Projet analyse de données - 2e partie

Olivier Berthier

2024-08-30

La deuxième partie du projet concerne les données génétiques de 50 patientes (fictives) atteintes d'un cancer de l'ovaire.
Les variables correspondent au nombre de copies d'ADN à une position donnée sur le chromosome 7.

Importons les données:

```
load("C:/Users/olivi/Downloads/data_cnv.Rdata")
d  <- Y
dim(d)
```

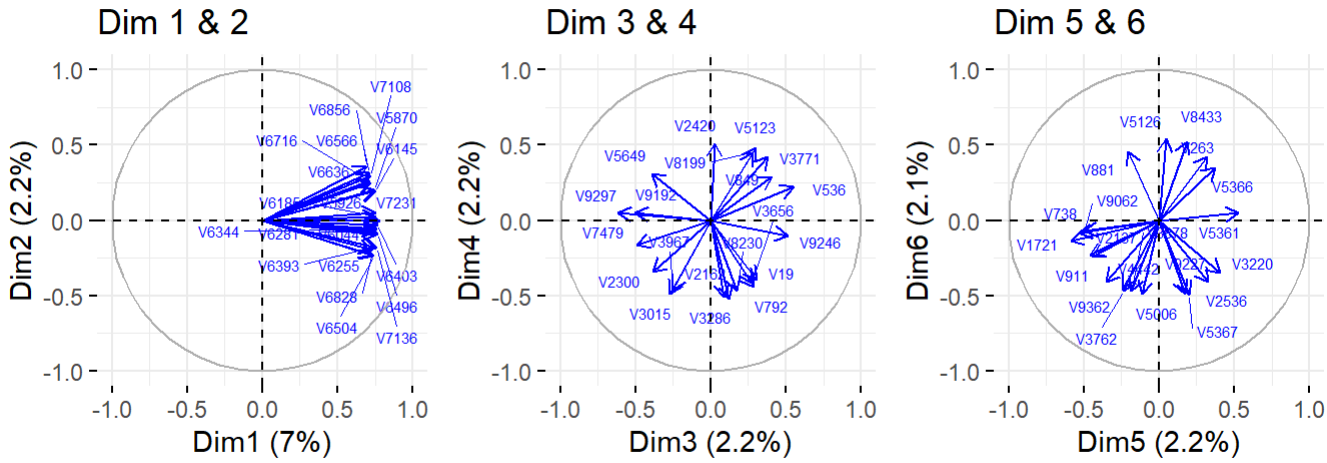
```
## [1]    50 10000
```

Le jeu de données est composé de 50 observations et 10000 variables.
Par rapport au cas précédent (sur la qualité des vins), nous nous trouvons ici dans le cas où c'est le nombre de variables qui est "grand" par rapport au nombre d'observations.

ACP

Appliquons une ACP et affichons le cercle des corrélations des premières dimensions:

```
acp_adn <- PCA(d, scale.unit = TRUE, ncp = 50, graph=FALSE)

acp_var_5_comp1 <- fviz_pca_var(acp_adn,
             select.var = list(contrib = 20),
             repel = TRUE,
             labelsize = 2,
             col.var = "blue",
             title = "Dim 1 & 2")


acp_var_5_comp2 <- fviz_pca_var(acp_adn,
             axes = c(3,4),
             select.var = list(contrib = 20),
             repel = TRUE,
             labelsize = 2,
             col.var = "blue",
             title = "Dim 3 & 4")

acp_var_5_comp3 <- fviz_pca_var(acp_adn,
             axes = c(5,6),
             select.var = list(contrib = 20),
             repel = TRUE,
             labelsize = 2,
             col.var = "blue",
             title = "Dim 5 & 6")

grid.arrange (acp_var_5_comp1, acp_var_5_comp2, acp_var_5_comp3, ncol = 3, nrow = 1)
```



C’est uniquement selon la composante 1 que les variables les plus contributives se regroupent de manière remarquable.

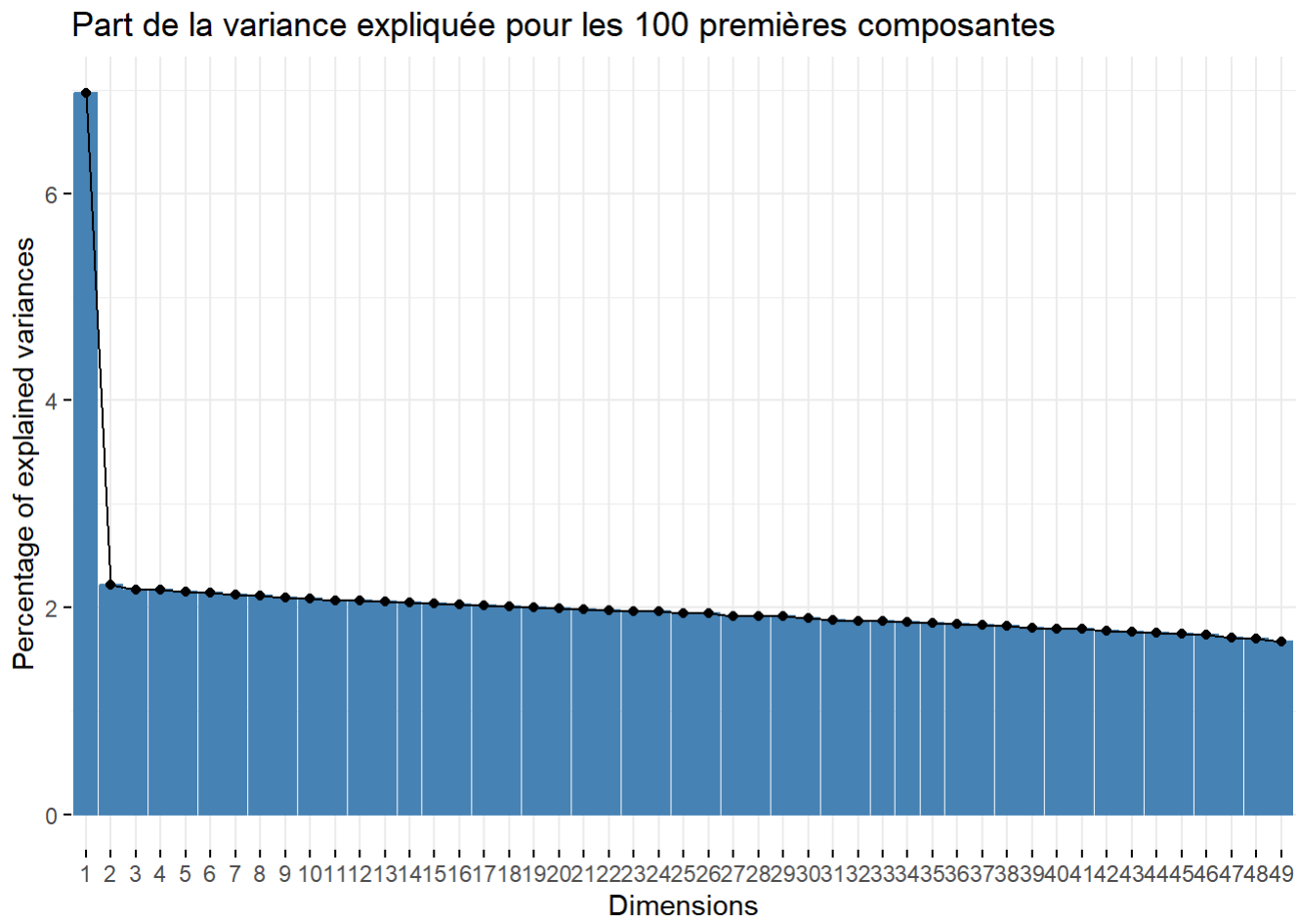
```
acp_adn$eig[1:10, "percentage of variance"]
```

```
##   comp 1   comp 2   comp 3   comp 4   comp 5   comp 6   comp 7   comp 8
## 6.974415 2.217789 2.172185 2.164921 2.150747 2.145109 2.125076 2.116671
##   comp 9  comp 10
## 2.091919 2.084083
```

La composante 1 explique environ 7% de la variance totale, elle se détache largement des autres qui expliquent chacune environ 2 % de la variance.

