Ingénierie de la fouille et de la visualisation de données massives (RCP216)

Réduction du volume de données

Michel Crucianu
(prenom.nom@cnam.fr)
http://cedric.cnam.fr/vertigo/Cours/RCP216/

Département Informatique
Conservatoire National des Arts & Métiers, Paris, France

23 septembre 2024

Plan du cours

Méthodes de réduction de la complexité

A. Réduction du volume de données (sans changer l'ordre de complexité)

1. Calculs sur échantillon : N données $\rightarrow$ échantillon de $n \ll N$ données
2. Réduction de dimension : réduire fortement le nombre de variables tout en conservant au mieux l'« information utile »

- Les résultats sont en général des *approximations* de ceux obtenus sur les données complètes
- Difficultés si les données sont à faible densité d'information :
  - Échantillon trop petit : les « régularités » recherchées ne se manifestent pas suffisamment pour être détectables...
  - Réduction trop forte de dimension : les « régularités » sont incomplètes, les capacités prédictives sont insuffisantes...

Méthodes de réduction de la complexité (2)

B. Réduction de l'ordre de complexité (sans diminuer le volume de données)
- Principe : les modèles sont souvent locaux, la décision est locale → les calculs impliquant des données éloignées peuvent être évités
- Plusieurs types de problèmes de modélisation et de décision
  - $O(N) \rightarrow O(\log(N))$ ou $O(cst.)$
  - $O(N^2) \rightarrow O(N \log(N))$ ou $O(N)$
- Solutions basées sur l'utilisation d'index multidimensionnels ou métriques
- Méthodes exactes ou approximatives
- Difficultés
  - Malédiction de la dimension (*curse of dimensionality*) : méthodes inefficaces si les variables sont nombreuses
  - Mise en œuvre efficace sur une plate-forme distribuée

- Possible de combiner plusieurs méthodes, par ex. réduction de l'ordre de complexité après réduction de dimension (diminuer l'impact de la malédiction de la dimension)

Plan du cours

Méthodes d'échantillonnage

- Méthodes non aléatoires : par quotas, au jugé, volontaires, etc. ; qualité discutable des résultats ; plutôt pour exploration à faible coût
- Méthodes aléatoires (voir par ex. [4]) :
  1. Échantillonnage simple : tirages sans remise, indépendants, chaque donnée a la même probabilité d'être sélectionnée, $p_s = \frac{n}{N}$
  2. Échantillonnage stratifié : l'ensemble de données est constitué de sous-ensembles (strates, par ex. département, tranche d'âge, tranche de revenus…) d'une certaine *homogénéité* interne par rapport à l'étude ; l'échantillonnage simple est appliqué dans chaque strate
     - Exemple : pour une étude des pratiques des clients du commerce en ligne, partitionnement en tranches de revenus (1 tranche = 1 strate) et échantillonnage simple avec une même p_s appliqué dans chaque tranche ; avec un échantillonnage simple sans strates, plus la population d'une tranche de revenus est faible, plus son taux de présence dans l'échantillon s'éloignerait de p_s
     - Par rapport à l'échantillonnage simple, l'échantillonnage stratifié augmente la précision pour une même valeur de n (ou conserve la précision avec n plus faible)

Méthodes d'échantillonnage (2)

- Méthodes aléatoires (suite) :
  3. Échantillonnage en grappes : l'ensemble de données est constitué de sous-ensembles (grappes) présentant une *hétérogénéité* interne par rapport à l'étude ; on choisit au hasard des « grappes » et on examine tous les individus de chaque grappe sélectionnée
     - Exemple : pour l'étude des pratiques sportives des élèves de troisième en zone urbaine, choix aléatoire d'un certain nombre de collèges et sondage auprès de *tous* leurs élèves de troisième (1 collège = 1 grappe)
     - Méthode adéquate si l'hétérogénéité intra-grappe est plus forte que l'hétérogénéité inter-grappes
     - Choisie pour des considérations pratiques liées au recueil des données
  - D'autres méthodes existent, voir par ex. [4]

- L'échantillon doit être de taille assez grande pour que les régularités supposées (qui seront recherchées) y trouvent un support suffisant

Échantillonnage aléatoire dans Spark

- Échantillonnage simple

  `sample(withReplacement : Boolean, fraction : Double, seed : Long) : Dataset[T]`

  - Peut s'appliquer à tout Dataset (ou DataFrame), retourne un Dataset de même type
  - Sans remise (`withReplacement : false`) ou avec remise (`withReplacement : true`)
  - `fraction` indique la probabilité de choisir une donnée

- Échantillonnage stratifié, peut s'appliquer à tout DataFrame, une valeur de clé définit un strate ; retourne un DataFrame de même type

  `sampleBy(col : String, fractions : Map[T, Double], seed : Long) : DataFrame`

  - Méthode de base, respecte de façon *approximative* la taille d'échantillon souhaitée
  - `fractions` indique, pour chaque clé k, la probabilité f_k de choisir une donnée

Plan du cours

Réduction du nombre de variables : objectifs

N données dans $\mathbb{R}^m \rightarrow N$ données dans $\mathbb{R}^k$, $k \ll m$

1. Réduire le volume de données à traiter, tout en conservant au mieux l'information « utile » $\leftarrow$ définir ce qu'est information utile
2. Améliorer le rapport signal / bruit en supprimant des variables non pertinentes $\leftarrow$ définir ce qu'est une variable non pertinente
3. Améliorer la « lisibilité » des données
   - Mettre en évidence des relations entre variables ou groupes de variables
   - Permettre la visualisation $\leftarrow$ définir ce qu'il faut mettre en évidence
4. Répondre à la « malédiction de la dimension » (*curse of dimensionality*)

