TP 3 - AFC : mise en oeuvre des calculs avec R et interprétations

L2 MIASHS - Université Lyon 2 - 2020/2021

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

1 Objectif du TP et étude de cas

L'objectif de la séance est de mettre en pratique, à l'aide du langage R, la méthode AFC 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'AFC d'une table de contingence. Il s'agit à la fois de mieux maîtriser le langage R et de connaître les commandes permettant de mettre en ouvre les formules du cours.

Au-delà des calculs et code R, le deuxième objectif du TP est d'aborder plus précisément les interprétations que nous pouvons faire une fois les différents axes factoriels déterminés et les nuages de profils projetés dans l'espace réduit. Pour nous aider dans cette analyse, nous pouvons calculer plusieurs mesures permettant d'apprécier la qualité de représentations des profils dans le plan factoriel. De plus, il est possible de projeter des éléments supplémentaires dits illustratifs dans l'espace réduit. Ces éléments permettent d'ajouter de nouvelles informations ce qui conduit à une interprétation des résultats de l'AFC plus riche.

L'ensemble des calculs consistent à coder des équations que nous avons vues en CM. Il est donc important de faire le TP accompagné du CM.

Le cas auquel nous nous intéressons est extrait de la référence suivante : *D. Busca et S. Toutain. (2009). "Analyse factorielle simple en sociologie, Méthodes d'interprétation et études de cas". De Boeck.* Cet exemple s'inscrit dans le champ disciplinaire de la sociologie de l'électorat et présente les données recueillies lors d'une enquête concernant l'élection présidentielle de 2007. Deux variables qualitatives nominales sont étudiées :

— `csp` : la catégorie socio-professionnelle possédant 8 modalités.
— `cand` : les candidats au 1er tour de l'élection avec 12 prétendants.

Les données recueillies concernent les intentions de votes sur un échantillon d'environ 4,493,000 individus. Les 8 catégories socio-professionnelles sont : Agriculteurs ; Artisans, commerçants et chef d'entreprise ; Professions libérales, cadres supérieurs ; Professions intermédiaires ; Employés ; Ouvriers ; Etudiants ; Chômeurs. Les 12 candidats sont : Schivardi ; Laguiller ; Besancenot ; Buffet ; Bové ; Royal ; Voynet ; Nihous ; Bayrou ; Sarkozy ; de Villiers ; Le Pen.

L'objectif global de l'étude est donc d'étudier les relations entre ces deux variables et de permettre d'appréhender des questions du type "qui a tendance à voter pour qui" ?

2 Chargement du fichier de données RData

1. Téléchargez le fichier `afc.RData` sur votre disque dur et chargez celui-ci dans votre espace de travail.
2. Identifiez les différentes variables à votre disposition. Dans le tableau de contingence, quelle variable qualitative est en ligne et quelle est celle qui est en colonne ?

3 Test du χ^2

Nous appliquons en premier lieu un test du χ^2 afin de vérifier le rejet de l'indépendance entre les deux variables à l'étude ce qui justifierait une analyse plus profonde par le biais d'une AFC.

3. Déterminez à partir de la variable `N` stockant la table de contingence, la table des fréquences relatives que vous stockerez dans une variable `F`.
4. Déterminez les marges des deux variables `cand` et `csp`, que vous stockerez respectivement dans `F.marges.cand` et `F.marges.csp`.
5. Entrez et exécutez les commandes suivantes :

   ```
   F.ind=matrix(F.marges.cand,ncol=1)%*%matrix(F.marges.csp,nrow=1)
   ```

   Que représente `F.ind` ?
6. A partir de `F` et `F.ind`, déterminez la valeur du ϕ^2. Que représente également cette grandeur ?
7. Déduisez-en la valeur de la statistique du χ^2 que vous stockerez dans une variable `chi2`.
8. Notons p et q le nombre de modalités des variables à l'étude. Sachant que dans notre cas, la statistique du χ^2 suit une loi du χ^2 à $(p-1)(q-1)$ degrés de liberté, déterminez le seuil de rejet avec une erreur de première espèce de 5%. Pour cela vous utiliserez la commande `qchisq` dont vous pourrez consulter l'aide si besoin.
9. Rejetez vous l'hypothèse nulle d'indépendance entre les deux variables à l'étude ?

