Analyse de données

L. BELLANGER

Master 1 Ingénierie Statistique
Dpt de Mathématiques - Université de Nantes

Plan

V. Conclusion

Quelle analyse factorielle pour quel type de tableau ?

Tableau de données	Transformation du tableau	Métrique	Pondération	Méthode
De type individus x variables où les variables sont quantitatives (soit continues, soit discrètes avec un nombre de modalités important, on réalise alors une ACP sur les rangs)	Centrage et réduction pour l'ACP normée, ou simple centrage pour l'ACP non normée	Euclidienne, après réduction (ou inverse des variances sans réduction préalable des données), euclidienne sans réduction des données (cas de l'ACP non-normée)	En fonction de l'échantillonnage des données sur lesquelles on travaille (plan de sondage), ou de la nature des individus (ex des départements que l'on souhaite pondérer par leur population).	**ACP (normée)**
Un tableau de contingence croisant 2 variables qualitatives	Passage par un tableau de fréquences, puis des profils-lignes et des profils-colonnes	Du Khi 2 (pondération des colonnes par l'inverse de leur fréquence marginale dans le calcul des distances interindividuelles – et réciproquement)	Lignes et colonnes sont pondérées par leurs fréquences marginales respectives.	**AFC**
De type individus x variables où les variables sont qualitatives ou mixtes (mise en classes des variables quantitatives)	Passage par un tableau disjonctif complet	Du Khi 2	Lignes dont pondérées selon l'échantillonnage, les colonnes selon leurs fréquences marginales relatives	**ACM**

+ *AFD !*

Classification non supervisée & Analyse factorielle

- Méthodes qui
  - s'inscrivent dans une même perspective
    - ➢ analyse exploratoire d'un tableau de données
  - diffèrent selon le mode de représentation
    - ➢ représentation par des cartes / par des groupes (hiérarchie indicée ou partition)

↪ Combiner les 2 approches en utilisant pour chaque méthode la même distance entre individus !

Classification non supervisée & Analyse factorielle

Complémentarité des approches

- Méthodes de classification non supervisée directement applicables sur les données brutes, qu'elles soient continues, qualitatives ou mixtes, à condition de **bien choisir la métrique**
- Mais souvent, elles sont associées en complément d'une analyse factorielle

Classification non supervisée & Analyse factorielle

Complémentarité des approches

- Ce que font les analyses factorielles :
  - Déterminer des associations entre des variables ou modalités de variables
  - Définir un nouvel espace orthonormé, réduit
  - Positionner les individus dans cet espace de dim réduite
- Ce que ne font pas les analyses factorielles :
  - mettre en évidence des individus qui se ressemblent (projections voisines sur des plans factoriels), mais sans créer de classes formellement (frontières ?)

Classification non supervisée & Analyse factorielle

Complémentarité

- **Avantage de l'enchaînement des 2 méthodes :**
  - Un nb limité de facteurs permet de représenter les données initiales
    => le calcul des distances entre points se fait dans un espace réduit (temps de calcul)
  - Variances décroissantes des facteurs : les premiers ont plus d'influence que les derniers dans les calculs de distances et/ou d'inertie, ce qui correspond à leur représentativité

Classification non supervisée & Analyse factorielle

Complémentarité

Avantage de l'enchaînement des 2 méthodes :

- Pour les données qualitatives, l'ACM préalable à la CAH permet de rendre quantitatives des variables qualitatives (les facteurs sont des variables continues) et donc possibilité d'appliquer la métrique euclidienne
- Propriété : si on utilise pour la CAH tous les facteurs issus de l'analyse factorielle ou les variables d'origine, on obtient la même partition.

Classification non supervisée & Analyse factorielle

- Analyse factorielle en amont d'une CAH
  - **Méthode** :
    - réaliser l'analyse factorielle du tableau $\mathbf{X}$:ACP, AFC ou AFCM selon la nature des données ;
    - effectuer une classification à partir du tableau $\mathbf{F}$ des composantes factorielles.
  - **Avantage** : permet de ne conserver pour la classification qu'une partie des composantes factorielles ; soit
    - celles qui apportent une réelle information. Suppression des dernière représentant du « bruit » ou
    - celles correspondant aux axes que l'on a pu interpréter

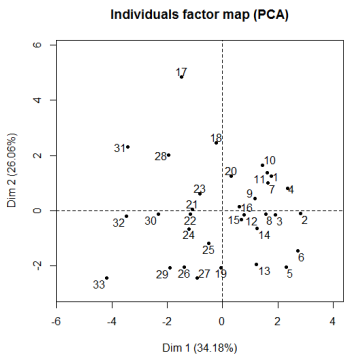
Classification non supervisée & Analyse factorielle

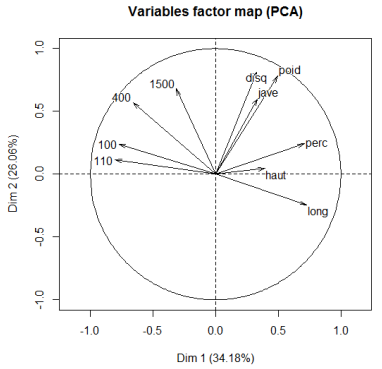
- Analyse simultanée d'un plan factoriel et d'une hiérarchie
  - Représentation sur le plan factoriel (en gal 1-2) des nœuds les plus haut de la CAH