Affichons la part de la variance expliquée par les 100 premières composantes:

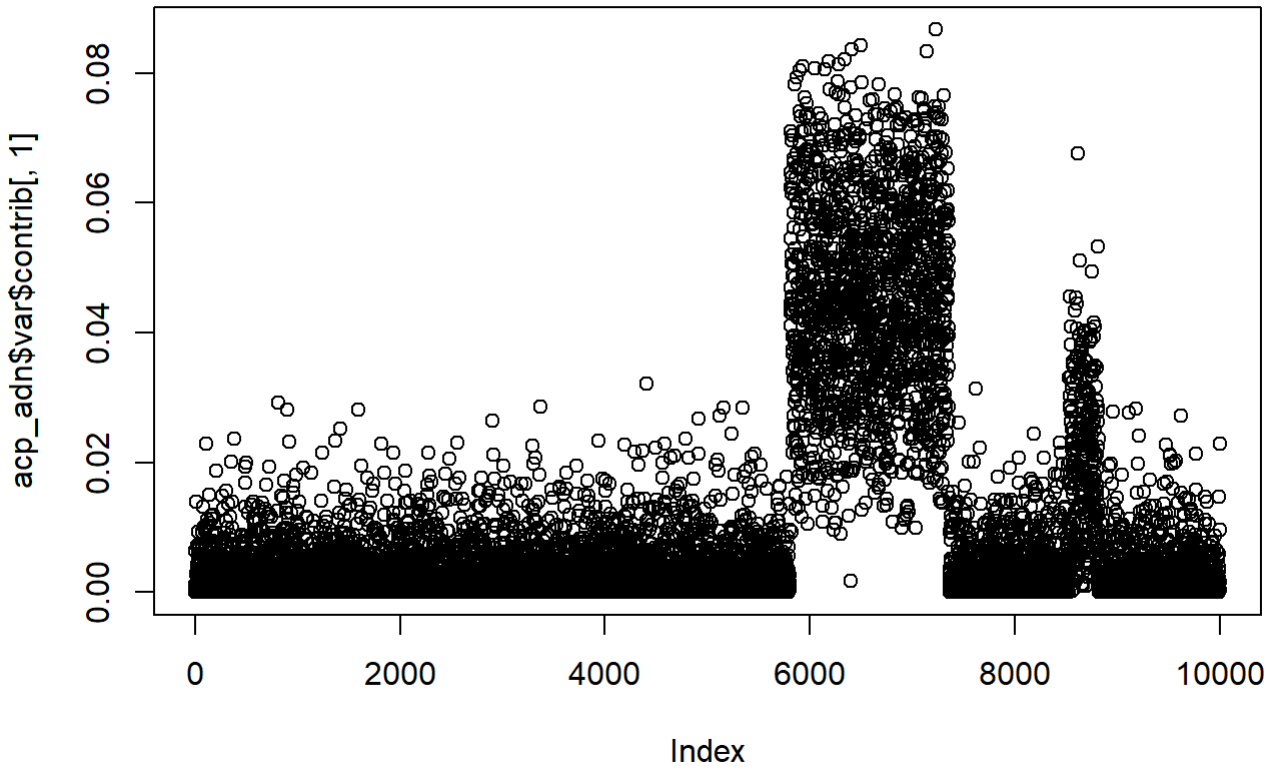
```
fviz_eig(acp_adn, ncp=100, title = "Part de la variance expliquée pour les 100 premières comp
osantes")
```



En dehors de la première composante, la variance expliquée diminue progressivement plus on avance dans les dimensions.

Contribution des variables à la première composante

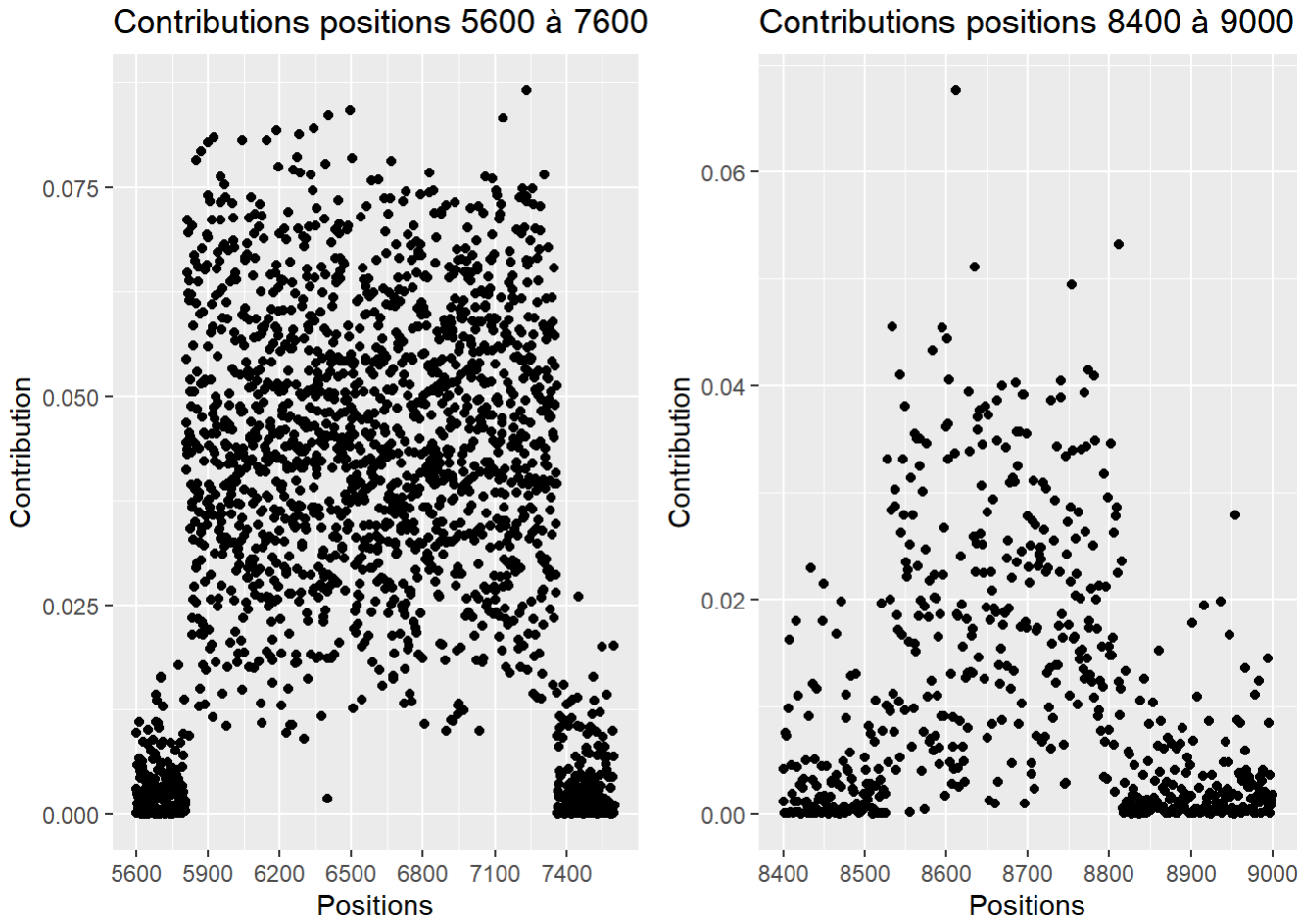
```
plot(acp_adn$var$contrib[,1])
```



Deux groupes se détachent, ce qui indique que les individus contribuant le plus à la variance expliquée par la première dimension se suivent dans le chromosome.
D'un point de vue génétique, cela signifie que ces régions du chromosome 7 subissent des modifications du nombre de copies d'ADN plus importantes que les autres régions.

