Analyse de regroupements

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

Algorithmes pour l'analyse de regroupements

L'analyse de regroupements cherche à créer une division de n observations de p variables en regroupements.

1. méthodes basées sur les centroïdes et les médoïdes (k-moyennes, k-médoides)
2. mélanges de modèles
3. méthodes basées sur la connectivité (regroupements hiérarchiques agglomératifs et divisifs)
4. méthodes basées sur la densité

Algorithme de partition autour des médoïdes (PAM)

1. Initialisation: sélectionner K des n observations comme médoïdes initiaux.
2. Assigner chaque observation au médoïde le plus près.
3. Calculer la dissimilarité totale entre chaque médoïde et les observations de son groupe.
4. Pour chaque médoïde $(k = 1, \dots, K)$:
   - considérer tous les $n - K$ observations à tour de rôle et permuter le médoïde avec l'observation.
   - calculer la distance totale et sélectionner l'observation qui diminue le plus la distance totale.
5. Répéter les étapes 2 à 4 jusqu'à ce que les médoïdes ne changent plus.

Algorithme CLARA (1/2)

L'algorithme CLARA, décrit dans Kaufman & Rousseeuw (1990), réduit le coût de calcul et de stockage.

On répète S fois les instructions suivantes:

- Tirer un sous-échantillon aléatoire de taille n_S
  - typiquement $K \ll n_S < 1000$
- Utiliser l'algorithme PAM sur ce sous-échantillon pour obtenir les médoïdes.
- Assigner le reste des observations de l'échantillon au regroupement du médoïde le plus près.

Pour chacune des S segmentations, on calcule la distance moyenne entre les médoïdes et les observations.

La meilleure segmentation est retournée: c'est celle qui a la plus petite distance moyenne parmi les S.

PAM et CLARA dans R

Disponible depuis le paquet `cluster`.

```r
set.seed(60602)
kmedoide5 <- cluster::clara(
    x = donsmult_std,
    k = 5L, # nombre de groupes
    sampsize = 500, #taille échantillon pour PAM
    metric = "euclidean", # distance l2
    #cluster.only = TRUE, # ne conserver que étiquettes
    rngR = TRUE, # germe aléatoire depuis R
    pamLike = TRUE, # même algorithme que PAM
    samples = 10) #nombre de répétitions aléatoires
```

Valeurs initiales et paramètres

Même hyperparamètres que K-moyennes (dissemblance, nombre de regroupements, initialisation et séparation).

Comme les K-moyennes, on fera plusieurs essais pour trouver de bonnes valeurs de départ. On peut tracer le profil des silhouettes.

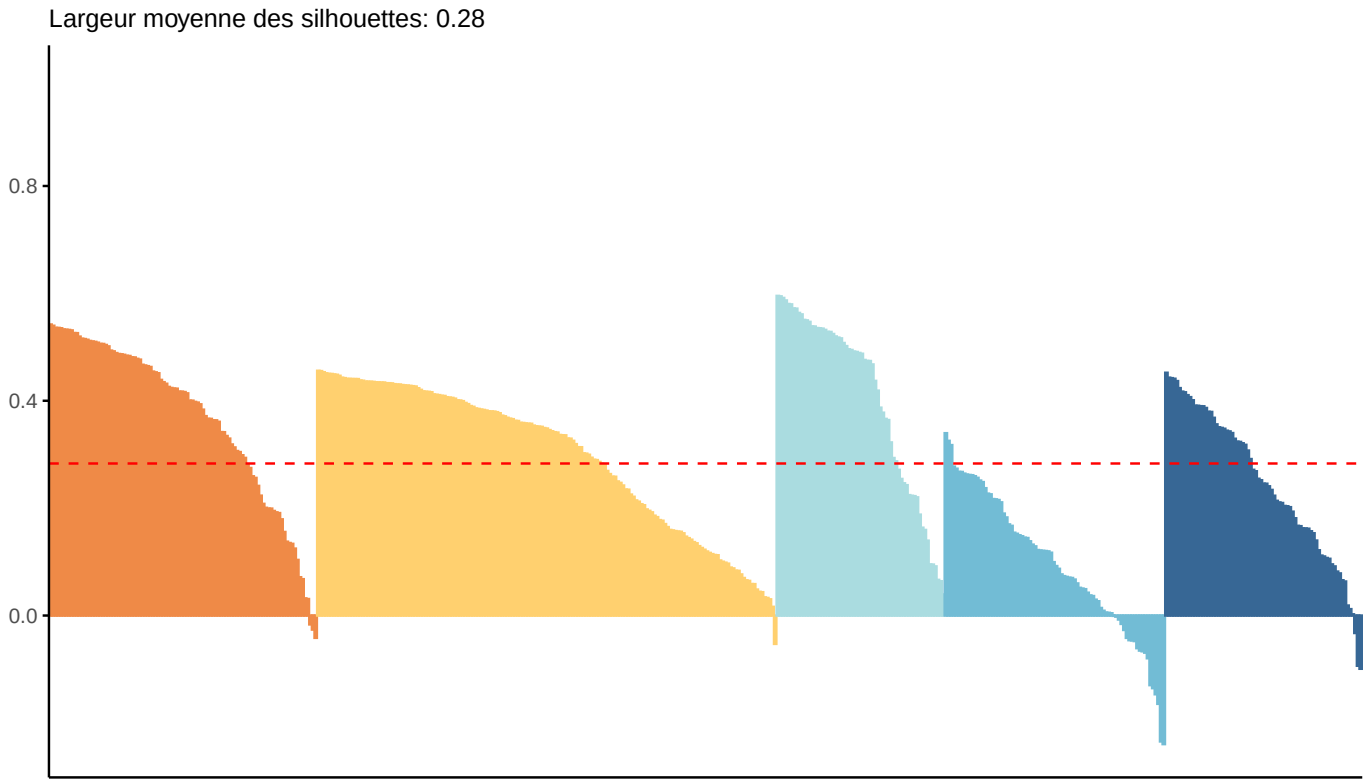


Figure 1: Silhouettes pour les données de dons multiples avec

Prototypes

Puisque les prototypes (médoïdes) sont des observations, on peut simplement extraire leur identifiant.

```
medoides_orig <- donsmult[kmedoide5$i.med,]
# Taille des regroupements
kmedoide5$clusinfo
```

Avantages et inconvénients des K-médoïdes

- $(+)$ les prototypes sont des observations de l'échantillon.
- $(+)$ la fonction objective est moins impactée par les extrêmes.
- $(-)$ le coût de calcul est prohibitif avec des mégadonnées (problème combinatoire). PAM fonctionne avec maximum 1000 observations.
- $(-)$ solution approximative pour grand échantillons avec CLARA

Mélange de modèles

On suppose qu'on a K groupes, chacun caractérisé par une densité de dimension p, soit $f_k(X_i; \theta_k)$ si X_i provient du groupe $k = 1, \dots, K$.

