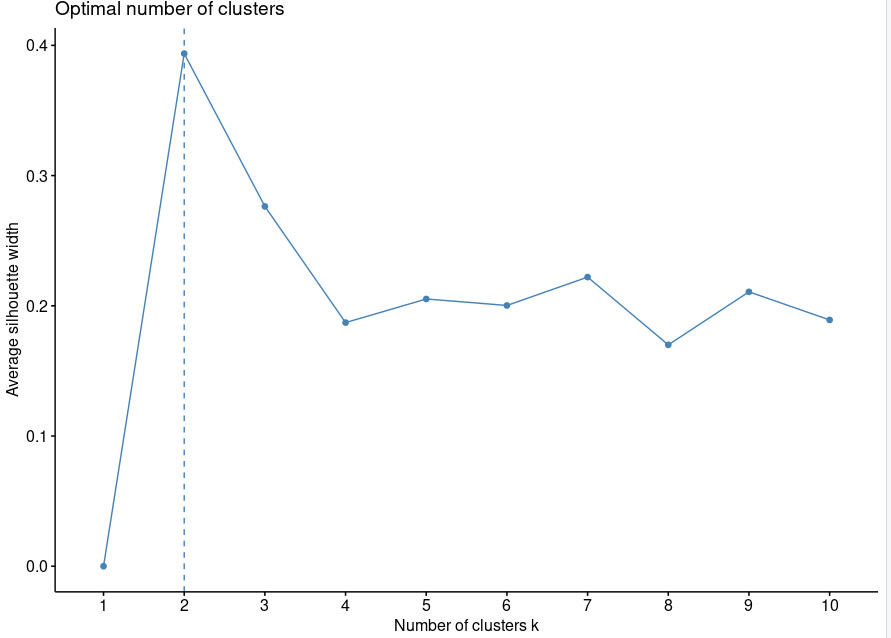
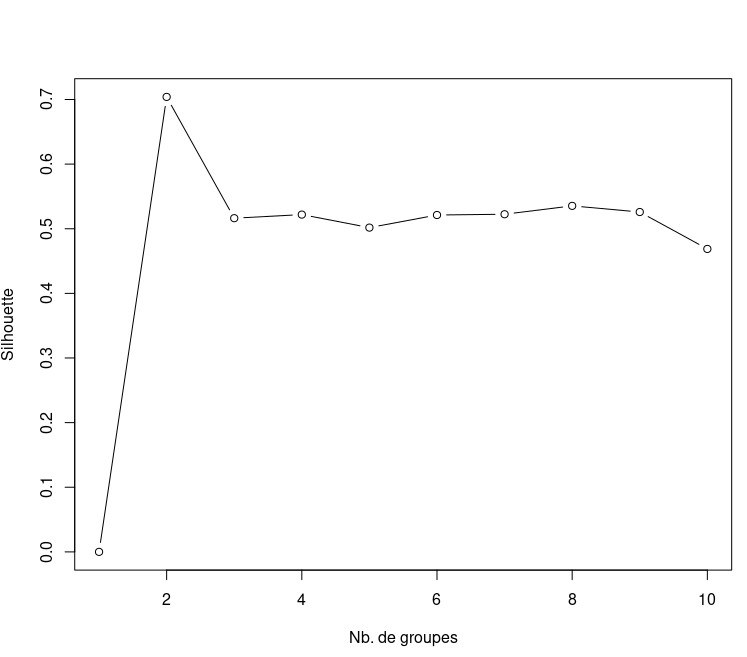
A priori, d’après notre problématique on devrait avoir 2 groupes: un pour les personnes atteintes de la maladie de Parkinson et un autre pour les personnes en bonne santé.

KMEANS

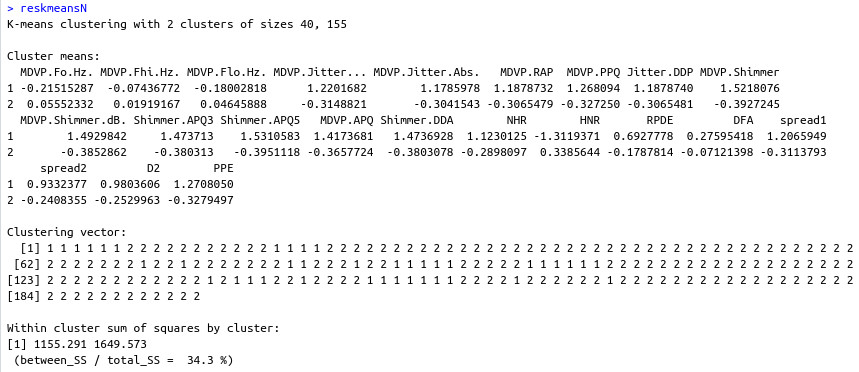
Cherchons le nombre optimal de classes pour le kmeans.

