Analyse de données avec R - Complémentarité des méthodes d'analyse factorielle et de classification

François Husson, Julie Josse, Jérome Pagès

Analyse de données avec R
Complémentarité des méthodes d'analyse factorielle et de classification

François Husson, Julie Josse & Jérôme Pagès

Laboratoire de mathématiques appliquées - 65 rue de St-Brieuc - 35042 Rennes cedex
husson@agrocampus-ouest.fr

Mots-clés : Analyse factorielle, ACP, ACM, AFM, classification, partitionnement
Keywords: Exploratory Data Analysis, PCA, MCA, MFA, classification, clustering

1 Introduction

L'objectif de cet exposé est de présenter les fonctionnalités disponibles sur R en analyse de données. R est un langage [3] gratuit en pleine expansion et de plus en plus utilisé tant par le monde de l'entreprise, de l'enseignement et de la recherche. La spécificité de l'analyse des données à la française est présente dans ce logiciel grâce à plusieurs bibliothèques de fonctions. Dans une première partie de l'exposé, on montrera comment mettre en œuvre, sous R, des méthodes d'analyses factorielles classiques (ACP, AFC ou ACM) et plus avancées (Analyse factorielle Multiple ou AFM Hiérarchique) à partir du package FactoMineR. Dans une seconde partie, on se focalisera sur la complémentarité des méthodes d'analyses factorielles et de classification pour visualiser des données.

2 L'analyse de données avec R

2.1 Le package FactoMineR

Le package FactoMineR est dédié à l'analyse de données. Les méthodes les plus classiques d'analyse de données y sont programmées: Analyse en Composantes Principales (fonction `PCA`), Analyse Factorielle des Correspondances (`CA`), Analyse Factorielle des Correspondances Multiples (`MCA`) et construction ascendante d'une hiérarchie (`HCPC`). Des méthodes plus avancées sont également disponibles et permettent de prendre en compte une structure sur les variables ou sur les individus: Analyse Factorielle Multiple (`MFA`), Analyse Factorielle Multiple Hiérarchique (`HMFA`) ou Analyse Factorielle Multiple Duale (`DMFA`).

Dans chaque méthode, il est possible d'ajouter des éléments supplémentaires: individus, variables quantitatives et/ou qualitatives. Pour chacune de ces analyses, de nombreuses aides à l'interprétation sont fournies: qualité de représentation, contribution pour les individus et les variables. Les représentations graphiques sont au centre de chacune des analyses et de nombreuses options graphiques sont disponibles: colorier les individus en

fonction d'une variable qualitative, ne représenter que les variables les mieux projetées sur les plans factoriels, etc. Des fonctions permettent d'aider à l'interprétation des dimensions factorielles (`dimdesc` ou à l'interprétation des classes `catdes`).

Le site `http://factominer.free.fr` est dédié à FactoMineR et le livre « Analyse de données avec R » [2] détaille ces méthodes et présente de nombreuses études de cas.

Le menu déroulant. Toutes les méthodes peuvent être mises en œuvre à l'aide d'une interface graphique qui s'installe grâce à la ligne de code suivante:

```
> source("http://factominer.free.fr/install-facto.r")
```

Ce menu déroulant permet une mise en œuvre instantanée des méthodes, sans connaissances préalables du logiciel R. Il permet aussi de s'approprier le langage en fournissant les lignes de code associées à chaque action.

