Analyse de données multidimensionnelles

Contents

ACP **1**
1. Données - Problématique . . . 1
2. Étude des individus . . . 2
3. Étude des variables . . . 2
4. Aides à l'interprétation . . . 2
5. Gestion des données manquantes . . . 3

AFC **3**
1. Données . . . 3
2. Modèle d'indépendance . . . 4
3. Les nuages et leur ajustement . . . 4
4. Pourcentages d'inertie et inertie en AFC . . . 4
5. Représentation simultanée des lignes et des colonnes . . . 5
6. Aides à l'interprétation . . . 5

ACM : analyse des correspondances multiples **5**
1. Données - Problématique . . . 5
2. Analyse des individus . . . 6
3. Étude des modalités . . . 7
4. Aide à l'interprétation . . . 7

Classification **8**
1. Classification ascendante hiérarchique (CAH) . . . 8
2. Exemple et choix du nombre de classes . . . 9
3. Méthode de partitionnement . . . 9
4. Compléments . . . 9
5. Caractérisation des classes d'individus . . . 10
6. Conclusion : . . . 11

AFM **11**
1. Données - Problématique . . . 11
2. Équilibre et ACP globale . . . 11
3. Étude des groupes de variables . . . 12
4. Compléments . . . 13

ACP

1. Données - Problématique

- type de données pour ACP : tableaux rectangulaires avec **individus** en lignes, et **variables quantitatives** en colonnes
- étude des individus :
  - quand dit-on que 2 individus se ressemblent du point de vue de l'ensemble des variables ?
  - si beaucoup d'individus, peut-on faire un bilan des ressemblances ?
  - construction de groupes d'individus
- étude des variables :
  - recherche de ressemblance entre variables : on parle plutôt de liaisons
  - souvent liaisons linéaires $\rightarrow$ coeff. de corrélation
  - visualisation de la matrice des corrélations

  - recherche d'un petit nombre d'indicateurs synthétiques pour résumer beaucoup de variables
- objectifs de l'ACP :
  - descriptif / exploratoire : visualisation de données par graphiques simples
  - synthèse / résumé de grands tableaux individus $\times$ variables

2. Étude des individus

- on cherche à étudier les distances entre individus au sein du nuage
- on étudie la forme du nuage réduit à 2 dimensions
- centrage - réduction des données : prétraitement
  - toujours centrer le nuage ne change pas sa forme
  - réduire les données (= standardiser, = normer) est nécessaire si les variables sont exprimées dans des unités différentes
  - la réduction permet de donner la même importance à chaque variable : plus la variance d'une variable est grande, plus cette variable aura d'importance
- l'ACP cherche un sous-espace qui résume au mieux les données (i.e. la forme originale du nuage)
- pour ajuster le nuage, on cherche l'image qui maximise la distance entre individus (qui éclate le plus le nuage), donc la variabilité sur plusieurs dimensions = l'**inertie**
  - trouver le meilleur axe (facteur) qui déforme le moins le nuage : maximiser la somme des carrés coordonnées projetées sur l'axe (on veut maximiser l'**inertie** $\sum_i (OH_i)^2$ avec $O =$ centre gravité du nuage, $H =$ coordonnée du point i sur l'axe)
  - ensuite trouver le meilleur plan qui contient le meilleur axe, avec le second axe orthogonal au premier, et maximisant toujours l'inertie $\sum_i (OH_i)^2$
  - on peut chercher d'autres axes ensuite, sachant qu'il faut toujours que le nouvel axe soit orthogonal à tous les autres et maximise l'inertie
- interprétation du graphe des individus grâce aux variables :
  - on va considérer les coordonnées des individus sur les 2 premiers axes, donnant 2 nouveaux vecteurs $F_{.1}$ et $F_{.2}$ avec I valeurs rassemblant les coordonnées des I individus
  - on calcule la corrélation entre chaque variable $x_{.k}$ et $F_{.1}$ et $F_{.2}$: le cercle des corrélations

3. Étude des variables

- on représente les variables par des flèches partant de l'origine
- quand les variables sont centrées, le cosinus de l'angle entre 2 variables k et l correspond à la corrélation entre celles-ci : $cos(\theta_{kl}) = r(x_{.k}, x_{.l})$
- si les variables sont réduites, les extrémités des flèches des variables se situent sur une hypersphère de rayon 1
- comme pour les individus, on cherche les axes orthogonaux qui maximisent la somme des carrés des corrélations entre l'axe et chacune des variables
- **attention** : seules les variables bien projetées (avec une flèche proche sur cercle des corrélations) peuvent être interprétées quant à leur corrélation

4. Aides à l'interprétation

- Pourcentage d'inertie (d'information) expliqué par chaque dimension
- Information supplémentaire : variables quanti. ou quali. supplémentaires
  - on peut directement superposer les variables quantitatives
  - on va projeter chaque modalité des variables qualitatives au barycentre des individus qui prennent cette modalité
- Qualité de représentation : les éléments bien projetés sont interprétables
  - d'une variable : cos^2 entre une variable et sa projection
  - d'un individu : cos^2 entre O_i et OH_i
  - on additionne les cos^2 calculés sur les 2 premiers axes pour avoir la qualité de la projection sur le premier plan

