Analyse de regroupements

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

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

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 $(k = 1, \ldots, K)$:
   - 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.

- Diviser l'échantillon en S sous-échantillons de taille approximativement égale de taille n_S (typiquement $K \ll n_S < 1000$)
- Utiliser l'algorithme PAM sur chaque sous-échantillon.

Une fois les médoïdes obtenus, le reste de toutes les observations de l'échantillon sont assignées au regroupement du médoïde le plus près.

Algorithme CLARA (2/2)

La qualité de la segmentation pour chacune des S segmentations est calculée en obtenant la distance moyenne entre les médoïdes et les observations.

On retourne la meilleure segmentation parmi les S (celle qui a la plus petite distance moyenne).

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.

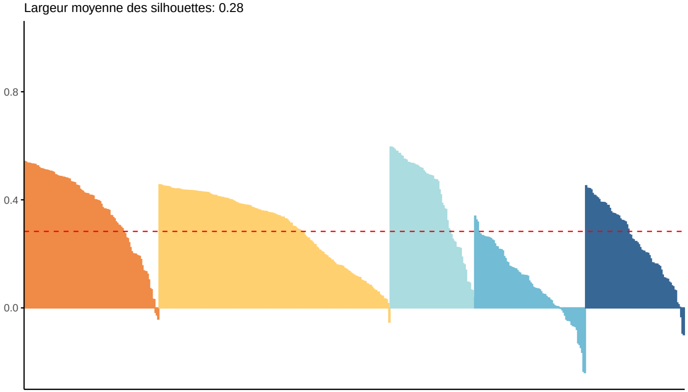

Figure 1: Silhouettes pour les données de dons multiples avec l'algorithme CLARA pour $K = 5$ regroupements.

Prototypes

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

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

Avantages et inconvénients des K-médoïdes

- ($-$) solution approximative pour grand échantillons
- ($+$) 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) avec complexité $\mathrm{O}(n^2)$. PAM fonctionne avec maximum 1000 observations.

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 $k = 1, \ldots, K$.

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

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

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 $\mathbf{X}_i$ tombe dans le groupe k,

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

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 $\mathbf{A}$ qui contrôle les variances de chaque observation et
- $\mathbf{D}$ 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`.

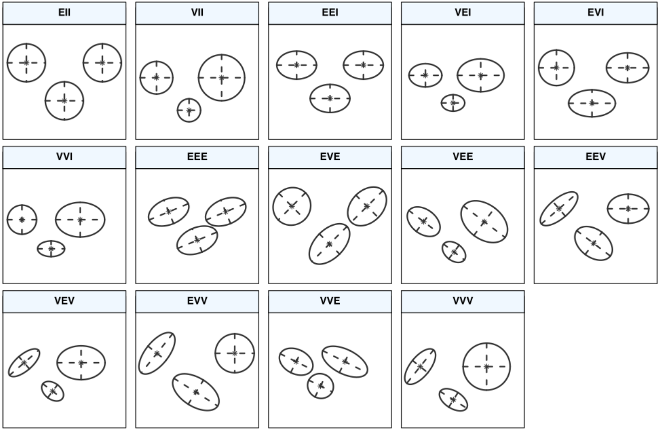

Figure 2: Forme des ellipsoïdes pour le mélange de modèle selon la forme de la structure de covariance. Tirée de `mclust5`, licence CC BY 4.0.

