effet de ces deux traitement sur l'expression des genes.

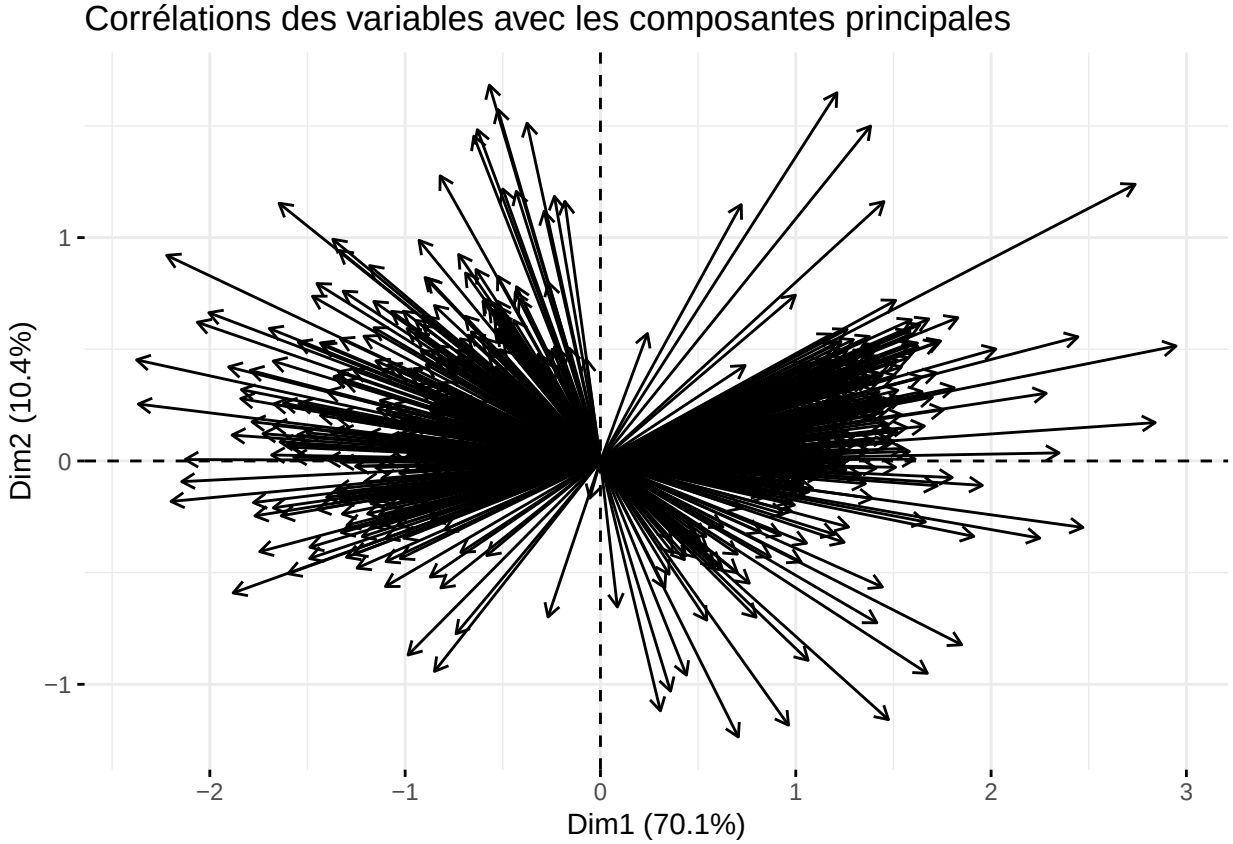
Analyse en composantes principales où les Tt sH Rr sont les individus décrits par les gènes.

L'ACP permet d'obtenir une vision synthétique des données en réduisant leur complexité tout en préservant l'information essentielle. Elle facilite l'identification des tendances et profils d'individus, ainsi que la construction de méta-variables, plus simples à interpréter, tout en évitant les redondances entre les variables. Pour faire cela, nous devons transposer la matrice de données originale qui elle décrivait les gènes (individus) en fonction des Tt sH Rr. Nous décidons de faire directement une ACP car, comme mentionné plus tôt lors de l'analyse de la figure , nous n'avons pas besoin de centrer et réduire le jeu de données.



Ce graphique représente la participation de chaque valeur propre de la matrice de corrélation du jeu de données dans l'inertie totale. L'inertie totale étant la somme des valeurs propres (qui elles sont les inerties axiale associées à l'axe de vecteur directeur le vecteur propre associé), chaque valeur propre est donc une fraction de l'inertie totale. On voit qu'on dépasse 80% de l'inertie totale rien qu'avec les deux premieres valeurs propres, on en prend donc les vecteurs propres associés comme axes principaux de l'analyse.





Contexte : les relevés aux heures sont décrits par les gènes (les gènes sont considérés comme les variables).

- les genes proches d'un axe sont très représentés par celui-ci
- les genes dont l'angle entre eux est petit sont corrélés entre eux

Interprétation globale du couple de graphes

On voit que les genes se polarisent principalement sur l'axe 1 dans un sens ou l'autre. Certaines flèches sont d'une longueur presque du rayon du cercle, indiquant une participation très forte des genes dans la variance expliquée par ces dimensions. Il n'y a pas de tendance particulière sur la direction selon l'axe 2 des flèches : Dans chaque "polarité" de fleches selon l'axe 1, il y a des fleches dont la direction est negative d'autres positive selon l'axe 2. Bien que l'on dénote une quantité plus grande de gènes corrélés positiviement à la dimension 2.

Le traitement 1 est entièrement groupé sur des valeurs très negatives de l'axe 1. On remarque dans ce groupement la présence des T2 et T3 à la première heure de relevés d'expression des gènes.

Pour le traitement 2 et 3, on les retrouves formant 2 groupements, 1 en haut à droite du graphe contenant les relevés à 2 et 3h puis un groupement s'étalant sur la droite du graphe du centre jusqu'en bas contenant les relevés à partir de 4h.

