Cours d'Analyse des Données Multidimensionnelles

Chapitre 4 : Classification non supervisée (clustering)

Zaineb Smida

5ème année GEA
INSA Lyon

7 janvier 2025

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Rappels de statistique bivariée

On considère une variable quantitative Y et une variable **qualitative** X à k modalités définissant k groupes (classes). On note

- $\bar{y}_r$: moyenne de Y dans le groupe r
- σ_r^2 : variance de Y dans le groupe r

Exemple

Voir les notes de cours : ***Tableau : SD1****.*

Variance inter-groupes et variance intra-groupes

Variance inter-groupes : on appelle variance inter-groupes la quantité :

$$Var_{INTER} = \frac{1}{n}\sum_{r=1}^{k} n_r(\bar{y}_r - \bar{y})^2$$

Elle correspond à la variance des moyennes des groupes (pondérée par les effectifs des groupes).

$\hookrightarrow$ mesure la dispersion de Y "entre les groupes" = dispersion due au facteur X

Cette quantité est d'autant plus grande que les moyennes conditionnelles de Y sachant X diffèrent. Elle est petite si Y a sensiblement la même moyenne dans tous les groupes définis par les modalités de X.

Exemple

Voir les notes de cours : ***Résultat : SD2***.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Variance intra-groupes : on appelle variance intra-groupes la quantité :

$$Var_{INTRA} = \frac{1}{n} \sum_{r=1}^{k} n_r \sigma_r^2$$

C'est la moyenne pondérée des variances conditionnelles de Y sachant X.

$\hookrightarrow$ mesure la dispersion de Y "à l'intérieur des groupes"
= dispersion due au hasard

Cette quantité est d'autant plus petite que les valeurs prises par Y sont homogènes dans les différents groupes définis par les modalités de X.

Exemple

Voir les notes de cours : ***Résultat : SD3****.*

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Décomposition de la variance totale de Y

Proposition

La variance ***totale*** *de Y est* ***la somme*** *des variances inter-groupes et intra-groupes :*

$$Var(Y) = \frac{1}{n}\sum_{r=1}^{k} n_r(\bar{y}_r - \bar{y})^2 + \frac{1}{n}\sum_{r=1}^{k} n_r\sigma_r^2.$$

Remarque

Cette formule s'appelle l'équation d'analyse de la variance :

Variance ***TOTALE*** *de Y = Variance INTER + Variance INTRA*

Exemple

Voir les notes de cours : ***Résultat : SD4***.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Rapport de corrélation

Le rapport de corrélation η^2 est le rapport entre la variance inter-groupes et la variance **totale** de Y :

$$\eta^2 = \frac{\frac{1}{n}\sum_{r=1}^{k} n_r(\bar{y}_r - \bar{y})^2}{Var(Y)} \in [0,1]$$

Interprétation

- η^2 proche de 0 : les moyennes conditionnelles de Y diffèrent très peu selon groupes définis par les modalités de X
  $\hookrightarrow$ absence de dépendance entre X et Y
  $\hookrightarrow$ pas d'impact de X sur Y (ou de Y sur X)

- η^2 proche de 1 : les moyennes conditionnelles de Y sont très différentes selon les groupes définis par les modalités de X
  $\hookrightarrow$ forte liaison entre X et Y
  $\hookrightarrow$ impact de X sur Y (ou de Y sur X)

Exemple

Voir les notes de cours : Résultat : SD5.

Applications et données

Applications

Segmentation des fichiers clients par les banques ou autres entreprises, études de marchés en Marketing, ...

Données

Il existe des techniques de classification non supervisée (analyse typologique ou clustering) adaptées à tout type de tableaux de données. Cependant, ici, nous nous limitons aux tableaux *individus*/*variables*.

Plus précisément, on suppose donc que l'on dispose de n observations décrites par p variables quantitatives i.e. d'un tableau de taille $n \times p$.

Remarque

Lorsque les variables sont très hétérogènes (variances ou échelles très différentes), on peut être amené, comme en ACP, à centrer et réduire les données.

Notations : on note $X_1, \ldots, X_n$ les n individus et $X^1, X^2, \ldots, X^p$ les p variables numériques.

Exemple

En marketing, on peut chercher à mettre en évidence ***des groupes de clients*** *dont les comportements d'achat sont* ***homogènes****. Les variables quantitatives à considérer peuvent être les quantités achetées pour différents produits.*

Objectifs

L'analyse typologique ou analyse classifiante ou classification est une technique d'**analyse de données** permettant de construire des groupes d'individus tels que :

- chaque groupe soit **homogène** selon certaines caractéristiques, c'est-à-dire que les observations d'un groupe se ressemblent le plus possible,
- chaque groupe **soit différent des autres** selon les mêmes caractéristiques, c'est-à-dire que les observations d'un groupe sont les plus différentes possible de celles des autres groupes.

