Analyse de données
M1 Statistique et économétrie
V. Monbet

Analyse Factorielle des Correspondances Multiples

1 ACM avec R

Plusieurs packages fournissent des outils permettant de réaliser une analyse factorielle des correspondances. On peut citer :

- dans le package MASS : `mva`
- dans le package ade4 : `dudi.acm`
- dans le package FactoMineR (cf. `http://factominer.free.fr`) : `MCA`

Les résultats les plus complets semblent être ceux fournis par la procédure MCA. Nous allons donc principalement utiliser FactoMineR.

1.1 Les données et les tableaux

Nous nous intéressons à un jeu de données fictif. Il comporte les réponses de 10 personnes aux trois questions suivantes :

(a) Êtes-vous un homme ou une femme ?
(b) Quel est votre niveau de revenus : moyen ou élevé ?
(c) Choisissez le dessert que vous préférez parmi les trois suivants : un fruit (A), une crème glacée (B), du chocolat (C) ?

1. Créer les données dans R en exécutant les instructions suivantes.

```
Sexe <- rep(c("F", "M"), c(5, 5))
Revenu <- rep(c("M", "E", "M"), c(2, 5, 3))
Pref <- c("A", "A", "B", "C", "C", "C", "B", "B", "B","A")
Resultats <- data.frame(cbind(Sexe, Revenu, Pref))
print(Resultats)
```

2. Sous quelle forme les données sont-elles stockées dans l'objet `Resultats` ? Obtenir la table de contingence `Resultats.cont`

```
Resultats.cont = as.data.frame(table(Resultats))
```

puis le tableau de Burt `Resultats.burt` et enfin le tableau disjonctif complet `Resultats.disj`. Pour le tableau de Burt et le tableau disjonctif, vous pouvez utiliser les fonctions du programme `divers_afcm.R` qui est sur la page web du cours ou les fonctions `acm.burt` et `acm.disjonctif` du package `ade4`.

1.2 AFCM

1. Réaliser l'AFCM du tableau de données `Resultats` des deux manières élémentaires suivantes :
   (a) L'analyse factorielle des correspondances du tableau de Burt.
   (b) L'analyse factorielle des correspondances du tableau de disjonctif complet.

   Dans les deux cas, vous pouvez utiliser la fonction `CA` du package `FactoMineR`.

2. Quels commentaires pouvez-vous formuler sur les liens entre les résultats des deux analyses précédentes. Il y a 7 modalités pour 3 variables. Vous vous intéresserez donc, en premier lieu, au lien entre les (7-3=4) quatre premières valeurs propres, les seules qui ont un sens statistique. Puis aux liens entre les deux représentations graphiques obtenues.

   Vous pourrez remarquer que les valeurs propres non-nulles de l'analyse factorielle des correspondances du tableau de Burt sont égales au carré des valeurs propres non-nulles de l'analyse factorielle des correspondances du tableau disjonctif complet.

   Comparez aussi les résultats obtenus avec les instructions suivantes :

   ```
   (rapcol <- (Resultats.burt.acm$col$coord/Resultats.disj.acm$col$coord))
   (sqvptdc <- sqrt((Resultats.disj.acm$eig$eig)[1:4]))
   rapcol %*% diag(1/sqvptdc)
   ```

   On observa ainsi que les coordonnées des colonnes, donc celles des modalités des variables, obtenues par AFC du tableau de Burt sont donc égales l'image de celles obtenues par AFC du tableau disjonctif complet par l'application linéaire diagonale égale aux racines carrées des valeurs propres obtenues lors l'AFC du tableau disjonctif complet. Les représentations graphiques des modalités des variables sur les axes factoriels ou sur les plans factoriels sont donc image l'une de l'autre par une affinité. Les graphiques obtenus ne seront donc pas la plupart du temps identiques.

3. Réaliser l'analyse factorielle des correspondances multiples du tableau initial avec la fonction `MCA` du package `FactoMineR`. Quelle est l'approche utilisée par cette fonction : AFC du tableau de Burt ou AFC du TDC ?
4. Quels résultats obtient-on si on fait l'analyse en composantes principales du tableau disjonctif complet ?

2 Fait-on confiance à la science ?

1. Importer le fichier `sciences.csv`. Ce fichier contient des données issues d'une enquête réalisée en 1993 auprès de n = 871 enquêtés sur leur avis sur le rôle de la science. Les enquêtés doivent répondre s'ils sont d'accord ou non avec les déclarations suivantes :
   — A : nous croyons trop souvent en la science et pas assez à l'intuition (aux sentiments et aux croyances).
   — B : globalement les sciences font plus de mal que de bien.
   — C : tous les changements que les hommes font à la nature, aussi scientifiques soient-ils, sont susceptibles d'empirer les choses.
   — D : la science va résoudre les problèmes d'environnement sans affecter notre faccon de vivre.

