FactoMineR

Rindra LUTZ / William ROBACHE

17/12/2020

FactoMineR est un package R dédié à l'analyse exploratoire multidimensionnelle de données. Il a été développé et il est maintenu par François Husson, Julie Josse, Sébastien Lê, d'Agrocampus Rennes, et J. Mazet.

Pourquoi FactoMineR ?

- Il permet de mettre en oeuvre des méthodes analyses de données très utilisées telles que l'analyse en composantes principales (ACP), l'analyse des correspondances (AC), l'analyse des correspondances multiples (ACM), ainsi que des analyses plus avancées.
- Il permet l'ajout d'informations supplémentaires telles que des individus et/ou des variables supplémentaires.
- Il fournit un point de vue géométrique et de nombreuses sorties graphiques.
- Il fournit de nombreuses aides à l'interprétation (description automatique des axes, nombreux indicateurs, …).
- Il peut prendre en compte diverses structures sur les données (structure sur les variables, hiérarchie sur les variables, structure sur les individus).
- Beaucoup de matériels pédagogique (MOOC, livres, etc.) est disponible pour expliquer aussi bien les méthodes que la façon de les mettre en oeuvre avec FactoMineR.
- Il gère les données manquantes avec missMDA.
- Il a une interface Shiny qui permet de construire des graphes de façon interactive avec Factoshiny.
- Il propose une interprétation automatique des résultats obtenus avec FactoMineR grâce à FactoInvestigate.

Utilisation pour l'ACP

Nous allons vous montrer l'utilisation du package FactoMineR à travers une ACP (Analyse en Composantes Principales). En effet, c'est un exemple courrant et très utilisé dans le monde de la data, et il serait bien trop long d'expliquer chaque utilisation de ce package pour chacune d'entre elles. D'autant plus que la démarche à suivre est souvent la même !

Il est nécessaire d'installer FactoMineR via la commande :

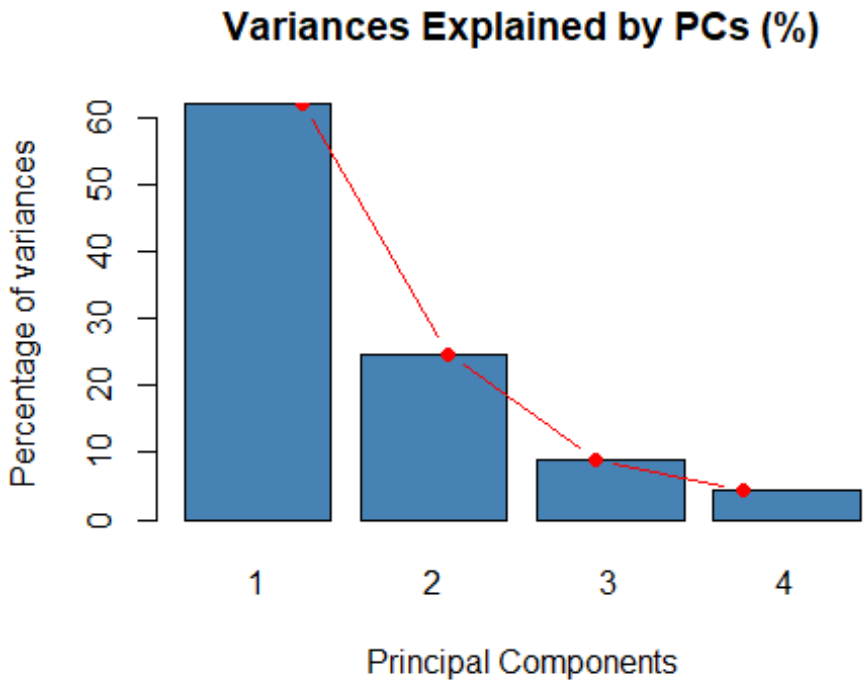
```
install.packages("FactoMineR")
```

Ensuite, nous allons chercher à calculer l'ACP en utilisant les données de démonstration USArrests. Le jeu de données contient des statistiques sur les arrestations par 100 000 habitants pour assaut, meurtre et viol dans chacun des 50 États américains en 1973.

```
library(FactoMineR)
data("USArrests")
res.pca <- PCA(USArrests, graph = FALSE)
```