Clustering

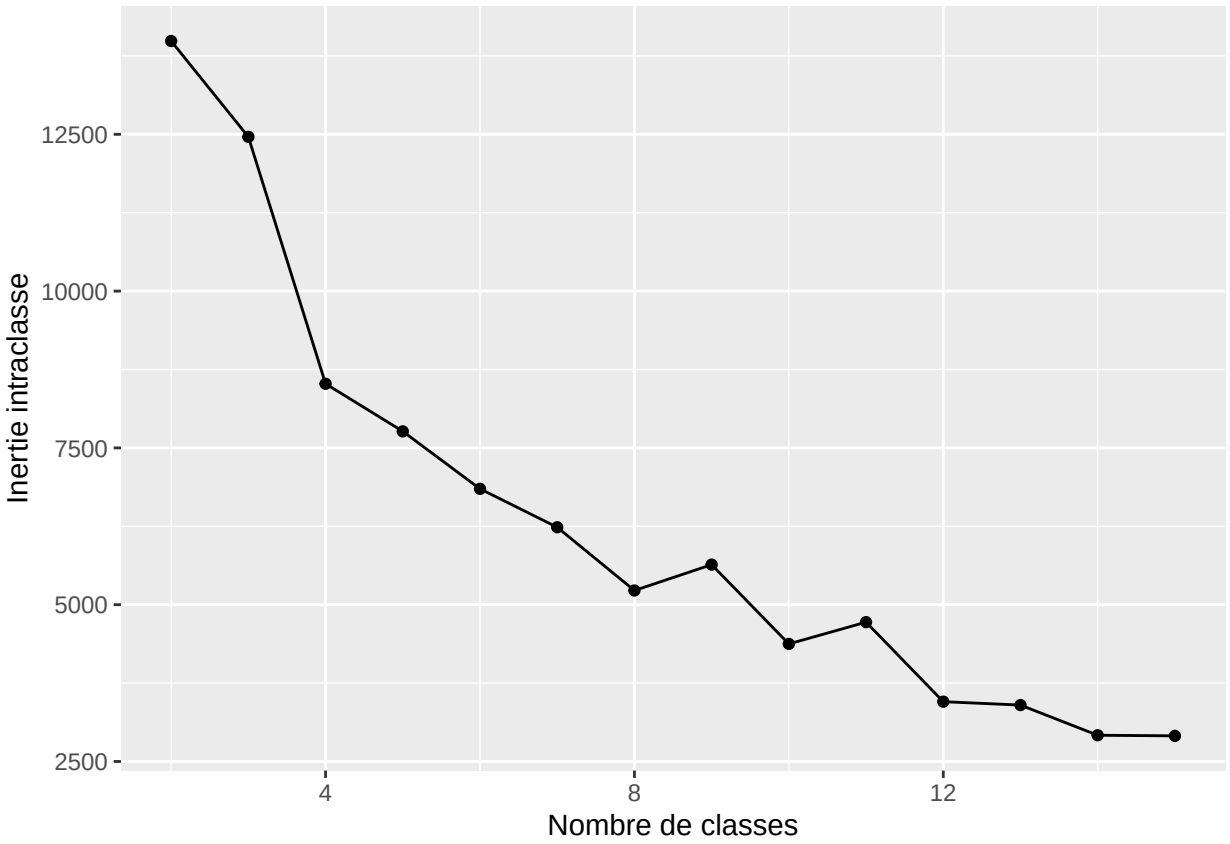
L'objectif du clustering est de regrouper des individus (ici, les Tt_sH_Rr) en groupes homogènes selon leurs similarités, sans connaissance préalable des catégories. Cela permet d'identifier des profils ou des comportements communs, facilitant l'interprétation des données et la mise en évidence de structures sous-jacentes.

Nous avons commencé par effectuer un clustering avec la méthode K-means basée sur des estimations de nombre de clusters optimaux avec Silhouette et Calinski-Harabasz ainsi que l'inertie intra-classe, pour finir par effectuer un clustering avec la méthode CAH dont le nombre de classes est déterminé par l'indice de Calinski-Harabasz calculé sur des coupures du dendogramme.

Pour le dendogramme, nous avons utilisé la mesure d'agrégation de Ward car il y a trop de points pour utiliser la mesure euclidienne avec efficacité.

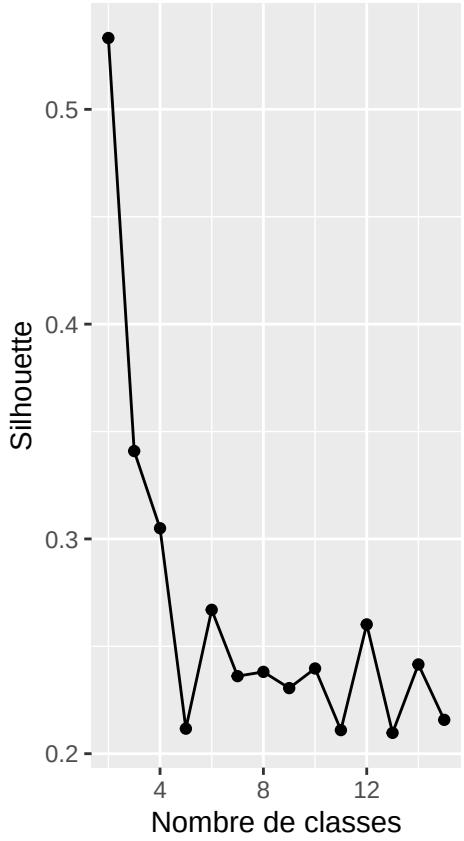
Clustering k-means

On commence par afficher l'inertie intra classe en fonction du nombre de classe pour le clustering avec k-means, le but étant de la minimiser tout en gardant un nombre de classes raisonnable.



On dénote un coude dans la courbe d'inertie intraclasse aux alentours de 4 clusters.

À présent on va afficher l'évolution de l'indicateur Silouhette en fonction du nombre de classe ainsi que



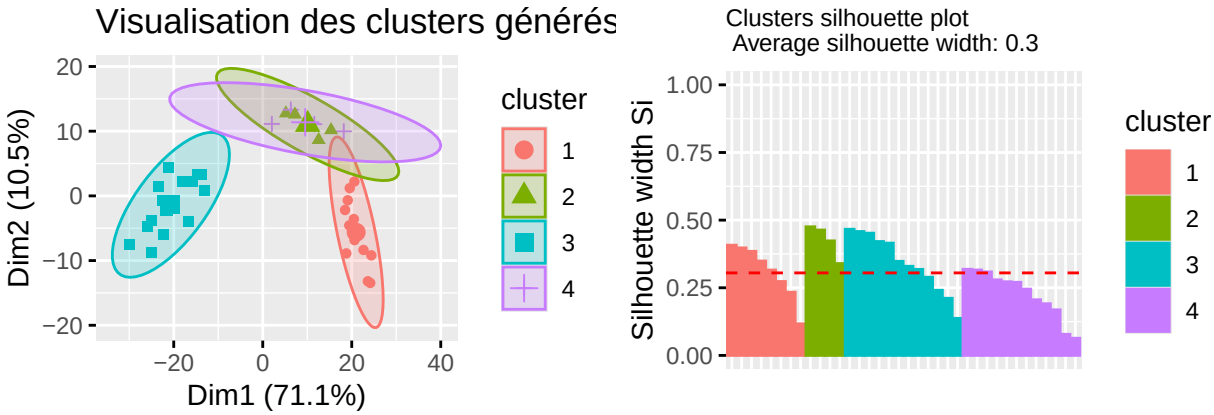
l'évolution de l'indice de Calinski-Harabasz en fonction du nombre de classe.