Généralement, on choisit une loi normale multidimensionnelle pour le ke groupe G,

$$X \mid G = k \sim \mathrm{No}_p(\mu_k, \Sigma_k)$$

La probabilité qu'une observation soit tirée du groupe $G = k$ est π_k.

Estimation du mélange de modèle

La vraisemblance est une fonction des paramètres μ_k, Σ_k et de la probabilité π_k qu'une observation X_i tombe dans le groupe k,

$$L_i(\{\mu_k, \Sigma_k, \pi_k\}_{k-1}^K; x_i) = \sum_{k=1}^{K} \pi_k f_k(X_i \mid \mu_k, \Sigma_k).$$

Le maximum de vraisemblance est obtenu à l'aide de l'algorithme d'espérance-maximisation en augmentant les observations avec un indicateur de groupe.

- Étape E: assignation aux groupes (multinomiale).
- Étape M: estimation des probabilités, des moyennes et variances.

Le mélange de modèle nous donne accès à la probabilité π_k qu'une observation appartiennent au groupe G_k (assignation probabiliste).

Fléau de la dimension

Chacune des K matrice de covariance contient $p(p+1)/2$ paramètres!

En paramétrisant ces dernière, on peut réduire le nombre de paramètres à estimer.

- compromis entre simplicité (d'estimation) et nombre de paramètres

Paramétrisation des matrices de covariance

La matrice de covariance dans `mclust` est paramétrisée en fonction de

- λ, qui contrôle le volume,
- une matrice diagonale A qui contrôle les variances de chaque observation et
- D une matrice orthogonale qui permet de créer de la corrélation entre observations.

Un index k spécifie que cette composante varie d'un regroupement à l'autre.

Paramétrisation des variables

Voir `mclust.options("emModelNames")` et la documentation dans le Tableau 3 de l'article sur `mclust`.

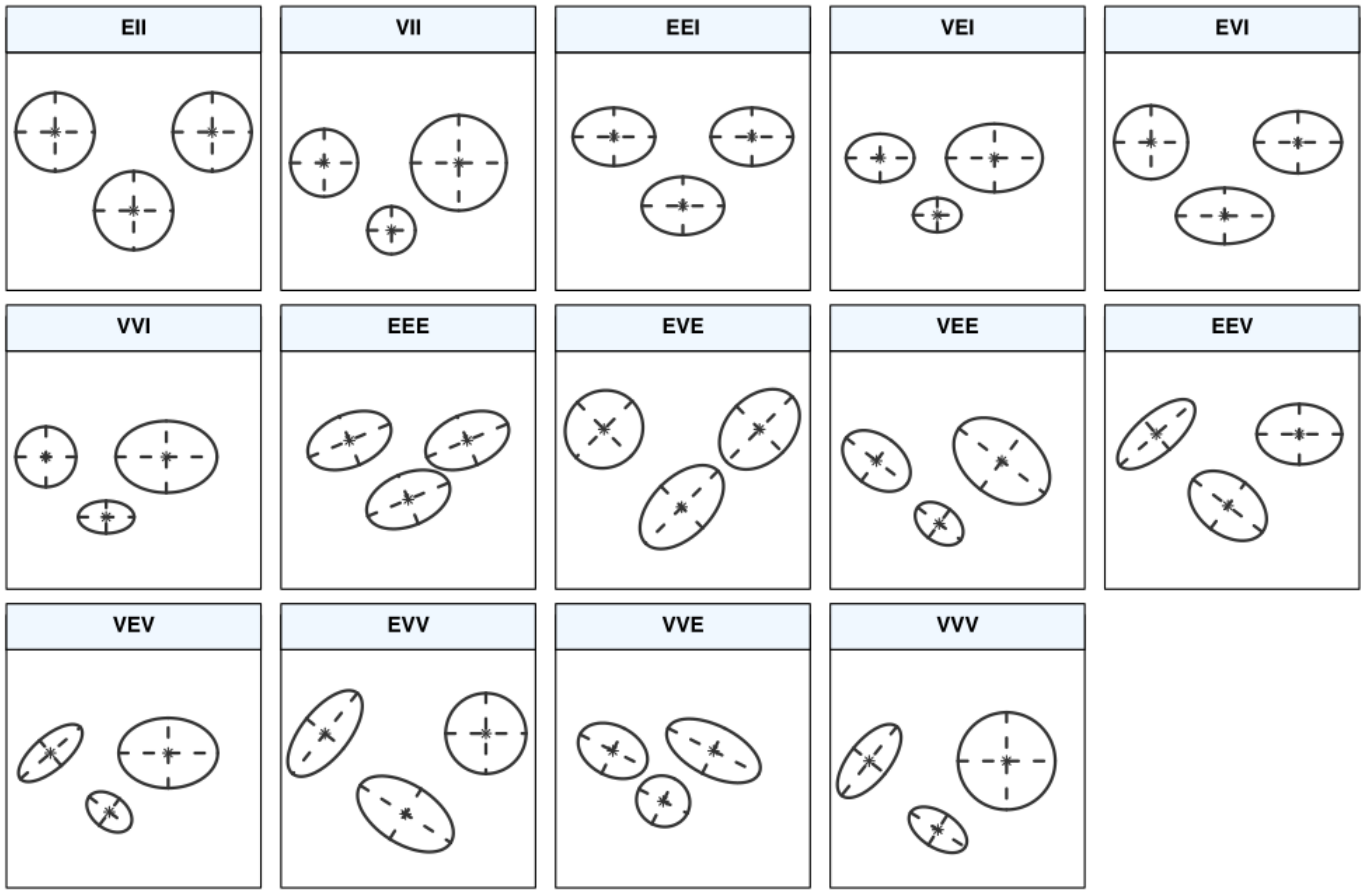


Figure 2: Forme des ellipsoïdes pour le mélange de modèle selon la forme de la structure de covariance. Tirée de `mclust5`, licence CC BY

Paquet `mclust`

```
## Mélanges de modèles gaussiens
set.seed(60602)
library(mclust)
mmg <- Mclust(data = donsmult_std,
       G = 1:10,
       # Matrice de covariance (par défaut, tous choix)
       # modelNames = mclust.options("emModelNames"),
       ## Ajouter composante uniforme
       ##  pour bruit (aberrances)
       initialization = list(noise = TRUE))
# Résumé de la segmentation
summary(mmg)
```

On peut obtenir les étiquettes (avec `0` pour le bruit) avec `mmg$classification`.