Objectifs différents, critères différents $\rightarrow$ méthodes différentes

Réduction du nombre de variables : approches

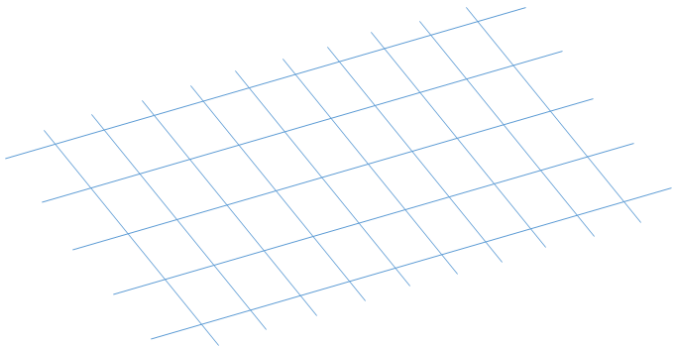
1. Sélection de variables
   - Sélection d'un sous-ensemble de k variables parmi les m variables initiales
   - Les variables sélectionnées gardent leur signification initiale
   - Solution potentiellement sous-optimale car cas particulier de la construction de nouvelles variables
2. Réduction de dimension
   - Modélisation à partir d'un nombre plus faible de variables obtenues par construction de *nouvelles* variables à partir des variables initiales
   - Plus de flexibilité par rapport à la sélection
   - Les nouvelles variables sont rarement interprétables

Sélection de variables

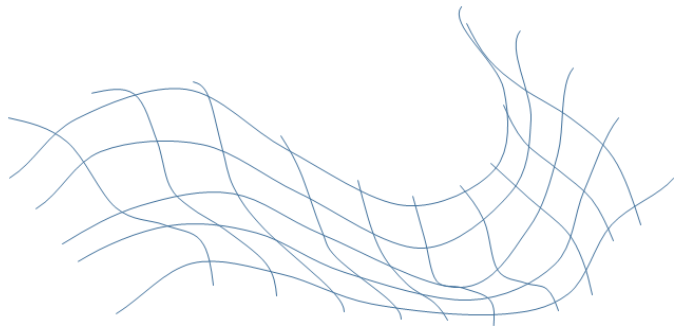
- Approches (voir par ex. [3]) :
  - Méthodes de filtrage : basées sur des critères (par ex. minimisation de la redondance entre variables, maximisation de l'information mutuelle avec la classe à prédire...) qui ne tiennent pas compte du modèle prédictif ultérieur
  - *Wrappers* : basées sur des mesures des performances du modèle prédictif qui emploie les variables sélectionnées
  - Méthodes intégrées (*embedded*) : l'opération de sélection est indissociable de la méthode d'apprentissage statistique
- Choisir k variables parmi $m \Rightarrow$ espace de recherche de l'ordre de $C_m^k \rightarrow$ solutions *approximatives* préférées :
  1. Tri des m *variables initiales* par rapport à un critère de « pertinence » exprimable par variable individuelle, puis sélection des k premières (par ex. `ChiSqSelector` dans Spark : sélection sur la base du test du χ^2)
  2. Construction incrémentale de l'ensemble de k variables : à chaque itération on ajoute la variable qui forme le meilleur ensemble avec celles déjà sélectionnées aux itérations précédentes

Réduction de dimension

- La dépendance entre les nouvelles variables et les variables initiales peut être
  - Linéaire : trouver un sous-espace linéaire de dimension k dans l'espace initial $\mathbb{R}^m$
  - Non linéaire : trouver un sous-espace non linéaire de dimension faible (voir la figure)



sous-espace linéaire



sous-espace non linéaire

- Nous rappellerons ici trois méthodes factorielles linéaires (voir par ex. [2, 1])
  - Analyse en composantes principales (ACP) : exploratoire, variables numériques
  - Analyse factorielle discriminante (AFD) : pour la discrimination, variables numériques
  - Analyse factorielle des correspondances binaires (AFCB) ou multiples (ACM) : exploratoire, variables nominales

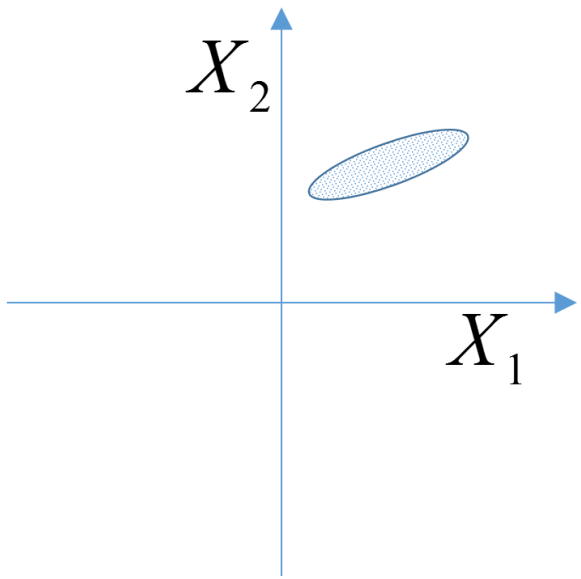
ACP : données, objectifs

- Données : N observations caractérisées par m variables quantitatives (matrice de données brutes $\mathbf{R}$)
- Objectif général : résumer les variables initiales par un petit nombre k de variables synthétiques (les composantes principales) obtenues comme des combinaisons linéaires des variables initiales
- Utilisations (dans un contexte de données massives)
  - Condenser les données en conservant au mieux leur organisation globale
  - Visualiser en faible dimension l'organisation prépondérante des données
  - Interpréter les corrélations ou anti-corrélations entre multiples variables
  - Interpréter les projections des *prototypes de classes* d'observations par rapport aux variables
  - Préparer des analyses ultérieures en éliminant les sous-espaces dans lesquels la variance des données est très faible (assimilés, parfois à tort, à des sous-espaces de bruit, sans information utile)

ACP générale

Sur les valeurs directement recueillies

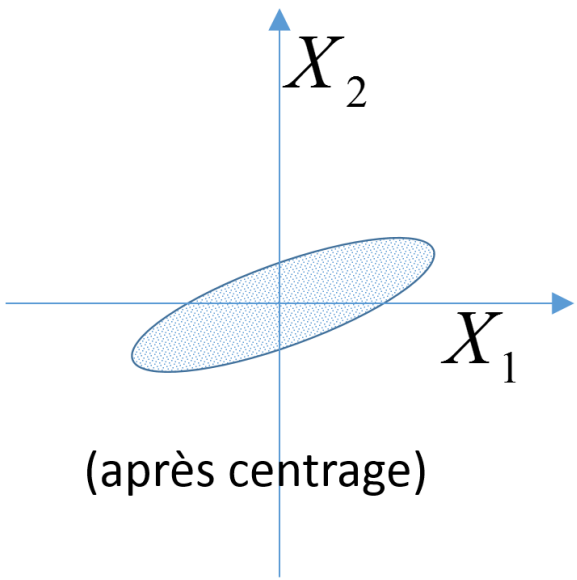
- Interviennent dans l'analyse à la fois la position du nuage d'observations par rapport à l'origine des axes et la forme du nuage
- Utilisation : analyse tenant compte du zéro naturel de certaines variables