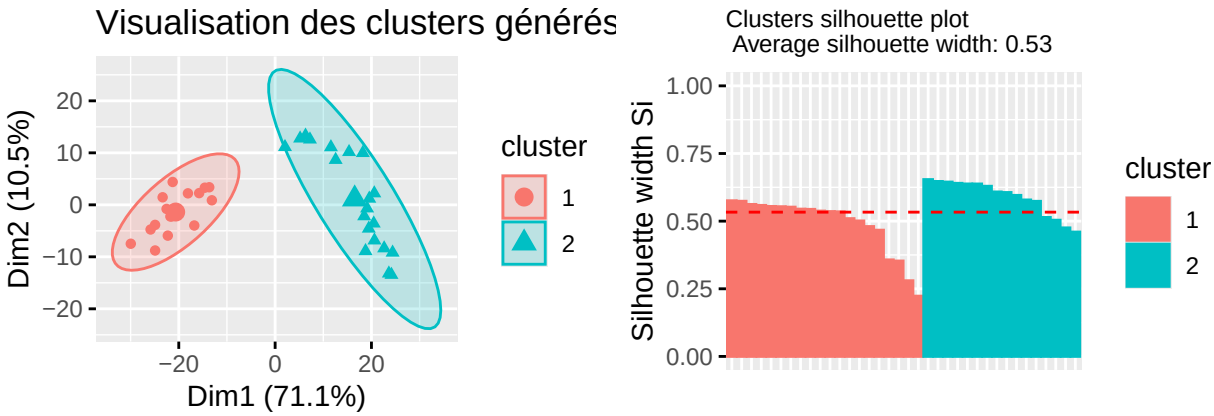
Silhouette et l'indice de Calinski-Harabasz ont un pic à 2, mais cela est normal sachant que Silhouette et l'indice de Calinski-Harabasz ont tendance à sous-estimer le nombre de clusters.

visualisation du clustering

```
##   cluster size ave.sil.width
## 1       1    8          0.31
## 2       2    4          0.43
## 3       3   12          0.34
## 4       4   12          0.23
```

```
##   cluster size ave.sil.width
## 1       1   20          0.49
## 2       2   16          0.59
```

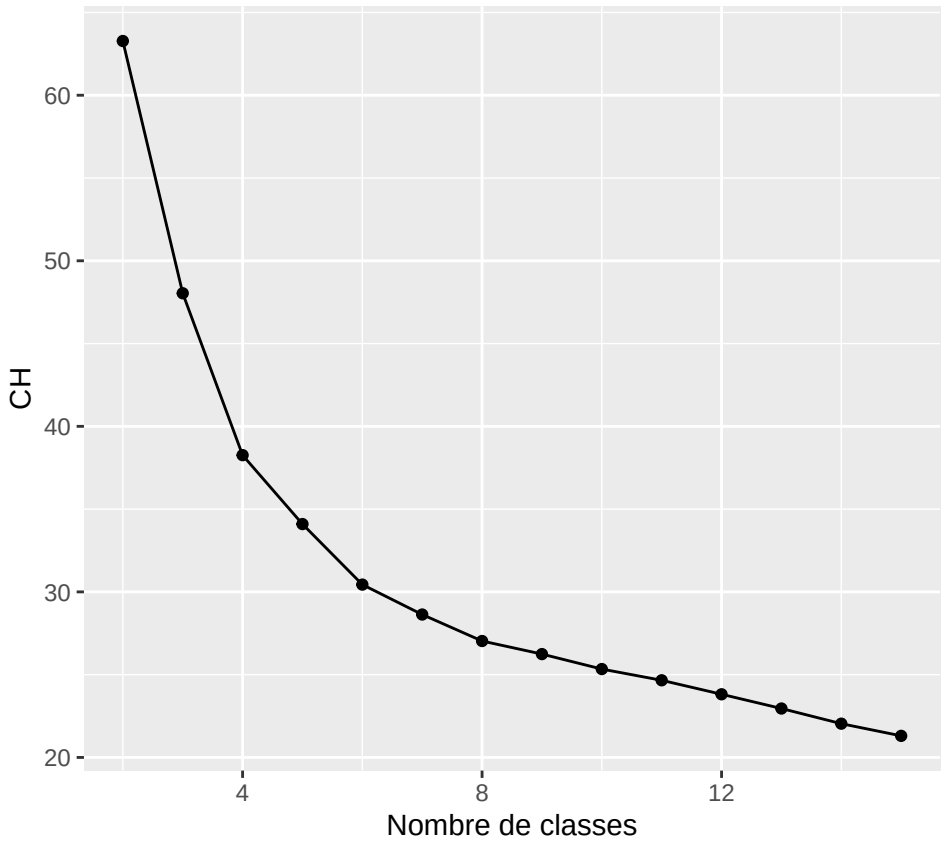




Après comparaison, on choisit 2 classes car cela nous semble plus optimal que 4.