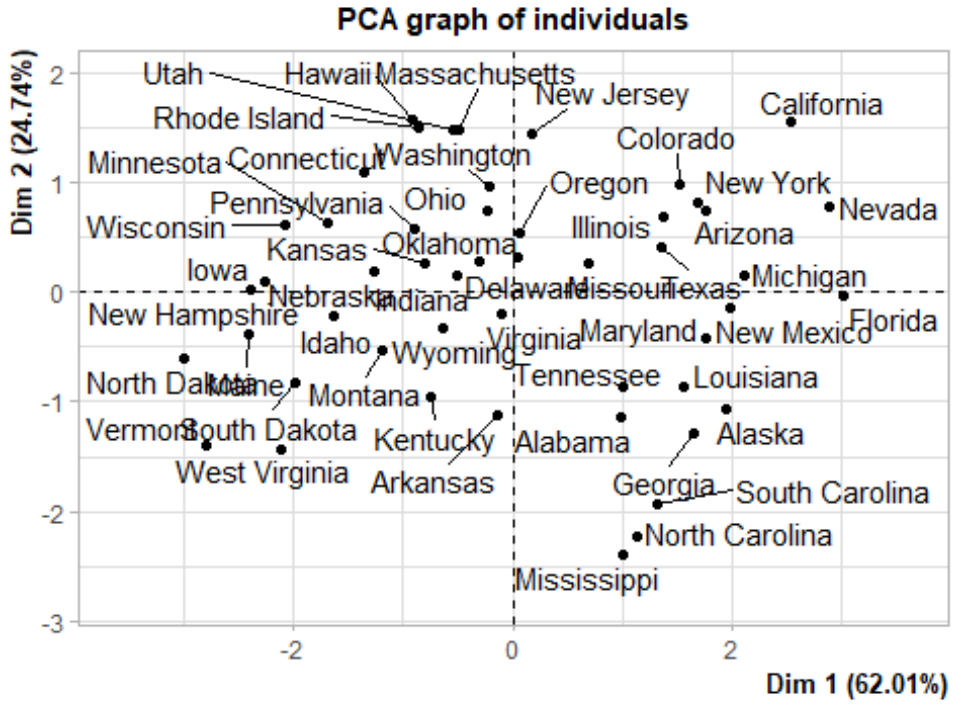
Afin d'avoir une meilleure idée des données, il peut être utile de visualiser les valeurs propres. Le code ci-dessous va nous permettre de montrer le pourcentage de variances expliquées par chaque axe principal.

```
eig.val <- res.pca$eig
barplot(eig.val[, 2],
        names.arg = 1:nrow(eig.val),
        main = "Variances Explained by PCs (%)",
        xlab = "Principal Components",
        ylab = "Percentage of variances",
        col ="steelblue")
# Add connected line segments to the plot
lines(x = 1:nrow(eig.val), eig.val[, 2],
      type = "b", pch = 19, col = "red")
```



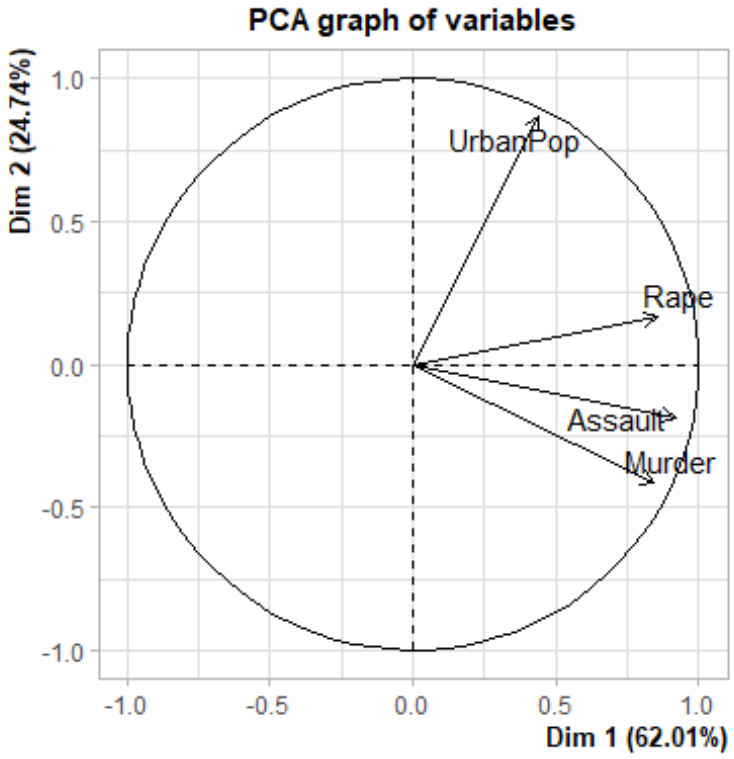
A présent, pour établir un grapahique des individus, il nous faut utiliser la commande plot, qui va également automatiquement rassembler les individus similaires ensemble.

```
plot(res.pca, choix = "ind", autoLab = "yes")
```



