TD - Analyse de données

3A CI - 2AD - 2015
ENSAE ParisTech
Thomas Merly-Alpa (thomas.merly-alpa@insee.fr) -
Antoine Rebecq (antoine.rebecq@insee.fr)

1 Analyse en Composantes Principales

Optionnel : installation de RStudio

Rstudio est un IDE libre et gratuit pour R qui présente de nombreuses fonctionnalités très pratiques : visualisation des graphiques, des dataframe, gestion des objets du namespace, auto-complétion, etc. Pour les utilisateurs avancés, RStudio intègre quelques fonctionnalités incontournables, notamment en ce qui concerne la création de *packages*.

Pour la suite de ce TD (ainsi que pour vos travaux futurs de statistiques sous R), il peut être intéressant de prendre le temps d'installer RStudio. Le logiciel peut être téléchargé depuis le lien suivant : `https://www.rstudio.com/products/rstudio/download/`

Quizz de culture générale

Pour cet exercice, nous allons nous intéresser à des données issues d'un site Internet sur lequel sont organisés des quizz de culture générale et des compétitions entre inscrits, qui s'appelle `http://www.learnedleague.com/`. Vous n'avez évidemment pas besoin de vous y inscrire ni d'y participer, mais il convient pour la suite de connaître quelques unes des règles de ce site. En particulier, les questions auxquelles doivent répondre les inscrits appartiennent toutes à l'une des 18 catégories de questions définies par les administrateurs du site, qui vont des Mathématiques au Cinéma en passant par la Géographie. Nous avons à notre disposition les résultats de 100 participants, et en particulier leur score relatif (c'est à dire le pourcentage de questions correctement répondues) pour chacune des catégories.

Question 1. *Dans ce cadre, peut-on réaliser une analyse en composantes principales ? Si oui, que peut-on s'attendre à obtenir ?*

Le jeu de données à utiliser est td1_donnees.csv. Celui-ci compte 101 lignes (une par participant, plus les noms de colonnes), et 19 colonnes, une par catégorie plus une dernière colonne comportant le sexe du participant récupéré à partir des informations complétées sur son profil. Pour faire lire ce fichier à R, il convient

d'utiliser deux fonctions (présentées dans le TD de R). Tout d'abord, on utilise setwd() pour indiquer à R dans quel dossier il doit chercher le fichier. Il faut remplacer dans l'instruction suivante le chemin par celui du dossier dans lequel vous avez placé le fichier td1_donnees.csv. Attention : ne pas oublier d'écrire les barres obliques comme / et non à l'envers.

```
> setwd("C:/Users/XXXXX/Desktop/TD ADD")
```

Ensuite, il suffit d'utiliser la fonction read.csv() pour transformer le fichier en un objet R. La fonction head() permet d'afficher les premières lignes du data.frame obtenu pour connaître sa structure, les noms des variables, etc.

```
> trivia <- read.csv("td1_donnees.csv")
> head(trivia)
  amer_hist      art bus_econ class_music curr_events     film food_drink games_sport
1  0.864865 0.627586 0.884956    0.549020   0.7395830 0.646154   0.730496    0.745946
2  0.738318 0.605505 0.619048    0.363636   0.5303030 0.942857   0.884615    0.348148
3  0.391304 0.227273 0.500000    0.117647   0.0714286 0.562500   0.206897    0.312500
4  0.559322 0.459016 0.533333    0.434783   0.4285710 0.530864   0.353846    0.432099
5  0.826087 0.409091 0.555556    0.411765   0.5714290 0.375000   0.827586    0.500000
6  0.361111 0.230769 0.107143    0.250000   0.3809520 0.780000   0.465116    0.750000
  geography  language lifestyle literature      math pop_music  science television
1  0.896000 0.8222220  0.808081   0.708333 0.8857140  0.783333 0.806897   0.737705
2  0.474860 0.5230770  0.767123   0.806818 0.3529410  0.872180 0.584112   0.875969
3  0.195122 0.0833333  0.142857   0.351351 0.0909091  0.310345 0.181818   0.583333
4  0.368932 0.5675680  0.615385   0.708738 0.9000000  0.236842 0.672131   0.472222
5  0.585366 0.5000000  0.785714   0.675676 0.3636360  0.344828 0.568182   0.416667
6  0.363636 0.5238100  0.458333   0.555556 0.5263160  0.479167 0.346667   0.555556
   theatre world_hist sexe
1 0.722222   0.772532    H
2 0.911765   0.522472    F
3 0.214286   0.378378    H
4 0.600000   0.676768    H
5 0.714286   0.648649    F
6 0.458333   0.354839    H
```