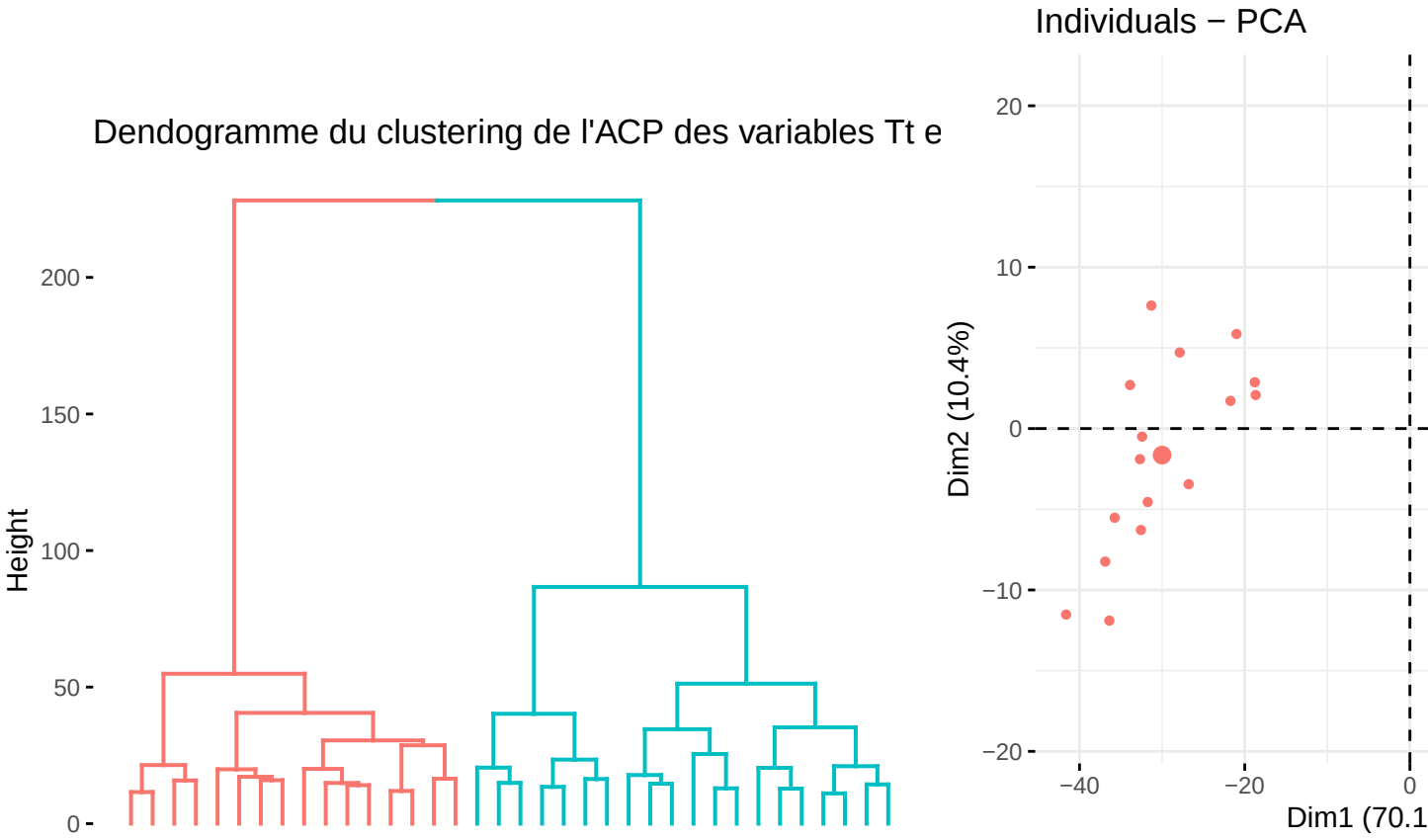
clustering CAH

À présent on va afficher l'évolution de l'indice de Calinski-Harabasz en fonction du nombre de classes utilisées pour découper le dendogramme.



On choisit alors de prendre 2 classes par observation du maximum du graphe.