Paquet mclust

```r
## Mélanges de modèles gaussiens
set.seed(60602)
library(mclust)
mmg <- Mclust(data = donsmult_std,
        G = 1:10,
        # 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

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

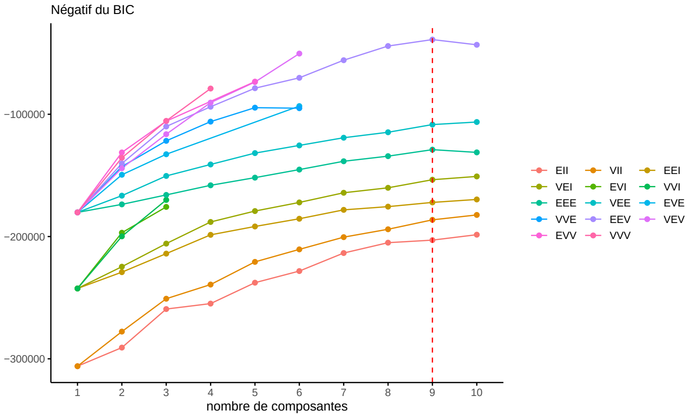

Figure 3: Valeur du négatif du BIC pour les mélanges de modèles gaussiens selon le nombre de regroupements et la structure de covariance.

Représentation graphique des regroupements

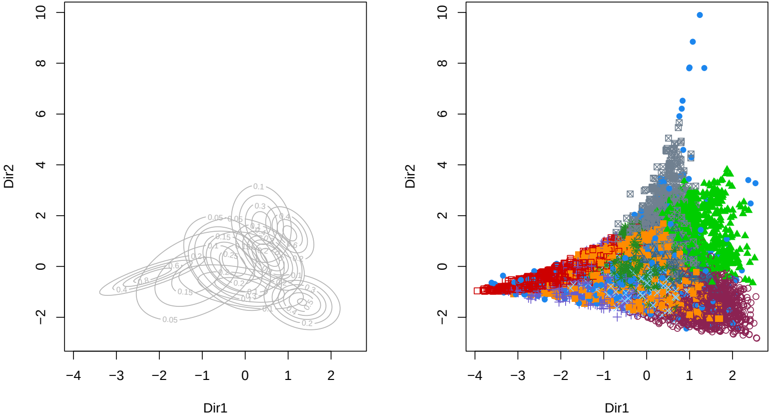

Figure 4: Projection des observations, colorées par regroupement (gauche) et structure des regroupements avec ellipsoides de confiance (droite).

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 minimum 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

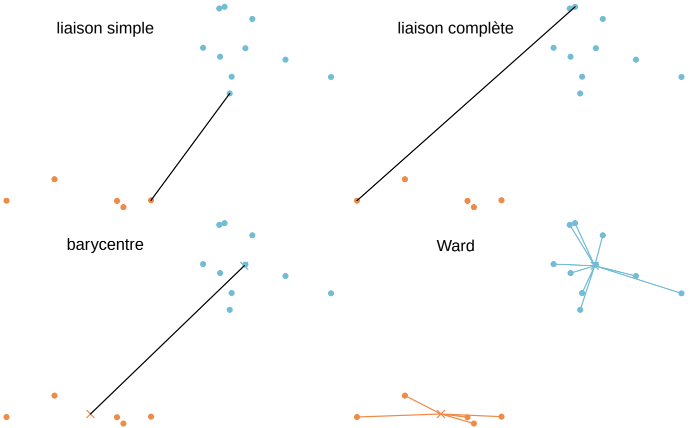

Figure 5: Distances entre regroupements selon la liaison (simple, complète, barycentre, homogenéité de Ward).

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 SCD_k ($k = 1, \ldots, M$).

On calcule ensuite l'homogénéité globale en faisant la somme, $\mathsf{H}^{(M)} = \mathsf{SCD}_1 + \cdots + \mathsf{SCD}_M$.

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

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

Alternative de Gagolewski (2016) qui 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 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.

L'implémentation **R** dans le paquet `genieclust` est nettement plus rapide que les autres alternative.

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.

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

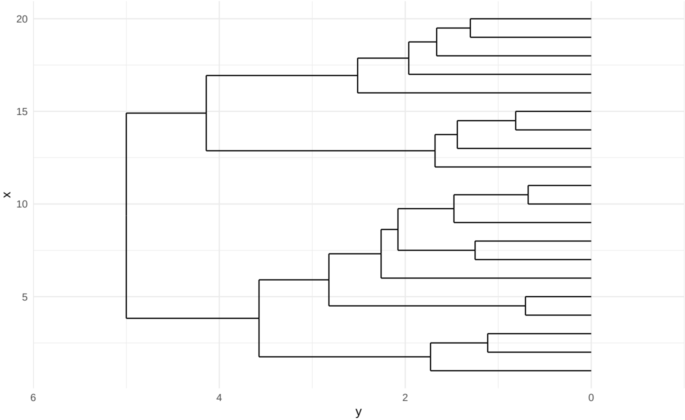

Figure 6: Dendrogramme pour l'exemple de regroupement hiérarchique avec la méthode de Ward et 20 observations.

Critères pour Ward

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

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

où $\mathsf{H}_{(1)}$ 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

$$R^2_{\mathsf{sp}(M)} = \frac{\mathsf{H}_{(M)} - \mathsf{H}_{(M-1)}}{\mathsf{H}_{(1)}},$$

mesure la perte d'homogénéité (relative) en combinant ces deux groupes.

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

Critères d'homogénéité (Ward)

Avantages et inconvénient, regroupements hiérarchiques

- (+) la solution du regroupement hiérarchique est toujours la même (déterministe)
- (—) 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
- (+) les méthodes d'arborescence sont faciles à expliquer
- (—) 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 $\mathrm{O}(n^2)$ pour la méthode de liaison simple et autrement $\mathrm{O}(n^3)$ pour la plupart des autres fonctions de liaison.