On peut utiliser la commande suivante pour afficher le nom des colonnes du tableau obtenu :

```
> colnames(trivia)
 [1] "amer_hist"   "art"         "bus_econ"    "class_music" "curr_events"
 [6] "film"        "food_drink"  "games_sport" "geography"   "language"
[11] "lifestyle"   "literature"  "math"        "pop_music"   "science"
[16] "television"  "theatre"     "world_hist"  "sexe"
```

Nous allons utiliser le package R appelé *FactoMineR* pour réaliser l'analyse en composantes principales de ces données. La documentation de ce package pour être récupérée à l'adresse suivante `https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf`. N'hésitez pas à vous y référer, directement ou en utilisant la fonction help() de R, tout au long du TD. La fonction permettant de réaliser des analyses en composantes principales s'appelle, sans surprise, PCA(). Pour installer et charger le package, on utilise les commandes suivantes :

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

Question 2. *Que se passe t-il si l'on demande à R de renvoyer PCA(trivia) ? Est-ce que c'est cohérent avec le fonctionnement de l'analyse en composantes principales ?*

Question 3. *Résoudre le problème soulevé à la question précédente, en utilisant les différents types de variables présentés en cours. On pourra s'aider de la documentation de la fonction PCA, en utilisant help(PCA).*

Lorsque l'on applique la fonction PCA() à un tableau de données, R réalise une analyse en composantes principales. On remarque directement que dans le module graphique, deux graphiques s'affichent.

Question 4. *À quoi correspondent les deux graphiques ?*

Mais ce n'est pas le seul résultat obtenu par la fonction PCA(). En effet, on obtient un objet multiple comportant plusieurs informations relatives à l'analyse en composantes principales. En effet, R nous renvoie :

```
**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 2689 individuals, described by 19 variables
*The results are available in the following objects:

   name                 description
1  "$eig"               "eigenvalues"
2  "$var"               "results for the variables"
3  "$var$coord"         "coord. for the variables"
4  "$var$cor"           "correlations variables - dimensions"
5  "$var$cos2"          "cos2 for the variables"
6  "$var$contrib"       "contributions of the variables"
7  "$ind"               "results for the individuals"
8  "$ind$coord"         "coord. for the individuals"
9  "$ind$cos2"          "cos2 for the individuals"
10 "$ind$contrib"       "contributions of the individuals"
11 "$quali.sup"         "results for the supplementary categorical variables"
12 "$quali.sup$coord"   "coord. for the supplementary categories"
13 "$quali.sup$v.test" "v-test of the supplementary categories"
14 "$call"              "summary statistics"
15 "$call$centre"       "mean of the variables"
16 "$call$ecart.type"   "standard error of the variables"
17 "$call$row.w"        "weights for the individuals"
18 "$call$col.w"        "weights for the variables"
```

Pour utiliser chacun des éléments de cet objet, il convient tout d'abord de le nommer :

```
> trivia_acp <- PCA(trivia,quali.sup = 19)
```

Question 5. *Quels sont les critères qui peuvent être utilisés pour déterminer le nombre d'axes à retenir ? En utilisant une ou plusieurs informations inclus dans trivia_acp, déterminer combien d'axes retenir ici. Quelle est la somme des valeurs propres ?*