Avec une méthode de clustering, chaque observation est classée dans un groupe donné.

Mesure de la similarité et inertie

Mesure de la similarité

Les notions de similarité ou de différences entre individus sont formalisées en mesurant des **distances** entre les individus. On choisit, dans la suite, de travailler avec la distance euclidienne usuelle :

$$d(X_i, X_{i'}) = \sqrt{\sum_{l=1}^{p} (X_i^l - X_{i'}^l)^2}.$$

Inertie

En clustering, nous considérons également l'inertie, qui est liée à la distance entre les observations et le vecteur **moyen** $\overline{X} = (\overline{X^1}, \overline{X^2}, \cdots, \overline{X^p})$:

$$Inertie = \sum_{j=1}^{p} Var(X^j) = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} (X_i^j - \overline{X^j})^2 = \frac{1}{n} \sum_{i=1}^{n} d^2(X_i, \overline{X})$$

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Décomposition de l'inertie

On considère une partition en k groupes, notés $G_1, G_2, \ldots, G_k$, d'effectifs $n_1, n_2, \cdots, n_k$ (avec $\sum_{r=1}^{k} n_r = n$).

Rappel : Une partition des observations $X_1, \ldots, X_n$ est un ensemble de k groupes $A_1, \ldots, A_k$ tel que :

- $A_1 \cup A_2 \cup \ldots \cup A_k = \{X_1, \ldots, X_n\}$
- $\forall i, j \in \{1, \ldots, k\}, i \neq j, , A_i \cap A_j = \emptyset$

L'inertie totale est écrite comme suit :

$$\text{Inertie} = \sum_{r=1}^{k} \sum_{i \in G_r} \frac{1}{n} d^2(X_i, \overline{X})$$

On peut décomposer l'inertie en :

Proposition

Inertie= *Inertie intra-groupes* + *Inertie inter-groupes*

Où,

- Inertie intra-groupes $= \sum_{j=1}^{p} Var_{INTRA}(X^j) = \sum_{r=1}^{k} \frac{n_r}{n} I_r$

  avec $I_r = \frac{1}{n_r} \sum_{i \in G_r} d^2(X_i, \overline{X}^{(r)})$: inertie dans le groupe G_r

- Inertie inter-groupes $= \sum_{j=1}^{p} Var_{INTER}(X^j) = \sum_{r=1}^{k} \frac{n_r}{n} d^2(\overline{X}^{(r)}, \overline{X})$

  avec $\overline{X}^{(r)}$: point moyen du groupe G_r.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Objectifs en termes d'inertie

Objectif : Afin de remplir les objectifs, les méthodes de classification visent à trouver la partition telle que **l'inertie inter-groupes soit suffisamment élevée** (et que l'inertie intra-groupes soit suffisamment faible) avec un nombre de groupes raisonnable (pas trop élevé).

Présentation des techniques de classification

Pour l'essentiel, les techniques de classification font appel à **une démarche algorithmique** et non aux calculs formalisés usuels en Analyse Factorielle (ACP, AFC).

Alors que les valeurs des composantes principales, par exemple, sont la solution d'une équation pouvant s'écrire sous une forme très condensée (même si sa résolution est complexe), la définition des **classes** ne se fait qu'à partir d'une *formulation algorithmique* : une série d'opérations définies de façon **répétitive**.

Il existe 2 grandes familles de méthodes :

- les méthodes de *partitionnement*
- les méthodes *hiérarchiques*.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Une classification par partitionnement consiste à rechercher directement une *partition* des individus.

Les groupes obtenus visent à maximiser l'inertie inter-classes et minimiser l'inertie intra-classes.

Il existe différentes méthodes de partitionnement. Nous choisissons de présenter une méthode populaire dite ***d'agrégation autour de centres mobiles (ou k-means)***.

- Avantage : cette méthode de type partitionnement est très **rapide** (fonction `kmeans()` sous `R`) et utilisable même pour de très grands tableaux.
- Inconvénient : le nombre de clusters doit être **fixé à l'avance** (il est généralement inconnu).