Zoomons sur ces 2 régions:

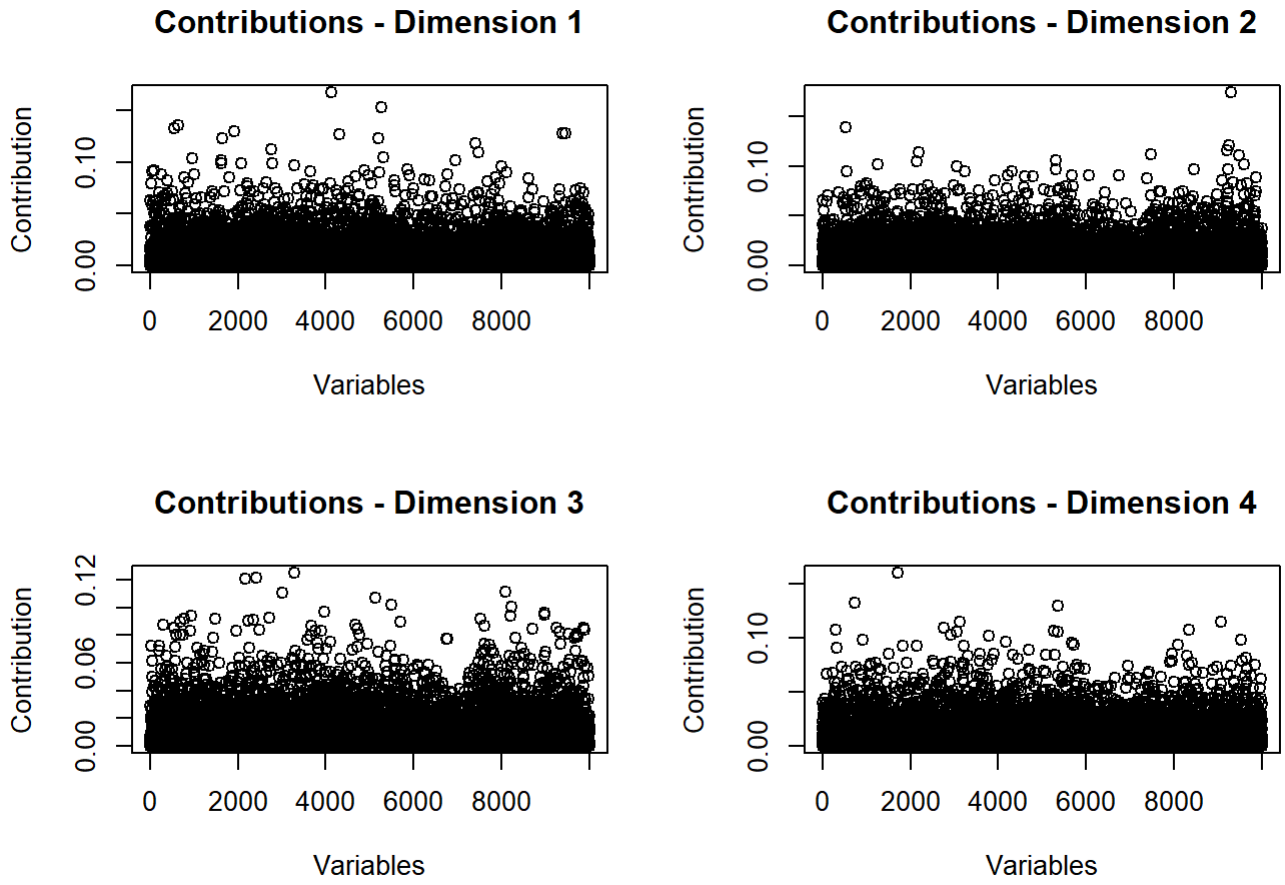
```
subset_contrib_1 <- acp_adn$var$contrib[5600:7600, 1]
df1 <- data.frame(Indices = 5600:7600, Contribution = subset_contrib_1)

plot_1 <- ggplot(df1, aes(x = Indices, y = Contribution)) +
  geom_point() +
  scale_x_continuous(breaks = seq(5600, 7600, by = 300)) +
  labs(title = "Contributions positions 5600 à 7600",
       x = "Positions", y = "Contribution")

# Retient les variables avec des indices entre 5600 et 7600
subset_contrib_2 <- acp_adn$var$contrib[8400:9000, 1]
df2 <- data.frame(Indices = 8400:9000, Contribution = subset_contrib_2)

plot_2 <- ggplot(df2, aes(x = Indices, y = Contribution)) +
  geom_point() +
  scale_x_continuous(breaks = seq(8400, 9000, by = 100)) +
  labs(title = "Contributions positions 8400 à 9000",
       x = "Positions", y = "Contribution")

grid.arrange(plot_1, plot_2, ncol = 2)
```



Attention, les échelles des deux graphiques sont différentes.
Le groupe des positions entre approximativement 5800 et 7300 se détache plus nettement que celui entre approximativement 8500 et 8800, aussi bien en terme d'amplitude que de nombre d'individus.

Affichons les contributions des variables sur les autres dimensions de l'ACP:

```
par(mfrow = c(2, 2))
# Créer les 4 graphiques avec des titres personnalisés
plot(acp_adn$var$contrib[,2], main = "Contributions - Dimension 1", xlab = "Variables", ylab
= "Contribution")
plot(acp_adn$var$contrib[,3], main = "Contributions - Dimension 2", xlab = "Variables", ylab
= "Contribution")
plot(acp_adn$var$contrib[,4], main = "Contributions - Dimension 3", xlab = "Variables", ylab
= "Contribution")
plot(acp_adn$var$contrib[,5], main = "Contributions - Dimension 4", xlab = "Variables", ylab
= "Contribution")
```