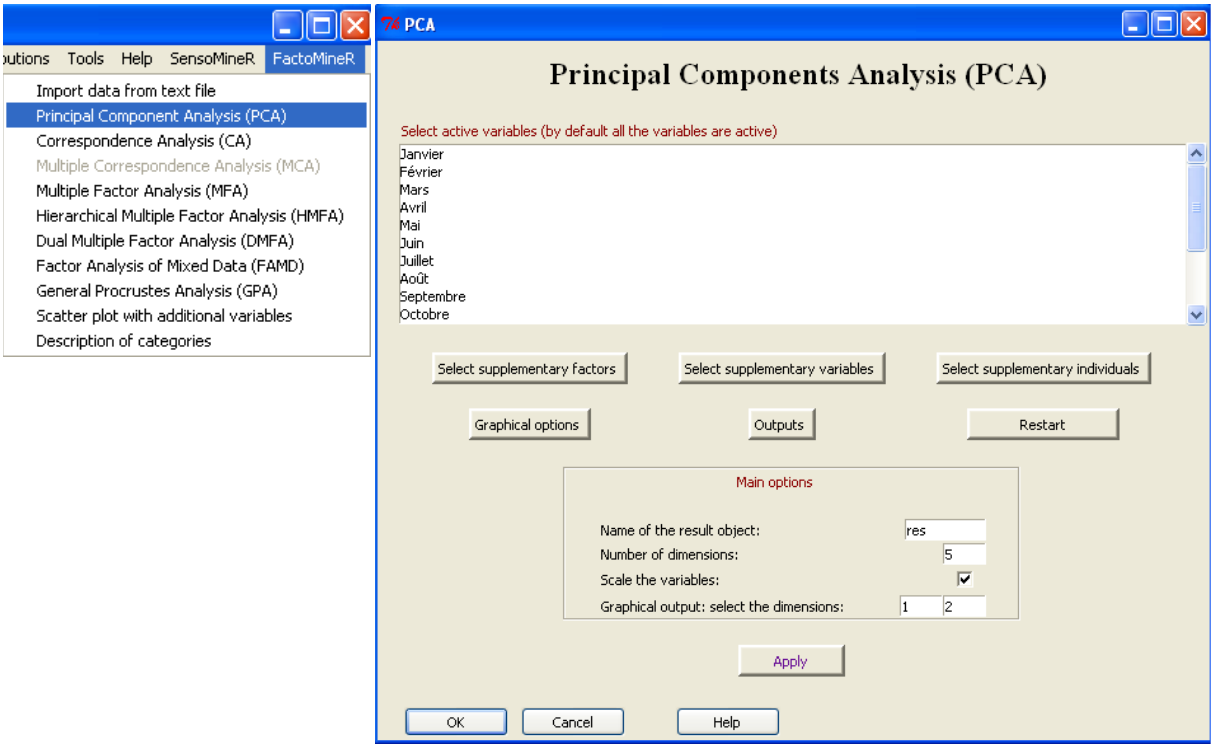

Figure 1: Menu de FactoMineR et fenêtre principale de l'ACP.

La figure 1 présente le menu de FactoMineR et celui de l'ACP. Des choix par défaut sont faits, mais ils peuvent être modifiés: aucun variable qualitative et quantitative supplémentaire, aucun individu supplémentaire, les résultats sur les 5 premières dimensions sont fournis (objet `res`), les variables sont normées et les graphiques sont fournis pour

le plan principal (axes 1 et 2). L'onglet `Apply` lance l'analyse en conservant la fenêtre ouverte, ce qui permet de modifier certaines options sans refaire le paramétrage (cf. [1] ou [2] pour plus de détails). De nombreuses options graphiques sont également disponibles.

2.2 Quelques autres packages

De nombreux packages permettent de faire de l'analyse des données. Ils sont listés ainsi que leurs principales caractéristiques à l'adresse suivante: `http://cran.r-project.org/web/views/Multivariate.html`. Citons les principaux:

- ade4 contient de nombreuses fonctions pour le traitement de données écologiques;
- ca est focalisée sur l'analyse des correspondances;
- cluster est dédiée aux méthodes de classification.

3 Classification et analyse factorielle

Classification automatique et analyse factorielle s'inscrivent dans une même perspective (l'analyse exploratoire d'un tableau rectangulaire) et diffèrent selon le mode de représentation (nuage euclidien, hiérarchie indicée ou partition). D'où l'idée de combiner les approches pour obtenir une méthodologie riche, qualité essentielle en statistique exploratoire car le fait de disposer de plusieurs points de vue ne peut que renforcer la solidité des conclusions. Dans ce cas, on utilise pour chaque méthode la même distance (euclidienne) entre individus.

3.1 L'analyse factorielle comme prétraitement de la classification

Cas de variables quantitatives. L'ACP permet d'obtenir des composantes principales qui sont des variables synthétiques (dimensions) orthogonales. Elle peut aussi être présentée comme la décomposition des données en un signal plus du bruit, les premières dimensions correspondant au signal et les dernières au bruit. De ce fait, l'ACP peut servir de prétraitement à la classification: seules les premières dimensions sont conservées pour calculer de nouvelles distances entre individus. Sans le bruit, la classification est plus robuste que celle obtenue sur les distance initiales. Ainsi, la forme du haut de l'arbre hiérarchique est plus stable.