Méthode d'Agrégation autour de Moyennes Mobiles (AMM)

Présentation de l'algorithme de l'AMM

- La méthode consiste tout d'abord **à fixer à priori un nombre** k de groupes et à représenter chaque groupe par un individu (centres initiaux).
- On affecte chacun des individus restants au groupe le plus proche.
- On calcule alors les centres (ou moyennes) de chaque groupe et les individus sont éventuellement réaffectés au groupe dont ils sont le plus proche.
- Cette dernière étape est répétée (on parle **d'itérations**) **jusqu'à ce que les centres de groupes soient peu ou pas modifiés**.

On montre d'un point de vue théorique que l'algorithme précédent converge sous des hypothèses peu restrictives. De plus, à chaque itération, l'inertie intra-classes diminue.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Choix des centres de classes initiaux

Il existe plusieurs façons de choisir les k individus représentant chacun des groupes pour **l'initialisation de l'algorithme**. On peut choisir entre autres :

- les k premiers individus
- k individus **tirés aléatoirement parmi les** n (ce qui est fait par la fonction `kmeans()`)
- k individus "suffisamment" éloignés les uns des autres ...

Or, ce choix n'est pas neutre. Il peut conditionner le résultat final.

Exemple

OBS	*REV*	*EDUC*
1	*5*	*5*
2	*6*	*6*
3	*15*	*14*
4	*16*	*15*
5	*25*	*20*
6	*30*	*19*

Voir les notes de cours :

*voir **Figure : Données**. Il y a 3 groupes très nettement différenciés avec bas-moyens-hauts revenus et niveaux d'études.*

Le choix des centres initiaux influence la partition finale. Ainsi, dans l'exemple, si on choisit les observations 1, 2 et 3 comme centres initiaux, on ne trouve pas la partition $\{\{1,2\},\{3,4\},\{5,6\}\}$ *mais la partition* $\{\{1\},\{2\},\{3,4,5,6\}\}$ *(voir **Figure CL0**).*

Remarque

- *La fonction* `kmeans()` *de* `R` *ne présente pas la possibilité de choisir des centres initiaux éloignés. Néanmoins, avec l'option* `nstart` *de cette fonction, l'algorithme choisit aléatoirement* `nstart` *ensembles de centres initiaux et donne en sortie la "meilleure" classification. Dans la pratique, il est conseillé de* ***fixer*** `nstart=100`.

Étapes de la classification par AMM

1. Choix du nombre de groupes

Pour utiliser une méthode de partitionnement, il faut se fixer un nombre k de groupes (défaut de ce type de méthodes). Si on n'a pas d'indication préalable, il est conseillé de faire plusieurs essais avec différentes valeurs de k.

$\hookrightarrow$ **Analyse du rapport de corrélation global**.

Pour une partition donnée en k groupes, on examinera le rapport inertie inter-classes/ inertie (R^2 **global**) qui doit être suffisamment élevé.

Exemple

Voir les notes de cours : ***Tableau CL5*** *et* ***Tableau CL6****.*

Choix du nombre de groupes : Coefficient silhouette

Pour une partition en k groupes de l'échantillon et i un individu de l'échantillon, on note :

- $a(i)$: distance moyenne entre i et tous les autres points du groupe auquel i appartient.
- $b(i)$: plus petite distance moyenne de i à tous les groupes auxquels i n'appartient pas.

Définition

On définit le coefficient silhouette pour l'individu i par :

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

Proposition

$$s(i) \in [-1, 1]$$

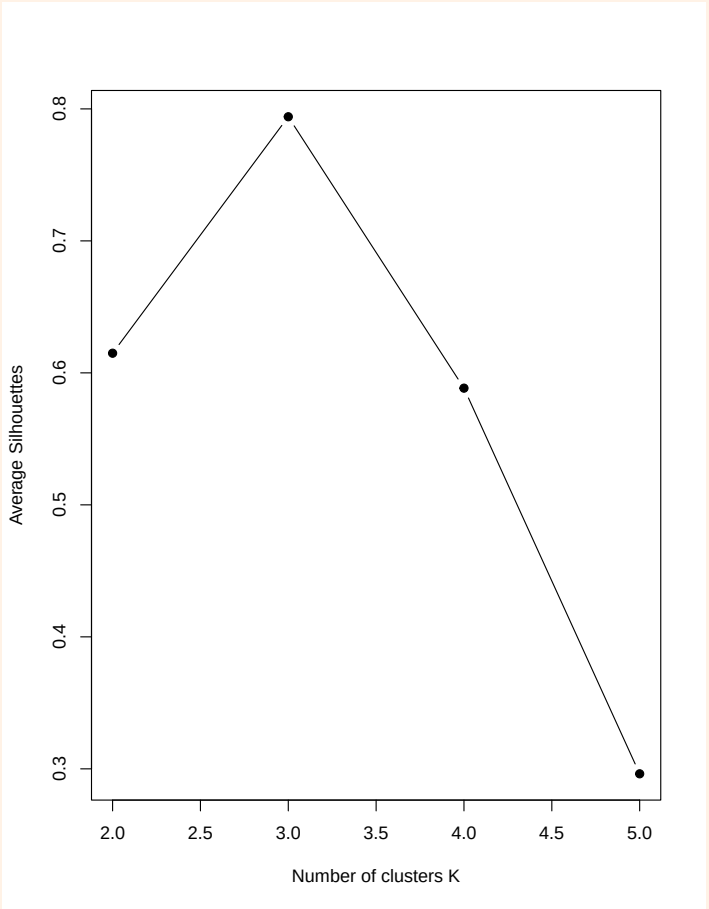
INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