- Contribution à la construction d'un axe :
  - contribution d'une variable : carré de la corrélation entre la variable et l'axe par rapport à la somme des carrés des corrélations de toutes les variables avec l'axe (exprimé en pourcentage)
  - contribution d'un individu : carré de sa coordonnée sur l'axe par rapport à la somme des carrés des coordonnées de tous les individus (exprimé en %)
  - la contribution permet de savoir si un axe a été construit surtout à cause d'une certaine variable ou d'un certain individu
  - si un individu contribue très fortement, on peut chercher à expliquer pourquoi, et à refaire l'analyse en retirant cet individu très particulier
- Description des dimensions :
  - par les variables quantitatives :
    * calcul des corrélations entre chaque variable et la dimension s
    * tri des coefficients de corrélation significatifs
  - par les variables qualitatives :
    * analyse de variance des coordonnées des individus sur l'axe s expliqués par la variable qualitative
    * 1 test F par variable pour voir si la liaison est significative entre les coordonnées des individus sur l'axe et la variable
    * 1 test t de Student par modalité pour comparer le moyenne de la modalité avec la moyenne générale des modalités

5. Gestion des données manquantes

- package `missMDA`
- ne pas faire :
  - supprimer les NA (on élimine de l'information)
  - imputation par la moyenne des données présentes (entraîne une distorsion des liaisons entre variables)
- ACP itérative
  - si 2 individus x et y se ressemblent sur l'ensemble des variables, on peut imputer une donnée manquante pour une variable de l'individu x par la valeur prise par l'individu y pour cette même variable
  - pour l'ACP itérative, on impute la valeur manquante par la moyenne prise par tous les individus pour la variable, puis on fait une 1ère ACP sur ce tableau complété, on met à jour la valeur manquante par la coordonnée de sa projection sur la droite d'ACP, et on réitère le processus jusqu'à convergence
  - faiblesse de la méthode : surajustement possible si on a une trop grande croyance dans la liaison entre les variables
  - on utilise alors l'**ACP itérative régularisée**
  - conséquences : on renforce les liaisons entre variables, donc on **surestime les % d'inertie** associés aux premières dimensions de l'ACP
- imputation multiple :
  - on impute plusieurs fois le tableau en générant chaque fois une valeur plausible mais différente de chaque valeur manquante
  - cela va ajouter une incertitude sur l'imputation
  - se visualiser par une ellipse autour des individus et des variables pour lesquels on a des données manquantes qui ont été imputées
  - fonction `MIPCA` dans `missMDA`

AFC

1. Données

- tableau de contingence ou tableau croisé : n individus représentés par 2 variables qualitatives V_1 et V_2 dont les modalités sont représentées en colonnes pour l'une et lignes pour l'autre. A l'intersection de la ligne i et de la colonne j, on trouve le nombre d'individus concernés par à la fois la modalité i de V_1 et la modalité j de V_2

- le tableau de contingence est transformé en tableau de probabilités : $f_{ij} = \frac{x_{ij}}{n}$
- ainsi pour couple de modalités (i, j) on a la probabilité d'obtention de ce couple
- les probabilités marginales s'obtiennent par la somme en ligne ou colonne des probabilités
- la liaison entre V_1 et V_2 est l'écart entre les données et le **modèle d'indépendance**

2. Modèle d'indépendance

- rappel : 2 événements indépendants : $P(A \cap B) = P(A) \times P(B)$
- si V_1 et V_2 sont indépendantes, alors $f_{ij} = f_{i.} \times f_{.j}$: la probabilité conjointe est le produit des probabilités marginales
- autrement dit $\frac{f_{ij}}{f_{i.}} = f_{.j}$ et $\frac{f_{ij}}{f_{.j}} = f_{i.}$: la probabilité conditionnelle = probabilité marginale
- on fait un test du χ^2
- $\chi^2 = n\phi^2$: intensité de la liaison = écart entre probabilités théoriques et observées
- on a une information sur la nature de la liaison (association entre modalités), mais pas sur la significativité
- l'AFC va par exemple comparer les "profils lignes" de chaque modalité i (= distribution conditionnelle de la variable V_2 sachant que l'on possède la modalité i de V_1) avec le "profil ligne moyen" (= distribution marginale V_2, profil de l'ensemble des individus étudiés)

3. Les nuages et leur ajustement

- nuage des profils lignes : I points de coordonnées $f_{ij}/f_{i.}$ sur l'axe j dans un espace à J dimensions
- nuage des profils colonnes : J points de coordonnées $f_{ij}/f_{.j}$ sur l'axe i dans un espace à I dimensions
- si indépendance entre les lignes : pour tout i on a $f_{ij}/f_{i.} = f_{.j} \Rightarrow$ les profils lignes sont confondus avec le profil moyen, le nuage N_I se réduit au point G_I (centre de gravité, origine) $\Rightarrow$ l'inertie du nuage est nulle
- ainsi, plus les données s'écartent de l'indépendance, plus les profils s'écartent de l'origine
- l'inertie du nuage N_I par rapport à son centre de gravité G_I est calculé par le ϕ^2 (= χ^2/n), qui est donc bien une mesure de l'intensité de la liaison - idem pour le nuage N_J
- on a une dualité importante : Inertie(N_J/G_J) = Inertie(N_I/G_I) c'est-à-dire que les lignes et les colonnes jouent des rôles symétriques
- l'AFC cherche à décomposer l'inertie du nuage N_I par analyse factorielle : elle fait une projection du nuage sur une série d'axes orthogonaux d'inertie maximum
- visualisation sur un graphe des 2 premières dimensions :
  - un point au centre du graphe (intersection des axes) est une ligne / colonne dont le profil est proche du profil ligne/colonne moyen
  - l'interprétation de la projection des lignes et des colonnes se fait de la même manière sur les axes, en vertu de la dualité