Nous avons doonc ici notre graphique des individus, mais qu'en est-il du graphique des variables ? Grâce à la commande suivante, nous pouvons avoir un rapide aperçu de ces données. Ici les variables corrélées positivement sont du même côté du graphique, et les variables corrélées négativement sont sur des côtés opposés du graphique.

```
plot(res.pca, choix = "var", autoLab = "yes")
```



Désormais, nous avons les deux graphiques nécessaires résultant de notre ACP. Afin d'accéder aux résultats de cette ACP plus en détail, voici la liste des commandes que nous pouvons utiliser.

```
# Valeurs propres
res.pca$eig
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1  2.4802416              62.006039                           62.00604
## comp 2  0.9897652              24.744129                           86.75017
## comp 3  0.3565632               8.914080                           95.66425
## comp 4  0.1734301               4.335752                          100.00000
```

```
# Résultats des variables
res.var <- res.pca$var
res.var$coord          # Coordonnées
```

```
##              Dim.1      Dim.2      Dim.3       Dim.4
## Murder   0.8439764 -0.4160354  0.2037600  0.27037052
## Assault  0.9184432 -0.1870211  0.1601192 -0.30959159
## UrbanPop 0.4381168  0.8683282  0.2257242  0.05575330
## Rape     0.8558394  0.1664602 -0.4883190  0.03707412
```

```
res.var$contrib        # Contributions aux axes
```

```
##              Dim.1     Dim.2     Dim.3     Dim.4
## Murder   28.718825 17.487524 11.643977 42.149674
## Assault  34.010315  3.533859  7.190358 55.265468
```

```
## UrbanPop  7.739016 76.179065 14.289594  1.792325
## Rape     29.531844  2.799553 66.876071  0.792533

res.var$cos2           # Qualité de représentation

##              Dim.1     Dim.2      Dim.3       Dim.4
## Murder   0.7122962 0.1730854 0.04151814 0.073100217
## Assault  0.8435380 0.0349769 0.02563817 0.095846950
## UrbanPop 0.1919463 0.7539938 0.05095143 0.003108430
## Rape     0.7324611 0.0277090 0.23845544 0.001374491

# Résultats des individus
res.ind <- res.pca$var
res.ind$coord          # Coordonnées

##              Dim.1      Dim.2      Dim.3       Dim.4
## Murder   0.8439764 -0.4160354  0.2037600  0.27037052
## Assault  0.9184432 -0.1870211  0.1601192 -0.30959159
## UrbanPop 0.4381168  0.8683282  0.2257242  0.05575330
## Rape     0.8558394  0.1664602 -0.4883190  0.03707412

res.ind$contrib        # Contributions aux axes

##              Dim.1     Dim.2     Dim.3     Dim.4
## Murder   28.718825 17.487524 11.643977 42.149674
## Assault  34.010315  3.533859  7.190358 55.265468
## UrbanPop  7.739016 76.179065 14.289594  1.792325
## Rape     29.531844  2.799553 66.876071  0.792533

res.ind$cos2           # Qualité de représentation

##              Dim.1     Dim.2      Dim.3       Dim.4
## Murder   0.7122962 0.1730854 0.04151814 0.073100217
## Assault  0.8435380 0.0349769 0.02563817 0.095846950
## UrbanPop 0.1919463 0.7539938 0.05095143 0.003108430
## Rape     0.7324611 0.0277090 0.23845544 0.001374491
```

Il est à noter que nous n'avons ici traiter que le cas d'une ACP lambda. De toutes les méthodes d'analyses de données, cela restait la plus simple à traiter. Il en existe cependant bien d'autres que nous pouvons utiliser grâce à ce package.

Interprétation

En règle générale, l'interprétation des méthodes d'analyses telle que l'ACP se fait de la même manière. Voici quelques points à retenir :

Une contribution trop importante d'un des points à un axe doit être regardée avec prudence (~25% d'inertie). Il est également nécessaire de s'assurer que les points contribuant le plus à l'axe sont bien représentés sur l'axe (sinon il faut les mettre en éléments supplémentaires.)

La contribution est juste une aide à l'interprétation :

- La contribution de certains points peuvent être très légèrement inférieures au seuil et conforter l'interprétation de l'axe que l'on aurait faite sans eux. On les inclut alors dans l'interprétation.

- Inversement, lorsqu'une contribution est très forte par rapport à d'autres qui sont pourtant en dessus du seuil, le point détermine l'axe presque exclusivement.

- L'interprétation des nouvelles variables (des axes factoriels) se fera à l'aide des individus et des variables contribuant le plus à l'axe avec la règle suivante : si une variable a une forte contribution positive à l'axe, les individus ayant une forte contribution positive à l'axe sont caractérisés par une valeur élevée de la variable.