ACP centrée

Sur les variables centrées (moyennes nulles)

- Analyse de la *forme* du nuage d'observations (par rapport à son centre de gravité)
- Utilisation : variables directement comparables

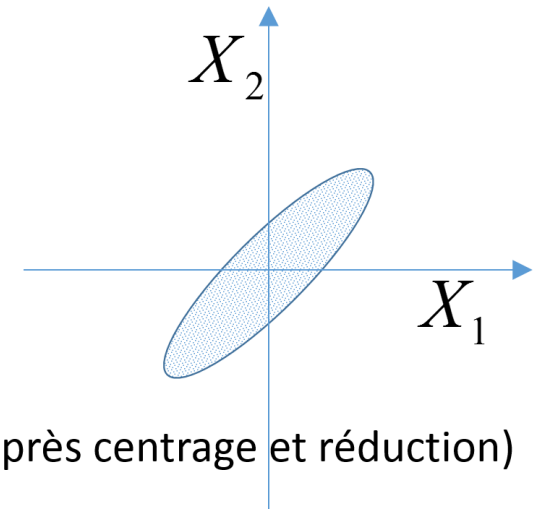


(après centrage)

ACP normée

Sur les variables centrées et réduites (moyennes nulles et écarts-types unitaires)

- Analyse de la forme du nuage après normalisation
- Utilisation : variables non directement comparables, ayant des unités de mesure différentes et/ou des intervalles de variation très différents



(après centrage et réduction)

ACP : solution

- Lignes de $\mathbf{R}$ décrivent les observations dans l'espace des variables initiales, colonnes de $\mathbf{R}$ décrivent les variables dans l'espace des observations $\Rightarrow$ deux analyses possibles : du nuage des observations ou du nuage des variables
- Analyse du nuage des observations : le sous-espace de dimension k recherché est généré par les k vecteurs propres $\mathbf{u}_\alpha$ associés aux k plus grandes valeurs propres λ_α de la matrice $\mathbf{X}^T\mathbf{X}$: $\mathbf{X}^T\mathbf{X}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$, $\alpha \in \{1, \ldots, k\}$
- La matrice $\mathbf{X}$ est :
  - ACP générale : la matrice des données brutes $\mathbf{X} = \mathbf{R}$
  - ACP centrée : la matrice des données centrées (de chaque variable on retire la moyenne empirique) $\Rightarrow \mathbf{X}^T\mathbf{X}$ est la matrice des *covariances* empiriques
  - ACP normée : la matrice des données normées (on divise chaque variable centrée par son écart-type) $\Rightarrow \mathbf{X}^T\mathbf{X}$ est la matrice des *corrélations* empiriques
- Rappel : si $\mathbf{u}_\alpha$ satisfait $\mathbf{X}^T\mathbf{X}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$ alors pour tout $\beta \in \mathbb{R}$, $\beta\mathbf{u}_\alpha$ satisfait la même relation

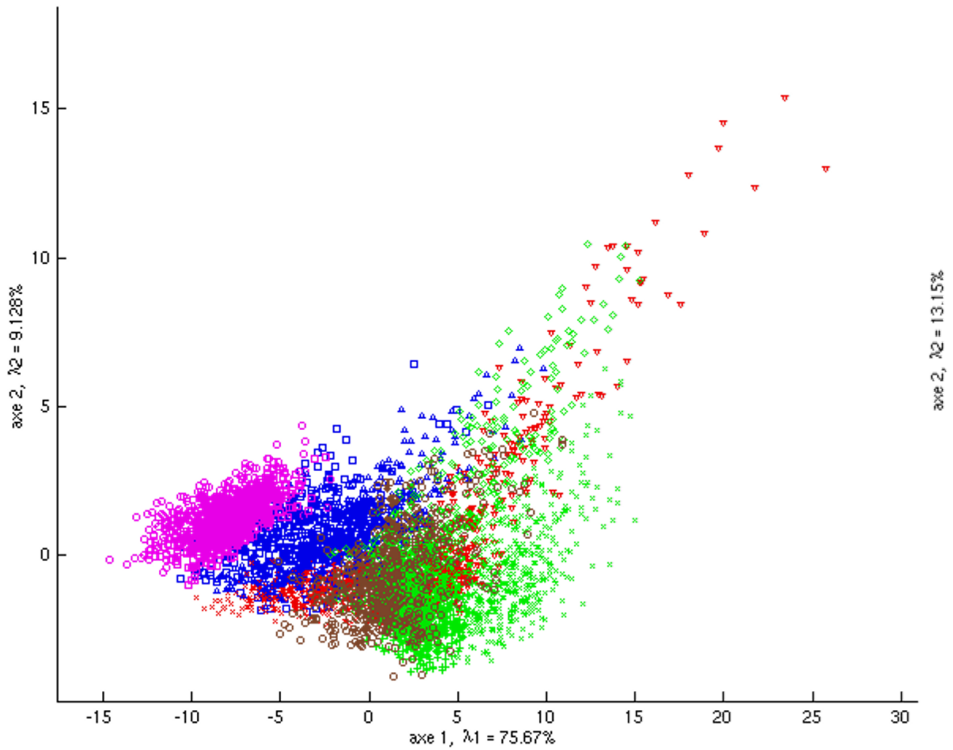
ACP : solution (2)

- La matrice $\mathbf{X}^T\mathbf{X}$ est symétrique et (semi-)définie positive, donc toutes ses valeurs propres sont réelles, positives (certaines nulles si matrice semi-définie) et les vecteurs propres associés à des valeurs propres différentes sont orthogonaux
  - Projection de l'observation $\mathbf{o}_i$ sur l'axe factoriel α : $\psi_{\alpha i} = \sum_j x_{ij} u_{\alpha j}$
- Analyse du nuage des variables : la matrice à diagonaliser est $\mathbf{X}\mathbf{X}^T$, ses valeurs propres non nulles sont les mêmes que celles de $\mathbf{X}^T\mathbf{X}$ et il y a des relations de transition entre les deux analyses
  - Projection de la variable $\mathbf{X}_j$ sur l'axe factoriel α : $\phi_{\alpha j} = \sum_i x_{ij} v_{\alpha i}$
- Comme en général $N \gg m$ (nombre d'observations $\gg$ nombre de variables initiales), on préfère traiter la matrice $\mathbf{X}^T\mathbf{X}$ de dimension $m \times m$ plutôt que $\mathbf{X}\mathbf{X}^T$ de dimension $N \times N$