Cas de variables qualitatives ou mixtes. La classification d'individus décrits par des variables qualitatives est un vaste domaine d'étude et de nombreuses mesures de ressemblance existent. Cependant, ces méthodes sont difficiles à mettre en œuvre et en pratique il est courant d'utiliser l'ACM comme prétraitement: les variables qualitatives

sont transformées en composantes principales, c-à-d en variables quantitatives. La classification est construite à partir des (premières) dimensions factorielles. De même, si des variables quantitatives et qualitatives sont disponibles simultanément, elles peuvent être prétraitées par l'Analyse Factorielle de Données Mixtes.

Prise en compte d'une structure sur les variables. Dans certaines situations, les variables sont structurées en groupes ou il existe une hiérarchie sur les variables. C'est le cas, par exemple, des questionnaires structurés en thèmes (et sous-thèmes). L'Analyse Factorielle Multiple (AFM) et l'AFM Hiérarchique (AFMH) permettent d'équilibrer l'influence des groupes (voire des sous-groupes à l'intérieur d'un groupe). Une classification effectuée sur les composantes factorielles de l'AFM ou de l'AFMH prend alors en compte l'équilibre des groupes dans le calcul de la distance entre individus.

3.2 Choix du nombre de classes à partir de l'arbre hiérarchique

Un arbre hiérarchique peut être considéré comme une séquence de partitions emboîtées, de la plus précise (un individu par classe) à la plus grossière (une seule classe). Ainsi, une hiérarchie est extrêmement utile pour déterminer le nombre de classes. Ce choix peut être fait à partir de l'allure générale de l'arbre, du niveau des nœuds (ces niveaux peuvent être représentées par un diagramme en barres), du nombre de classes (qui doit être ni trop faible ni trop grand) et de l'interprétabilité des classes. La fonction `HCPC` propose un niveau de coupure "optimal" de l'arbre hiérarchique (Fig. 2).

3.3 Complémentarité des trois méthodes pour la visualisation

La complémentarité consiste à représenter la hiérarchie ou la partition sur le plan factoriel. Avec une partition, on se limite à la projection du centre de gravité des classes. Une telle représentation enrichit l'interprétation principalement sous deux aspects:

- on dispose à la fois d'une vision continue (les « tendances » matérialisées par les axes factoriels) et discontinue (les classes de la classification) du même ensemble de données, le tout dans un cadre unique;
- le plan factoriel ne fournit aucune information sur la position des points dans les autres dimensions; les classes, établies à partir de l'ensemble des dimensions, apportent sur le plan un peu d'information « extérieure au plan »; deux points proches sur le plan pouvant être dans la même classe (et donc pas trop éloignés l'un de l'autre le long des autres dimensions) ou dans deux classes différentes (parce qu'ils sont éloignés l'un de l'autre le long des autres dimensions).

Dans la présentation de l'arbre hiérarchique (Fig. 2), le premier axe de l'ACP permet de trier les individus ou les groupes d'individus. Cela facilite la lecture de l'arbre car les individus sont ainsi triés selon leur principale dimension de variabilité.

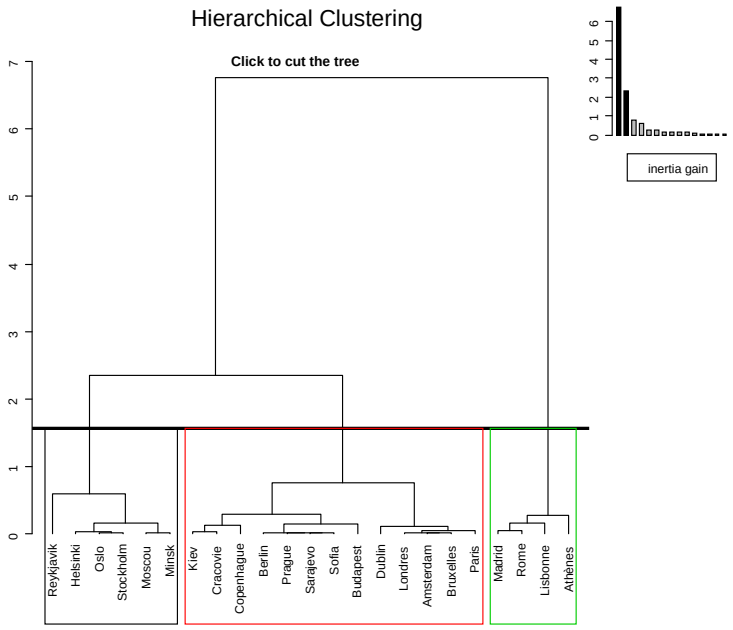

Figure 2: Arbre hiérarchique (individus triés selon leur coordonnée sur la 1^{e} composante).