4. Pourcentages d'inertie et inertie en AFC

- qualité de représentation de N_I par un axe : inertie projetée sur cet axe / inertie totale
- en général exprimée en % = **valeur propre**
- les % d'inertie mesurent l'écart à l'indépendance
- les % d'inertie projetées s'additionnent, la somme totale = inertie totale de N_I = ϕ^2
- $n\phi^2 = \chi^2 \Rightarrow$ comparer au χ^2 théorique pour trouver la probabilité critique
- on peut faire un histogramme des valeurs propres et observer la décroissance en fonction du rang de l'axe, ce qui permet de savoir combien d'axes conserver
- si une valeur propre = 1, on a parfaite opposition, 2 catégories mutuellement exclusives
- en AFC, il faut d'abord regarder les valeurs propres
- quand une valeur propre d'un axe est très inférieure à 1, on est loin d'une association exclusive entre une ligne et une colonne
- quand ϕ^2 est très inférieur au nombre d'axes (qui est le maximum de la somme des inerties des axes si chacune valait 1), on est loin d'une liaison parfaite (*i.e.* loin d'une association exclusive) entre les modalités des 2 variables

5. Représentation simultanée des lignes et des colonnes

- relation de transition = propriétés barycentriques
- une ligne est du côté des colonnes auxquelles elle s'associe le plus
- une colonne est du côté des lignes auxquelles elle s'associe le plus
- attention : on peut comparer les distances des points lignes entre eux ou des points colonnes entre eux, mais pas entre points lignes et points colonnes (seule la direction est interprétable)

6. Aides à l'interprétation

- **qualité de représentation** d'un point (ou du nuage) : inertie projetée sur l'axe / inertie totale = $\cos^2$
- pour démarrer l'interprétation d'une AFC, on commence par les points remarquables : éloignés du centre du centre le long d'un axe, avec une bonne qualité de représentation sur cet axe
- **contribution** d'un point : inertie projetée du point sur l'axe (brute, ou en % de l'inertie totale de l'axe)
- on peut additionner les les contributions de plusieurs point sur un axe pour savoir à quel point cet axe est déterminé par ces points
- pour de grands tableaux, on peut sélectionner un sous-ensemble d'éléments à interpréter s'ils ont une forte contribution (conjointement à leur qualité)
- important : si dans le tableau on a des effectifs marginaux très différents d'une ligne (ou colonne) à une autre, il faudra impérativement examiner les contributions (les points extrêmes ne sont pas forcément les plus contributifs)
- on peut visualiser des lignes ou colonnes supplémentaires
- propriété intéressante de l'AFC, l'**équivalence distributionnelle** : si plusieurs lignes (ou colonnes) ayant le même profil sont regroupées en une seule, les résultats de l'AFC sont strictement équivalents
- on peut utiliser cette propriété et regrouper sous un terme unique 2 termes strictement équivalents
- le nombre maximum d'axes d'inertie non nulle est égal au minimum de ($I-1$, $J-1$) ; de même $\phi^2 < min(I-1, J-1)$
- **V de Cramer** : $V = \frac{\phi^2}{min(I-1,J-1)}$ (compris entre 0 et 1), indicateur borné de la liaison entre les 2 variables

ACM : analyse des correspondances multiples

1. Données - Problématique

- tableau de données rectangulaire : I individus en lignes J variables qualitatives en colonnes v_{ij} = modalité de la variable j possédée par l'individu i
- exemple : enquête sur I personnes, avec J questions pouvant prendre des modalités distinctes (ex : j = statut marital, modalités = {célibataire, marié, veuf, divorcé,…}
- à partir du tableau de données, on construit un **tableau disjonctif complet** ("TDC") : toujours I lignes, mais les modalités sont "splittées" en colonnes (si K modalités, on a $K \times J$ colonnes au total) ; dans le tableau, on met "1" si l'individu possède la modalité, "0" s'il ne la possède pas
- une colonne est appelée **"fonction indicatrice"**, ou **"indicatrice"**
- à partir du TDC, on calcule la marge-colonne pour chaque individu : elle vaut J ; la somme totale du tableau vaut IJ
- étude des individus :
  - 1 individu = 1 ligne = ensemble de ses modalités
  - ressemblance des individus
  - variabilité des individus
  - principales dimensions de la variabilité des individus (en relation avec les modalités)
- étude des variables :
  - liaisons entre variables qualitatives (en relation avec les modalités)
  - visualisation d'ensemble des associations entre modalités
  - variable synthétique (chercher un indicateur quantitatif fondé sur des variables qualitatives)
- problématique proche de l'ACP

- on a parfois 2 types de variables dans un même tableau, par exemple : variables liés à l'activité (cinéma, restaurant…) + variables signalétiques (âge, sexe de l'individu)
- dans ce cas on peut souhaiter traiter les variables "signalétiques" comme des variables supplémentaires, ou inversement, ou traiter les 2 en actif
- dans le premier cas, un individu est représenté uniquement par son profil d'activité, son poids est de $1/I$
- la valeur d'une cellule du TDC : $y_{ik} = 1$ ou 0
- si un individu possède une modalité k rare, cela le caractérise beaucoup plus qu'une modalité fréquente, et inversement si une modalité est partagée par tous les individus, elle n'en caractérise aucun, d'où : $x_{ik} = y_{ik}/p_k$, qui est équivalent à la donnée centrée réduite de l'ACP