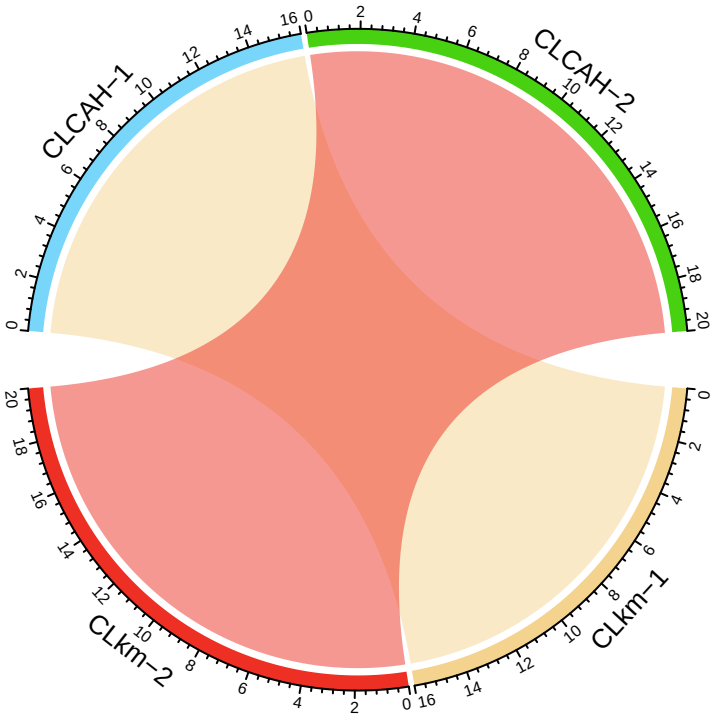
```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Comparaison des clusterings

A 2 classes nous obtenons une classification qui ne change pas entre chaque méthode. Nous décidons donc qu'il s'agit donc d'un bon choix de nombre de classes.

La classification obtenue est en accord avec les observations faites lors de l'ACP, on y retrouve plus ou moins les mêmes groupements : celui majoritairement composé des relevés de T1 avec une majorité de gènes sans changement d'expression relative et celui composé des relevés de T2 et T3.



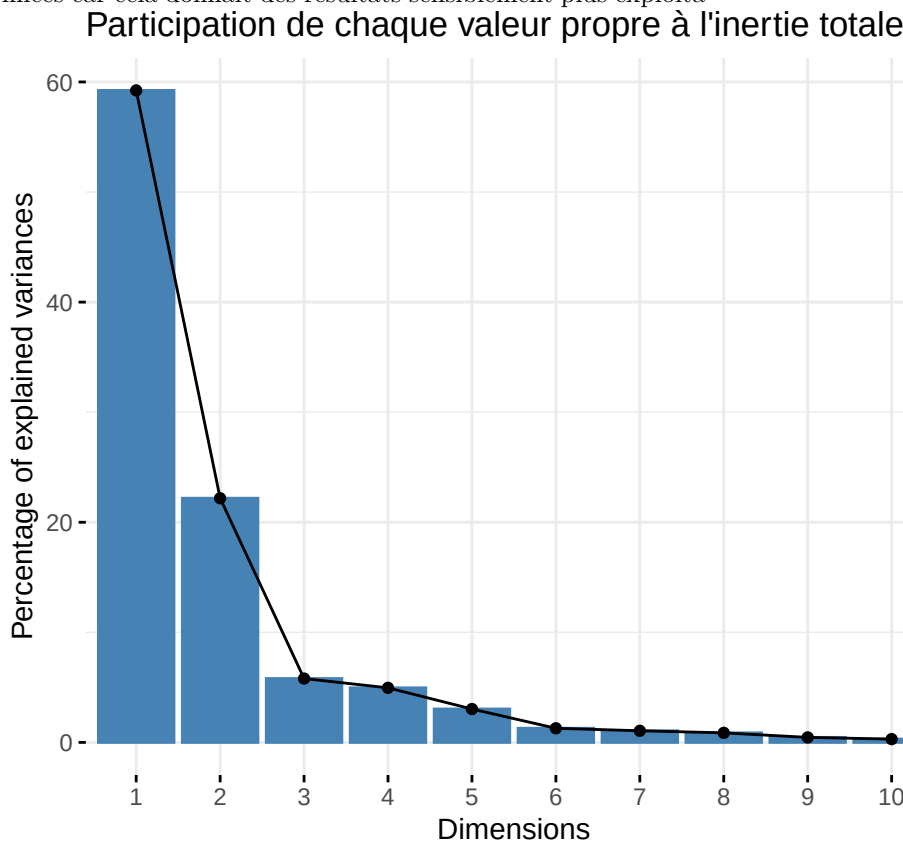
ANALYSE DES GENES

Generation de dataExpMoy

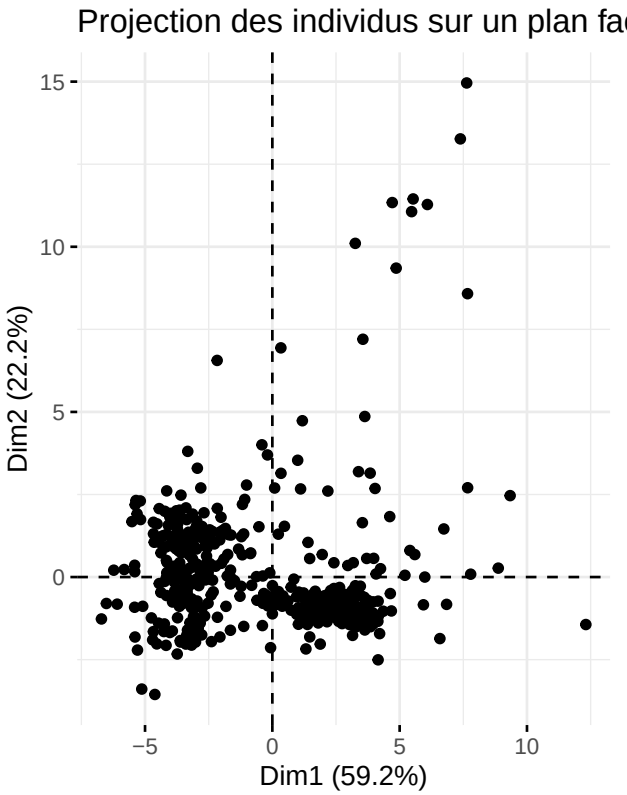
Nous construisons le jeu de données DataExpMoy contenant la moyenne des expressions sur les réplicats de chaque gène, pour chaque traitement et chaque heure. DataExpMoy est donc une matrice de taille 542× 18.

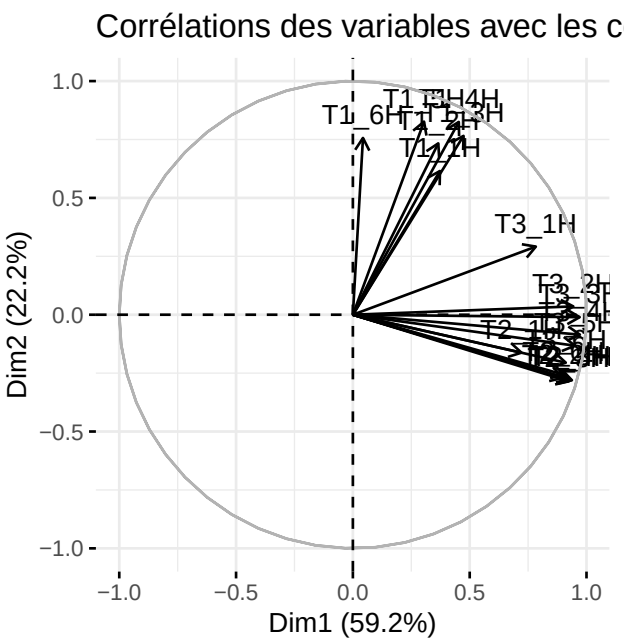
ACP de DataExpMoy

A partir d'ici, nous avons centré et réduit les données car cela donnait des résultats sensiblement plus exploitables au niveau des indicateurs pour le clustering.



On voit qu'on dépasse 80% de l'inertie totale avec les deux premières valeurs propres, on prend donc les vecteurs propres associés à ces deux valeurs propres comme composantes principales de notre ACP.





Il semble y avoir deux groupes principaux de gènes (individus), ce qui pourrait refléter une différenciation claire entre les niveaux d'expression sous différents traitements. Les points plus éloignés du centre vers le haut représentent des gènes ayant des comportements plus spécifiques ou atypiques.

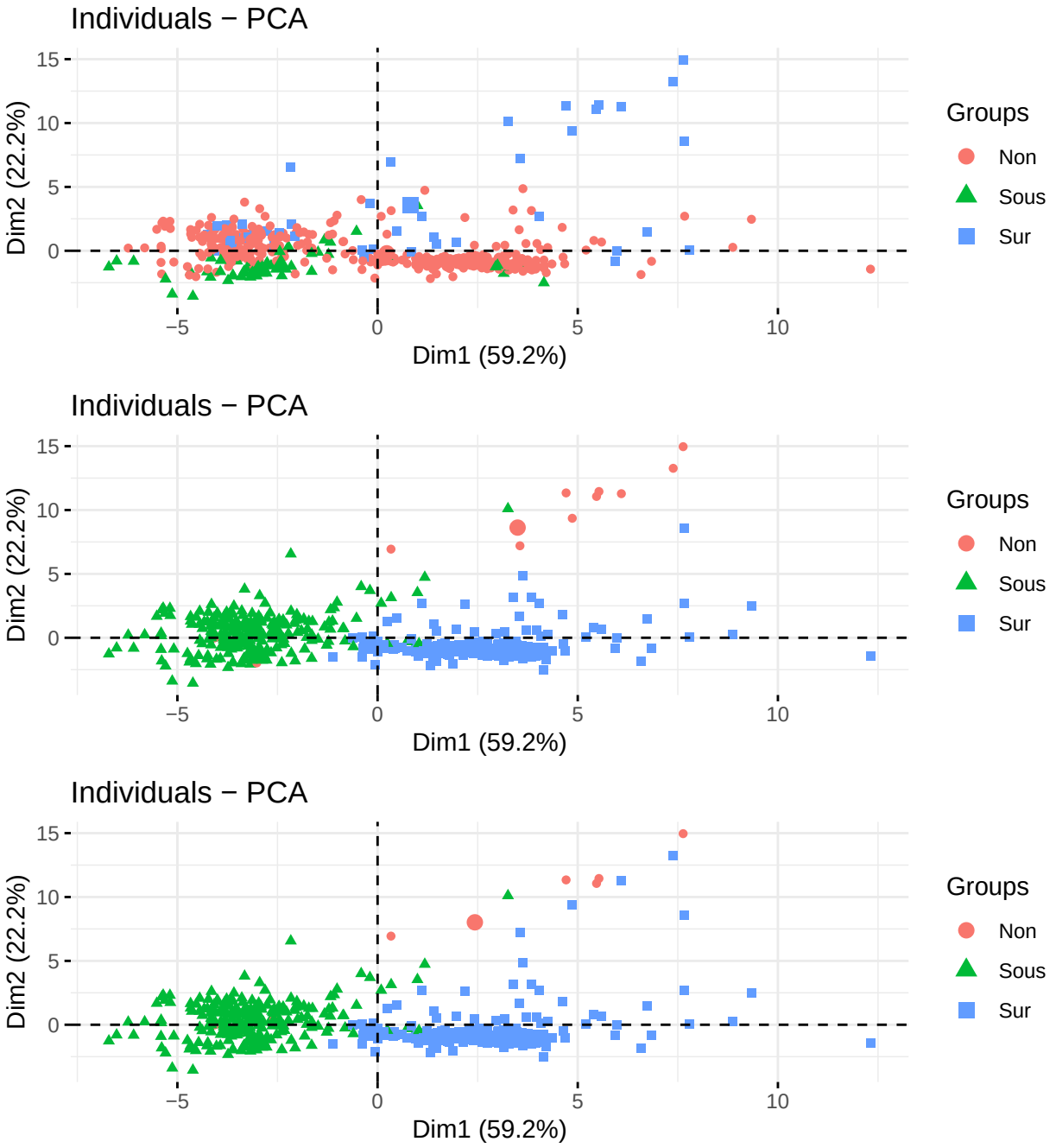