ACP : nuage des observations

Données mises à disposition par le Laboratoire de Traitement d'Image et de Reconnaissance de Formes (LTIRF) de l'Institut National Polytechnique de Grenoble (INPG) dans le cadre du projet ESPRIT III ELENA (No. 6891) et du groupe de travail ESPRIT ATHOS (No. 6620)

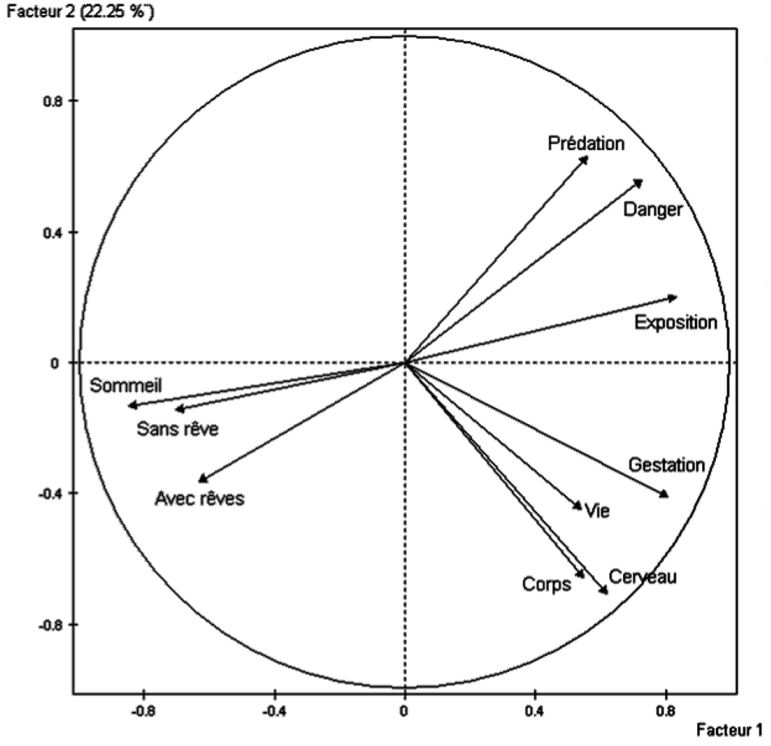
- Visualisation : projection (d'un échantillon) des observations sur les premiers plans factoriels
- $N = 5500,\ m = 40$



ACP : nuage des variables

Données issues de *Allison T., Cicchetti D., Sleep in mammals : ecological and constitutional correlates, Science, vol. 194, pp. 732-734.*

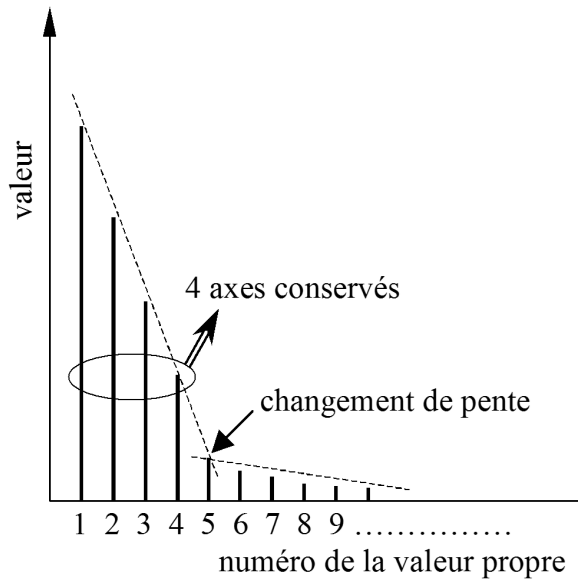
- Visualisation : projection des variables sur le premier plan factoriel
- Dans l'exemple considéré ($N = 62$, $m = 10$) :
  - Groupes : « sommeil », « danger », « corps, cerveau, vie, gestation » (CCVG)
  - Forte opposition entre le groupe « sommeil » et le groupe « danger »
  - Opposition plus faible entre le groupe « danger » et le groupe CCVG



ACP : choix du nombre de composantes

Suivant l'objectif de l'analyse :

- Analyse descriptive avec visualisation : à partir de quel ordre les différences entre les pourcentages d'inertie expliquée ne sont plus significatives ?
- Réduction du volume de données : qualité d'approximation mesurée par le taux d'inertie expliquée
- Prétraitement avant méthodes décisionnelles : simple critère de conditionnement de la matrice des covariances empiriques (ou corrélations si analyse normée), ou nombre d'axes comme paramètre de la méthode décisionnelle



Algorithmes pour l'ACP

Objectif : calculer toutes les valeurs propres de $\mathbf{M} = \mathbf{X}^T\mathbf{X}$ (de dimension $m \times m$) ou seulement les k plus grandes, avec les vecteurs propres unitaires associés