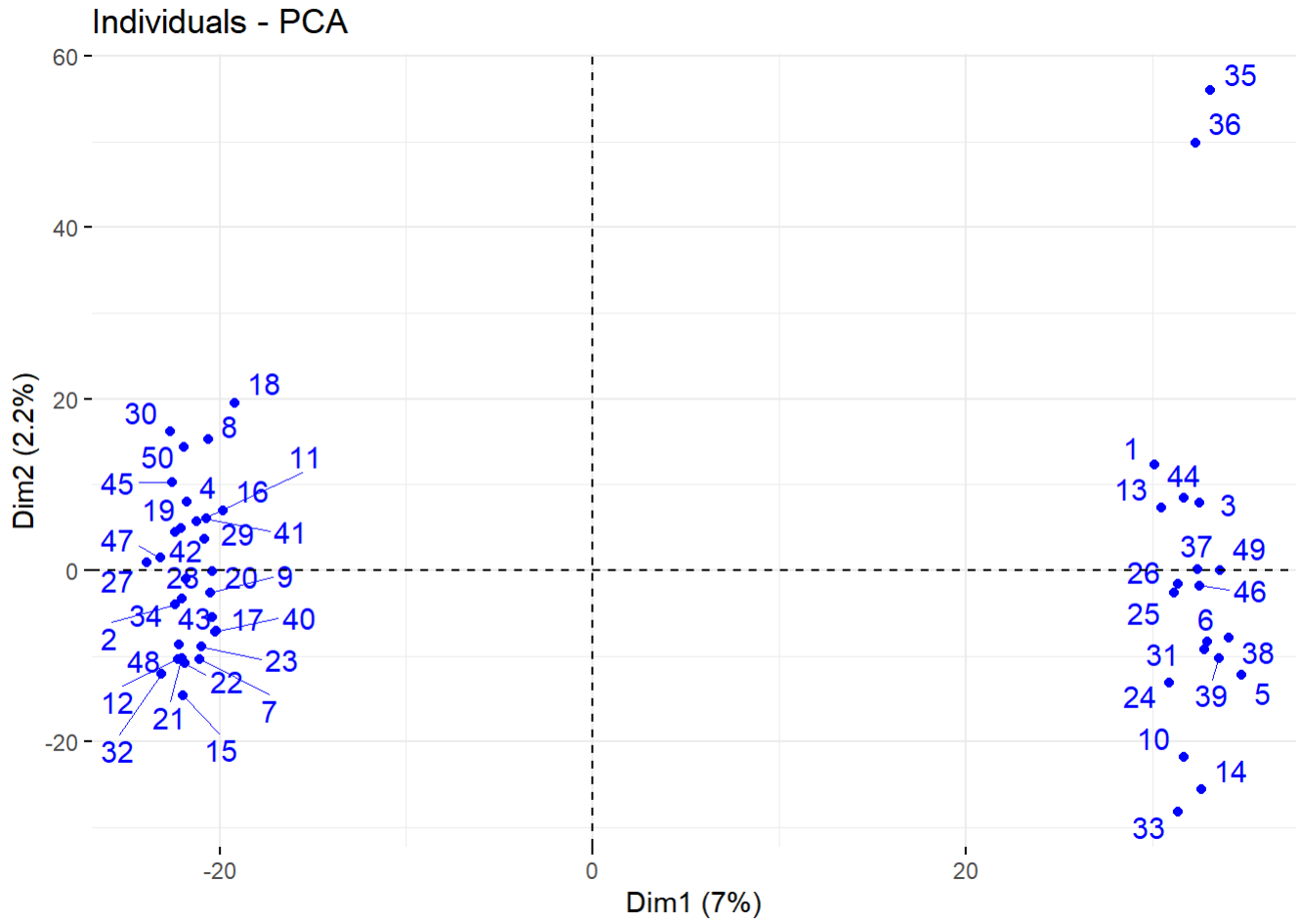



```
par(mfrow = c(1, 1))
```

Aucun groupe marqué n’apparaît nettement sur les autres dimensions.

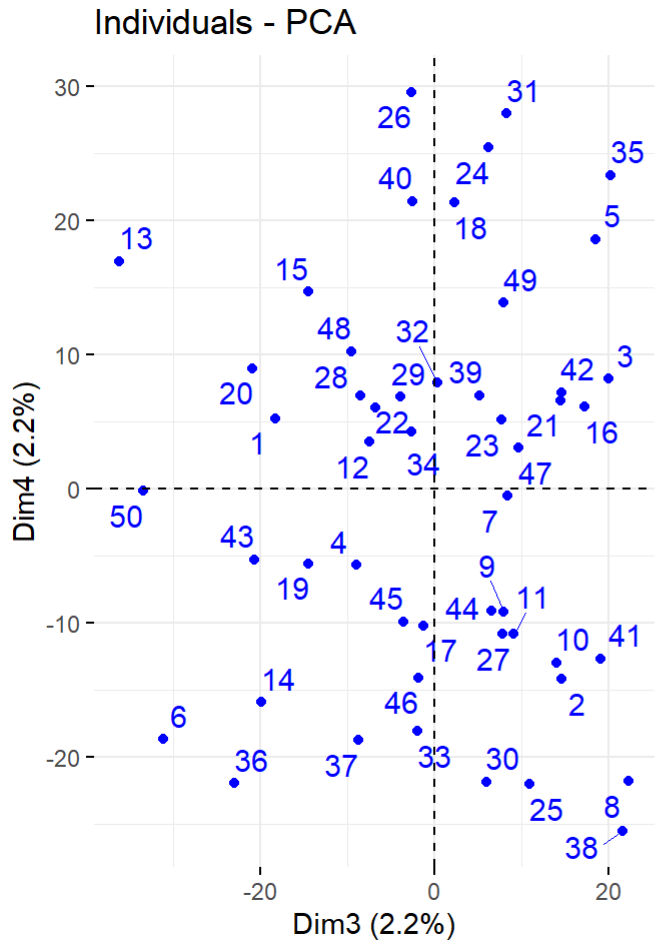
Projection des individus sur les deux premières composantes principales

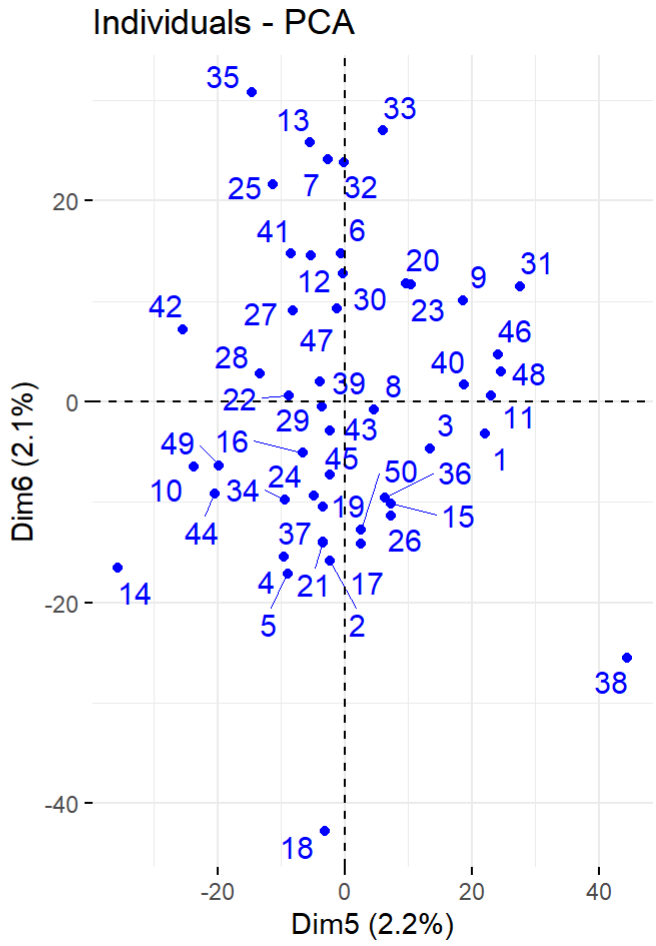
```
fviz_pca_ind(acp_adn, repel = TRUE, label = "all", col.ind = "blue")
```



Deux groupes bien distincts apparaissent selon la première dimension, celle qui nous intéresse le plus.
Ces individus ont donc des copies d'ADN plus importantes que les autres sur les positions correspondantes (les variables contribuant fortement à la composante 1).
La deuxième dimension sépare surtout deux *outliers* des autres individus.

```
plot_ind_34 <- fviz_pca_ind(acp_adn, repel = TRUE, axes = c(3,4), label = "all", col.ind = "b
lue" )
plot_ind_56 <- fviz_pca_ind(acp_adn, repel = TRUE, axes = c(5,6), label = "all", col.ind = "b
lue")
grid.arrange(plot_ind_34, plot_ind_56, ncol = 2)
```