Nous utilisons l’indice «silhouette moyenne» selon deux approches, et on obtient les résultats suivants



On peut donc soupçonner que «2 est le bon nombre de profils type».

L’application du kmeans avec 2 groupes donne



On a 2 groupes dont les effectifs semblent se rapprocher de ceux du dataset de départ. Seulement, l’inertie est mauvaise; et d’après le classement proposé, les 2 groupes ne correspondent pas à ceux qu’on a dans la réalité. Par exemple, les 31 premiers individus sont malades -donc de classe 2 d’après la classification proposée par kmeans- mais il se trouve qu’on a plusieurs individus de la classe 1 entre le 1er et le 31e.

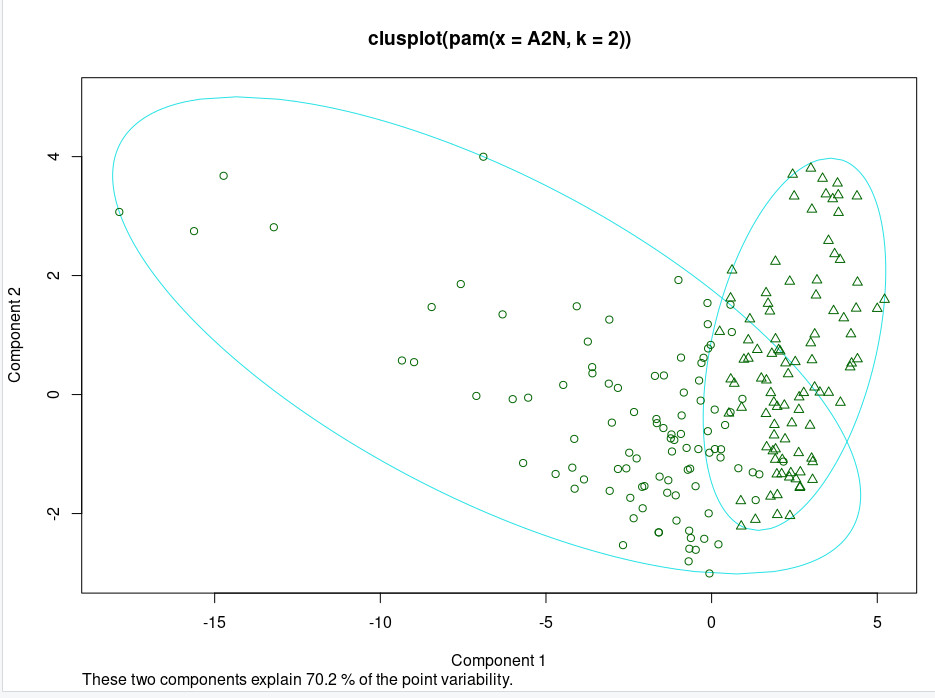
La représentation graphique nous permet de confirmer cela malgré qu’elle nous donne 2 groupes distincts et bien séparés



En conclusion on a 2 groupes séparés mais leur structure est différente de celle du jeu de données initial; la méthode kmeans ne nous donne donc pas un résultat intéressant.

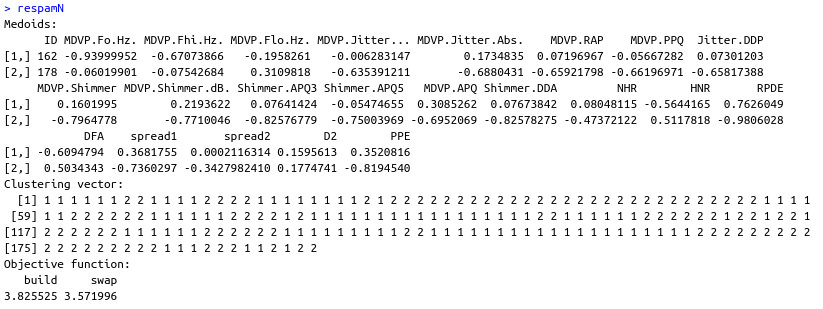
PAM.