2. Analyse des individus

- nuage des lignes = des individus
- 1 individu = 1 point représenté dans un espace à K dimensions (il y a K valeurs, car il y a K modalités)
- chaque modalité a un poids proportionnel à son effectif = p_k/J
- l'individu est représenté par un point M_i de coordonnées x_{ik} et a un poids de $1/I$
- l'ensemble des points individus forme un nuage N_I de centre de gravité $G_I = O$ (origine)
- la distance entre 2 individus est la somme des carré des écarts sur chaque axe (modalité) pondérée par le poids de la modalité correspondante, p_k/J
- si 2 individus prennent les mêmes modalités, ils ont le même profil : distance = 0
- s'ils ont en commun beaucoup de modalités, leur distance sera petite
- s'ils ont en commun beaucoup de modalités, mais qu'il diffèrent sur 1 modalité relativement rare, alors la distance sera plus grande, pour refléter cette spécificité
- si 2 individus ont en commun une modalité rare, la distance sera plus petite pour refléter cette spécificité commune
- inertie d'un individu = carré de la distance au centre de gravité pondéré par son poids
- l'inertie du nuage N_I = somme des inerties des individus $= (K/J) - 1$
- cette inertie totale dépend donc uniquement du format du tableau (nombre de modalités et de variables), contrairement à l'analyse des correspondances (AFC) où l'inertie mesurée par le ϕ^2 mesure l'écart à l'indépendance ; elle se rapproche davantage de l'inertie en ACP qui dépend uniquement du nombre de variables
- on ajuste le nuage des individus par projection séquentielle sur des axes orthogonaux d'inertie maximale
- on commence par regarder la décroissance des inerties
- regarder le nuage pour voir s'il y a des représentations particulières :
  - regarder si on voit des groupes distincts
  - un nuage en "V" ("effet Guttman") : souvent le premier axe ségrège les individus selon une valeur allant de la - à la + élevée, et le second axe oppose les individus les plus extrêmes (les - et le +) aux individus moyens
  - colorier les individus selon certaines modalités (ex. : jardinage oui/non) pour voir s'ils ont tendance à se répartir selon cette modalité
- on peut visualiser les **modalités** sur le graphe des individus en les plaçant au barycentre des individus qui les possèdent ; on peut interpréter la distance des modalité au point d'origine (barycentre du nuage) en fonction de la taille des effectifs : moins une modalité est possédée par des individus, plus elle s'éloigne du point d'origine
- représentation des **variables** : l'idée est de considérer les coordonnées projetées des individus sur un axe et de calculer un indicateur de liaison entre ces coordonnées et chaque variable qualitative :
  - on calcule le rapport de corrélation entre la variable j et la composante s : $\eta(v_{.j}, F_s)$
  - on examine le carré η^2 qui donne le pourcentage de variabilité de la variable quantitative expliquée par la variable qualitative
  - on fait un graphe de ces valeurs pour les variable = "graphe du carré des liaisons"
  - ce graphe s'inscrit dans un carré (0,1) sur les 2 premières dimensions
  - attention, on ne peut pas interpréter la significativité numérique car la construction de l'analyse se fait en maximisant les les corrélations (sauf pour les variables supplémentaires) : l'axe s est orthogonal à tout axe t, et est l'axe le plus le plus lié aux variables qualitatives au sens du η^2

3. Étude des modalités

- nuage des modalités = nuage des colonnes
- ensemble de I valeurs numériques représentées par un point M_k dans un espace à I dimensions
- la variance d'une modalité k est le carré de la distance entre cette modalité et le point d'origine qui est le centre de gravité du nuage N_k : $Var(k) = d^2(k, O) = ... = (1/p_k) - 1$
- ainsi, la distance entre un point M_k et l'origine est d'autant plus grande que cette modalité est rare : en ACM, les modalités rares sont plus éloignées de l'origine
- attention, le poids diminue avec la rareté, il y a antagonisme poids / distance
- ce qui compte en ACM, c'est l'inertie d'une modalité $k = (1 - p_k)/J$
- quand une modalité est rare, elle a une forte inertie et va influencer les résultats de l'ACM
- une modalité rare (1/10) aura autant d'influence qu'une modalité très rare (l'inertie diminue fortement au début, mais de moins en moins avec la rareté)
- plus il y a d'individus qui possèdent une seule des 2 modalités, plus la distance entre ces 2 modalités est grande
- inertie d'une modalité : Inertie$(k) = (1 - p_k)/J$
- inertie d'une variable : Inertie$(j) = (K_j - 1)/J$ (elle est proportionnelle au nombre de modalités de la variable)
- principe de double propriété barycentrique : si on dilate les nuages des modalités (ou des individus) d'un facteur $1/\sqrt{\lambda}$, on obtient une représentation optimale à la fois des individus et des modalité : un individu est du côté des modalités qu'il possède, et réciproquement une modalité est du côté des individus qui la possèdent