   Il y a cinq réponses possibles : 1 tout à fait d'accord ; 2 d'accord ; 3 pas d'avis ; 4 pas d'accord ; 5 absolument pas d'accord. Puis des questions de signalétique leur ont été posées :
   — Sexe : Homme (sex1), Femme (sex2)
   — Age (6 classes) : 16-24 (age1), 25-34, 35-44,45-54,55-64, 65 et plus (age6)
   — Niveau d'éducation (6 niveaux) : enseignement primaire (edu1), niveau collège, niveau lycée, niveau Bac + 2, niveau Bac + 5, niveau Bac + 8 (edu6)

   Le but de l'étude est de caractériser, de décrire la population de cette enquête, de faire une typologie des enquêtés en fonction de leur attitude vis-à-vis des sciences.
2. Réaliser et commenter l'analyse du jeu de données sciences.csv, tous les coups sont permis !
3. Interpréter finement la proximité des modalités "C5" et "B5".
4. Y-a-t'il un lien entre le niveau d'éducation et le comportement face à la science ?
5. Comment sont positionnées les modalités Homme et Femme ?
6. Comment peut-on faire pour étudier le comportement des femmes qui ont un haut niveau d'éducation et des hommes qui ont un haut niveau d'éducation ? Que concluez-vous ?
7. Comment interpréter la distance entre deux individus ? Quelle distance est utilisée en AFCM ?
8. Comment peut-on expliquer le faible rapport de corrélation entre la 4ème variable D et les deux premières dimensions ? Ou comment peut-on interpréter la position des modalités de la variables D ?
9. Peut-on avoir un test de significativité des contributions des variables ? Effectuer l'analyse

3 Pour ceux qui vont vite, un autre exemple

3.1 Les données

Nous considérons une partie des données issues de l'enquête "Les étudiants et la ville" effectuée en 2001 par des étudiants de sociologie sous la direction de S. Denèfle à l'Université de Tours. Cet exemple est décrit dans : Crucianu M., Asselin de Beauville J-P., Boné R., Méthodes factorielles pour l'analyse des données, Hermès-Lavoisier 2004.

L'analyse porte sur cinq questions en rapport avec le logement étudiant. L'ensemble des individus statistiques est ici un échantillon de 383 étudiants. Les questions sont les suivantes reportées dans la table ci-dessous.

Question	**Réponse possible**	**Abbréviation**
Habitez-vous (variable "mode d'occupation")	Seul	*Seul*
	Colocataires	*Coloc*
	En couple	*Couple*
	Avec les parents	*Parents*
	Non réponse	*NR1*
Quel type d'habitation occupez-vous ? (variable "type d'habitation")	Cité universitaire	*Cité*
	Studio	*Studio*
	Appartement	*Appart*
	Chambre chez un particulier	*Chambre*
	Autre	*Autre*
	Non réponse	*NR2*
Si vous vivez en dehors du foyer familial, depuis combien de temps ? (variable "ancienneté")	Moins de 1 an	*< 1 an*
	De 1 à 3 ans	*1-3 ans*
	Plus de 3 ans	*> 3 ans*
	Non applicable	*NA*
	Non réponse	*NR3*
A quelle distance approximative de la Fac vivez-vous ? (variable "éloignement")	Moins de 1 km	*< 1 km*
	De 1 à 5 km	*1 à 5 km*
	Plus de 5 km	*> 5 km*
	Non réponse	*NR4*
Quelle est la superficie de votre logement ? (variable "superficie")	Moins de 10 m^2	< 10 m^2
	De 10 à 20 m^2	10 à 20 m^2
	De 20 à 30 m^2	20 à 30 m^2
	plus de 30 m^2	>30 m^2
	Non réponse	NR5

Pour l'ACM, il est quasiment indispensable de regrouper les modalités dont la fréquence est trop faible (inférieure à 5% par exemple) avec d'autres modalités. Aussi, dans les données qui suivent, les modalités "Parents" et "NR1" ont été regroupées pour la variable "mode", de même que "NA"

et "NR3" pour la variable "ancienneté" et ">5km" et "NR4" pour la variable "éloignement". Il reste donc 22 modalités distinctes.

3.2 Analyse

1. Lire les données à partir du fichier `Etudiants-ville.csv`.
2. Identifier les modalités à faible effectif (par exemple inférieur à 5%).
3. Réaliser l'analyse des correspondances multiples sur les modalités.
4. On peut adapter les commandes de la question précédente pour réaliser l'analyse des correspondances aussi sur les individus en ajoutant la variable identifiant les individus.
5. Projeter les individus et les modalités sur le premier plan factoriel et commenter le graphique.
6. Les variables `Type` et `Superf` sont redondantes. Comment peut-on faire pour pour définir la variable `Superf` en variable supplémentaire ? Refaire les graphiques. L'interprétation des axes est-elle rendue plus aisée ? Vous semble t'il intéressant de passer d'autres variables en variable supplémentaire ?