Classification & Analyse factorielle

- Exemple sous R
  - Reprenons le fichier **olympic** traité dans le chapitre ACP

```
library(ade4) ; library(FactoMineR)
data(olympic)
olympic.pca<-PCA(olympic$tab, ncp=2)# ACP normée avec PCA
```

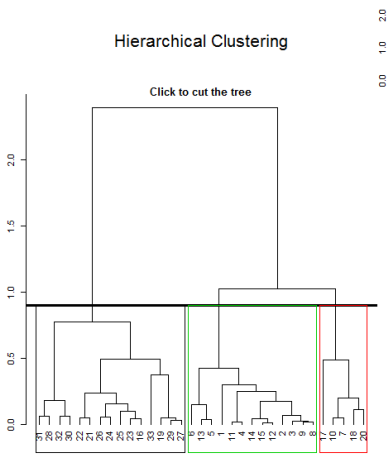


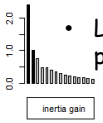


Classification & Analyse factorielle

- Exemple sous R : **olympic**

```
res.pcah<-HCPC(olympic.pca)# CAH Ward à partir des 2 1eres CP
```





- La forme du dendrogramme suggère une partition des athlètes en **3 classes** :
  - niveau de coupure optimal calculé par ***HCPC***
- ***res.pcah*$call$t** contient les résultats de la CAH
  - **$tree** : sorties de `cluster(agnes)`
  - **$nb.clust** : nb de classes optimal
  - **$quot** : rapport d'inertie intra à minimiser $quot^K = \frac{I_W^{K+1}}{I_W^K}$

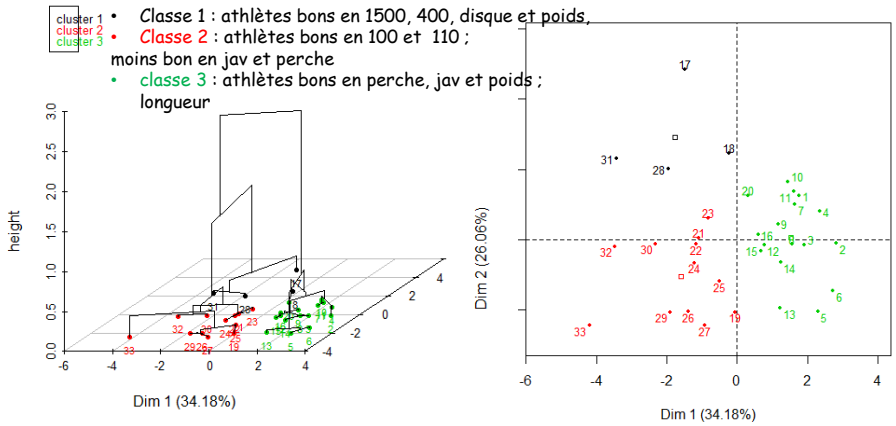
Classification & Analyse factorielle

- Exemple sous R : **olympic**

```
res.pcah<-HCPC(olympic.pca)# CAH Ward à partir des 2 1eres CP
```

Hierarchical clustering on the factor map

Factor map



- Classe 1 : athlètes bons en 1500, 400, disque et poids,
- Classe 2 : athlètes bons en 100 et 110 ; moins bon en jav et perche
- classe 3 : athlètes bons en perche, jav et poids ; longueur

Classification & Analyse factorielle

- Exemple sous R : **olympic**

```
res.pcah<-HCPC(olympic.pca)# CAH Ward à partir des 2 1eres CP
```

Les 3 classes sont décrites dans **res.pcah$desc.var**

Pour plus détails voir Husson et al. (2009)

Références bibliographiques

- L. Bellanger, R. Tomassone, *Exploration de données et méthodes statistiques : Data analysis & Data mining avec R. Collection Références Sciences, Editions* Ellipses, Paris, 2014.
- B.S. Everitt, S. Landau, L. Morven. *Cluster Analysis*, 4th ed., Oxford University Press Inc., Oxford, 2001..
- A.D., Gordon, A. D., *Classification*. 2nd Edition. London: Chapman and Hall / CRC, 1999.
- F. Husson, S. Lê & J. Pagès, *Analyse de données avec R*. PUR, Rennes, 2009.
- L. Lebart, A. Morineau, M. Piron, *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 2006.
- J.-P. Nakache, J. Confais, *Approche pragmatique de la Classification*. Editions Technip, Paris, 2005.
- G. Saporta, *Probabilités, Analyse des données*. Editions Technip, Paris, 2006.
- Statistics with R : http://zoonek2.free.fr/UNIX/48_R/all.html
- S. Tufféry, *Data mining et statistique décisionnelle : L'intelligence dans les bases de données*. Editions Technip, Paris, 2005.