4. Aide à l'interprétation

- l'**inertie d'un axe** en ACM : $\lambda_s = 1/J \times \Sigma(\eta^2(F_s, v_{.j}))$ = moyenne des carrés des rapports de corrélation (donc toujours comprise entre 0 et 1)
- ainsi, l'inertie d'un axe **mesure la relation entre cet axe et les variables considérées**
- souvent les inerties sont plus grandes qu'en ACP ou AFC, car les individus évoluent dans un espace de plus grandes dimensions ($dim = K - J$)
- le pourcentage d'inertie maximum sur une dimension s est inférieur ou égal à $J/(K - J) \times 100$
- la moyenne des valeurs propres non nulles est $= 1/J$
- donc, on va interpréter uniquement les dimensions dont l'inertie est $> 1/J$
- comme il y a beaucoup de dimensions, la qualité des projections (cos^2) des individus tend à être relativement faible
- les contributions des modalités doivent être regardées dans le tableau des contributions, et ne sont pas déduites du graphe : une contribution importante ne correspond pas forcément à une position extrême sur le graphe (car une modalité est associée à un poids qui dépend de sa fréquence)
- la contribution d'une variable est la somme des contributions de toutes ses modalités = rapport de corrélation entre l'axe et la variable J
- la contribution relative d'une variable : $\frac{\text{contribution de la variable}}{\text{inertie de l'axe } \lambda_s}$
- utilisation des relations de transition pour les éléments (individus, modalités) supplémentaires
- en ACM on travaille avec des données qualitatives, donc les variables quantitatives sont forcément traitées en variables supplémentaires
- parfois on peut calculer une variable quantitative à partir du tableau, à partir d'une hypothèse, et l'utiliser comme variable quantitative supplémentaire à analyser, analysée sur le graphe des corrélations comme en ACP
- on peut aussi discrétiser une variable quantitative en classes, et l'analyser en variable qualitative active
- description des dimensions : par les variables qualitatives actives (test de Fisher), les modalités (test de Student) et les variable quantitatives (corrélation)
- tableau de Burt : tableau qui croise toutes les modalités et variables entre elles 2 à 2 (analogue à une matrice des corrélations entre variables quantitatives)
- on peut réaliser une AFC sur le tableau de Burt : valeurs propres de Burt = (valeurs propres du TDC)^2
- l'ACM ne dépend que des liaisons entre variables prises 2 à 2 (analogie avec l'ACP qui ne dépend que de la matrice des corrélations)

- conclusion : l'ACM pour analyser tableaux individus $\times$ variables qualitatives
- c'est une méthode très générale, notamment utile pour analyser des enquêtes
- interprétation des valeurs propres comme une moyenne des rapports de corrélation au carré : les facteurs de l'ACM sont des variables quantitatives qui synthétisent les variables qualitatives
- il peut être utile de valider les interprétations en revenant aux données initiales : l'ACM suggère des liaisons entre des variables qualitatives,on peut alors construire un tableau croisé de 2 variables qu'on analyse par AFC
- l'ACM peut être vue comme un pré-traitement en vue d'une analyse de classification

Classification

1. Classification ascendante hiérarchique (CAH)

- construire des groupes (ou classes) d'individus ayant des traits ou caractères communs
- 2 types de classification :
  - **hiérarchique** : arbre, CAH
  - méthode de **partitionnement** : partition
- tableau de données : I individus en lignes $\times K$ variables quantitatives en colonnes
- l'objectif est de produire une structure (arborescence) afin de :
  - mettre en évidence des liens hiérarchiques entre individus ou groupes d'individus
  - détecter un nombre de classes "naturel" au sein de la population
- critères de classification : ressemblance entre individus
  - distance euclidienne
  - indices de similarité (Jacquard, …)
- on s'intéresse aussi à la ressemblance entre groupes :
  - saut minimum ou lien simple (plus petite distance)
  - lien complet (plus grande distance)
  - critère de Ward
- algorithme :
  - on part de 8 points individuels, on calcule la distance euclidienne entre chaque individu 2 à 2
  - on regarde quelle est la distance la plus petite
  - on regroupe ces individus en 1 nouveau groupe
  - on calcule la distance entre chaque ce groupe et chaque autre individu, en utilisant la mesure du **saut minimum** (plus petite distance entre un des membres du nouveau groupe et un autre individu)
- une fois un arbre réalisé, on peut définir un niveau de coupure (distance) pour construire une partition
- critères de qualité d'une partition :
  - quand les individus d'une même classe sont proches = variabilité intra-classe réduite
  - quand les individus de 2 classes différentes sont éloignés = variabilité inter-classes élevée
  - or, inertie totale = inertie intra-classe + inertie inter-classe : théorème de Huygens (avec inertie = somme des carrés des écarts)
  - donc pour améliorer la qualité, il suffit de s'intéresser à un seul des 2 critères : minimiser la variabilité intra va mathématiquement maximiser la variabilité inter, et réciproquement
  - la qualité de la partition peut se mesurer au final par le rapport (inertie inter) / (inertie intra) qui sera compris entre 0 et 1
  - le rapport vaut 0 quand les classes ont les mêmes moyennes pour toute variable k, on ne peut donc pas classifier
  - le rapport vaut 1 si tous les individus d'une classe sont indentiques pour toute variable k, c'est donc idéal pour classifier
  - attention, ce critère dépend du nombre d'individus et du nombre de classes, sa valeur ne peut pas être jugée dans l'absolu
- méthode de Ward :
  - on commence par établir 1 classe par individu : inertie inter-classe = inertie totale
  - à chaque étape, on agrège les classe a et b qui minimisent la diminution de l'inertie inter-classe
  - permet d'éviter l'effet de chaîne en regroupant les objets de faible poids et en regroupant des classes

ayant des centres de gravité proches

2. Exemple et choix du nombre de classes

- exemple des températures mensuelles moyennes sur 30 ans (12 variables) pour 15 villes de France + latitude et longitude en données supplémentaires
- questions : quelles villes ont des profils similaires ? comment caractériser des groupes de villes ?
- sur le graphique on voit l'histogramme des inerties, montrant la perte d'inertie si on regroupe des classes
- la somme des pertes d'inertie = inertie totale
- on peut calculer le pourcentage $\frac{\text{inertie inter}}{\text{inertie totale}} \times 100$ pour voir quel pourcentage d'information est expliqué par les classes ainsi faites
- il faut s'assurer que l'on peut interpréter les regroupements en classes