L’application de la méthode PAM à 2 classes à notre jeu de données donne:

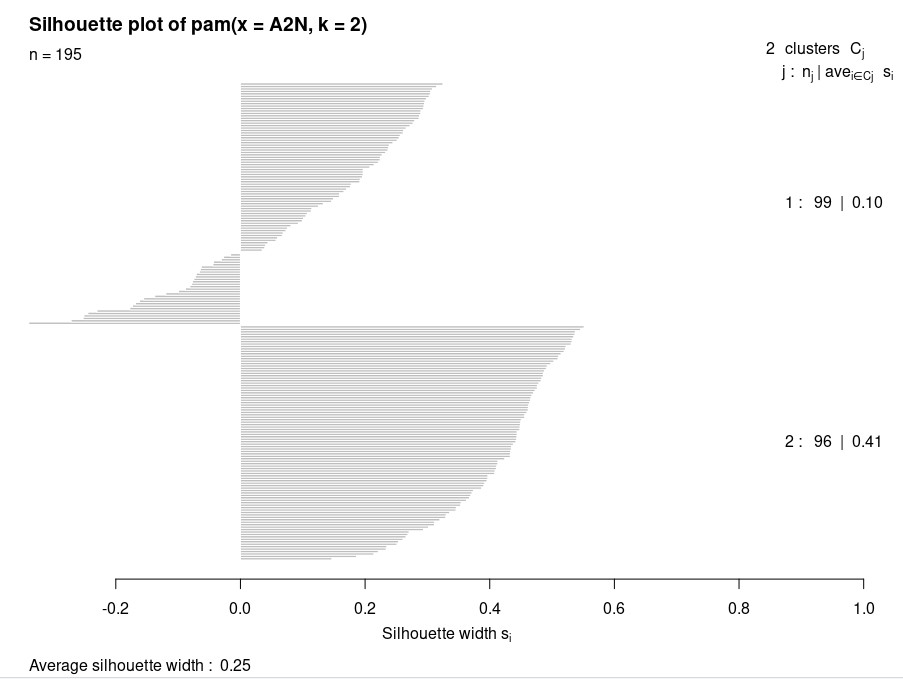


Ce graphique nous montre qu’on a clairement 2 groupes qui ne sont pas isolés malgré un fort taux d’explicabilité.

La répartition des individus dans les classes vient illustrer doublement cela comme on peut le voir:



L’intuition qu’on a ici est que 2 n’est pas le bon nombre; Ceci est confirmé par la visualisation des valeurs silhouettes sur la figure suivante.



En effet nous avons:

- des valeurs sihouettes très faibles

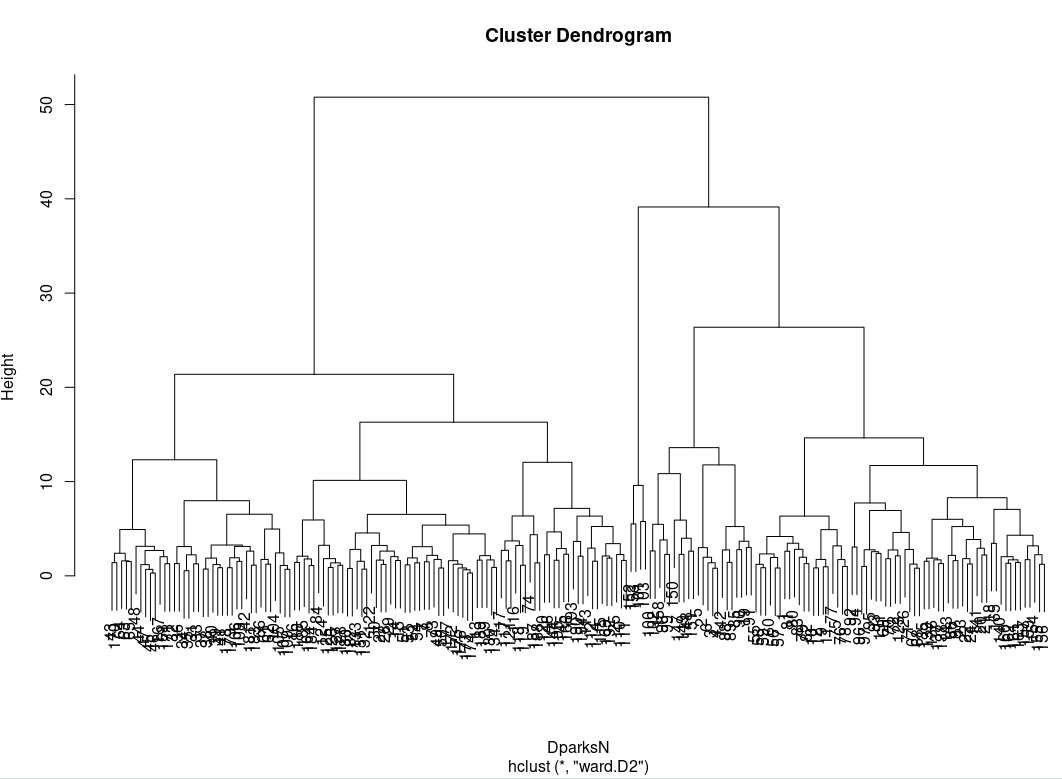
- des points (de la classe 1) avec des valeurs silhouettes négatives (donc mal classés)