Le premier axe (DIM 1) est fortement corrélé avec les traitements T2, T3 à des moments précis, en particulier pour les temps plus avancés (par exemple, T3 4H, T3 5H, T3 6H). Cela suggère que Dim1 reflète les différences d'expression liées à l'effet des traitements T2 au fil du temps, qui a la particularité de faire exprimer ses gènes sur les temps plus longs. Cela explique aussi la corrélation des variables de relevés de T3 à cette dimension, car T3 est un mélange de T1 et T2. Le deuxième axe (DIM2) semble corrélé avec des effets à court terme, typiques de T1. Cela pourrait indiquer des gènes qui réagissent rapidement mais dont l'effet s'atténue à long terme.

En regardant la correlations des variables qui représentent les relevés sur T1, comme elles sont très positivement corrélées avec la dimension 2, on en déduis que les gènes vers les valeurs élevées de la dimension 2 sont ceux ciblés par le traitement 1, et 3.

Comme T3 est une combinaison des traitements T1 et T2, il semble structurer fortement l'axe principal (Dim1) pour des temps intermédiaires et avancés, mais est notablement corrélé positivement à la dimension 2 pour les premières heures de traitement, correspondant bien avec le fait que T1 fait exprimer les gènes qu'il vise de manière très rapide. On retrouve bien le fait que T3 est la somme des deux autres traitements.

Ajout des variables qualitatives

Nous allons maintenant afficher les points en dans les dimensions de l'ACP, colorés en fonction des modalités des 3 variables qualitatives ExpT1, ExpT2 et ExpT3. Cela peut nous permettre d'affiner notre interprétation des ré-

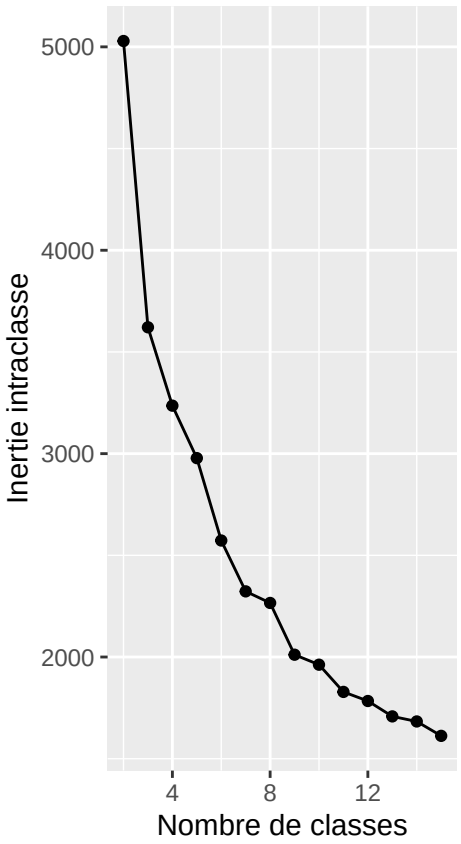


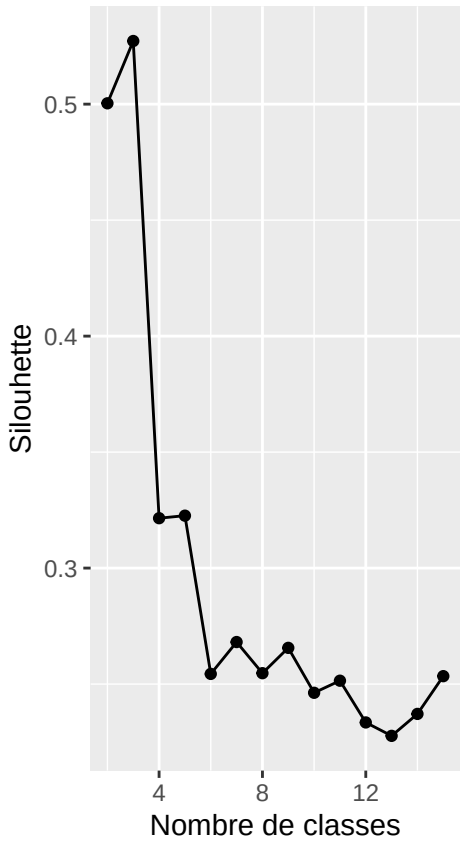
sultats de l'ACP.