On calcule alors le coefficient silhouette moyen s (moyenne de tous les $s(i)$ de l'échantillon)

- **La meilleure valeur de** s est **1** et la pire est -1,
- Des valeurs de s proches de 0 indiquent que des groupes se chevauchent.

Pour le calcul de s avec `R`, utiliser le package `cluster`.

Exemple



Voir les notes de cours : *pour les codes associés. La figure illustre la silhouette, dont le maximum pour* $k = 3$ *indique un regroupement optimal en* **3 clusters**.

2. Interprétation des groupes (typologie)

L'objectif d'une classification est d'obtenir une **typologie** c'est-à-dire des groupes ou types d'individus avec une description de chacun de ces groupes. Pour obtenir cette interprétation, on doit revenir aux variables initiales et calculer des moyennes par groupes.

On interprétera uniquement les variables qui diffèrent suffisamment selon les groupes. Pour repérer ces variables, on doit examiner les rapports de corrélation de chaque variable.

Si, pour la variable quantitative considérée, le rapport de corrélation est $>0{,}5$ alors on interprétera la variable (car variance-inter $>$ variance-intra).

Exemple

*Voir les notes de cours : **Tableau CL8**.*

On procède alors à la description des groupes à l'aide du tableau des moyennes par groupe.

Exemple

Voir les notes de cours : ***Tableau CL3***.

3. Graphiques

Un fois repérées les variables importantes et si elles ne sont pas en nombre trop important, on peut représenter les individus sur ces variables en tenant compte du groupe d'affectation à l'aide de boîtes à moustaches juxtaposées par exemple.

Exemple

Voir les notes de cours : ***Figure CL7***.

4. Cas des très petits groupes

Souvent, pour de gros fichiers, on trouve des classes regroupant beaucoup d'individus et d'autres classes très petites (parfois un seul individu par classe). C'est le cas en particulier lorsqu'un individu est très atypique. Une fois repérés de tels individus, il peut être intéressant de les enlever ou de refaire l'analyse avec davantage de groupes.

Classification ascendante hiérarchique (CAH)

Principe

La méthode consiste à considérer comme typologie initiale autant de classes que d'individus (n) et à regrouper les classes par étapes successives jusqu'à l'obtention d'une seule classe contenant tous les individus.

Algorithme

Initialisation : on calcule toutes les distances entre individus (tableau de taille $n \times n$). Chaque individu représente une classe.

Itérations :

- on agrège les 2 classes les plus proches en une nouvelle classe,
- on met à jour le tableau des distances en tenant compte de la nouvelle classe (calcul des distances de cette nouvelle classe avec les autres classes).

Les itérations s'arrêtent lorsque l'on ne dispose plus que d'une seule classe.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Calcul des distances

Dans la présentation de l'algorithme ci-dessus, on utilise la notion de **"distance" entre classes**, c'est-à-dire entre groupes d'individus. Cette notion doit être précisée et, selon la distance choisie, on obtient différents algorithmes.

Dans la suite, on propose de se concentrer sur un choix particulier qui consiste à utiliser **une distance euclidienne** entre les centres de classes.

Si on dénote par c_a (respectivement c_b) le centre de la classe a de taille n_a (respectivement b de taille n_b), la distance utilisée est le **saut de Ward** définie par :

$$D^2_{ward}(a,b) = \frac{n_a n_b}{n_a + n_b} d^2(c_a, c_b).$$

Le saut de Ward est la distance la plus utilisée **en pratique**. Il permet de **minimiser**, à chaque itération, la décroissance de l'inertie **inter-groupes**.