Question 6. *Que peut-on dire de l'axe 1 de l'analyse en composantes principales ? On pourra utiliser les commandes suivantes pour justifier sa réponse.*

```
> trivia_acp$var$coord
> trivia_acp$var$cor
> trivia_acp$var$cos2
> trivia_acp$var$contrib
```

Pour choisir quels axes on veut afficher dans les différents graphiques, on rajoute la variable *axes=* en entrée de la fonction PCA() :

```
> trivia_acp_axes23 <- PCA(trivia,axes=c(2,3),quali.sup = 19)
```

Question 7. *Expliquer à quoi correspondent les axes 2 et 3 de l'analyse en composantes principales. On utilisera les mêmes commandes que pour la question 6.*

Intéressons-nous maintenant aux individus. Vous avez à disposition après avoir calculé trivia_acp et trivia_acp_axes23 les graphiques relatifs aux individus selon ces deux jeux d'axes.

Question 8. *Que dire des individus numéros 74, 2 puis 31 ?*

Question 9. *Comment interpréter les résultats liées à la variable qualitative supplémentaire, sexe ?*

Enfin, on peut également rajouter des individus supplémentaires. On peut par exemple rajouter un joueur moyen, qui répond juste à la moitié des questions, peu importe leur catégorie.

```
> joueur_moyen <- c(rep(0.5,18),NA)
> trivia_moyen <- rbind(trivia,joueur_moyen)
> trivia_moyen_acp <- PCA(trivia_moyen,quali.sup = 19, ind.sup = 101)
```

Question 10. *Dresser le profil de l'individu moyen. Où s'attend on à le retrouver sur les graphiques ? Est-ce que cela se vérifie ? Pourquoi ?*
De quel individu se rapproche t-il le plus ? Le vérifier.

Question 11. *Pourquoi un message d'erreur est-il renvoyé par R ?*

Le graphe relatif aux individus a désormais un 101ème cercle, en bleu, mais si vous ne le repérez pas bien, il est possible d'utiliser la fonction plot.PCA(), comme suit :

```
> plot.PCA(trivia_alea_acp,axes=c(2,3),select = "101",invisible="ind")
```

On se référera à l'aide help(plot.PCA) pour comprendre les différentes variables utilisées ici. Très rapidement, *axes=* a déjà été vue, *select=* permet de choisir quel individu afficher en priorité, *invisible=* permet d'indiquer quels types de points on veut faire disparaître du graphe, ici tous les individus (sauf le 101, grâce à l'instruction *select=*). On peut aussi utiliser la variable *cex*, qui modifie la taille du texte sur le graphe.

2 Analyse factorielle des correspondances

Élections présidentielles de 2012

Dans un premier temps, on s'intéresse uniquement aux départements de France métropolitaine.

Question 12. *Importer les données de la table* elections_2012_resultats.csv *dans un dataframe, dont les noms de ligne sont les libellés des départements.*

Question 13. *Construire le tableau de contingence pour les* ***suffrages exprimés****.*

Question 14. *Effectuer un test du khi-deux en utilisant la fonction* chi.square*. Donner le tableau des fréquences attendu en cas d'indépendance en utilisant l'argument $expected. Le test du khi-deux possède-t-il une réelle signification statistique ?*

Question 15. *Effectuer l'AFC avec le nombre d'abstentions par département comme colonne supplémentaire*

Question 16. *Combien d'axes peut-on retenir pour l'étude ? On s'aidera d'une représentation graphique des valeurs propres*

Question 17. *Retrouver l'inertie totale de deux façons :*
- *à partir de l'inertie de chaque colonne dans les résultats de l'AFC*
- *à partir du test du khi-deux.*

Question 18. *Représenter les plans factoriels pour les axes choisis, séparément pour les candidats et les départements.*

Question 19. *Effectuer une interprétation des axes choisis. S'aider pour cela des attributs de l'objet renvoyé, à la manière de ce qui avait été fait lors du TD1.*

Question 20. *Calculer la distance de chaque département au centre de gravité du nuage. Quels sont les départements ayant le plus contribué à l'inertie totale ?*