- les effectifs des 2 classes sont sensiblement les mêmes, ils ne reflettent en aucune manière les proportions du jeu de données initial.

Au vu de ce qui précède, on peut conclure que la méthode PAM ne permet pas d’obtenir un bon clustering de notre jeu de données.

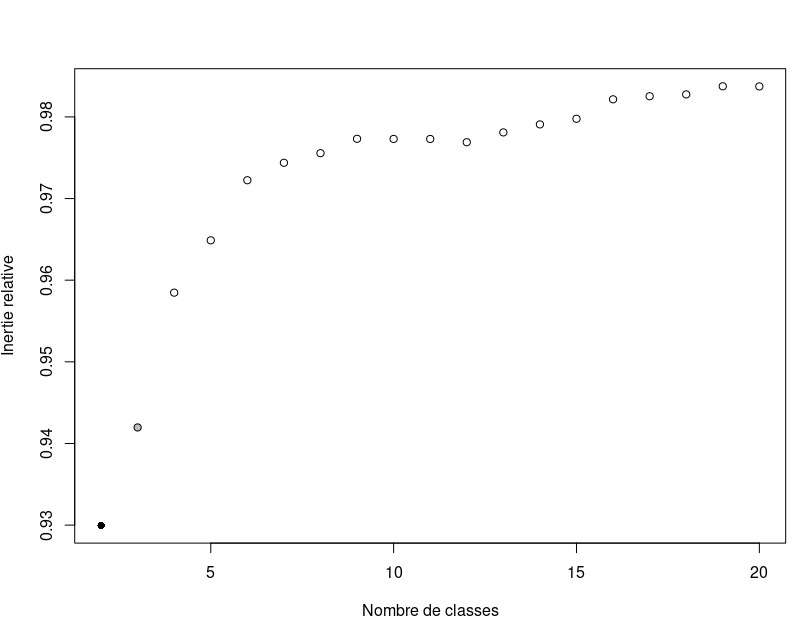
CAH

Le calcul du dendogramme donne la figure suivante

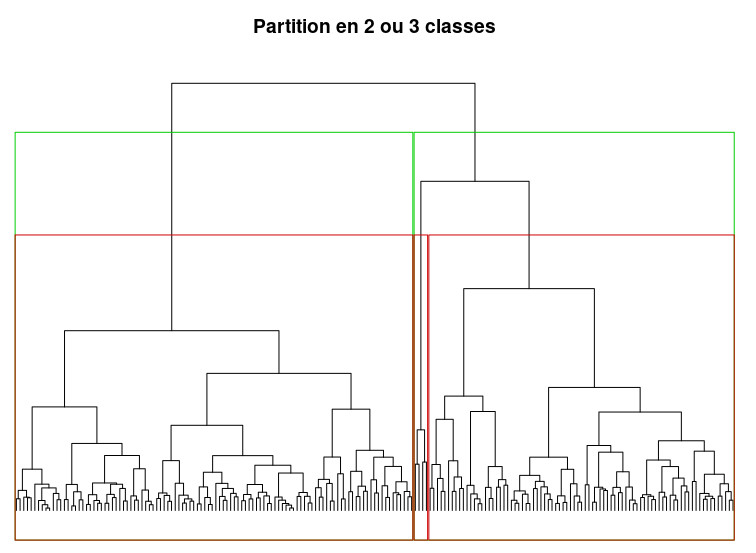


A la vue de cette figure on peut envisager 2 classes. Pour le confirmer, regardons le graphique des pertes relatives d’inertie.

La meilleure partition selon ce critère est représentée par un point noir et la seconde par un point gris.



Un découpage en deux classes minimise ce critère.



Au vu de cette figure, une analyse à 3 classes ne semble pas pertinente car il n’y aurait pas un supplément conséquent d’informations

Nous allons donc retenir un découpage en deux classes.