4 Analyse du nuage des profils lignes

10. Stockez dans les variables `D.cand` et `D.csp` les matrices diagonales des marges des deux variables.
11. Stockez dans la variable `L` la matrice des profils lignes. Vérifiez que pour chaque profil ligne la somme de ses composantes fait 1.
12. Stockez dans la variable `S`, la matrice $\mathbf{S}$ dont les vecteurs propres sont les axes factoriels permettant de représenter les profils lignes dans un espace réduit.
13. Procédez à la décomposition en éléments propres de $\mathbf{S}$ et vous stockerez le résultat de cette décomposition dans la variable `S.eigen`.
14. Vérifiez que la 1ère valeur propre vaut 1. Commentez.
15. Entrez et exécutez les commandes suivantes :

    ```
    S.eigen$values=S.eigen$values[-1]
    S.eigen$vectors=S.eigen$vectors[,-1]
    ```

    A quoi correspondent ces opérations ?
16. Tracez l'histogramme des valeurs propres de `S`. Combien d'axes proposez vous de garder ?
17. Stockez dans deux variables `u1` et `u2`, les deux premiers axes factoriels $\mathbf{u}_1$ et $\mathbf{u}_2$. Attention ! R normalise les vecteurs propres de sorte à ce que leurs normes soient unitaires mais cela est au sens de la métrique canonique classique et non pas au sens de la métrique appropriée en AFC (ie $\mathbf{D}_Q^{-1}$ pour $\mathbb{R}^q$). Ainsi, il faut calculer $\langle \mathbf{u}_m, \mathbf{u}_m \rangle_{\mathbf{D}_Q^{-1}}$ la norme du vecteur propre au sens de $\mathbf{D}_Q^{-1}$ et multiplié le résultat obtenu avec R par la racine carrée de cette valeur. Dans notre cas, $\mathbf{D}_Q$ est stockée dans `D.csp`.
18. Calculez les composantes factorielles associées aux deux premiers axes. Il s'agit des 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 $\mathbf{D}_Q^{-1}$.
19. Vérifiez la propriété suivante : $\sum_{i=1}^{p} f_{i.} f_i^1 = 0$.

20. Calculez la somme des valeurs propres de $\mathbf{S}$ que vous stockerez dans la variable `int`. A quoi correspond cette valeur ? Faites le lien avec une question précédente.
21. Calculez le pourcentage de l'inertie associée à l'axe $\mathbf{u}_1$. Faites de même pour l'axe factoriel $\mathbf{u}_2$.
22. Représentez sur le premier plan factoriel les profils lignes.

5 Représentation des profils colonnes dans le premier plan factoriel des profils lignes

23. En AFC et contrairement à l'ACP, on peut représenter de façon cohérente les profils colonnes dans le premier plan factoriel du nuage des profils lignes. Pour cela calculez les facteurs $\mathbf{g}_1$ et $\mathbf{g}_2$ des profils colonnes à l'aide des relations barycentriques. Vous stockerez ces vecteurs dans les variables `g1` et `g2`.
24. Entrez et exécutez les commandes suivantes :

```
points(g1,g2,pch=2,col="blue",cex=0.75)
text(cbind(g1,g2),labels=liste_csp,pos=1,cex=1,offset=0.3,col="blue")
```

Que fait la commande `points` ainsi que les différents paramètres associés ?

6 Interprétations dans le premier plan factoriel

Nous allons restreindre notre étude au premier plan factoriel des profils lignes engendré par $\mathbf{u}_1$ et $\mathbf{u}_2$, au sein duquel nous avons également représenter les profils colonnes. Cependant, les calculs qui seront mis en oeuvre ci-dessous peuvent également être appliqués au plan factoriel des profils colonnes engendrés par $\mathbf{v}^1$ et $\mathbf{v}^2$. Je vous encourage à le faire chez vous.