Question 21. *Un institut de marketing d'opinion souhaite effectuer un sondage pré-électoral dans un département "représentatif". Quel département devrait-il retenir ?*

Question 22. *Peut-on interpréter facilement le point supplémentaire "Abstensions" ?*

Question 23. *Ajouter les départements d'outre-mer comme points supplémentaires à l'analyse factorielle des correspondances. Analyser.*

Question 24. *On souhaite vérifier que l'analyse statistique effectuée pour les données de 2012 restent valable pour l'élection de 2007. Effectuer une AFC suivant les mêmes étapes que pour l'analyse de l'élection présidentielle de 2012 sur la table des données 2007 (*elections_2007_resultats.csv*). Commenter.*

3 Analyse des correspondances multiples

Où faire jouer Valbuena ?

On dispose d'une table (*td3_donnees.csv*) des joueurs des deux premières division du championnat de France de football (Ligue 1 et Ligue 2).

Pour chaque joueur de la table, on dispose des informations suivantes :
- *nom* : Le nom du joueur
- *pied* : Le pied préféré du joueur (droit ou gauche)
- *position* : La position du joueur sur le terrain (voir figure 1)
- *championnat* : Le championnat auquel participe le joueur (Ligue1 ou Ligue2)
- et différentes variables techniques, décrites plus loin.

Les variables techniques peuvent prendre 4 modalités, notant le niveau ou le classement du joueur realtivement à la caractéristique étudiée [1].

1. Faible
2. Moyen
3. Fort
4. Très fort

Ces variables peuvent être regroupées en plusieurs catégories :
- Caractéristiques physique : "age","taille"
- Le niveau global du joueur : "general"
- Capacités attaque : "centre", "finition", "passes_courtes", "volees"
- Aptitudes techniques : "dribbles", "effets", "passe_longue", "controle_balle"
- Aptitudes physiques : "acceleration", "vitesse", "agilite", "reactions", "equilibre", "puissance_tir", "force", "tirs_lointains"
- Mental : "aggressivite", "interceptions", "positionnement_def", "vision", "penalties"
- Capacités défense : "marquage", "tacle_debout", "tacle_glisse"
- Capacités gardien : "plongeon", "prise_balle", "degagement", "positionnement_goal", "reflexes"

La répartition des postes sur le terrain se fait comme indiqué sur la figure 1 (Précision : aucune connaissance préalable en football n'est nécessaire pour pouvoir profiter de ce TD !).

Question 25. *Charger les données dans un data.frame, et transformer les colonnes en facteur (utiliser pour cela les fonctions* factor *et* apply*).*

1. Ces variables proviennent des notes attribuées à chaque joueur dans le jeu vidéo FIFA 15

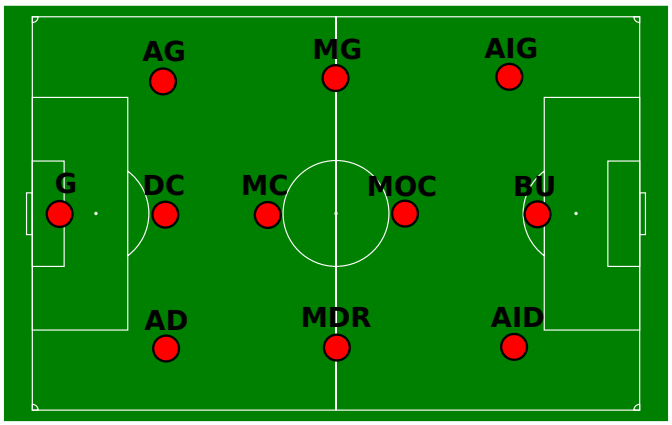

FIGURE 1 – Positionnement des joueurs de football sur le terrain.