Hyperparamètres

- le nombre de regroupements K
- la forme des ellipsoïdes (structure de covariance)
- les valeurs pour l'initialisation.

Les mêmes considérations pratiques qu'avec les K-moyennes s'appliquent.

En pratique, on ajuste le modèle avec différent nombre de regroupements et différentes structures de covariance et on prend le modèle avec le plus petit BIC.

Sélection des hyperparamètres

```
1 plot(mmg, what = "BIC")
```

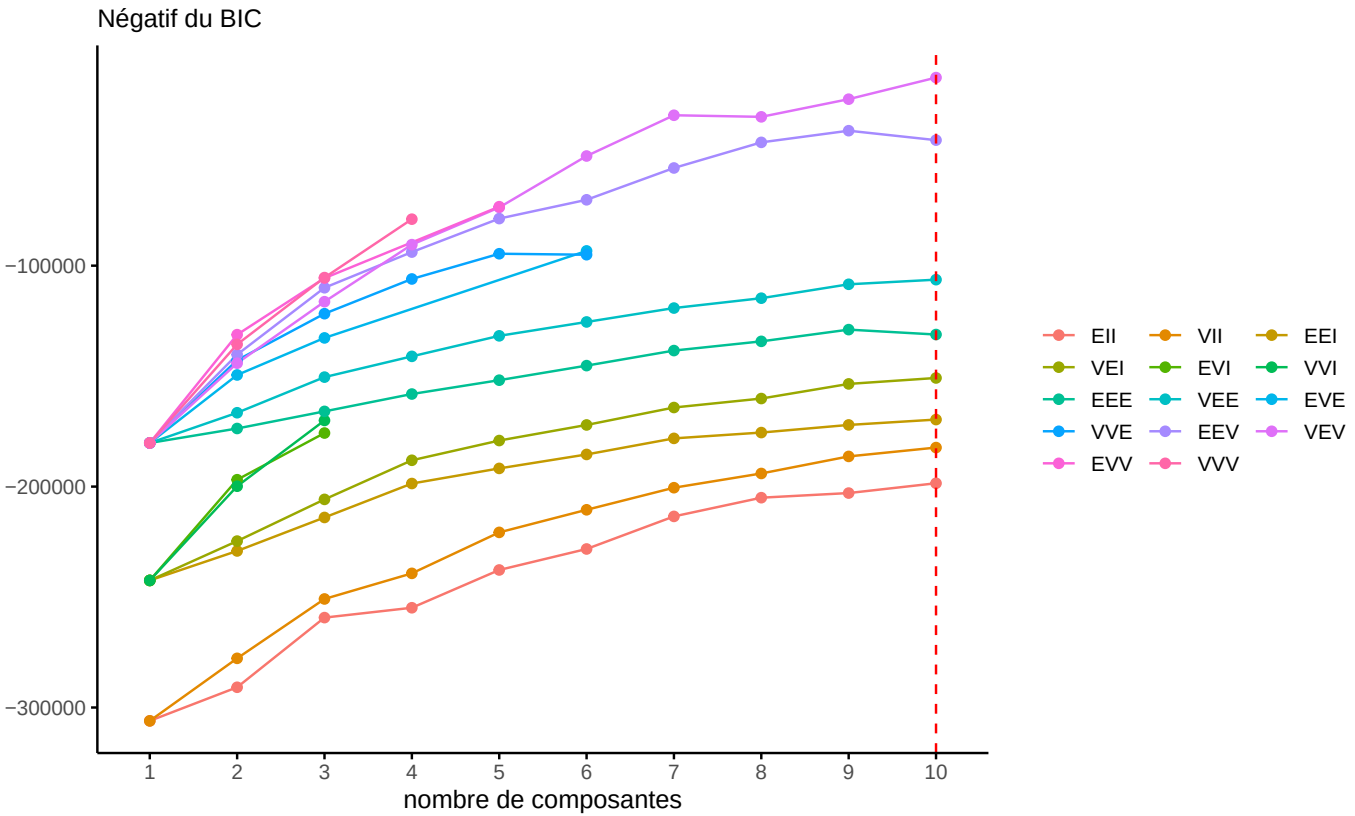


Figure 3: Valeur du négatif du BIC pour les mélanges de modèles gaussiens selon le nombre de regroupements et la structure de covariance.

