Analyse de données

Gwenlake

Présentation

Objectifs du cours

- Introduction et présentation de l'Analyse en Composantes Multiples (ACM).
- Points clés mathématiques et liens avec l'ACP.
- Choix du nombre d'axes principaux
- Interprétations

Analyse en Composantes Multiples

Introduction & Structure des Données

- L'ACM est une méthode factorielle adaptée aux tableaux **Individus x Variables Qualitatives**. Son but est de résumer l'information et visualiser les associations.
- **La Transformation Clé :** On ne peut pas calculer de moyenne sur "Bleu" ou "Rouge". On transforme donc le tableau brut en **Tableau Disjonctif Complet**
- **Tableau Disjonctif Complet :** Codage binaire (0 ou 1). Si une variable a k modalités, elle génère k colonnes. Réduire la dimensionnalité : Passer d'un grand nombre de variables à quelques composantes clés pour simplifier l'analyse.
- L'ACP cherche les facteurs qui font varier les chiffres, l'ACM cherche à détecter les "Patterns" (Associations) : Son but premier est de révéler quelles modalités vont ensemble.
- Les variables qualitatives sont résumées en valeurs numériques.

individu	pays	sport
Paul	France	Football
Marie	Angleterre	Athlétisme

individu	pays_France	pays_Angleterre	Sport_Football	Sport_Athlé
Paul	1	0	1	0
Marie	0	1	0	1

Analyse en Composantes Multiples

Le Rôle Central du Chi2

- En ACM, on utilise une nouvelle métrique qui remplace la distance euclidienne de l'ACP. Le principe est de mesurer en plus la rareté pour créer des liens.

- $\text{distance}^2(i,l) = \sum_{k=1}^{K} \frac{1}{p_k} (x_{ik} - x_{lk})^2$

- x_{ik} ,x_{lk} : valent 1 ou 0 si l'individu à la modalité k
- Pk: Proportion de personnes avec la modalité k (rareté). Poids elevé si rareté elévée
- Similaire à une distance euclidienne mais on divise par la rareté de la modalité

Exemple

- 2 personnes ont un Chien (Banal) : Lien faible (Distance proche de 0).
- 2 personnes ont un Iguane (Rare) : Lien très fort (Poids élevé).

Analyse en Composantes Multiples

L'ACM est une ACP sur le nouveau tableau des modalités

- Les résultats vus pour l'ACP sont similaires ici :
- Valeurs Propres : Elles représentent l'inertie portée par chaque axe
- **Critère du Coude :** Pour choisir le nombre d'axes, on regarde le graphique des valeurs propres.

On retient les axes situés avant la "cassure" (le coude) de la courbe.

Remarque : en ACM, des taux d'inertie faibles sont normaux. Avoir 10% sur le premier axe est souvent un très bon résultat.

- On obtient un graphique pour représenter les individus, les variables mais aussi les modalités.

Concepts généraux sur la distance:

- 2 individus prennent les mêmes modalités : distance = 0
- 2 individus ont en commun beaucoup de modalités : distance petite
- 2 individus dont l'un des 2 possède une modalité rare : distance grande pour prendre en compte la spécificité d'un des 2
- 2 individus ont en commun une modalité rare : distance petite pour prendre en compte leur spécificité commune

Analyse en Composantes Multiples

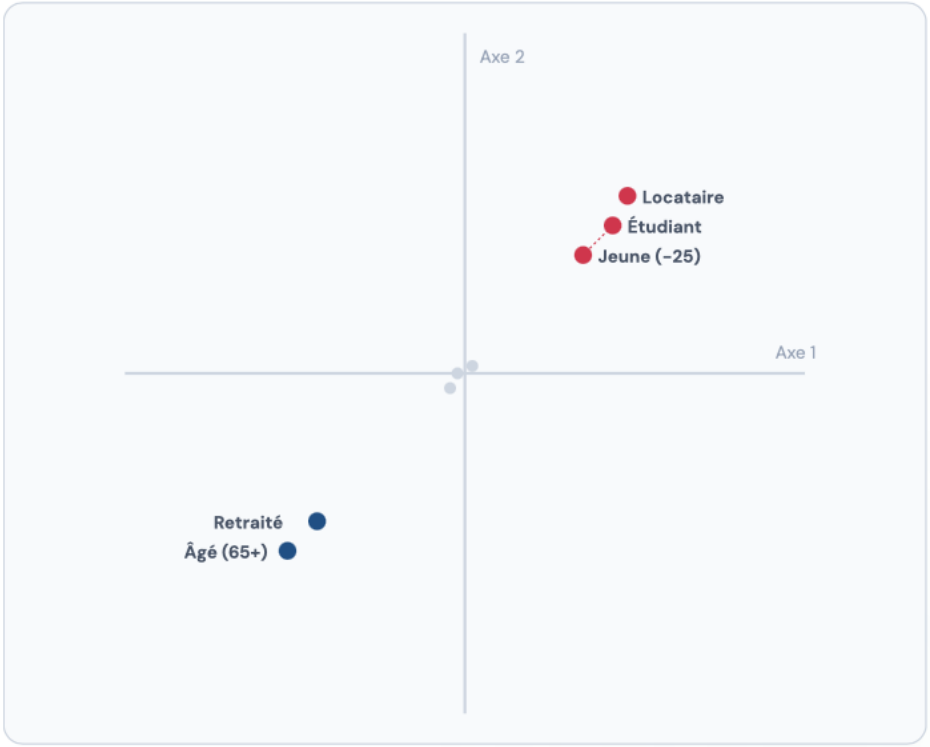
Représentation des variables pour interpréter les dimensions

- **Différence majeure :** En ACM, les variables (modalités) sont des **points**, pas des FLÈCHES. Il n'y a pas de cercle des corrélations.

- Proximité = Association : Ici, "Jeune" est proche de "Étudiant". Ces deux modalités vont souvent ensemble.
- **Éloignement :** Les points au centre (gris) sont banals. Les points éloignés sont discriminants.

Méthode d'Interprétation:

- Prends l'Axe 1 (Horizontal).
- Regarde qui est tout à gauche vs tout à droite.
- Trouve le thème qui crée cette opposition (ex: Riche vs Pauvre, Urbain vs Rural, Moderne vs Traditionnel).



Analyse en Composantes Multiples

Représentation des individus

- Les points proches se ressemblent.
- On peut identifier des clusters ou des familles d’individus.
- Les points isolés sont des points atypiques
- On peut observer des points sous forme de trajectoire : effet Guttman.