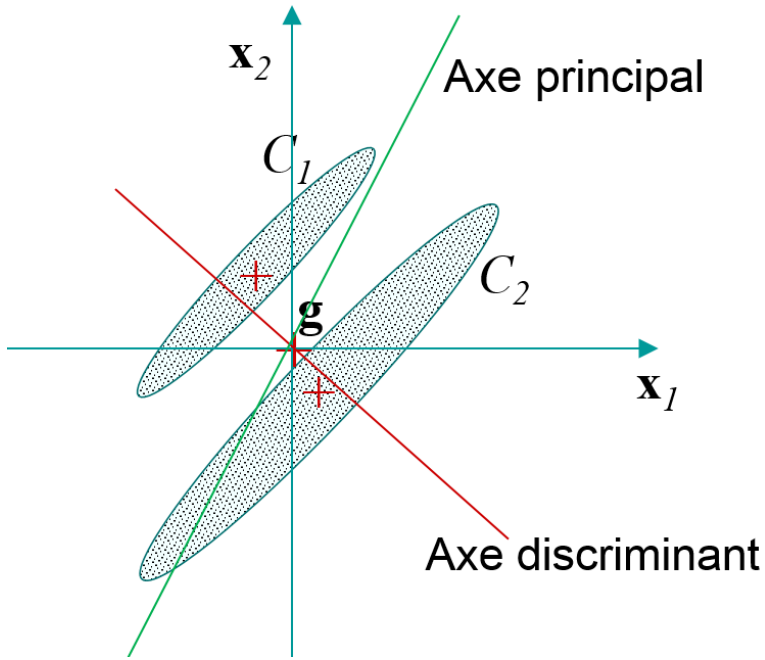
1. Toutes les valeurs propres : calculer le déterminant de $(\mathbf{M} - \lambda\mathbf{I}_m)$ pour trouver les valeurs propres, ensuite résoudre pour chaque valeur propre un système linéaire de m équations avec m inconnues pour trouver le vecteur propre $\rightarrow$ complexité $O(m^3)$!
2. Les k plus grandes valeurs propres : algorithme itératif de complexité $O(Nmk)$
   - Intéressant surtout pour variables « creuses » : si chaque variable prend des valeurs non nulles pour au maximum p observations, $p \ll N$, alors p remplace N dans la complexité
   - On obtient les valeurs propres une par une, à partir de la plus grande, et le vecteur propre associé
   - Pour cela, on itère $\mathbf{x}_{i+1} = \frac{\mathbf{M}\cdot\mathbf{x}_i}{\|\mathbf{M}\cdot\mathbf{x}_i\|}$ (à partir d'un vecteur $\mathbf{x}_0$ quelconque non nul)
   - A convergence, $\lambda_1 = \mathbf{x}^T \cdot \mathbf{M} \cdot \mathbf{x}$ sera la plus grande valeur propre et le $\mathbf{x}$ obtenu le vecteur propre associé
   - On calcule $\mathbf{M}_2 = \mathbf{M} - \lambda_1\mathbf{x}\cdot\mathbf{x}^T$ et on applique le même processus itératif sur $\mathbf{M}_2$ pour obtenir λ_2, etc.

AFD : données, objectifs

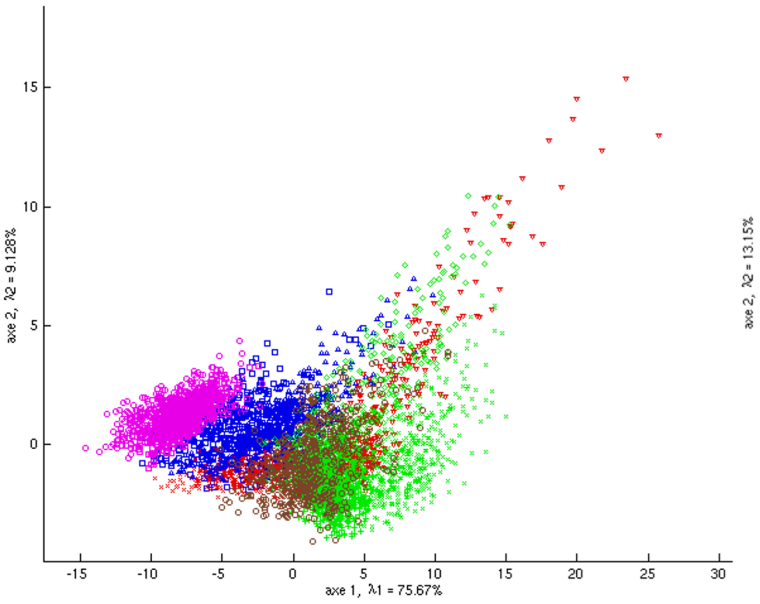
- Données : N observations caractérisées par m variables quantitatives (matrice de données $\mathbf{X}$) et une variable nominale « classe » $Y \in \{1, \ldots, q\}$
- Objectifs :
  - Étape descriptive : identifier des « facteurs discriminants » (combinaisons linéaires de variables explicatives) qui permettent de différencier au mieux les classes
  - Étape décisionnelle : sur la base des valeurs prises par les variables explicatives, décider à quelle classe affecter une nouvelle observation
- Utilisations :
  - Descriptive : condenser la représentation des données en conservant au mieux la séparation entre les classes
  - Décisionnelle : classer de nouvelles observations à partir du sous-espace linéaire qui optimise la séparation

AFD *versus* ACP

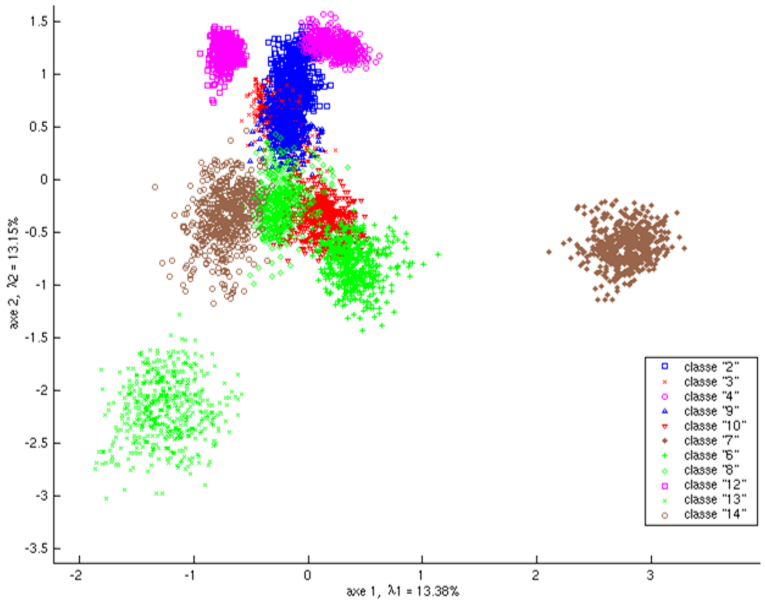
- L'ACP maximise la variance des projections sur le sous-espace
- L'AFD maximise la différenciation entre les classes dans le sous-espace