Représentation graphique des regroupements

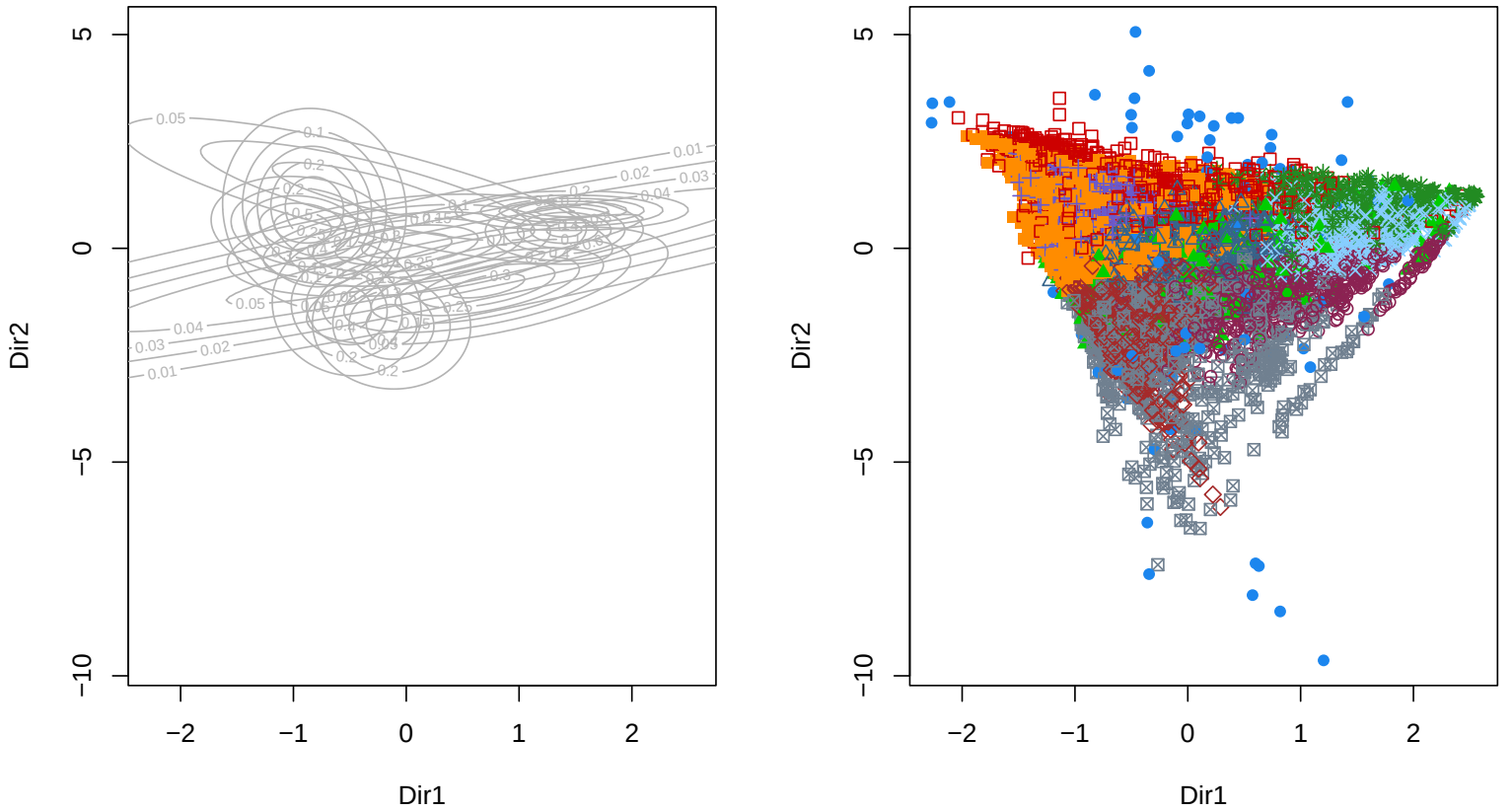


Figure 4: Projection des observations, colorées par regroupement (gauche) et structure des regroupements avec ellipsoides de confiance (droite).

Avantages et inconvénients des mélanges de modèles

- $(+)$ approche est plus flexible que les K-moyennes.
- $(+)$ l'ajout d'une composante uniforme permet de gérer les aberrances (supporté par `mclust`).
- $(+)$ l'algorithme EM garantie la convergence à un optimum local (comme pour les K-moyennes)
- $(+)$ on obtient une assignation probabiliste plutôt que rigide
- $(-)$ le coût de calcul est plus élevé que les K-moyennes
- $(-)$ le nombre de paramètre des matrices de covariance augmente rapidement avec la dimension p.

Regroupements hiérarchiques

Méthode déterministe de regroupement à partir d'une matrice de dissimilarité.

1. Initialisation: chaque observation est assignée à son propre groupe.
2. les deux groupes les plus rapprochés sont fusionnés; la distance entre le nouveau groupe et les autres regroupements est recalculée.
3. on répète l'étape 2 jusqu'à obtenir un seul regroupement.

Fonction de liaison

Il y a plusieurs façons de calculer la distance entre deux groupes d'observations de plusieurs observations, notamment

- liaison simple (`method = single`): plus proches voisins
- liaison complète (`method = complete`): voisins les plus éloignés
- liaison moyenne (`method = average`): utilise la moyenne des distances entre toutes les paires de sujets (un pour chaque groupe) provenant des deux groupes.
- méthode de Ward (`method = ward.D2`): calcul de l'homogénéité globale

Illustration des mesures de liaison

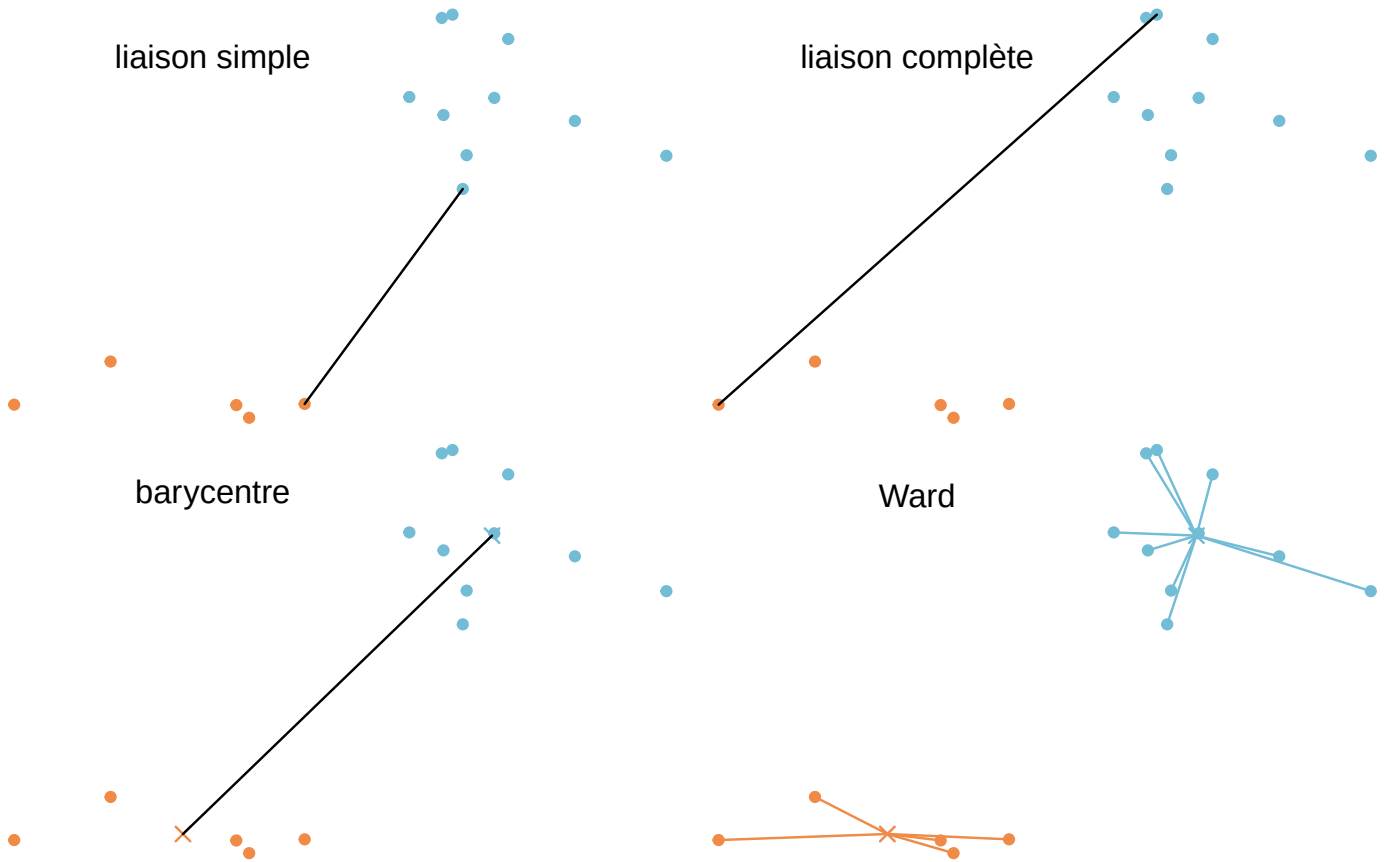


Figure 5: Distances entre regroupements selon la liaison (simple, complète, barycentre, homogenéité de Ward).