Question 26. *Pourquoi l'ACM est-elle la méthode privilégiée ici ? Pourquoi ne pas faire une ACP ?*

Question 27. *Effectuer l'ACM sur les données avec les colonnes "championnat", "age" et "position" en variables supplémentaires. Observer la représentation des individus. Qu'observez-vous ? Modifier le dataframe en conséquence.*

Question 28. *Combien d'axes peut-on retenir pour l'analyse ? Interpréter les axes de l'ACM, en s'intéressant notamment :*

- *aux variables et individus contribuant le plus à l'inertie de chaque axe*
- *aux points supplémentaires*

Le joueur Mathieu Valbuena vient d'intégrer le championnat de France de Ligue 1. En équipe de France, où il est régulièrement sélectionné, il joue au poste "AID". Créer ses caractéristiques à l'aide de la commande suivante :

```
> caracteristiques_valbuena <- c("Droit","AID","Ligue1",3,1
                                 ,4,4,3,4,3,4,4,4,4,4,3,4,4,3,3,1,3,2,1,3,4,3,1,1,1)
```

Question 29. *Intégrer Mathieu Valbuena au dataframe des joueurs du championnat de France et refaire l'ACM en l'intégrant comme individu supplémentaire.*

Question 30. *Un journaliste souhaite écrire un article sur Valbuena en imaginant sa carrière en championnat de France en se fondant sur l'analyse de la carrière du joueur "qui lui ressemble le plus". En utilisant les résultats de l'ACM, quel joueur pouvez-vous proposer à ce journaliste ? Donner une réponse graphique et une réponse par le calcul.*

Question 31. *Le président du club qui a acheté le joueur souhaite innover en faisant appel à une analyse statistique*[2] *pour déterminer le poste optimal auquel*

2. Ce n'est pas encore courant dans le football français, mais beaucoup de franchises de sport américaines disposent maintenant de leurs propres analystes statistiques. Voir par exemple le film "Moneyball" avec Brad Pitt.

faire jouer Valbuena. Quelle réponse apporteriez-vous ? Donner une réponse graphique et une réponse par le calcul.

Question 32. *Quelles sont les critiques et améliorations que l'on pourrait apporter à cette analyse ?*

Question 33. *Que penser de l'importance de la latéralité (variable "pied") dans la détermination des postes ?*

4 Classification

Chambre des représentants des États-Unis

On s'intéresse aux votes exprimés par les représentants élus au sein de la Chambre aux Etats-Unis, pendant l'année 1984, qui comptait 267 démocrates et 168 républicains. Le jeu de données, classique en *machine learning*, est décrit à l'adresse suivante : `https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records` et disponible légèrement modifié dans le fichier td4_donnees.csv.

Question 34. *Charger le jeu de données dans R. À quoi correspondent les modalités des différentes variables ?*[3]

Nous allons tout d'abord nous intéresser à l'application d'une classification ascendante hiérarchique à ce jeu de données. Pour cela, nous allons utiliser le package *cluster* de R.

Question 35. *Rappeler le principe de la classification ascendante hiérarchique, et expliquer le type de résultat attendu.*

Question 36. *Effectuer une CAH en affichant l'arbre de classification.*

Question 37. *À quoi s'attendre si l'on sélectionne deux classes ?*

Question 38. *Combien de classes faut-il retenir ? Pourquoi ?*

Question 39. *Décrire les classes obtenues.*

Question 40. *Que se passe t-il si l'on applique une CAH en omettant la variable "party" dans le jeu de données ?*

Intéressons-nous maintenant à une autre méthode de classification, la méthode des centres mobiles.

Question 41. *Rappeler le fonctionnement de cette méthode.*

Question 42. *Quels sont les résultats obtenus pour $k = 2$? Commenter le résultat. Trouve t-on la même chose si l'on omet la variable "party" ?*

Question 43. *Quels sont les résultats obtenus pour k égal au nombre optimal de classes obtenu pour la CAH ? Comparer avec les résultats de la CAH.*

Question 44. *On s'intéresse aux différences entre les méthodes d'analyse de données. Déterminer si l'on peut appliquer une ACP, une ACM et une AFC, puis comparer les résultats obtenus avec les deux méthodes de classification.*

3. Pour vous aider, un extrait de la page internet mentionnée précedemment : "voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition)."