4 Exemple: les données de températures

On s'intéresse aux températures moyennes mensuelles de 23 capitales européennes. L'ACP normée réalisée sur ces données résume 97 % de l'inertie totale sur les deux premières dimensions (voir graphe des variables, Fig. 3). La classification est réalisée grâce à la

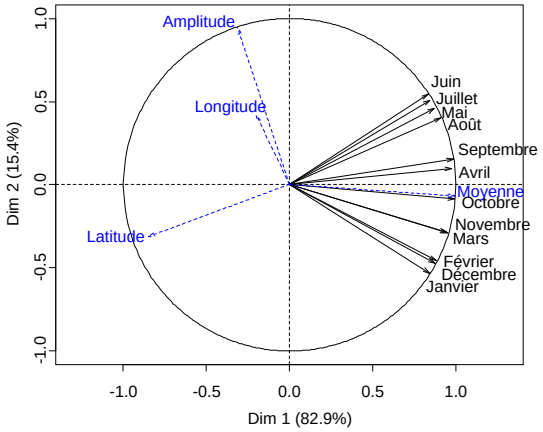

Figure 3: Représentation des variables sur le premier plan factoriel.

fonction `HCPC`:

```
> library(FactoMineR)
> temperature <- read.table("http://factominer.free.fr/livre/temperat.csv",
      header=TRUE, sep=";", dec=".", row.names=1)
> res.pca <- PCA(temperat[1:23,1:12], scale.unit=TRUE)
> HCPC(res.pca,conso=0)
```

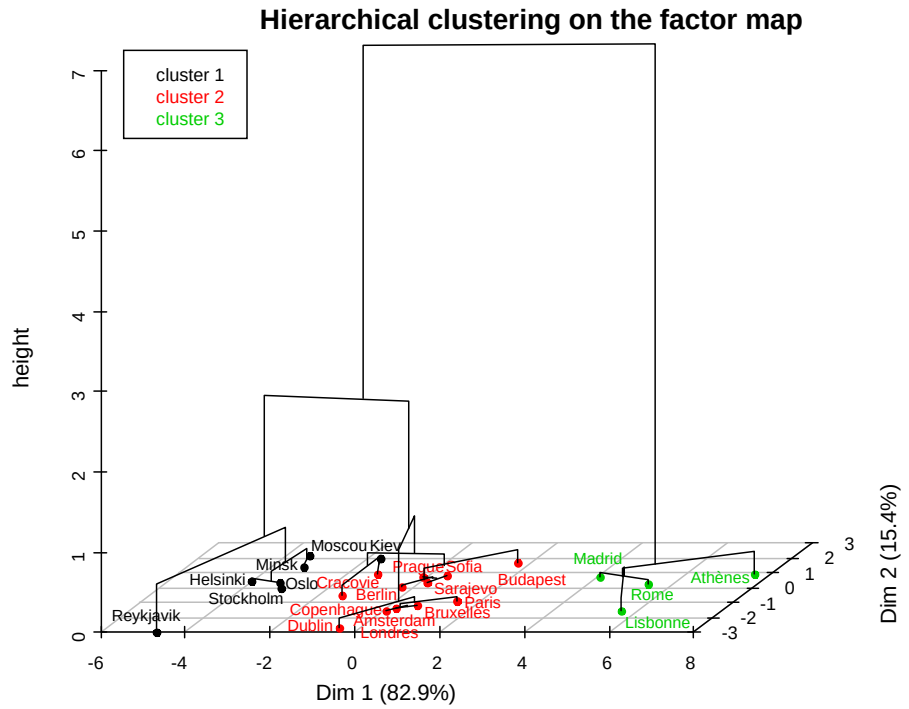

Figure 4: Représentation 3D de l'arbre hiérarchique sur le premier plan factoriel.

L'arbre hiérarchique est représenté en trois dimensions sur le plan principal de l'ACP (Fig. 4). La fonction ayant proposé un découpage en trois classes, les individus sont colorés en fonction de leur classe d'appartenance. Les classes sont ensuite décrites par les variables quantitatives et/ou qualitatives du jeu de données par la fonction `catdes`.

Bibliographie

[1] Cornillon, PA., Guyader, A., Husson, F., Jégou, N., Josse, J., Kloareg, M., Matzner-Løber, E. et Rouvière, L. (2008) *Statistique avec R*, Edition PUR, Rennes.

[2] Husson, F., Lê, S. et Pagès, J. (2009) *Analyse de données avec R*, Edition PUR, Rennes.

[3] R Core Team (2009) R: A Language and Environment for Statistical Computing.