TP 4 - ACM : mise en oeuvre des calculs avec R et interprétations et éléments supplémentaires

L2 MIASHS - Université Lyon 2 - 2020/2021

Responsable : Julien Ah-Pine, Stéphane Chrétien et Antoine Gourru

1 Objectif du TP

L'objectif de la séance est de mettre en pratique, à l'aide du langage R, la méthode ACM telle que vue en cours. A l'aide d'un exemple numérique, nous allons effectuer pas à pas les différents calculs permettant de réaliser l'ACM. Comme pour les précédentes séances, il est important de faire le TP accompagné du CM.

Nous rappelons que l'ACM est une AFC d'un tableau disjonctif complet. Le TP5 et TP6 pourront donc être une source intéressante pour la réalisation de ce TP. Nous rappelons également que le tableau disjonctif étant une matrice avec des propriétés particulières, les formules de l'AFC dans le cadre de l'ACM se simplifient et peuvent être exprimées en fonction du tableau disjonctif complet.

Le cas d'étude auquel nous nous intéressons est extrait de la référence suivante : *Husson F., Le S., Pagès, J.. (2016). "Analyse de données avec R". Des Presses Universitaires de Rennes.*

L'étude de cas concerne une véritable enquête à propos de la consommation de thé. 300 consammateurs de thè sont interrogés au travers d'un formulaire avec questions à choix fermés. L'objectif est de comprendre à partir de cet échantillon, les différents comportements et l'image que les consommateurs ont du produit.

Les variables actives sont au nombre de 18 au total. Elles peuvent êtres séparées en plusieurs groupes :

- Le moment de la consomation du thé :
  - breakfast (yes, no),
  - tea.time (à 16h) (yes, no),
  - evening (yes, no),
  - lunch (yes, no),
  - dinner (yes, no),
  - always (yes, no).
- Le lieu de la consommation du thé :
  - home (yes, no),
  - work (à 16h) (yes, no),
  - tearoom (yes, no),
  - friends (yes, no),
  - resto (yes, no),
  - pub (yes, no).
- Le type de thé consommé et la façon de le consommer :
  - Tea (black, Earl grey, green),
  - How (alone, lemon, milk, other),
  - sugar (yes, no),
  - how (tea bag, tea bag+unpackaged, unpackaged),
  - where (chain store, chain sotre+tea shop, tea shop),
  - price (p_branded, p_cheap, p_private label, p_unknown, p_upscale, p_variable).

Nous étudierons également deux variables illustratives :

- age (continue),

— sex (F, M).

2 Chargement du fichier de données RData

1. Téléchargez le fichier `acm.RData` sur votre disque dur et chargez celui-ci dans votre espace de travail.
2. Identifiez les différentes variables à votre disposition et faites le lien avec la présentation de l'étude donnée dans la section précédente. A l'aide des commandes `nrow` et `ncol`, stockez dans `n` et `p`, le nombre d'individus et de variables respectivement.

3 Tableau disjonctif complet

La première étape est de déterminer $\mathbf{Z} = (z_{ik})_{i=1,\ldots,n;k=1,\ldots,q}$ le tableau disjonctif complet.

3. Entrez, exécutez et commentez les commandes suivantes :

```
install.packages("ade4")#installation de la librairie
library(ade4)#chargement de la librairie
Z=acm.disjonctif(T)
Z=as.matrix(Z)#conversion d'un data frame en matrice
summary(Z)
```

   Prenez le temps d'appréhender la variable `Z`.

4. Stockez dans `q` le nombre total de modalités (toute variable confondue) et stockez `list.mod` la liste de toutes les modalités. Pour cela, utilisez la commande `colnames` que vous appliquerez à `Z`.

4 Analyse du nuage des individus

5. Calculez et stockez dans `F` la matrice $\mathbf{F}$ de terme général $f_{ik} = z_{ik}/(np)$.
6. Calculez les marges de `Z` que vous stockerez dans les variables `F.marges.ind` et `F.marges.mod`.
7. Stockez ensuite dans `D.ind` et `D.mod` les matrices diagonales des marges.
8. Stockez dans la variable `L` la matrice des points lignes (individus). Vérifiez que pour chaque profil ligne la somme de ses composantes fait 1.
9. Stockez dans la variable `S`, la matrice $\mathbf{S}$ dont les vecteurs propres sont les axes factoriels permettant de représentez les points lignes dans un espace réduit.
10. Procédez à la décomposition spectrale de `S`. Vous stockerez le résultat dans `S.eigen`. Remarque : en raison d'instabilités numériques, il est possible que les éléments propres contiennent des nombres complexes. La partie imaginaire est alors négligeable et pour se ramener à des rééls vous pourrez utiliser la commande `Re` comme suit :

```
S.eigen$values=Re(S.eigen$values)
S.eigen$vectors=Re(S.eigen$vectors)
```

11. Comme pour l'AFC, enlevez la première et le premier vecteur propre de `S.eigen`.
12. Tracez l'histogramme des valeurs propres de `S`. Combien d'axes proposez vous de garder ?
13. Stockez dans deux variables `u1` et `u2`, les deux premiers axes factoriels $\mathbf{u}_1$ et $\mathbf{u}_2$. Attention ! normez ces vecteurs au sens de la métrique $\mathbf{D}_M^{-1}$. Dans notre cas, $\mathbf{D}_M$ est stockée dans `D.mod`.
14. Calculez les coordonnées des individus sur $\mathbf{u}_1$ et $\mathbf{u}_2$. Vous stockerez ces deux vecteurs dans les variables `f1` et `f2`. Attention ! n'oubliez pas d'utiliser la bonne métrique pour le calcul des projections qui est également $\mathbf{D}_M^{-1}$.
15. Vérifiez que les coordonnées factorielles des individus sont centrées en 0.