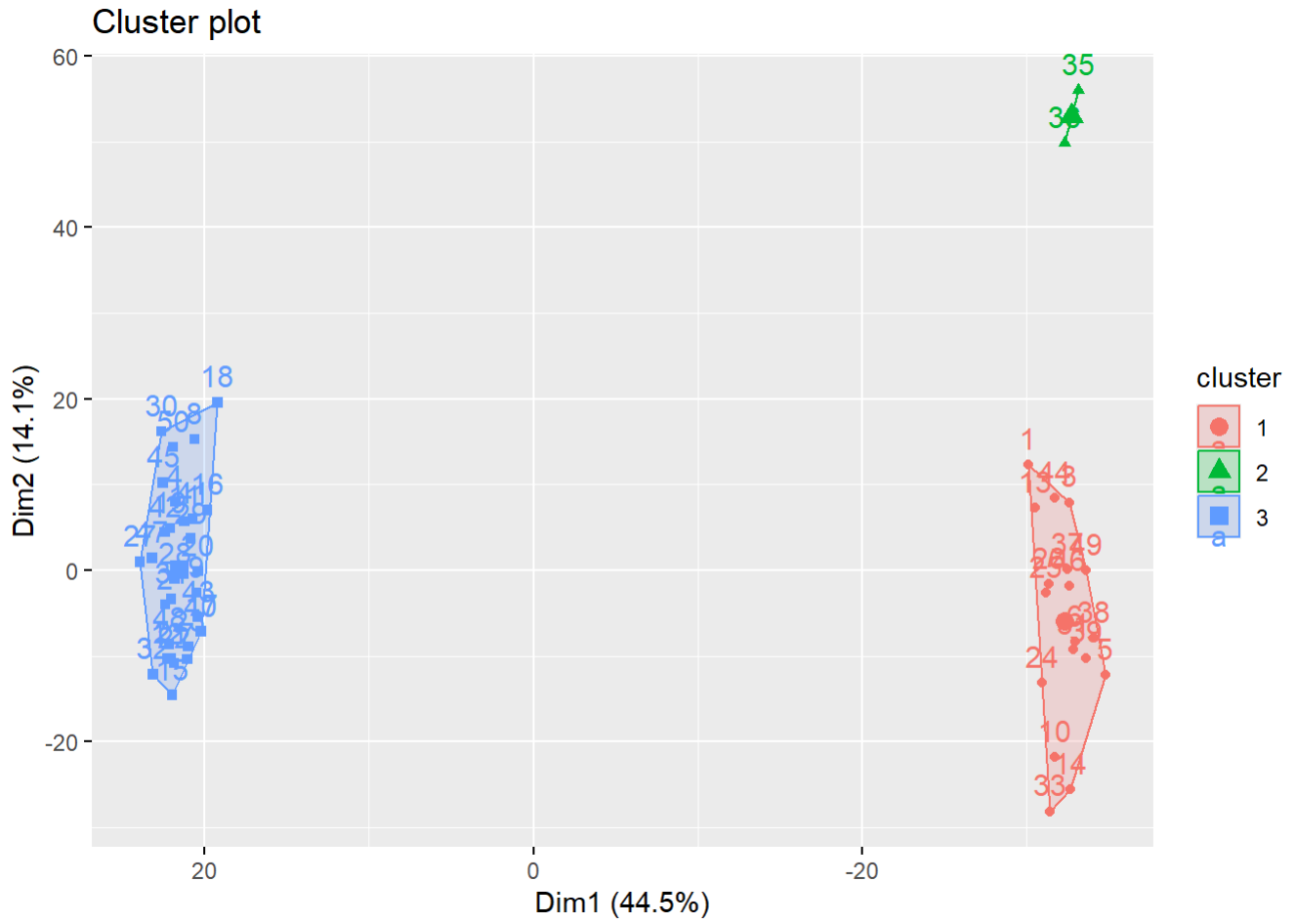


Aucun groupe n'apparaît réellement pour les dimensions 3,4 et 5. Seuls quelques "oultliers" se détachent.

Clustering sur les individus en utilisant les 5 premières composantes principales

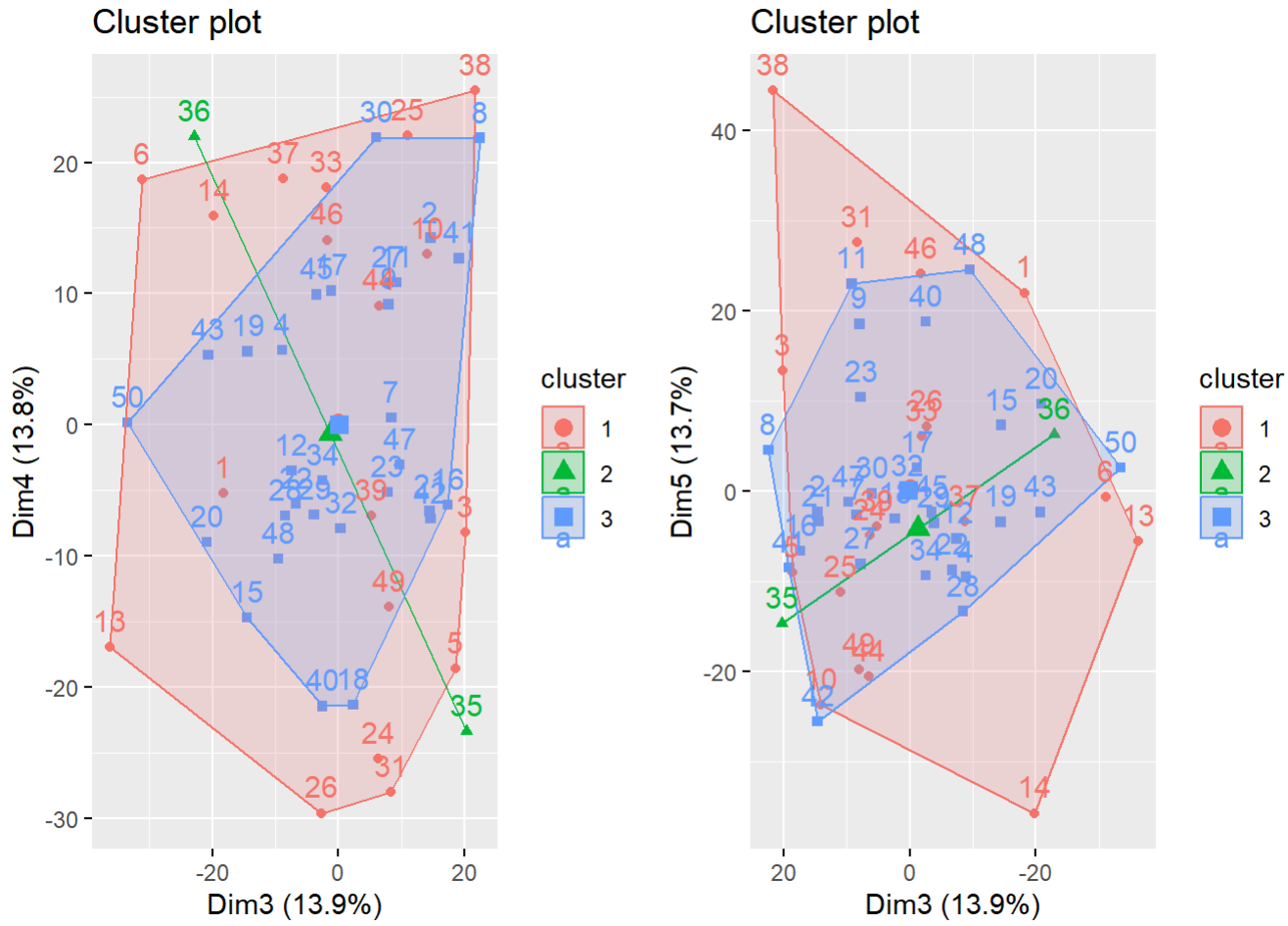
k-means

```
projection_adn <- acp_adn$ind$coord[, 1:5]

set.seed(106)  # 106 coupe bien en trois groupes....
kmeans_result <- kmeans(projection_adn, centers = 3, iter.max= 1000)  # 3 clusters

fviz_cluster(kmeans_result, data = projection_adn, stand = FALSE) +
        scale_x_reverse()  # Inverse l'axe des X, la fonction inverse naturellement le sens d
e la dim 1
```