AFD *versus* ACP : illustration



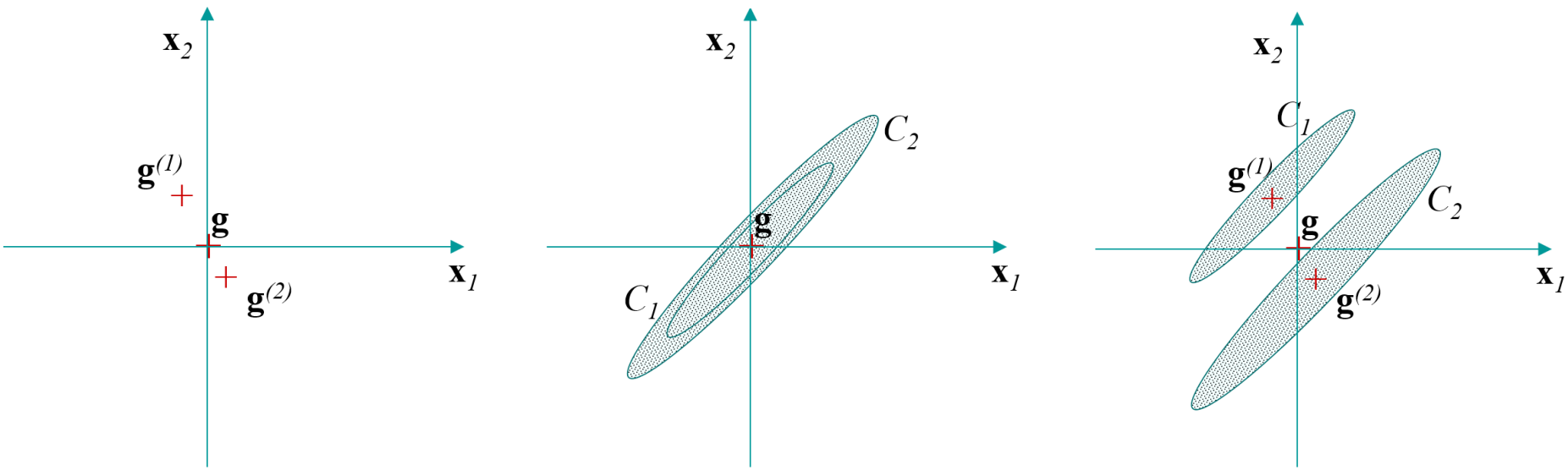
ACP



AFD

AFD : solution

- Covariances :
  1. Inter-classes : $\mathbf{E}$, calculée en considérant que les observations sont les centres de gravité des classes
  2. Intra-classes : $\mathbf{D}$, calculée sur les observations de départ, en centrant chaque classe sur son centre de gravité
  3. Totale : $\mathbf{S}$, calculée sur les observations de départ ; relation de Huygens $\mathbf{S} = \mathbf{E} + \mathbf{D}$

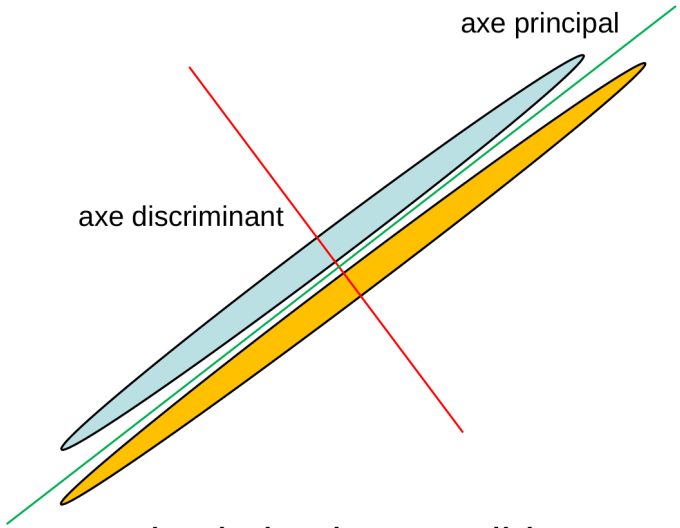


AFD : solution (2)

- Le sous-espace de dimension k recherché est généré par les k vecteurs propres $\mathbf{u}_\alpha$ associés aux k plus grandes valeurs propres λ_α de l'équation de valeurs et vecteurs propres généralisée $\mathbf{E}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{S}\mathbf{u}_\alpha$, $\alpha \in \{1, \ldots, k\}$
  - Il est possible de résoudre plutôt $\mathbf{E}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{D}\mathbf{u}_\alpha$ si le rang de $\mathbf{D}$ n'est pas inférieur à celui de $\mathbf{S}$
  - Si $\mathbf{S}$ est inversible, il est préférable de résoudre $\mathbf{S}^{-1}\mathbf{E}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$
- Approche fréquente si $\mathbf{S}$ est singulière : réduire la dimension avec une ACP pour rendre dans l'espace réduit $\mathbf{S}'$ de rang complet, ensuite résoudre ${\mathbf{S}'}^{-1}\mathbf{E}'\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$ dans l'espace réduit

AFD : solution (3)

- Si l'ACP est appliquée pour réduire la dimension au-delà de la nécessité de rendre $\mathbf{S}'$ de rang complet, risque d'élimination de variables discriminantes !