3. Méthode de partitionnement

- méthode directe : k-means = "méthode d'agrégation autour des centres mobiles"
- algorithme :
  - on définit arbitrairement le nombre Q de classes souhaité
  - on choisit au hasard Q individus, puis on attribue à ces classes les individus qui en sont le plus proche
  - on calcule le barycentre des Q classes
  - puis on réitère l'affectation des individus les proches au barycentre et on recalcule le barycentre
  - quand l'algorithme a convergé, on a trouvé les Q classes optimales
- 2 défauts :
  - il faut connaître a priori le nombre de classes Q
  - la partition finale est sensible à l'initialisation (solution : lancer plusieurs fois l'algorithme et garder la meilleure partition)

4. Compléments

- *Consolidation d'une partition obtenue par CAH*
  - on utilise la partition obtenue par CAH pour initialiser le partitionnement par k-means
  - avantage : on consolide le partitionnement car il minimise l'inertie
  - inconvénient : on perd l'information de hiérarchie (certains individus peuvent changer de classe lors du partitionnement k-means)
- *CAH en grandes dimensions :*
  - si trop grand nombre de variables :
    - la CAH devient difficile et plus longue à calculer (car besoin de calculer de nombreuse distances)
    - on peut faire une ACP pour ne garder que les premières dimensions
    - on réalise alors la CAH sur un tableau individus $\times$ dimensions factorielles (quantitatives, de taille plus raisonnable)
  - si trop grand nombre d'individus :
    - l'algorithme de CAH est trop long et peut ne pas aboutir
    - faire une partition par k-means en une centaine de classes
    - construire la CAH à partir de ces classes (en utilisant l'effectif des classes dans le calcul)
    - on obtient ainsi le "haut" de l'arbre hiérarchique
    - (en général si le nombre d'individus est très grand, il est inutile de commenter le bas de l'arbre)
- *CAH sur des variables qualitatives : 2 stratégies*
  1. se ramener à des variables quantitatives :
  - faire une ACM en ne conserver que les premières dimensions
  - faire la CAH à partir des composantes principales de l'ACM
  2. utiliser des mesures adaptées aux données qualitatives : indice de similarité, de dissimilarité de Jacquard, etc.
  - on obtient une matrice des distances
  - on construit la CAH sur cette matrice

- *Enchaînement analyse factorielle - classification*
  - données qualitatives : l'ACM renvoie des composantes principales qui sont quantitatives, sur lesquelles ont peut faire directement une classification
  - l'intérêt de réaliser une analyse factorielle en préalable à une classification, c'est de concentrer l'information sur les premières composantes, et éliminer les dernières qui contiennent le bruit $\rightarrow$ classification plus stable
  - on peut aussi visualiser l'arbre et les classes sur un plan factoriel, donnant une vue plus complète de l'information

5. Caractérisation des classes d'individus

- constitution des classes : édition des **parangons** (parangon = individu le plus proche du centre d'une classe)
- caractérisation des classes par les variables : *1.* trouver les variables les plus caractérisantes pour la partition, *2.* caractériser une classe (ou un groupe d'individus) par des variables quantitatives, et *3.* trier les variables qui caractérisent le mieux les classes
  1. *variables caractérisant le mieux la partition :*
  - on peut considérer la partition comme une variable qualitative $=$ variable de classe, avec autant de modalités qu'il y a de classes
  - pour chaque variable quantitative (colonne) : construire un modèle d'analyse de variance entre la variable quantitative expliquée et la variable de classe, puis faire un test de Fisher de l'effet de la classe
  - on garde les variables ayant une probabilité critique < 5%, et trier ces variables en fonction de leur probabilité critique
  - exemple des données villes / températures mensuelles : la variable "octobre" est celle qui caractérise le mieux les classes avec $P = 1.93e-05$
  - mais attention : comme ces variables sont actives et ont donc participé à la construction des classes, il faut interpréter ces probabilités avec prudence
  2. *caractériser une classe par les variables quantitatives :*
  - représenter graphiquement les variables colonnes en fonction de la variable quantitative et colorer les points en fonction des classes
  - *idée 1* : si les valeurs de la variable X pour la classe q semblent tirées au hasard parmi les valeurs de X, alors X ne caractérise pas la classe q
  - *idée 2* : plus l'hypothèse d'un tirage aléatoire est douteuse, plus X caractérise q
  - construire un test : tirage au hasard de n_q valeurs parmi N, quelles valeurs peut prendre $\overline{x}_q$ (i.e. quelle est la loi de $\overline{X}_q$?
  - valeur-test $= \frac{\overline{x}_q - \overline{x}}{\text{écart type de } (\overline{x}_q)}$ suit une loi Normale centrée réduite $\mathcal{N}(0,1)$
  - si |valeur-test| > 1.96 alors elle vient pas d'une loi Normale et la variable X caractérise donc la classe q
  - plus |valeur-test| est grande, mieux la variable X caractérise la classe q
  - on peut classer les variables par |valeur-test| décroissante et chercher les valeurs soit les plus positives (moyenne dans classe supérieure à la moyenne générale) ou la plus négative (moyenne dans la classe plus basse que la moyenne générale)
  3. *caractériser des classes par les variables qualitatives :*
  - pour chaque variable qualitative, construire un test du χ^2 entre la variable et la variable de classe
  - trier les variables par probabilité critique croissante
  - si une variable qualitative caractérise significativement une classe, on peut ensuite voir si elle se caractérise par une modalité particulière de cette variable qualitative
  - pour ce faire, on construit un tableau réduit avec les effectifs de cette classe prenant la modalité ou non, et les effectifs des autres classes, et on compare les proportions
  - la variable aléatoire calculée suit une loi hypergéométrique et permet de mesurer si elle dévie significativement de l'hypothèse H_0 de non sur-représentation de la modalité dans la classe
  - on a alors une nouvelle valeur test centrée réduite : plus la valeur-test est $>$ (ou $<$) à 1.96, plus on a une sur- (ou sous) représentation de la modalité dans la classe considérée, et donc plus cette modalité caractérise la classe
  4. *on peut caractériser une classe par les axes factoriels :*

