L'analyse de données avec FactoMineR : les nouveautés

François Husson et **Julie Josse**

Laboratoire de Mathématiques appliquées
Agrocampus Ouest
65 rue de Saint-Brieuc - 35042 Rennes
husson@agrocampus-ouest.fr
josse@agrocampus-ouest.fr

Mots clefs : Analyse de données, module graphique, `FactoMineR`, données manquantes.

Plusieurs packages sont disponibles pour faire de l'analyse des données sous `R`, citons par exemple les packages `ade4`, `FactoMineR` et `ca`. Nous nous focalisons dans cet exposé sur le package `FactoMineR` [1] et nous présentons les dernières nouveautés, notamment le module graphique et la façon dont les données manquantes sont gérées.
Les tableaux de données incomplets sont parfois gérés de façon très succincte dans les logiciels : la donnée manquante est remplacée par la moyenne pour les données quantitatives ou encore la donnée manquante génère une modalité que l'on peut appeler "manquante" quand les données sont qualitatives. Ceci est une première approche d'imputation des données manquantes mais [2] ont proposé de nouvelles méthodes d'imputation qui sont disponibles dans le package `missMDA` [3]. Ce package peut être utilisé en complément du package `FactoMineR` pour faire des analyses en composantes principales (ACP), des analyses des correspondances multiples (ACM), des analyses factorielles de données mixtes (AFDM) ou des analyses factorielles multiples (AFM) avec données manquantes.
En ce qui concerne le module graphique de FactoMineR, il possède maintenant un algorithme qui positionne de façon "optimale" les libellés des éléments (individus, variables, modalités, fréquences, groupes de variables ... selon la méthode utilisée, ACP, analyse des correspondances (AFC), ACM, AFDM, AFM). L'algorithme calcule et minimise un taux de recouvrement entre libellés ce qui permet d'avoir des graphes plus lisibles. Il est également possible de colorier les libellés en fonction d'une variable qualitative, de ne mettre les libellés que de certains points : par exemple ceux qui ont le plus contribué à la construction du plan factoriel, ceux qui sont projetés avec une qualité de représentation suffisante, ou encore de les sélectionner par leur nom. Là encore, cela permet d'améliorer la lisibilité des graphes. La figure 1' montre un graphe d'ACP obtenu en coloriant les individus en fonction d'une variable qualitative à 2 modalités (athlètes participant à un décathlon lors d'un Decastar ou lors de Jeux Olympiques). Seuls les individus ayant une bonne qualité de représentation dans le plan (cosinus carré supérieur à 0.6) sont coloriés, les autres individus sont positionnés mais dessinés avec une certaine transparence et sans libellé). Le graphe 1 correspond au graphe directement obtenu avec les lignes de code suivantes :

```
> library(FactoMineR)
> data(decathlon)
> res.pca <- PCA(decathlon, quanti.sup = 11:12, quali.sup=13)
> plot(res.pca,habillage="Competition",select="cos2 0.6")
```

Cette sélection des libellés les plus intéressants qui permet de conserver la vision de l'ensemble du nuage de points est très utile pour les graphes ayant beaucoup d'éléments, comme par

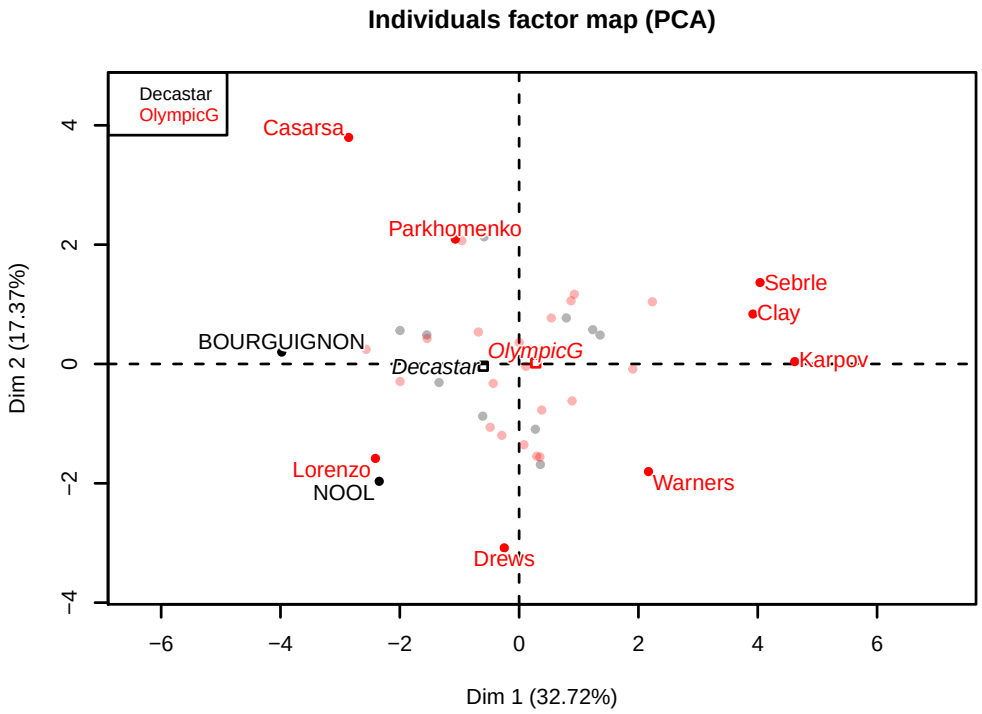

FIGURE 1 – Graphe des individus obtenu directement par `FactoMineR`. Les individus sont coloriés en fonction d'une variable qualitative à 2 modalités et seuls les individus ayant une qualité de projection suffisante (cosinus carré supérieur à 0.6) ont un libellé.

exemple pour l'analyse d'une enquête par ACM ou par l'analyse de données textuelles par AFC. Des vidéos disponibles sur Youtube en français ou en anglais permettent de voir comment utiliser le module graphique, comment gérer les données manquantes, etc.

Références

[1] Husson, F., Josse, J., Lê, S. & Mazet, J. (2013). FactoMineR : Multivariate Exploratory Data Analysis and Data Mining with R, R package version 1.24, `http://factominer.free.fr`.
[2] Josse, J. & Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. Journal de la SFdS, 153(2), p 79-99.
[3] Husson, F. & Josse, J. (2013). missMDA : Handling missing values with/in multivariate data analysis (principal component methods), R package version 1.7.