Discrimination possible

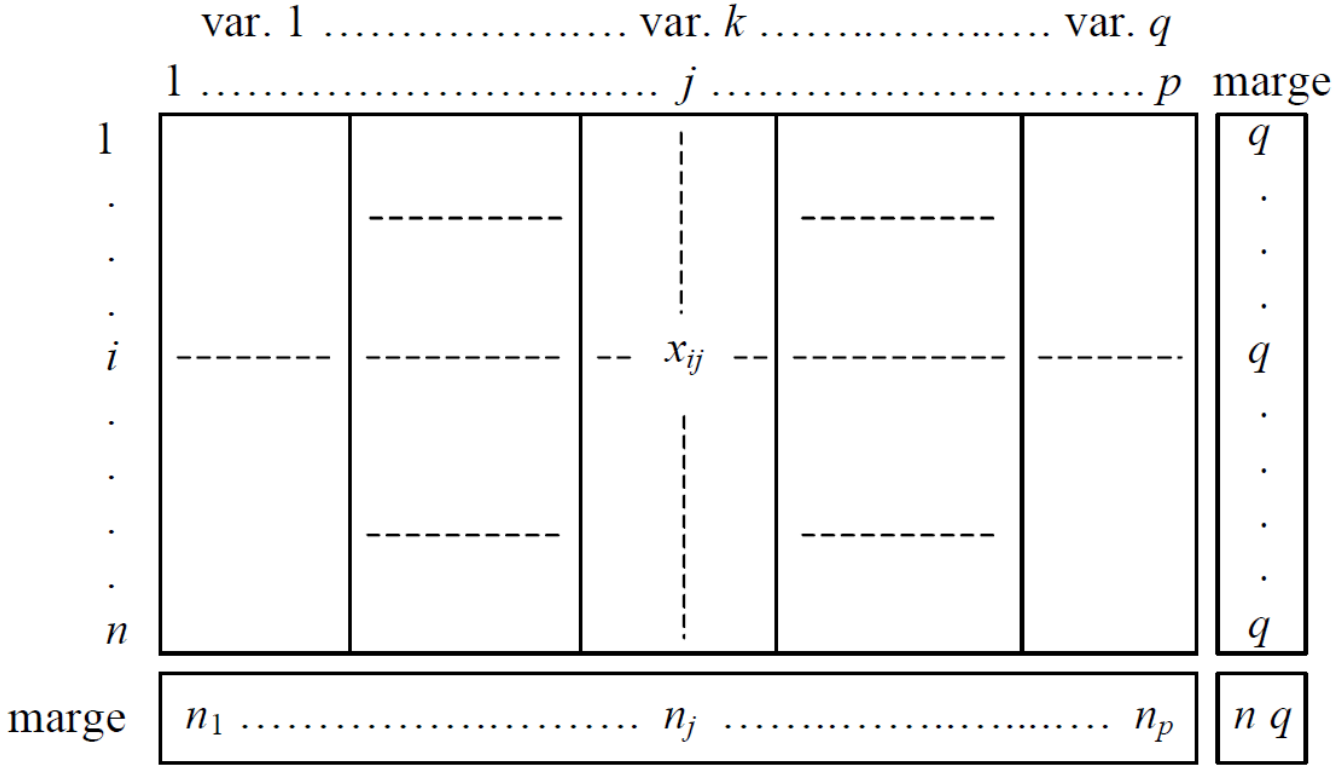
$\rightarrow$ Préférer une approche de régularisation : remplacer $\mathbf{S}$ par $\mathbf{S} + r\mathbf{I}_m$ pour r assez grand

AFD : choix du nombre d'axes

1. Tests statistiques (hypothèses : classes issues de lois normales) :
   1. Test de Rao : test d'égalité à 0 de la i-ème valeur propre
   2. Test du Lambda de Wilks : apport significatif des axes au-delà du i-ème ?
   3. Test incrémental : apport significatif du $i+1$-ème axe ?
2. Méthode de l'échantillon-test (si utilisation décisionnelle directe ou couplage avec un modèle décisionnel) :
   - Extraire (par tirages aléatoires) un échantillon-test
   - Répéter pour différentes valeurs du nombre d'axes : appliquer l'AFD sur les données restantes, développer le modèle décisionnel sur ces mêmes données, évaluer le modèle décisionnel sur l'échantillon-test
   - Choisir les paramètres (dont nombre d'axes d'AFD) qui donnent le meilleur résultat
   - La capacité de généralisation du modèle retenu sera estimée à partir de données non utilisées pour l'AFD ou pour le choix des paramètres

Analyse des correspondances : données

- Données : N observations caractérisées par m variables *nominales* (ou à modalités), représentées par un *tableau disjonctif complet* (TDC) ; chaque observation possède exactement une modalité pour chaque variable



Analyse des correspondances : objectifs, utilisations

- Objectif général : mettre en évidence les relations dominantes entre modalités des variables nominales initiales
- Utilisations
  - Traitement d'enquêtes basées sur des questions fermées à choix multiples
  - Résumer un grand nombre de variables nominales par un faible nombre de variables quantitatives
  - Lorsque le nombre d'observations est faible, peut mettre en évidence des relations entre observations et variables
  - Possibilité d'inclure des variables quantitatives dans l'analyse, après leur transformation en variables nominales

ACM : exemple

Données issues de l'enquête « Les étudiants et la ville » réalisée en 2001 sous la direction de S. Denèfle à l'Université de Tours, voir [1]

- Questions (variables nominales) choisies :
  - Habitez-vous : seul(e), en colocation, en couple, avec les parents
  - Quel type d'habitation occupez-vous : cité U, studio, appartement, chambre chez l'habitant, autre
  - Si vous vivez en dehors du foyer familial, depuis combien de temps : moins d'1 an, de 1 à 3 ans, plus de 3 ans, non applicable
  - A quelle distance de la fac vivez-vous : moins d'1 km, de 1 à 5 km, plus de 5 km
  - Quelle est la surface habitable de votre logement : moins de $10m^2$, de $10m^2$ à $20m^2$, de $20m^2$ à $30m^2$, plus de $30m^2$
- Observation : sur les 5 variables, 3 sont issues de variables quantitatives discrétisées !
- Nombre élevé de modalités $\Rightarrow$ le TDC (ou le tableau de Burt $=$ concaténation des tableaux de contingences par paires de variables) peu lisible $\Rightarrow$ synthèse par ACM nécessaire