analyse On observe donc 3 clusters de gènes : - Les gènes étant poussés à la sur-expression par T1, affichés comme non-exprimés durant T2 et partiellement sur-exprimés avec T3 - Les gènes étant sous-exprimés (verts) durant T2 et T3, non-exprimés durant T1. - Les gènes étant sur-exprimés (bleus) durant T2 et T3, non-exprimés durant T1.

On confirme donc bien notre analyse descriptive préliminaire, et nos suppositions de sens prétés aux dimensions 1 et 2 de l'ACP.

Clustering

L'objectif de ce clustering est de regrouper des individus (ici, des gènes) en groupes homogènes selon leurs similarités.





En lisant le graphe de l'inertie intra-classe, on observe le coude aux alentours de 3 classes. En observant l'indice de Silouhette, on remarque le pic recherché à 3 classes également. Les indices concordants, on en déduit que le choix optimal du nombre de classes pour le clustering est 3.

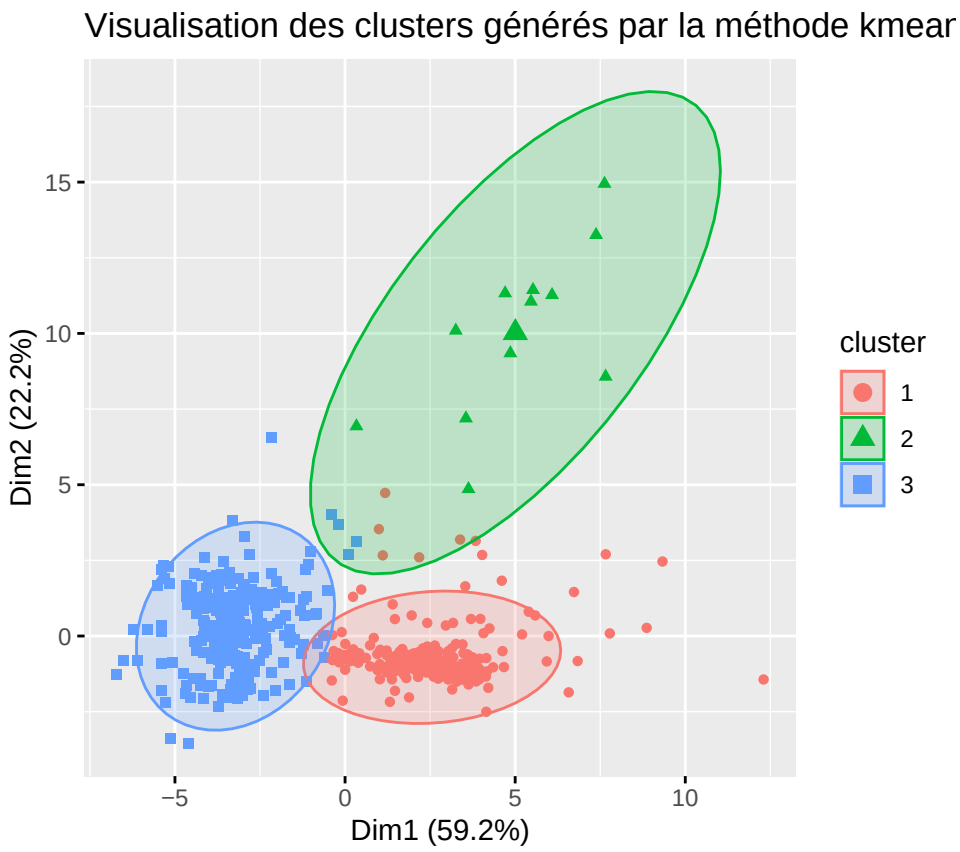
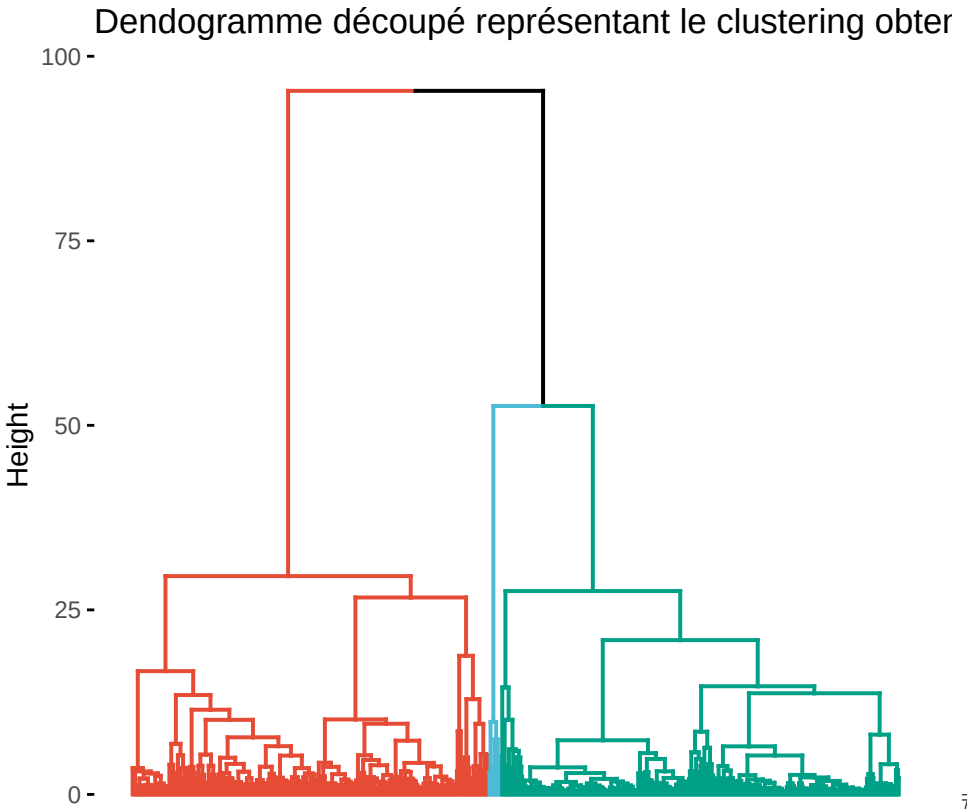


Figure 1: Visualisation des clusters générés par la méthode kmeans dnas le plan factoriel 1,2

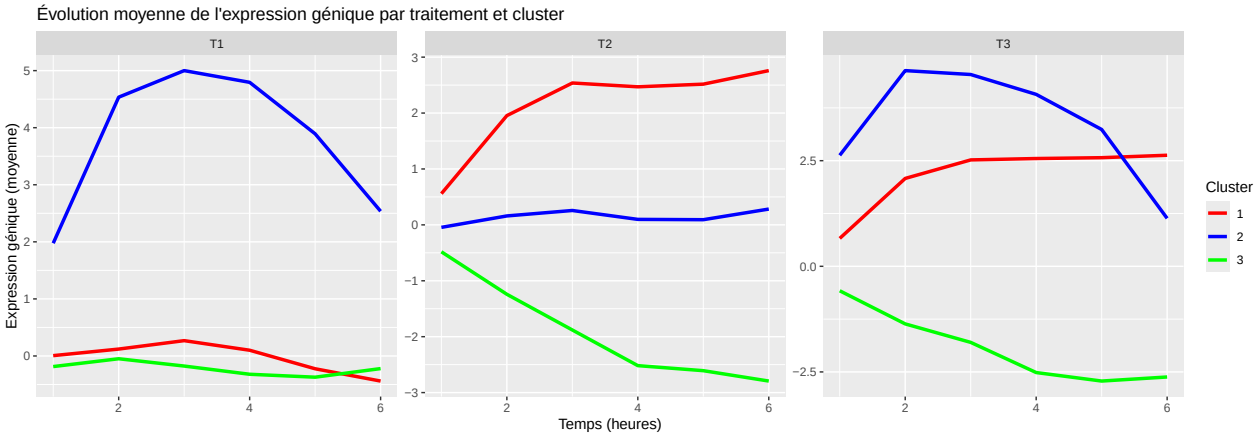


analyse des clusterings On remarque que les clusterings correspondent aux groupes que nous avions discernés sur la figure

Evolution de l'expression des genes en fonction de leur traitement et cluster

Nous avons généré 3 graphiques, représentant l'évolution de l'expression moyenne des gènes par cluster et par traitement. Cela pourra nous permettre d'affiner nos observations sur les traitements.

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



On lis sur les graphiques de la figure <> la liaison de toutes les observations que nous avons pu faire :

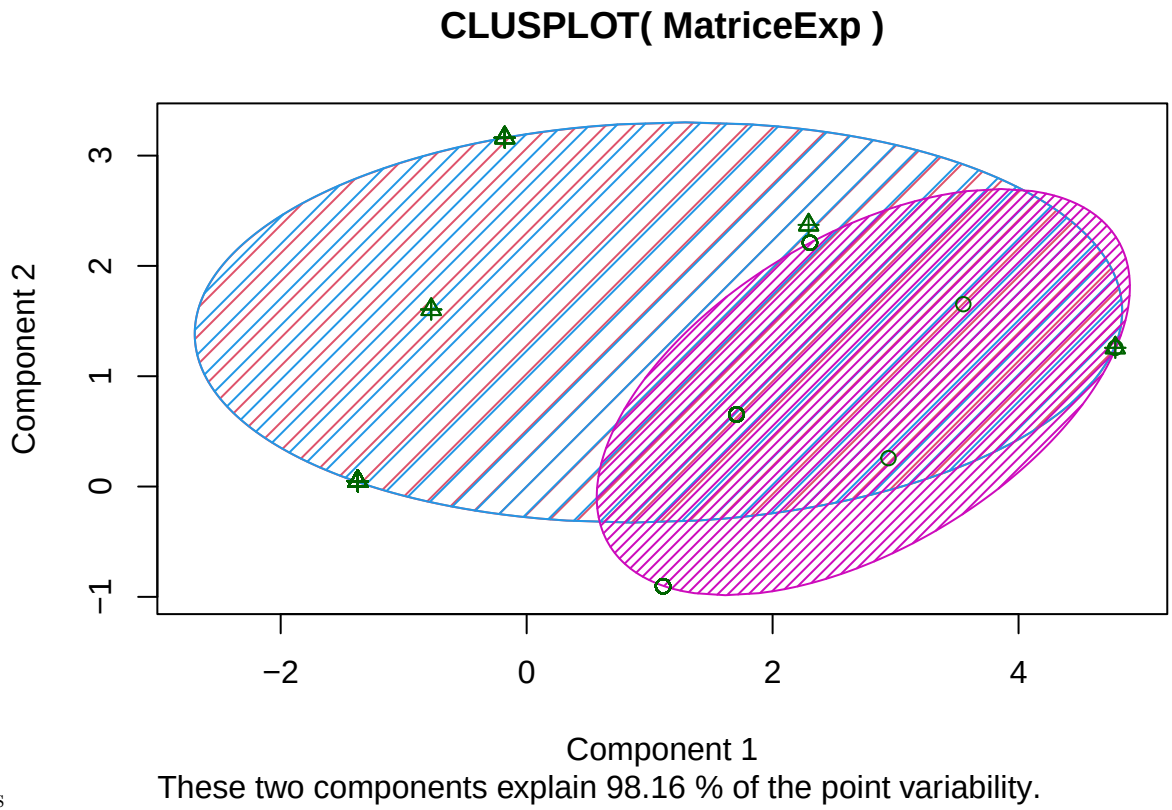