25. Afin de calculer la qualité de la représentation des profils lignes sur un axe, il nous faut dans un premier temps déterminer la norme de ces vecteurs centrés par rapport au barycentre (marge de la variable CSP) en utilisant la métrique appropriée. Dans cette perspective, entrez, exécutez et commentez les commandes suivantes :

```
L.cent=scale(L,center=F.marges.csp,scale = FALSE)
L.cent%*%solve(D.csp)%*%t(L.cent)
diag(L.cent%*%solve(D.csp)%*%t(L.cent))
cand.norm=diag(L.cent%*%solve(D.csp)%*%t(L.cent))
```

26. Calculez la qualité de la représentation de chaque profil ligne sur l'axe $\mathbf{u}_1$. Vous stockerez le résultat dans la variable `qlt.cand.u1`. Quels sont les candidats les mieux représentés sur ce premier axe factoriel (pour ce faire vous pourrez comparer la qualité d'une modalité par rapport à une tendance centrale telle que la médiane) ? Mêmes questions pour l'axe $\mathbf{u}_2$.
27. Calculez la contribution de chaque profil ligne à la constitution de l'axe $\mathbf{u}_1$. Vous stockerez le résultat dans la variable `ctr.cand.u1`. Quels sont les candidats contribuant le plus à $\mathbf{u}_1$? Mêmes questions pour l'axe $\mathbf{u}_2$.
28. Calculez la qualité de la représentation de chaque profil colonne sur l'axe $\mathbf{u}_1$. Vous stockerez le résultat dans la variable `qlt.csp.u1`. Quels sont les csp les mieux représentés sur ce premier axe factoriel ? Mêmes questions pour l'axe $\mathbf{u}_2$.
29. Calculez la contribution de chaque profil colonne à la constitution de l'axe $\mathbf{u}_2$. Vous stockerez le résultat dans la variable `ctr.csp.u1`. Quels sont les csp contribuant le plus à $\mathbf{u}_1$? Mêmes questions pour l'axe $\mathbf{u}_2$.
30. A l'issue des mesures calculées dans les questions précédentes, proposez une interprétation de chaque axe en répondant notamment aux questions suivantes :
    — Quels types de candidats oppose l'axe $\mathbf{u}_1$? Même question pour l'axe $\mathbf{u}_2$.
    — Quels types de csp oppose l'axe $\mathbf{u}_1$? Même question pour l'axe $\mathbf{u}_2$.
    — Quelles csp ont tendance à voter pour quels candidats ?

— Quelles csp ont tendance à ne pas voter pour quels candidats ?

31. Vous trouverez en section 9 un exemple d'interprétation des résultats de l'AFC. Comparez vos conclusions à cet exemple.

7 Ajout d'un candidat supplémentaire sur le premier plan factoriel

L'objectif est d'effectuer l'AFC en considérant un ensemble de candidats actifs et un ensemble de candidats supplémentaires. Nous allons supposer que le candidat "Le Pen" n'est pas actif. Il sera donc enlevé du tableau de contingence sur lequel s'effectueront les calculs, mais il sera stocker dans une variable a part et nous le projetterons a posteriori sur le plan factoriel.

32. Stockez dans une variable `N.cand.sup` la ligne du tableau de contingence relative au candidat supplémentaire.
33. Enlever de la variable `N` la ligne relative au candidat supplémentaire. Enlever également de la variable `liste_cand` l'élément correspond au candidat supplémentaire.
34. Identifiez dans votre implémentation précédente, les lignes de commandes essentielles permettant de mettre en oeuvre l'AFC du nuage des profils lignes. Réutilisez alors ces lignes en faisant du "copier-coller" et en apportant si nécessaire quelques modifications afin de déterminer le premier plan factoriel sur les candidats actifs uniquement. Le résultat attendu est la projection sur le premier plan factoriel de tous les éléments actifs qu'il s'agisse des profils lignes ou des profils colonnes.
35. Projetez le candidat supplémentaire sur le premier plan factoriel déterminé à la question précédente. Vous utiliserez la couleur rouge pour la représentation graphique de cet élément afin de le distinguer des autres.
36. Commentez le résultat obtenu.