Plusieurs essais ont été nécessaires concernant le choix de la graine de départ afin d'aboutir à un *clustering* qui nous paraît pertinent sur les deux premières dimensions.

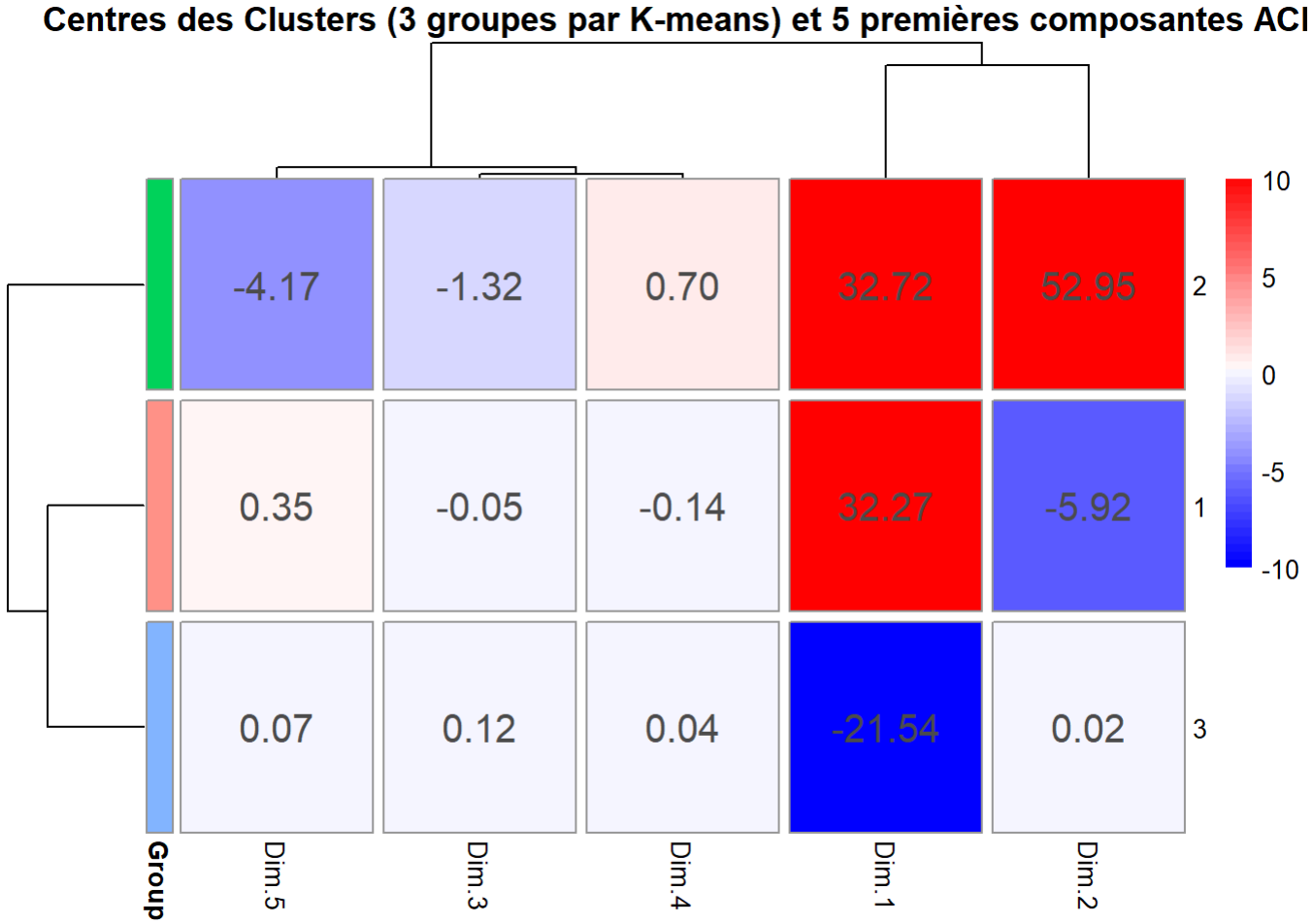
```
plot_34 <- fviz_cluster(kmeans_result, data = projection_adn, stand = FALSE, axes = c(3, 4))
plot_35 <- fviz_cluster(kmeans_result, data = projection_adn, stand = FALSE, axes = c(3, 5))
+  scale_x_reverse()
grid.arrange(plot_34, plot_35, ncol = 2)
```



Selon les dimensions 3, 4 et 5, aucun groupe n'apparaît. Les centres de classe se trouvent tous situés très proches de l'origine.
C'est assez logique car:
- Nous avons choisi la valeur de départ de la classification (la graine) pour privilégier un clustering pertinent pour les deux premières dimensions;
- Seule la dimension 1 explique de manière significative la variance totale (au moins relativement aux autres dimensions), les autres composantes expliquent très peu cette variance (la dimension 2 sépare principalement les individus 35 et 36 des autres).

Heatmap composantes / clusters

```
cluster_colors <- data.frame(Group = c("Green", "Blue", "Red"))
rownames(cluster_colors) <- c(1, 2, 3)
centers <- kmeans_result$centers

pheatmap(centers,
         cluster_cols = TRUE,
         cluster_rows = TRUE,
         show_rownames = TRUE,
         annotation_row = cluster_colors,
         cutree_cols = 5,
         cutree_rows = 3,
         display_numbers = TRUE,
         fontsize_number = 15,
         breaks = seq(-10, 10, length.out = 50),  # Limiter l'échelle des couleurs
         color = colorRampPalette(c("blue", "white", "red"))(50),
         main = "Centres des Clusters (3 groupes par K-means) et 5 premières composantes AC
P",
         annotation_legend = FALSE)
```



Dans ce heatmap, nous reprenons le code couleur des groupes du graphique des dimensions 1 et 2, afin de mieux pouvoir relier les deux représentations.

Le groupe vert (composé uniquement des individus 35 et 36) est élevé pour la dimension 1 et très élevé pour la dimension 2, il est relativement neutre dans les autres dimensions (quoique plus faible que les autres groupes dans la dimension 5).

Le groupe rouge est pareillement élevé selon la dimension 1 mais assez faible (légèrement négatif) pour la dimension 2. Il est neutre pour les autres dimensions.
(Notons ici que ce sont les centres des classes qui sont neutres, cela n'implique pas que les individus qui composent ces classes le sont. Voir pour illustration, la représentation des individus du graphique précédent selon les dimensions.)

