

Corrigé de l'examen d'Analyse des Données - ESSAI 1

Session de mai 2019

Exercice 1

1. (1pt) Quelle relation existe-t-il entre l'ACM et l'Analyse Factorielle des Correspondances (AFC) ?

Réponse : L'ACM est une AFC sur le tableau disjonctif complet (TDC).

2. (1pt) Calculer le taux d'inertie cumulé des 2 premiers axes de cette ACM.

Réponse : Si $m = 11$ est le nombre de variables et $p = 22$ est le nombre de modalités alors l'inertie totale de cette ACM est donnée par la formule $\frac{p}{m} - 1$ ce qui donne 1 et le taux d'inertie cumulé des 2 premiers axes est donné par $(0.335 + 0.129)/\text{Inertie totale} = 0.464$ soit 46.4%.

3. (2pts) Discuter le nombre d'axes à retenir en vous basant sur 3 critères différents.

Réponse : 1) Retenir les axes dont les valeurs propres sont supérieures à la moyenne soit $\frac{1}{m} = 0.09$ donc les 4 premiers axes. 2) Tracer le graphique des valeurs propres et noter que le coude se trouve au niveau du deuxième axe. 3) Raisonner sur le taux d'inertie cumulé et retenir les 2 premiers axes qui, sur les 11 axes de l'ACM, accumulent quasi la moitié de l'inertie. En conclusion, le choix des 2 premiers axes serait justifié.

4. (2pts) Donner une interprétation de la première carte des modalités. Indiquer la commande R qui permettrait de faciliter l'interprétation de cette carte ?

Réponse : Il suffisait d'identifier les groupes de modalités qui sont proches dans la carte. Par exemple, le groupe de modalités à droite, constitué des modalités *Cheese_n*, *Mayo_n*, *Diarrhea_n*, *Abdo_n* et *Fever_n* et dont l'interprétation est la suivante : les enfants qui n'ont pas mangé du fromage et de la mayonnaise et qui n'ont ni diarrhée ni douleur abdominale ni fièvre. Ce groupe est opposé sur le premier axe au groupe *Cheese_y*, *Mayo_y*, *Diarrhea_y*, *Abdo_y*, *Fever_n*, *Vomit_y* et *Nausea_y* ce qui signifierait que les symptômes diarrhée et vomissement seraient liés à la consommation de fromage et de la mayonnaise.

La commande R n'est autre que `dimdesc(res.mca)`.

5. (1.5pts) Donner une interprétation de la première carte des individus. On commentera uniquement les individus dont le numéro apparaît sur la carte.

Réponse : Pour interpréter cette carte, on se base sur l'interprétation de la carte des modalités (cf. question 4). Ainsi, le groupe d'enfants dont les numéros sont 38, 53, ..., 26 situé du côté droit de la carte serait caractérisé par les modalités *Cheese_n*, *Mayo_n*, *Diarrhea_n*, *Abdo_n* et *Fever_n* et qui sont donc les enfants sains. Les 2 autres groupes qui s'opposent à ce premier groupe sur le premier axe sont plutôt caractérisés

par *Cheese_y*, *Mayo_y*, *Diarrhea_y*, *Abdo_y*, *Fever_n*, *Vomit_y* et *Nausea_y*. Le fait que ces 2 groupes s'opposent sur le deuxième axe est lié au fait que le groupe 44, ..., 23 n'a pas mangé de courgette tandis que le groupe 6, ..., 37 n'a pas mangé de pomme de terre.

6. (1pt) Quelle démarche proposeriez vous pour une interprétation plus précise et plus détaillée de la carte des individus ?

Réponse : Il faudrait effectuer une classification des enfants sur les facteurs de l'ACM, déterminer la meilleure partition puis décrire les classes issues de cette partition à l'aide de la commande `catdes`.

Exercice 2

1. (1.5pts) Combien de composantes principales devrait-on retenir ?

Réponse : Le choix du nombre d'axes se fait selon trois critères. 1) Kaiser : 3 axes car 3 valeurs propres supérieures à 1. 2) Critère du coude : en tracant la courbe on remarque que le coude se trouve au niveau du deuxième axe. 3) Le taux d'inertie cumulé des deux premiers axes est de $(2.559 + 1.296 + 1)/6$ soit 80.9%. On va donc retenir les 3 premiers axes (il serait correct de retenir les 2 premiers aussi).

2. (2pts) Donner une interprétation des axes retenus.

Réponse : En consultant les coordonnées des variables sur les 3 premiers axes on déduit que : 1) Le premier axe oppose l'Inflation (0.91) la Supérficie (0.70) et le Chomage (0.64) au PNB (-0.77). 2) Le deuxième axe est essentiellement corrélé à la balance Commerciale (0.74). 3) Le troisième axe est essentiellemnt corrélé à la Population (0.91).

3. (2pts) Donner une interprétation de la première carte des individus.

Réponse : L'axe 1 oppose l'Algérie et dans une moindre mesure l'Ethiopie et l'Egypte caractérisés par une Inflation forte, un chômage élevé, une grande Superficie et un PNB faible au Koweit, l'Allemagne et la Finlande. Sur le deuxième axe on note que contrairement à la France, l'Arabie S. est caractérisée par d'importantes valeurs de la balance des échanges commerciaux.

4. (1pt) En vous basant sur la hiérarchie déterminer la meilleure partition des 10 pays.

Réponse : En coupant au niveau du plus haut saut entre 2 paliers successifs on obtient une partition en 2 classes. Toutefois, la partition en trois classes est plus fine et plus intéressante.

5. (2pts) Commenter les résultats de `res_nbclust`.

Réponse : En se basant sur les résultats de l'indice Silhouette donnés par la commande `NbClust`, il est clair que le meilleur nombre de classes est 4 associé à la valeur maximale de cet indice qui est 0.7827. C'est donc la partition en 4 classes qu'il faut retenir d'après ces résultats.

6. (3pts) Dans la suite on considère la partition en 3 classes. En utilisant la fonction `catdes` pour décrire les 3 classes, nous avons obtenu les résultats donnés ci-dessous.

En vous basant sur ces résultats, donner une interprétation des 3 classes.

Réponse : A partir des effectifs des classes, on peut facilement identifier la classe 1 dont l'effectif est 5 et qui contient les pays suivants : Af. du sud, Algérie, Tunisie, Egypte, Ethiopie, la classe 2 dont l'effectif est 3 et qui contient l'Allemagne, la France et la Finlande. La classe 3 contient l'Arabie S. et le Koweït.

La classe 1 est caractérisée par une inflation importante ($v.test > 2$) : la moyenne de l'inflation y est de 25.68 qui est largement supérieure à la moyenne globale qui est de 14.69. Elle est aussi caractérisée par un PNB faible ($v.test < -2$) : en effet, la moyenne dans cette classe y est de 1286 alors que la moyenne globale de cette variable est de 9871.8.

Pour les 2 autres classes il faudrait procéder de la même manière.