Méthode de Ward

La méthode de Ward utilise l'homogénéité comme critère.

Pour chaque groupe, on calcule la somme des carrés des distances par rapport à la moyenne du groupe, disons $\mathrm{SCD}_{k,M}$ ($k = 1, \dots, M$).

On calcule ensuite la somme des distances,

$$\mathrm{SCD}_{(M)} = \mathrm{SCD}_{1,M} + \cdots + \mathrm{SCD}_{M,M}.$$

La méthode de Ward va fusionner les deux groupes qui feront augmenter le moins possible l'homogénéité.

Performance relative

- Liaison simple: fonctionne bien si l'écart entre deux regroupements est suffisamment grand. S'il y a du bruit entre deux regroupements, la qualité des regroupements en sera affectée. Souvent quelques valeurs isolées et un seul grand regroupement
- Liaison complète: moins sensible au bruit et aux faibles écarts entre regroupements, mais a tendance à casser les regroupements globulaires.
- Homogénéité de Ward: le critère ressemble à celui des K-moyennes.

Voir la page scikit-learn pour une illustration.

Méthodes hiérarchiques et coût de calcul

Les algorithmes de regroupement hiérarchiques stockent une matrice de dissemblance $n \times n$: coût de stockage quadratique $\mathrm{O}(n^2)$.

Généralement, le coût de calcul est au mieux $\Omega(n^2)$ et au pire $\mathrm{O}(n^3)$.

Pour la méthode de liaison simple, un algorithme permet d'obtenir un coût de calcul quadratique de $\mathrm{O}(n^2)$ sans stocker la matrice de dissemblance, d'où un coût de stockage linéaire de $\mathrm{O}(n)$.

`stat::hclust` permet de faire des regroupements agglomératifs, mais le paquet `fastcluster` propose une version avec une empreinte mémoire inférieure (plus rapide!)

Genie

La fonction de liaison simple permet des calculs rapides, mais le résultat est rarement satisfaisant.

Une proposition de Gagolewski (2016), implémentée dans le paquet R `genieclust`, modifie la fonction de liaison simple en retenant son efficacité de calcul.

Plutôt que de simplement trouver la paire de regroupements à distance minimale, cette fusion n'est appliquée que si une mesure d'inéquité, le coefficient de Gini, est inférieur à un seuil spécifié par l'utilisateur.

Si les regroupements sont fortement inéquitables, la fusion survient entre les regroupements dont un de la taille minimale courante.

Hyperparamètres des méthodes hiérarchiques

1. choix de la fonction de liaison (et hyperparamètres associés)
2. mesure de dissemblance
3. nombre de regroupements

On peut représenter le modèle à l'aide d'un arbre, où les feuilles indiquent les regroupements à chaque étape jusqu'à la racine à la dernière étape (dendrogramme).

La distance entre chaque embranchement est déterminée par notre critère: cela nous permet de sélectionner un nombre de regroupements K après inspection visuelle du dendrogramme.

On élague l'arbre à la hauteur voulue.