En effet, lorsque l'on agrège 2 classes, l'inertie inter-classe diminue nécessairement mais on souhaite qu'elle diminue le moins possible car on veut séparer au mieux les groupes.

Pour utiliser ce critère dans la fonction `hclust()` de `R`, il faut mettre l'option *method=ward.D2*.

Autres méthodes d'agrégation

1- Méthode du **saut minimal**

$$D_{min}(a,b) = \inf\{d(i,j), i \in a, j \in b\}$$

Cette méthode peut entraîner des classes trop larges.

2- Méthode du **diamètre**

$$D_{diam}(a,b) = \sup\{d(i,j), i \in a, j \in b\}$$

3- Méthode de la **distance moyenne**

$$D_{moy}(a,b) = \frac{1}{n_a n_b} \sum_{i \in a, j \in b} d(i,j)$$

Cette méthode est un bon compromis entre les deux méthodes précédentes.

Les étapes de la CAH

1. Choix du nombre de groupes

Avantage de la CAH : il n'est pas nécessaire de fixer le nombre de groupes a priori.

Le dendrogramme : Il s'agit d'une représentation sous forme d'arbre des regroupements successifs obtenus par CAH.

Exemple

*Voir les notes de cours : **Figure CL11**.*

Le dendrogramme :

- Dans un arbre hiérarchique à grappes, deux individus de l'ensemble de données initial sont connectés à un certain niveau.
- La hauteur du lien (height) représente la distance entre les deux grappes qui contiennent ces deux individus. Elle correspond, dans notre contexte, à la perte en termes d'inertie inter-classe lors de l'agrégation de deux groupes.
  Si, en regroupant 2 groupes, la hauteur augmente trop, cela indique qu'ils sont trop différents. Il est donc préférable d'éviter ce regroupement.
- On coupe le dendrogramme en traçant une ligne horizontale. Les observations qui se rejoignent en dessous de la ligne sont dans le même groupe.

Pour choisir le nombre de groupes, on peut tracer le graphique des hauteurs en fonction du nombre de groupes. En partant de la droite, il faut repérer le premier grand saut de la hauteur et **s'arrêter juste avant ce saut**.

Exemple

*Voir les notes de cours : **Figure CL10**.*

Un fois choisi le nombre k de groupes, on obtient une partition particulière et on peut récupérer une variable d'affectation aux groupes pour cette partition.

Exemple

*Voir les notes de cours : **Figure CL12**.*

2. Interprétation des groupes

Comme pour l'AMM, on interprète les moyennes par groupes (uniquement pour les variables de départ dont le rapport de corrélation est suffisamment élevé).

3. Graphiques représentant les groupes

Exemple

*Voir les notes de cours : **Figure CL13**.*
Même figure obtenue en utilisant l'AMM.

Remarque

Ces deux méthodes de classification (AMM et CAH) ne donnent pas nécessairement les mêmes groupes.

Un exemple de classification mixte

- L'avantage de la CAH comparativement à la méthode d'Agrégation autour des Moyennes Mobiles AMM est qu'elle ne nécessite pas de connaître le nombre de groupes à l'avance.
- Le dendrogramme permet de nous guider dans le choix du nombre de groupes.
- Toutefois, cette méthode, contrairement à la méthode d'agrégation autour des moyennes mobiles, est très coûteuse en temps de calcul et ne peut-être utilisée si le nombre d'observations est grand.

Un exemple de classification mixte

Pour combiner les avantages des deux méthodes, on propose la stratégie suivante :

- Notons n le nombre d'observations du fichier de données (par exemple $n = 10\ 000$).
- On ne peut pas utiliser la CAH mais on peut utiliser l'AMM en choisissant un grand nombre k_{fast} de groupes (par exemple $k_{\text{fast}} = 500$).
- L'intérêt de choisir ce grand nombre de groupes est que la partition que l'on obtient n'a agrégé que les individus vraiment proches.
- Maintenant, il s'agit de considérer les 500 moyennes associées aux 500 groupes et d'exécuter une CAH sur ces moyennes.
- Avec 500 observations (qui sont en réalité des moyennes), la CAH peut maintenant fonctionner et nous pouvons choisir grâce au dendrogramme, un nombre k groupes.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Conclusion

Remarque

*Lorsque que l'on dispose de grands tableaux, on peut parfois commencer par une **ACP** et utiliser les premières composantes principales dans la classification.*

Justification de la classification non supervisée (clustering) :

- tableau de type individus / variables avec des variables quantitatives,
- recherche d'une classification des individus en vue de définir quelques typologies caractéristiques.