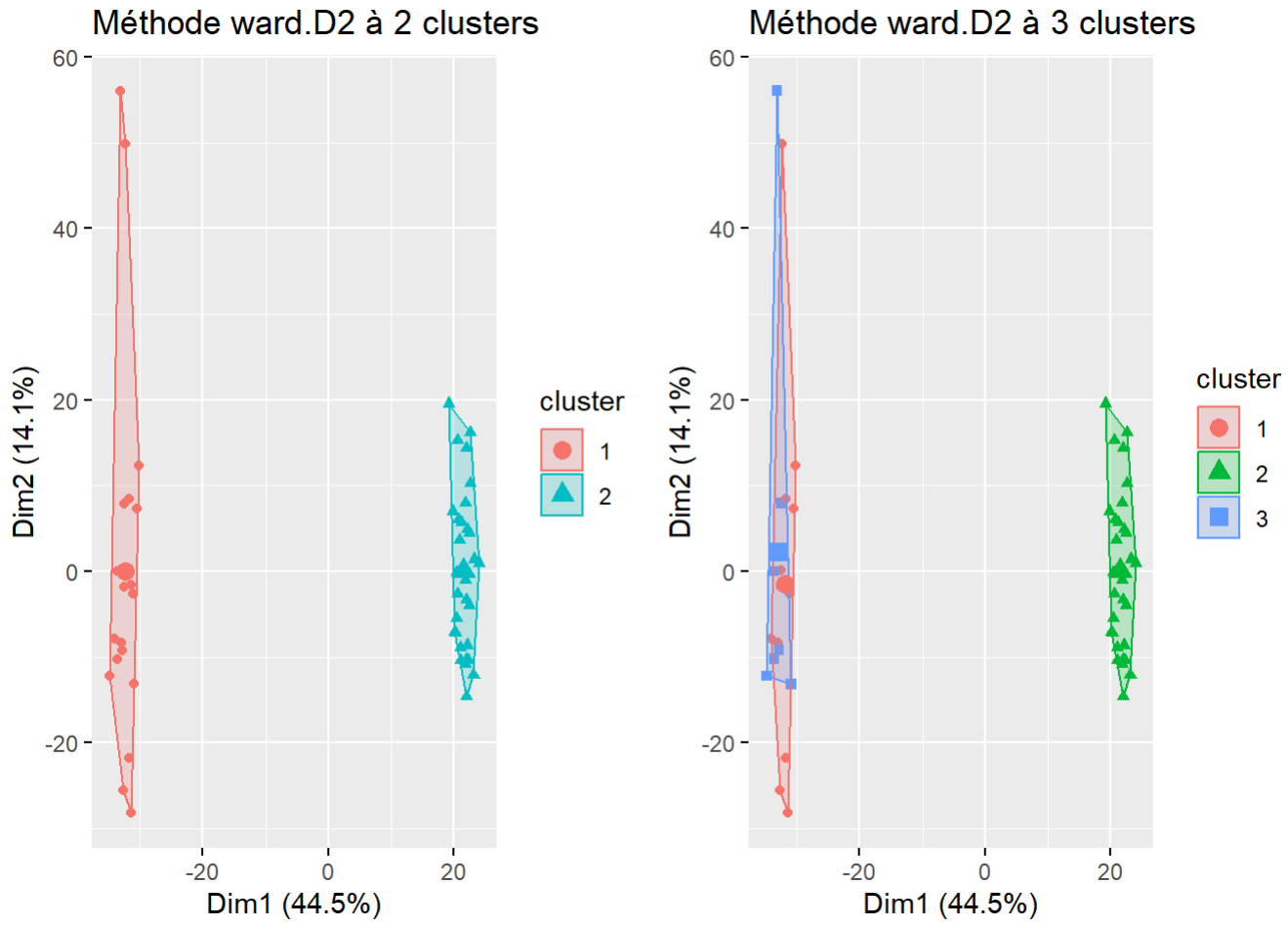
Le groupe bleu est très faible dans la dimension 1 (où il s'oppose donc aux deux autres groupes), il est neutre dans les autres dimensions.

Notons que la CAH sur les dimensions fusionne les dimensions 1 et 2 très haut, les autres dimensions sont elles sur une branche opposée, ce qui indique la distance entre les composantes.