- les axes sont des variables quantitatives, donc même méthodologie que 2. (dans le tableau d'analyse, la moyenne générale est toujours égale à 0 car les données ont été centrées-réduites sur les axes)

6. Conclusion :

- la classification s'applique aux tableaux individus $\times$ variables quantitatives $\rightarrow$ l'ACM permet de transformer des variables qualitatives en variables quantitatives (dimensions factorielles)
- la CAH donne un arbre hiérarchique qui montre les distances en individus et groupes d'individus, et une idée du nombre de classes
- on peut alors utiliser une méthode partitionnement comme le k-means pour consolider les classes
- le partitionnement peut être utilisé pour pré-traiter un jeu de données de grandes dimensions
- on peut caractériser les classes obtenues par des variables actives et supplémentaires, quantitatives ou qualitatives

AFM

1. Données - Problématique

- exemple : 10 vins de Loire en individus lignes, descripteurs sensoriels comme variables colonnes (dans chaque cellule, la note moyenne attribuée par des juges pour chaque vin et chaque variable) et une dernière colonne avec la variable cépage (qualitative : Vouvray ou Sauvigon)
- on pourrait faire une analyse par ACP, avec le cépage en variable qualitative supplémentaire
- cependant, il y a 3 catégories de juges : experts, consommateurs, étudiants
- questions : comment caractériser les vins ? sont-ils décrits de la même façon par les différentes catégories de juges ? y a-t-il des spécificités par catégorie de juges ?
- on peut ajouter une variable supplémentaire qui est l'appréciation globale de chaque vin par 60 consommateurs (note hédonique)
- structure du tableau de données :
  - I vins (individus) en lignes
  - J variables quantitatives et/ou qualitatives formant des groupes de K_j variables
  - la valeur prise dans une cellule du tableau est notée x_{ik}
- autres types d'exemples : génomique (mesures d'ADN, d'expression, de protéines), enquêtes avec plusieurs groupes de variables, économie (indicateurs économiques chaque année), …
- objectifs :
  - étudier les ressemblances entre individus du point de vue de l'ensemble des variables **et** les relations entre variables
  - étudier globalement les ressemblances/différences entre groupes de variables
  - étudier les individus pour voir s'ils sont décrits de la même manière par différents groupes de variables (ex : un vin est-il décrit de la même manière par plusieurs catégories de juges)
  - comparer les typologies issues des analyses séparées
- **important** : il faut veiller au bon équilibre d'influence de chaque groupe de variable dans l'analyse

2. Équilibre et ACP globale

- en ACP, le fait de normer équilibre l'influence de chaque variable (dans le calcul des distances entre individus i et i')
- en AFM, on équilibre les groupes
- *idée 1* de pondération : diviser chaque variable par l'inertie totale du groupe auquel elle appartient (ainsi, un groupe ayant moins de variables aura la même influence qu'un groupe qui aura plus) ; mais problème : cette méthode est sensible à la répartition des inerties au sein de chaque groupe
- *idée 2* : pondérer en fonction de l'inertie maximum d'une dimension, en divisant chaque variable par la (racine carrée de la) 1ère valeur propre du groupe auquel elle appartient
- l'AFM est un ACP pondérée :
  - calculer la 1ère valeur propre λ_1^j du groupe de variables j (avec $j = 1, ..., J$)

  - réaliser l'ACP globale sur le tableau pondéré où chaque valeur est divisée par la racine carrée de la valeur propre
- on donné un même poids à toutes les variables d'un même groupe, on préserve ainsi la structure (*i.e.* l'équilibre des variables) du groupe
- pour chaque groupe, la variance de la principale dimension (1ère valeur propre) est égale à 1
- aucun groupe ne peut donc générer à lui seul la première dimension
- un groupe multidimensionnel contribue à plus de dimensions qu'un groupe unidimensionnel
- les représentations graphiques sont les mêmes qu'en ACP :
  - étudier les ressemblances entre individus du point de vue de l'ensemble des variables
  - étudier les relations entre variables (graphe des corrélations)
  - décrire les individus à partir des variables
- mêmes sorties qu'en ACP (coordonnées, cosinus, contributions)
- ajout d'individus et variables supplémentaires (quantitatifs et qualitatifs)