Dendrogramme

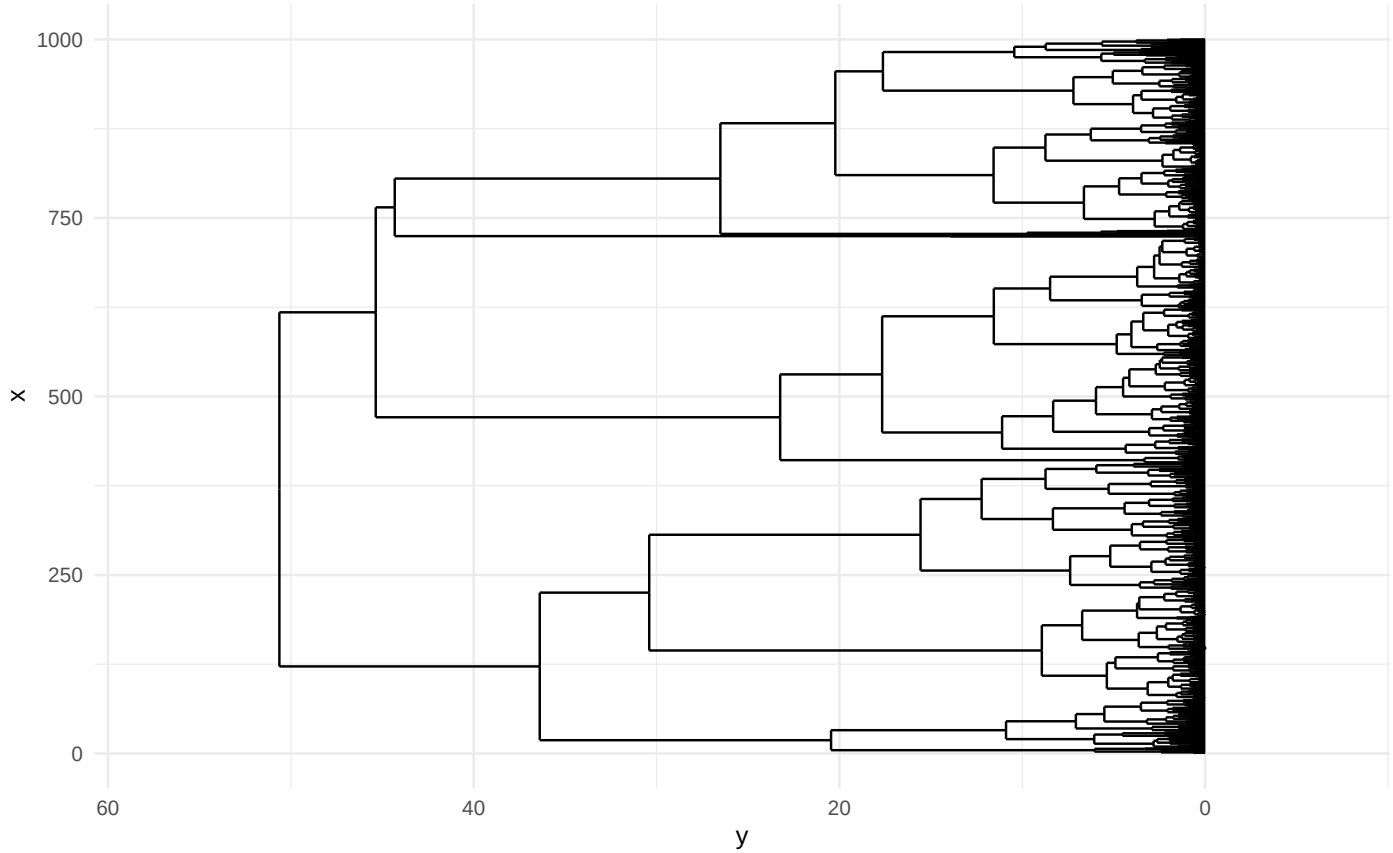


Figure 6: Dendrogramme pour l'exemple de regroupement hiérarchique avec la méthode de Ward et 100 premières observations.

Critères pour Ward

On peut choisir K à partir du pourcentage de variance expliquée, R^2 en calculant

$$R^2_{(M)} = 1 - \text{SCD}_{(M)}/\text{SCD}_{(1)},$$

où $\text{SCD}_{(1)}$ est l'homogénéité globale avec un seul groupe.

Le R-carré semi-partiel mesure la perte d'homogénéité d'une étape à l'autre, renormalisée par

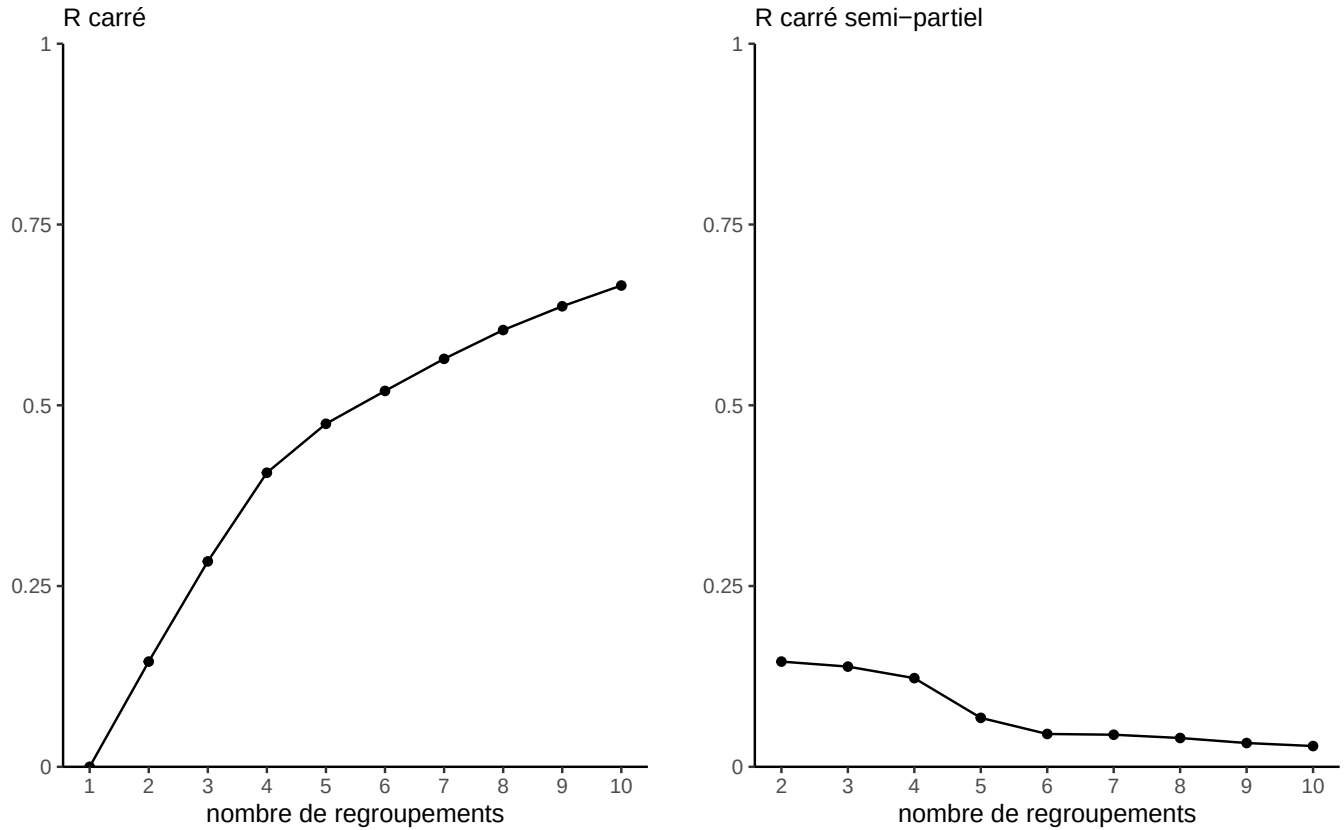
$$R^2_{\text{sp}(M)} = \frac{\text{SCD}_{(M-1)} - \text{SCD}_{(M)}}{\text{SCD}_{(1)}}, \quad M = 2, \ldots, n.$$

Critères d'homogénéité (Ward)

```
ward <- fastcluster::hclust(
  d = dist(donsmult_std),
  method = "ward.D2")
hecmulti::homogeneite(rhier = ward,
                      ngroupes = 10,
                      data = donsmult_std)
```

Graphiques du R^2 et du R^2 semi-partiel

On cherche un point d'inflection (un coude).



Avantages et inconvénient, regroupements hiérarchiques

- (+) la solution du regroupement hiérarchique est toujours la même (déterministe)
- (+) les méthodes d'arborescence sont faciles à expliquer
- (—) l'assignation d'une observation à un regroupement est finale
- (—) les aberrances ne sont pas traitées et sont souvent assignées dans des regroupements à part
- (—) le nombre de groupes n'a pas à être spécifié apriori (une seule estimation)
- (—) le coût de calcul est prohibitif, avec une complexité quadratique de $O(n^2)$ pour la méthode de liaison simple et autrement $O(n^3)$ pour la plupart des autres fonctions de liaison.

Méthodes basées sur la densité

L'algorithme DBSCAN (density-based spatial clustering of applications with noise) est une méthode de partitionnement basée sur la densité des points.

L'idée de base de l'algorithme est de tracer une boule de rayon ϵ autour de chaque observation et de voir si elle inclut d'autres observations.