16. Calculez la somme des valeurs propres de $\mathbf{S}$ que vous stockerez dans la variable `int`. Vérifiez que vous obtenez bien $q/p - 1$ comme vu en Cm.
17. Calculez le pourcentage de l'inertie associée à l'axe $\mathbf{u}_1$. Faites de même pour l'axe factoriel $\mathbf{u}_2$.
18. Représentez sur le premier plan factoriel les profils lignes.

5 Représentation des points colonnes (modalités) dans le premier plan factoriel des points lignes

Comme en AFC, on peut représenter de façon cohérente les points colonnes dans le premier plan factoriel du nuage des points lignes.

19. Calculez la matrice des points colonnes $\mathbf{C}$ que vous stockerez dans la variable `C`.
20. Calculez ensuite les facteurs des modalités $\mathbf{g}_1$ et $\mathbf{g}_2$ (coorodonnées sur les axes $\mathbf{v}^1$ et $\mathbf{v}^2$ -qu'il n'est pas nécessaire de déterminer-) à l'aide des relations barycentriques. Vous stockerez ces vecteurs dans les variables `g1` et `g2`.
21. Ajoutez ces points, et leur libellé, sur le premier plan factoriel où sont déjà représentés les individus. Pour cela utilisez les commandes `points` et `text`. Vous utiliserez des triangles bleus pour les modalités.
22. La représentation graphique simultanée étant très chargée, à l'aide de la commande `plot`, éditez un nouveau gaphique représentant uniquement les points colonnes, et leur libellé, sur le premier plan factoriel.

6 Interprétations des modalités dans le premier plan factoriel

23. Calculez la qualité de la représentation [1] de chaque point colonne sur les axes $\mathbf{v}^1$ et $\mathbf{v}^2$. Vous stockerez les résultats dans les variables `qlt.mod.u1` et `qlt.mod.u1`. Quelles sont les modalités les mieux représentées sur les parties positives et négatives des deux axes ?
24. Calculez la contribution de chaque point colonne à la constitution des axes $\mathbf{v}^1$ et $\mathbf{v}^2$. Vous stockerez les résultats dans les variables `ctr.mod.u1` et `ctr.mod.u1`. Quelles sont les modalités contribuant le plus aux axes ?
25. Ecrivez un texte donnant vos interprétations sur la représentation des modalités sur le premier plan factoriel.

7 Ajout d'une variable quantitative supplémentaire sur le premier plan factoriel

Comme en ACP, nous pouvons ajouter *a posteriori*, à un plan factoriel issu d'une AFC ou d'une ACM, une variable quantitative supplémentaire que l'on notera $\mathbf{x}$ et qui est un vecteur de $\mathbb{R}^n$. Dans ce cas, les coordonnées de $\mathbf{x}$ sur un axe $\mathbf{v}^m$ est simplement $r(\mathbf{x}, \mathbf{v}^m)$, le coefficient de corrélation entre la variable et l'axe.

26. Dans le CM, en AFC, nous avons vu la formule de transition suivante :

$$\mathbf{f}^m = \sqrt{\lambda_m}\mathbf{D}_P^{-1}\mathbf{v}^m$$

De cette équation nous inférons la relation suivante :

$$\mathbf{v}^m = \frac{1}{\sqrt{\lambda_m}}\mathbf{D}_P\mathbf{f}^m$$

où dans notre cas $\mathbf{D}_P$ est stocké dans `D.ind`.
Implémentez cette formule en R afin de déterminer $\mathbf{v}^1$ et $\mathbf{v}^2$ à partir des variables `f1` et `f2`. Vous stockerez les résultats dans `v1` et `v2`.

1. Comme pour l'AFC, nous pourrez calculer au préalable les normes au carré des modalités dans l'espace initial.

27. A l'aide de la commande `cor` (consultez l'aide si besoin), déterminez les coordonnées de la variable `age` (donnée dans la première colonne de `Tsup`) sur les axes $\mathbf{v}^1$ et $\mathbf{v}^2$. Vous stockerez les résultats dans `g1.age` et `g2.age`.
28. Ajoutez en rouge cette variable et son libellé, sur le plan factoriel et commentez sa position.

8 Ajout d'une variable quantitative supplémentaire sur le premier plan factoriel

Nous pouvons aussi ajouter *a posteriori*, au plan factoriel, les modalités d'une variable quanlitative. Nous utiliserons ici la variable `sex` donnée dans la seconde colonne de `Tsup`.

29. Entrez, exécutez et commentez les commandes suivantes :

```
Zsup=acm.disjonctif(as.data.frame(Tsup$sex))
```

30. Entrez et exécutez la commandes suivante :

```
g1.F=sum(Zsup[,1]*f1)/(sum(Zsup[,1])*sqrt(S.eigen$values[1]))
```

Cette commande est l'implémentation d'une formule barycentrique donnée en CM permettant de calculer les coordonnées d'une modalité supplémentaire sur les axes $\mathbf{v}^1$ et $\mathbf{v}^2$ à partir des facteurs $\mathbf{f}^1$ et $\mathbf{f}^2$. En particulier il s'agit ici de la coordonnée de la modalité `F` sur l'axe $\mathbf{v}^1$. Identifiez dans le CM de quelle formule il s'agit.

31. A l'aide de cette formule barycentrique, calculez les coordonnées factorielles restantes des modalités illustratives `F` et `M` sur les axes $\mathbf{v}^1$ et $\mathbf{v}^2$.
32. Ajoutez en rouge ces modalités, et leur libellé, sur le plan factoriel et commentez leurs positions.

9 Résultats graphiques attendus