3. Étude des groupes de variables

- la première composante principale de l'AFM est la variable v_1 qui maximise la liaison avec tous les groupes (au sens de l'inertie projetée de toutes les variables du groupe j sur v_1, $\mathcal{L}_g$)
- cette inertie $\mathcal{L}_g$ vaut 0 si toutes les variables du groupe K_j sont non-corrélées à v_1, et vaut 1 si la variable v_1 est confondue avec la 1ère composante principale du groupe K_j
- on peut représenter les groupes en utilisant cette mesure de liaison $\mathcal{L}_g$ entre les groupes et les 2 premières composantes principales v_1 et v_2, ce qui donne un graphique avec des axes x et y de coordonnés sur un plan $[0, 1]$
- un groupe j a pour coordonnées $\mathcal{L}_g(K_j, v_1)$ et $\mathcal{L}_g(K_j, v_2)$
- ce graphe donne une représentation synthétique des groupes : les positions relatives des individus sont-elles plus ou moins similaires d'un groupe à l'autre
- mesure de similarité entre groupes : $\mathcal{L}_g$ entre 2 groupes de variables
- mesurer le $\mathcal{L}_g$ entre un groupe de variables et lui-même, cela revient à mesurer la dimensionnalité de ce groupe
- comme $\mathcal{L}_g$ entre 2 groupes n'est pas borné, on peut le diviser par la dimensionnalité de chacun des 2 groupes pour obtenir le critère RV compris entre 0 et 1
- $RV(K_j, K_m) = 0$ si toutes les variables de K_j et K_m sont non-corrélées, ou $= 1$ si les 2 nuages de point sont homothétiques
- on peut aussi calculer ces indices pour le groupe "total" MFA et voir comment chaque groupe de variables se rapproche de la moyenne
- représentation des points partiels : il s'agit de représenter chaque individu uniquement par les variables d'un seul groupe
- dans ce cas on obtient un nuage avec autant de points que de groupes (par exemple si on a 3 groupes, un individu sera représenté par un nuage de 3 points correspondant chacun à la vue de l'individu par un des 3 groupes) et avec au barycentre le point global de l'individu
- pour que la représentation soit bonne, il faut en revanche dilater ce nuage de points en le divisant par le nombre de groupes J
- les points partiels permettent d'observer la concordance ou discordance de représentation des individus par les groupes de variables
- Les relations de transition de l'ACP s'appliquent aux points moyens en AFM, pour l'interprétation des individus grâce au graphe des variables : un individu se retrouve du côté des variables pour lesquelles il prend de fortes valeurs
- ratio d'inertie $= \frac{\text{inertie inter-individus sur l'axe } s}{\text{inertie totale sur l'axe } s}$ (avec inertie totale $=$ inertie inter. $+$ inertie intra.)
- permet d'estimer si les points partiels sont globalement proches des points moyens, dimension par dimension : si le ratio est proche de 1, alors les coordonnées des points partiels tendent à être proches des coordonnées du point moyen correspondant sur cette dimension
- l'inertie intra-individu peut aussi être décomposée par individu : on trie les individus par inertie intra. décroissante pour voir quel individus sont vus de manière moins concordante par les différents groupes de variables

4. Compléments

- Si on a des groupes de variables toutes qualitatives, même principe : on cherche à équilibrer les groupes dans une analyse globale
  - on réalise des graphes spécifiques de l'ACM pour analyser les différences / similitudes entre groupes
  - le graphe des individus, comme en ACM, contient les modalités qui se retrouvent au barycentre des individus qui la prennent
  - ensuite graphes spécifiques : groupes, représentation superposée, axes partiels des analyses séparées
- pour un tableau de données mixtes (groupes de variables quantitatifs et groupes de variables qualitatifs), l'AFM fonctionne comme une ACP pour les quantitatives, et comme une ACM pour les qualitatives
- la pondération de l'AFM permet d'analyser les deux types de variables conjointement
- cas particulier si chaque groupe est composé d'une seule variable, on réalise une AFDM (Analyse Factorielle de Données Mixtes) :
  - en AFDM, on fait le graphe des individus et des modalités de l'ACM, le graphe du cercle des corrélations de l'ACP, et le graphe du carré des liaisons de l'ACM
- l'AFM peut être utilisée sur des tableau de contingence (AFMTC) par exemple :
  - extension à l'analyse textuelle
  - enquête dans plusieurs pays (CSP $\times$ questions par pays, avec nombre de personnes ayant répondu oui à une question) chaque pays représente 1 tableau de contingence
  - en écologie (sites $\times$ espèces par année, avec la fréquence de l'espèce j dans le site i) chaque année représente 1 tableau de contingence
- on peut ajouter des individus ou variables (ou groupes de variables) supplémentaires
- indicateurs : contribution et qualité de représentation
  - pour individus et variables, mêmes calculs qu'en ACP
  - pour un groupe de variables, sa contribution à un axe est sa coordonnée sur l'axe divisée par la somme des coordonnées, $\times$ 100 (contribution relative)
  - qualité de représentation d'un groupe : cos^2 entre le point et sont projeté sur l'axe (on peut faire la somme des cos^2 pour la contribution dans un plan ou sous-espace
- description des dimensions :
  - par des variables quantitatives :
    - calcul de la corrélation entre chaque variable et la composante principale
    - tri des coefficients de corrélation pour ne conserver que les significatifs
  - par des variables qualitatives :
    - construire une analyse de variance avec les coordonnées des individus expliqués par la variable qualitative puis faire un test F par variable
    - pour chaque catégorie, on réalise un test de Student