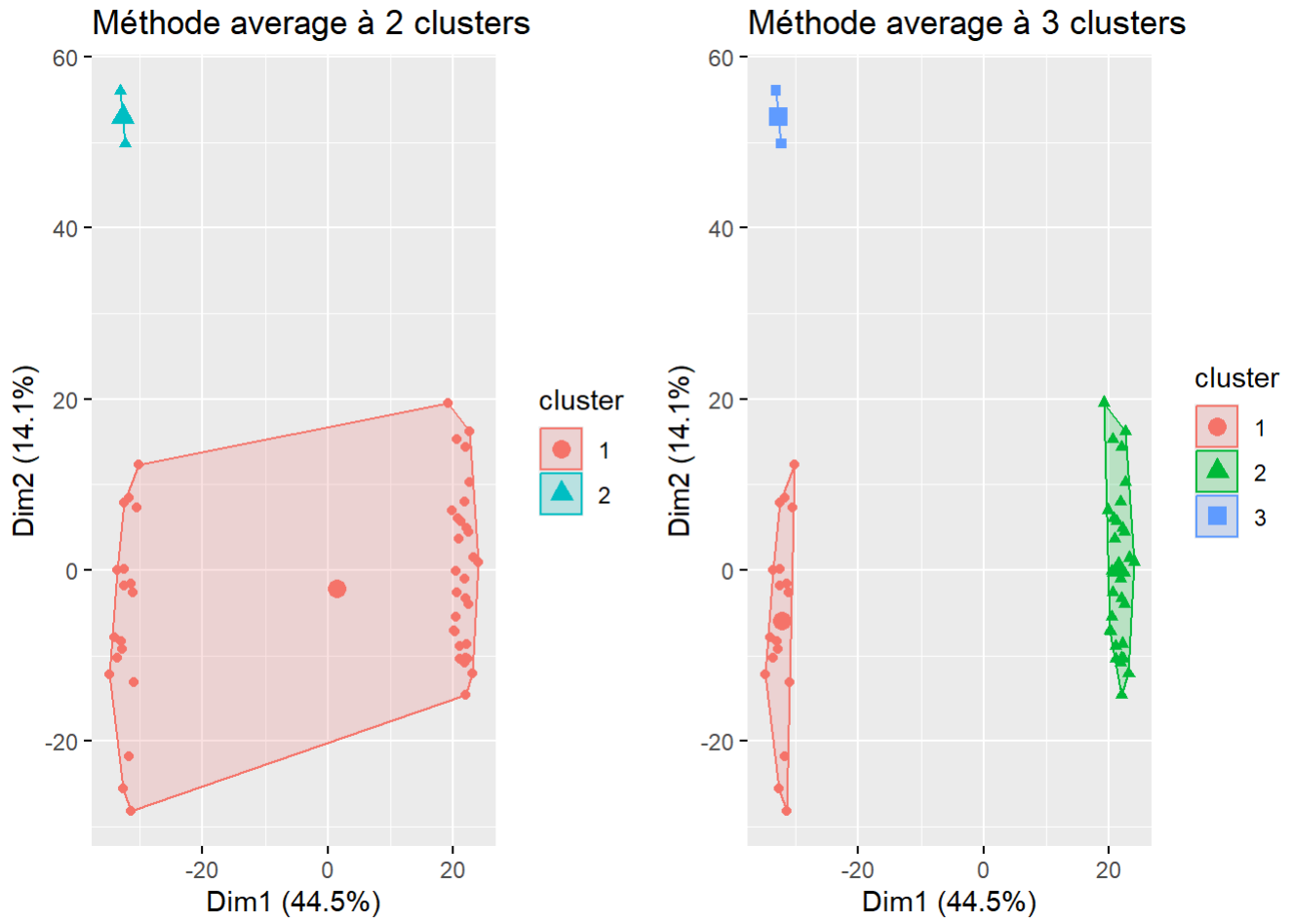
CAH

Voyons si une CAH parvient à classifier aussi efficacement les individus que la méthode des *k-means*:

```
hc_result_ward <- hclust(dist(projection_adn), method = "ward.D2")
clusters_2 <- cutree(hc_result_ward, k = 2)
clusters_3 <- cutree(hc_result_ward, k = 3)
par(mfrow = c(1, 3))
plot_ward_2 <- fviz_cluster(list(data = projection_adn, cluster = clusters_2), geom = "poin
t", stand = FALSE)+
  ggtitle("Méthode ward.D2 à 2 clusters")
plot_ward_3 <- fviz_cluster(list(data = projection_adn, cluster = clusters_3), geom = "poin
t", stand = FALSE)+
  ggtitle("Méthode ward.D2 à 3 clusters")
grid.arrange(plot_ward_2, plot_ward_3, ncol = 2)
```



```
hc_result_avg <- hclust(dist(projection_adn), method = "average")
clusters_2 <- cutree(hc_result_avg, k = 2)
clusters_3 <- cutree(hc_result_avg, k = 3)
par(mfrow = c(1, 3))
plot_avg_2 <- fviz_cluster(list(data = projection_adn, cluster = clusters_2), geom = "point",
stand = FALSE)+
  ggtitle("Méthode average à 2 clusters")
plot_avg_3 <- fviz_cluster(list(data = projection_adn, cluster = clusters_3), geom = "point",
stand = FALSE)+
  ggtitle("Méthode average à 3 clusters")
grid.arrange(plot_avg_2, plot_avg_3, ncol = 2)
```



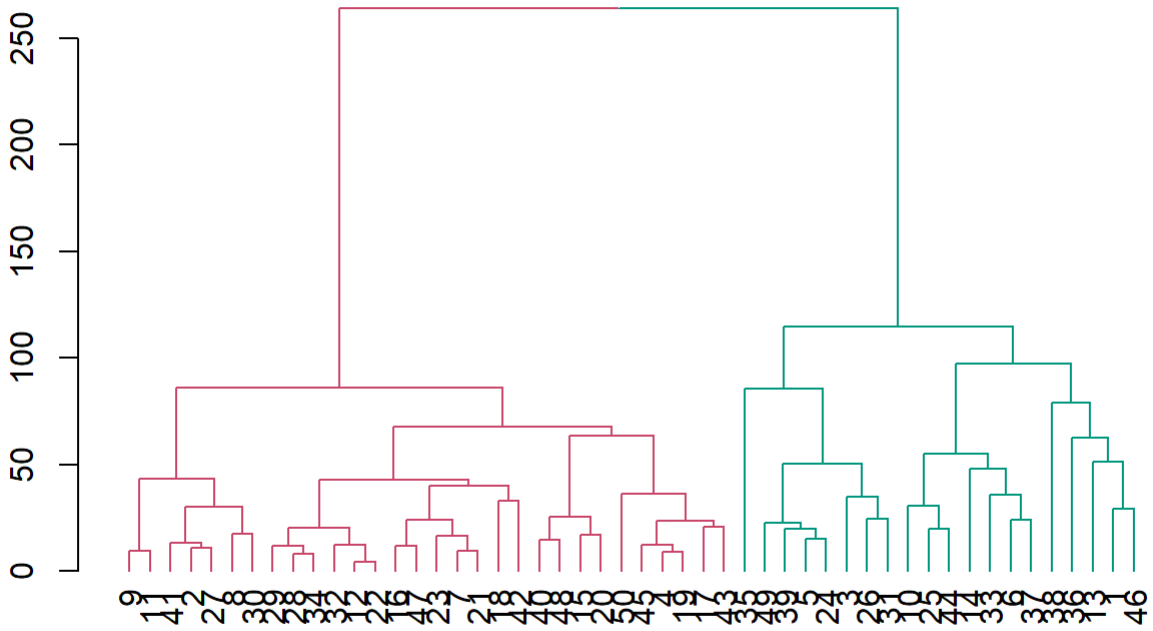
Avec la méthode *ward.2*, la CAH parvient bien à distinguer les 2 groupes selon la dimension 1 (c'est l'essentiel), mais si on applique une classification selon 3 clusters, elle ne fait pas apparaître le groupe particulier que représentent les individus 35 et 36 en dimension 2.

Avec la méthode *average*, à l'opposé, les 3 clusters correspondent à ceux obtenus par la méthode *k-means* mais avec 2 clusters, la CAH échoue à faire apparaître les 2 groupes principaux de la dimension 1, ce qui la disqualifie.

Les autres méthodes de CAH (non représentées ici) n'ont pas obtenu de meilleurs résultats.
La méthode *K-means* semble plus appropriée (au prix de plusieurs essais pour le choix de la graine).

Affichons le dendogramme de la CAH par la méthode *ward.D2*:

```
hc_result_ward %>%
  as.dendrogram() %>%
  set("branches_k_color", k = 2) %>%
  plot()
```



Le dendogramme permet de constater la grande distance entre les deux groupes les plus importants et, en dehors de cette fusion, que les individus sont assez proches.

Nous n'affichons pas ici le *heatmap* des clusters calculés par CAH, les deux groupes principaux étant les mêmes que par les *K-means*, la visualisation des relations de ces clusters avec les composantes principales des variables aboutiraient donc aux mêmes conclusions.

Conclusion

La dimension 1 semble capturer des modifications importantes du nombre de copies d'ADN pour certaines patientes atteintes du cancer des ovaires. Ces modifications se situent aux positions indiquées par le numéro des variables les plus contributives à cette dimension.
On peut formuler l'hypothèse suivante: les patientes présentant ce profil d'altérations génétique représenteraient un sous-groupe cliniquement et biologiquement distinct du reste de la population, et ces caractéristiques pourraient être reliées à la survenue du cancer.

Cependant, nos données ne nous permettent pas de valider ou invalider cette hypothèse.
Dans le cas d'une étude réelle, nous pourrions envisager une validation clinique afin de voir si les groupes d'individus identifiés par l'ACP et le clustering montrent des différences cliniques significatives (comme par exemple l'âge de déclenchement de la maladie, la vitesse d'évolution de la tumeur, etc).
On pourrait comparer également ces données avec celles d'individus sains et voir si l'on retrouve ou non les mêmes particularités aux mêmes positions sur le chromosome 7 (rien ne permet en effet de relier les éléments mis en avant ici à la prévalence du cancer des ovaires).
Nous pouvons simplement conclure qu'il y a là matière à approfondir.

D'un point de vue statistique:

- Nous avons hésité à ne pas standardiser les données, elles semblent en effet être de même unité et de même échelle.
  Nous avons essayé les deux approches et avons constaté que les résultats étaient similaires concernant l'essentiel, seuls les *outliers* différaient significativement entre les deux approches.
  Les résultats semblant plus lisibles et plus facilement interprétables avec standardisations, nous avons retenu cette méthode.
- Afin de pallier les difficultés rencontrées concernant le choix de la graine de départ, lors de la classification par *K-means*, une possibilité serait d'utiliser une contrainte de type *must-link* (qui impose que deux individus soient dans le même cluster, dans notre cas les individus 35 et 36).