Mécanisme d'assignation

L'algorithme classe les observations en trois catégories:

- Un point central est une observation qui possède $M - 1$ voisins à distance ϵ.
- Un point frontière est un point qui est distant de moins de ϵ d'un point central, sans en être un.
- Un point isolé est une observation qui n'est pas rattachée à aucun regroupement.

Algorithme DBSCAN

L'algorithme répète les étapes suivantes jusqu'à ce que chaque observation ait été visitée.

1. Choisir un point aléatoirement parmi ceux qui n'ont pas été visités.
2. Si le point n'est pas étiqueté, calculer le nombre de points voisins qui se trouvent dans un rayon ϵ: s'il y a moins de M observations, provisoirement étiqueter l'observation comme point isolé, sinon comme point central.
3. Si l'observation est un point central avec $M - 1$ voisins ou plus, créer un regroupement.
4. Étiqueter chaque point à distance ϵ créé et l'ajouter au regroupement, ainsi que tout point à distance ϵ de ces voisins.

Ce site web offre une visualisation interactive des différentes étapes de l'algorithme et de comparer la performance de DBSCAN selon le type de regroupements.

Hyperparamètres

- le rayon ϵ et
- le nombre minimal de points pour former un regroupement, M

Hyperparamètres M et ϵ corrélés: si on augmente le nombre minimal de point M par regroupement, il faudra également augmenter le rayon ϵ pour éviter d'avoir un nombre trop élevé de points isolés et d'aberrances.

Classification avec DBSCAN

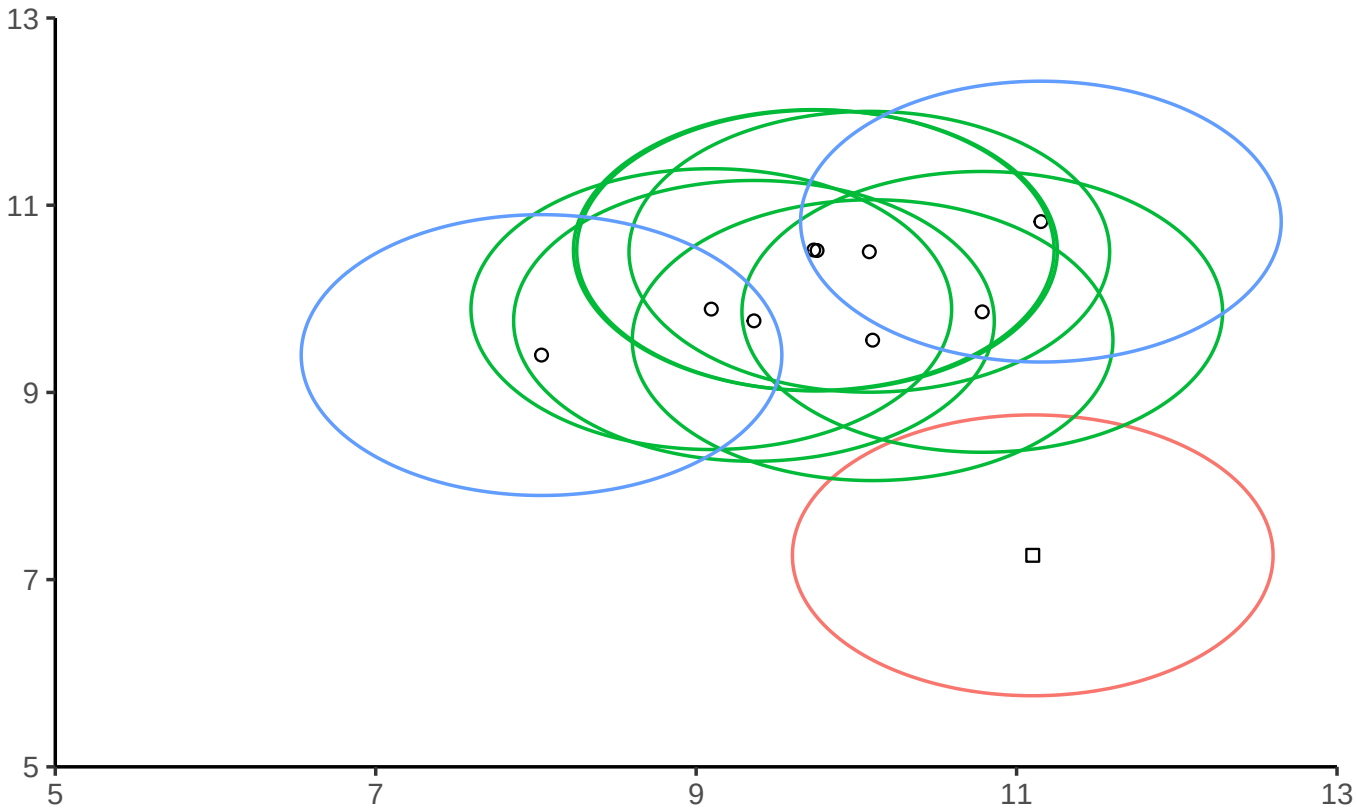


Figure 7: Illustration de la classification des points avec DBSCAN: toutes les observations sont assignées à un regroupement, moins une aberrance.

Choix des hyperparamètres

Avec p variables explicatives, on recommande $M > p + 1$.

Pour ε, considérer les M plus proches voisins.

La fonction `kNNdistplot` du paquet `dbscan` permet de tracer un graphique de la distance moyenne des k plus proches voisins pour chaque observation:

- en prenant $k = M - 1$, calculer la distance entre le k plus proche voisin de chaque observation.
- ordonner ces distances.
- choisir ϵ selon coude