On remarque sur T1 le fait que le cluster 2 correspond au peu de gènes qu'il influence et qui vont se sur-exprimer tôt dans le traitement, pour ensuite l'expression relative re-diminue. Lors du traitement 2, les

gènes de ce cluster ne vont pas/peu s'exprimer et lors du traitement 3, ils vont s'exprimer de la même façon que lors du traitement 1, ce qui est logique sachant que T3 est le mélange de T2 et T1.

Comme lors de l'analyse descriptive, on observe les deux autres clusters se sous-exprimer légèrement lors de T1 et se sur et sous exprimer pour T2 et T3, tout en conservant leur expression dans le temps. Le cluster 1 correspond aux gènes sur-exprimés et le cluster 3 aux gènes sous-exprimés pour T2 et T3.

Clustering des gènes à partir des variables ExpT1, ExpT2 et ExpT3.

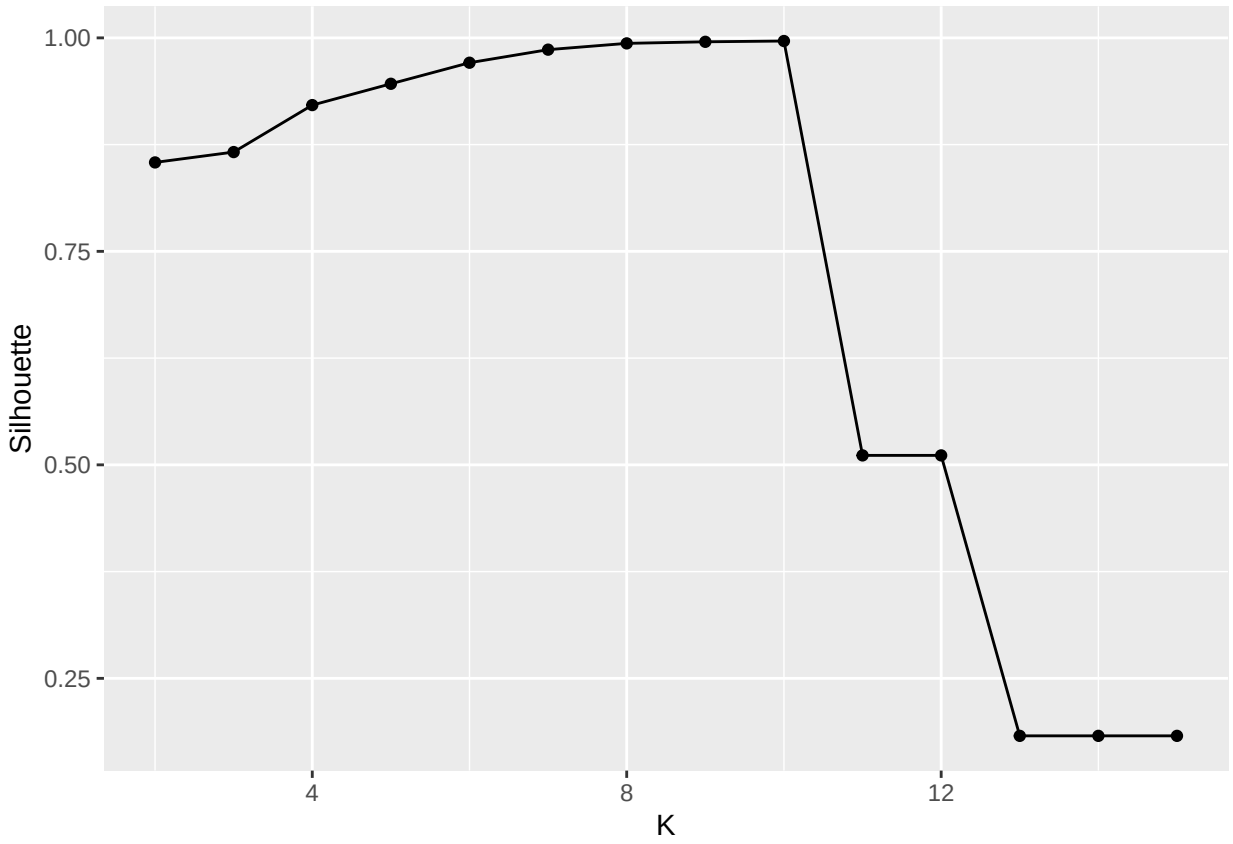
Comme ce sont des données qualitatives, on doit utiliser des méthodes alternatives de clustering comme les kmodes ou les kmeans mais sur les coordonnées de points projetés dans le plan d'une ACM ## generation de la matrice de données qualitatives



K-modes
K-means sur ACM

PAM

Pour effectuer un clustering PAM avec des données qualitatives, nous avons généré en amont une matrice de dissimilarité grâce à la métrique de gower, pour ensuite lancer l'algorithme PAM sur cette matrice.



TODO :

- refaire sans centrer-réduire
- interpréter par rapport aux méta-variable
- ne pas afficher le clustering dans les plans de l'acp pour les 2 dernières ACP
- présenter chaque méthode/algo et pourquoi on l'utilise avant le code
- analyser la seconde et troisième acp et les clustering qu'on en fait
- refaire l'analyse de la première acp
- choisir des graphes
- voir si la troisième acp est bien ce qu'il faut psq wtf ??