ACM : solution

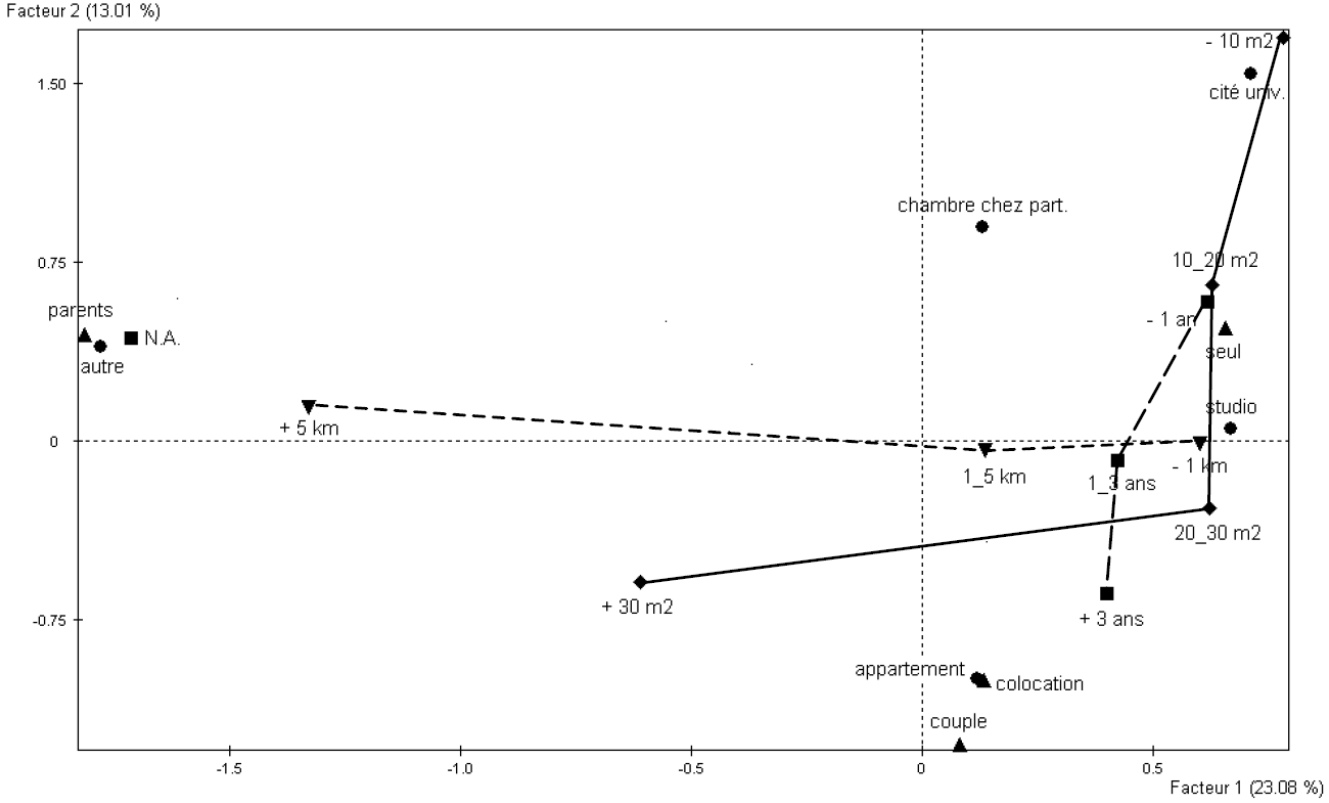
- Analyse des observations (lignes du TDC) et analyse des modalités (colonnes TDC)
  - Si le nombre d'observations (N) est grand alors on s'intéresse surtout à l'analyse des modalités
- *Pondération* de chaque modalité (colonne du TDC) par sa fréquence relative $\rightarrow$ les modalités rares ont un impact faible sur l'analyse
- Distance utilisée :
  - Emploi de la distance du χ^2 : l'influence de chaque composante est pondérée par l'inverse de son poids
  - Équivalence distributionnelle de la distance du χ^2 : si deux colonnes proportionnelles sont cumulées en une seule (fusion de deux modalités), les résultats de l'analyse ne changent pas
- Comme pour l'ACP, on cherche un espace de dimension k ($\ll$ nombre colonnes du TDC) qui résume le mieux la dispersion du nuage analysé

ACM : interprétation

- Sur la base de similarités (ou oppositions) entre projections de modalités :
  - De variables différentes : mêmes populations d'observations
  - D'une même variable (donc mutuellement exclusives) : populations similaires par rapport aux *autres* variables
- Observations importantes :
  - Le centre de gravité des modalités d'une même variable se confond avec le centre de gravité du nuage de toutes les modalités
  - Plus une modalité est rare, plus elle est éloignée du centre de gravité

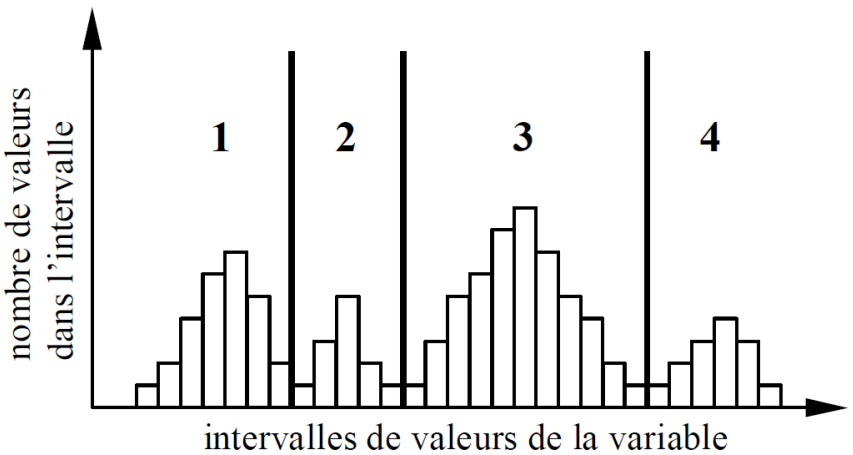
ACM : résultats sur l'exemple [1]

- Oppositions fortes entre modalités : « seul » *vs* « parents », « +5 km » *vs* « -1 km »
- Pour la lisibilité, les modalités successives de variables ordinales (issues de variables quantitatives discrétisées) sont reliées entre elles



ACM : inclusion de variables quantitatives

- Intérêt :
  - Trouver des relations entre modalités de variables qualitatives et *intervalles* de valeurs de variables quantitatives
  - Mettre en évidence des relations *non linéaires* entre intervalles de variables quantitatives
- Découper en intervalles le domaine de variation de la variable quantitative (sur la base de connaissances *a priori*, à partir de l'histogramme, etc.) ; chaque intervalle sera une modalité de la nouvelle variable qualitative (voir [1])



Références I

M. Crucianu, J.-P. Asselin de Beauville, and R. Boné.

Méthodes d'analyse factorielle des données : méthodes linéaires et non linéaires.

Hermès, Paris, 2004.

G. Saporta.

Probabilités, Analyse des Données et Statistique.

Technip, Paris, 2011.

J. Tang, S. Alelyani, and H. Liu.

Feature selection for classification : A review.

In *Data Classification : Algorithms and Applications*, pages 37–64. 2014.

Y. Tillé.

Théorie des sondages.

Dunod, Paris, 2001.