Critère du coude

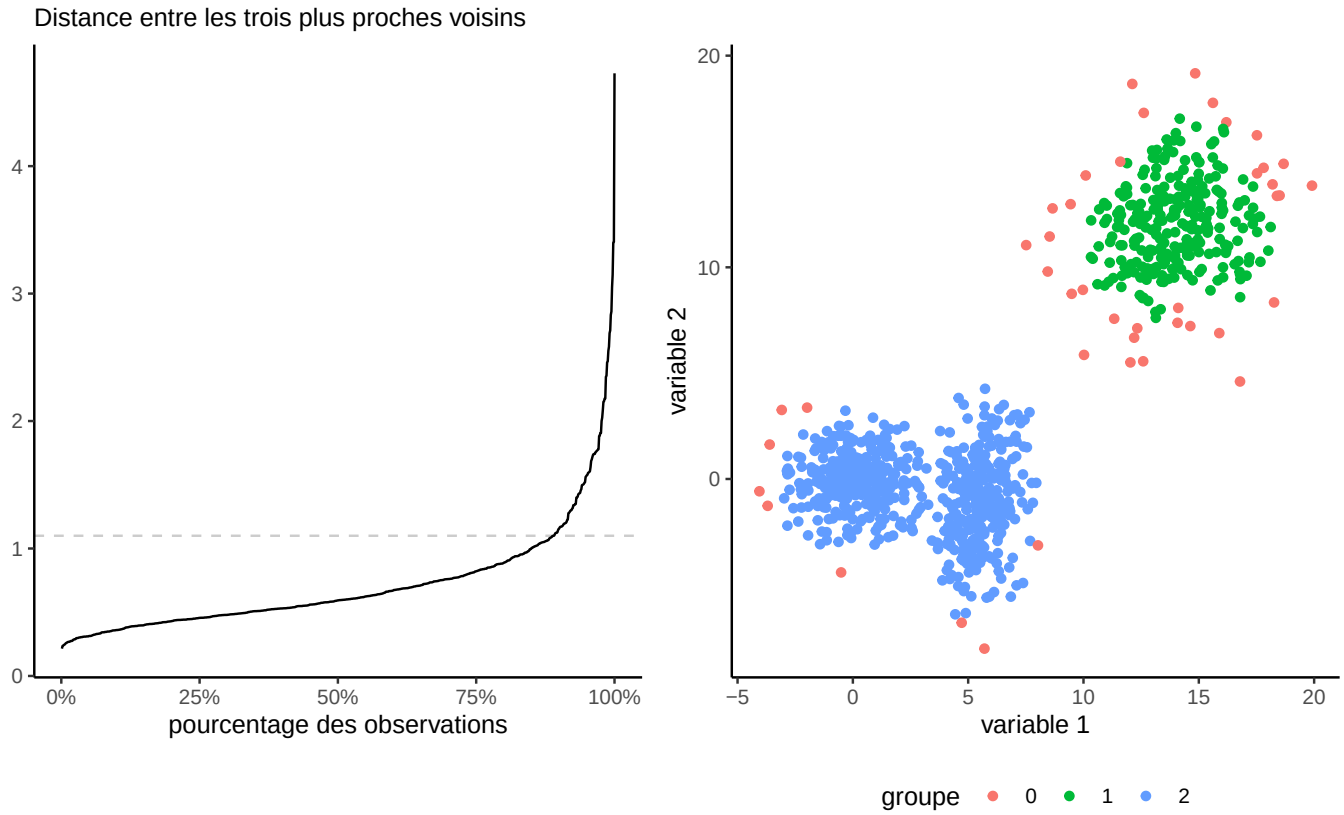


Figure 8: Graphique des distances entre chaque observation et son troisième plus proche voisin (gauche), en fonction du pourcentage d'observations à moins de cette distance et regroupements obtenus avec DBSCAN avec $M = 10$ et $\epsilon = 1.1$ (droite).

Avantages et inconvénients de DBSCAN (1/2)

- $(+)$ le traitement des aberrances est automatique et l'algorithme est robuste.
- $(+)$ le nombre de regroupements n'a pas à être spécifié apriori.
- $(+)$ la forme des regroupements est arbitraire, peut être non convexe et de taille différente.
- (+/-) la complexité de l'algorithme est d'au mieux $\mathrm{O}(n \ln n)$ (calcul) et $\mathrm{O}(n)$ pour le stockage puisque chaque point est visité à tour de rôle et comparé aux autres pour trouver les plus proches voisins.

Avantages et inconvénients de DBSCAN (2/2)

- (+) les hyperparamètres ont une interprétation physique
- (—) mais leur choix n'est pas aisé
- (—) DBSCAN ne permet pas de traiter le cas où la densité des regroupements change et risque de fusionner des regroupements s'il y a une série d'observations qui permet de relier deux regroupements.
- (—) comme la plupart des algorithmes, le voisinage des points devient épars quand p augmente en raison du fléau de la dimension.

Comparaison de segmentations

On veut parfois comparer les regroupements de différentes méthodes.

Les étiquettes ne sont pas nécessairement identiques même si les regroupements le sont (permutation des étiquettes).

Une mesure de similarité, l'indice de Rand, permet de comparer deux vecteurs d'observations catégorielles (étiquettes des regroupements).

- même longueur (même nombre d'observations)
- mais pas nécessairement le même nombre de modalités (nombre de regroupements potentiellement différents d'une partition à l'autre).

Idée: comparer à tour de rôle chacune des $n(n-1)/2$ paires d'observation.

Indice de Rand

On indique si les paires d'observations sont dans le même groupe (1) ou un groupe différent (O).

Exemple: deux partitions de 4 observations

- Étiquettes de la méthode A avec deux regroupements: (2, 2, 2, 1)
- Étiquettes de la méthode B avec trois regroupements: (3, 1, 3, 2)

{1,2}	{1,3}	{1,4}	{2,3}	{2,4}	{3,4}
10	11	OO	10	OO	OO

L'indice de Rand est le taux de bonne classification du tableau de contingence résultant. Parmi les six paires, quatre sont concordantes (OO ou 11), d'où un indice de Rand $= 0.66$

Une valeur de 1 indique que les deux partitions sont identiques.

Récapitulatif

Les étapes d'une analyse de regroupements

1. Choisir les variables pertinentes à l'analyse. Cette étape peut nécessiter de créer, transformer de nouvelles variables ou d'aggréger les données.
2. Décider quel méthode sera utilisée pour la segmentation.
3. Choisir les hyperparamètres de l'algorithme (nombre de regroupements, rayon, etc.) et la mesure de dissemblance.
4. Valider la qualité de la segmentation (interprétabilité, taille des groupes, homogénéité des regroupements).
5. Avec les étiquettes, calculer un prototype de groupe.
6. Interpréter les regroupements obtenus à partir des prototypes.

Récapitulatif

- L'analyste a une grande marge de manoeuvre.
- Il n'y a pas de vérité: la segmentation n'est utile que si elle a une valeur ajoutée.
- Aucun algorithme ne performe uniformément mieux, mais certains sont plus faciles à employer que d'autres.
  - avec des mégadonnées, la complexité est un facteur important pour choisir la méthode.
  - la plupart du temps, le choix des hyperparamètres nécessite un peu d'essai-erreur.
  - la segmentation peut être médiocre parce que les hyperparamètres sont mal choisis.

Récapitulatif

Le Diable est dans les détails:

- variables catégorielles et binaires vs standardisation
- valeurs manquantes
- aberrances

Le nombre de groupes peut être guidé par le contexte: les formules et indicateurs de qualité servent de balises.