8 Equivalence avec l'AFC des profils colonnes

L'objectif est de procéder à l'analyse du nuage des profils colonnes $\mathbb{NC}$ et de vérifier que celle-ci conduit exactement au même plan factoriel qui a été trouvé précédemment.

37. En utilisant au maximum les lignes de commande déjà écrites, procédez à l'analyse du nuage des profils colonnes et vérifiez que la représentation graphique sur le premier plan factoriel engendré par $\mathbf{v}_1$ et $\mathbf{v}_2$ est identique à celle engendra par $\mathbf{u}_1$ et $\mathbf{u}_2$.
Remarque : il arrive en calcul numérique que les résultats de la commande `eigen` soient des nombres complexes alors que cela ne devrait pas être le cas. Dans ce cas, vous pourrez remarquer que la partie imaginaire des nombres est nulle dans la très grande majorité des cas. En pratique, nous pouvez donc vous restreindre à la partie réelle et pour cela vous pouvez utiliser la commande `Re`.

9 Exemple d'interprétations selon les axes

Voici ci-dessous des extraits des interprétations issues de la référence [1] d'où est extrait cette étude de cas :

L'axe 1 oppose deux profils. Le premier profil (P1) associe les "Ouvriers" aux candidats Schivardi, Laguiller, Nihous et Le Pen. Les coordonnées négatives de Le Pen sur l'axe 2 invitent à s'interroger sur la spécificité de l'électorat potentiel de ce candidat. Il semble que les "Agriculteurs" pourraient constituer une composante de l'électorat de Le Pen. En effet, la csp "Agriculteurs" présente des coordonnées négatives sur l'axe 2 au même titre que le candidat. Cependant, le plan factoriel ne permet pas à lui seul d'associer de façon certaine Le Pen aux "Agriculteurs"...

1. *D. Busca et S. Toutain. (2009). "Analyse factorielle simple en sociologie, Méthodes d'interprétation et études de cas". De Boeck.*

Le profil P1 s'oppose aux "Professions intermédiaires", "Professions intellectuelles et cadres supérieurs" dont les intention de vote se portent sur Bayrou (profil P2). L'attachement des "Professions intermédiaires" au candidat Bayrou demande à être précisé par l'analyse d'autres plans factoriels...

L'axe 2 oppose deux profils. Le troisième profil montre que l'électorat potentiel de Sarkozy et de de Villiers se compose d'"Artisans, commerçants" et d'"Agrculteurs" (profil P3). Néanmoins les "Agriculteurs" semblent se déclarer davantage en faveur de de Villiers et les "Artisans, commerçants" pour Sarkozy, en raison de la plus grande proximité des coordonnées sur le plan entre ces candidats et ces csp. En parallèle, si le plan n'amène pas de façon catégorique à associer les "Professions libérales et cadres supérieurs" à l'électorat de Sarkozy, leurs coordonnées positives sur l'axe 1 laissent supposer que ses électeurs déclarés appartiennent à la csp "Professions libérales et cadres supérieurs".

Le quatrième profil souligne que les "Etudiants" et "Chômeurs" privilégient les candidats Besancenot et Buffet (profil P4). Néanmoins, ces candidats semblent attirer une intention de vote ouvrier du fait de leurs coordonnées négatives sur le premier axe.

Enfin, le plan isole la candidate Royal (profil P5). Une analyse des coordonnées sur les axes des fréquences actives conduit à intégrer les catégories "Chômeurs", "Etudiants", et "Professions intermédiaires" à son électorat potentiel.

Vous pouvez également consulter la référence en note de bas de page pour une interprétation de Bernard Denni (Prof. de science-po à l'IEP de Grenoble) plus ancrée dans le cadre des théories et concepts en sociologie sur ces questions de choix électoral.