TP ACP

 Par M. BEZARA

Année 2022-2023

**Packages R**

Plusieurs fonctions sont disponibles sur le logiciel *R* pour la construction d’une ACP:

* prcomp() et princomp() [fonction de base, package stats],
* PCA() [package FactoMineR], factoextra
* dudi.pca() [package ade4],
* et epPCA() [package ExPosition]

Peu importe celle que vous décidez d’utiliser, l’important est de sortir les livrables essentiels.

Exercice 1

On étudie les consommations annuelles de 8 denrées alimentaires chez les individus des 8 catégories socio-professionnelles.

Les données sont des moyennes par CSP :

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PAO | PAA | VIO | VIA | POT | LEC | RAI | PLP |
| AGRI | 167 | 1 | 163 | 23 | 41 | 8 | 6 | 6 |
| SAAG | 162 | 2 | 141 | 12 | 40 | 12 | 4 | 15 |
| PRIN | 119 | 6 | 69 | 56 | 39 | 5 | 13 | 41 |
| CSUP | 87 | 11 | 63 | 111 | 27 | 3 | 18 | 39 |
| CMOY | 103 | 5 | 68 | 77 | 32 | 4 | 11 | 30 |
| EMPL | 111 | 4 | 72 | 66 | 34 | 6 | 10 | 28 |
| OUVR | 130 | 3 | 76 | 52 | 43 | 7 | 7 | 16 |
| INAC | 138 | 7 | 117 | 74 | 53 | 8 | 12 | 20 |

AGRI = Exploitants agricoles PAO = Pain ordinaire

SAAG= Salariés agricoles PAA = Autre pain

PRIN = Professions indépendantes VIO = Vin ordinaire

CSUP = Cadres supérieurs VIA=Autre vin

CMOY= Cadres moyens POT= Pommes de terre

EMPL= Employés LEC=Légumes secs

OUVR = Ouvriers RAI=Raisin de tables

INAC = Inactifs PLP= Plats préparés

On demande d’identifier les groupes ainsi que leur mode ou style de consommation.

Travail préparatoire (données)

* Préparation d’une R-données (conception dirècte ou importation fichier).
* Installation package ( FactoMineR, factoextra).
* res= PCA( data, quali.sup=1), fviz\_eig (res), fviz\_biplot(res,col.var= « #2E9FDF » ,col.ind= » #696969 »,repel=T).
* (version ameliorée : fviz\_pca\_var(res, col.var= « contrib»,gradient.cols=c(« #00AFBB », »#E7B800 », »#FC4E07 »),repel=T)
* Etablir un corrélogramme des variables : ( package : corrplot, M =cor(data [,-1]) corrplot(M, type= « upper »,order= « hclust »).

ACP basique R

Cas 2 (à faire avec R et tanagra)

Voici une base de données sur quelques marques de voitures. Les variables prélevées sont : puissance, longueur, hauteur, poids, pollution.

L’objectif fixée est de voir la qualité des variables collectées puis de connaitre comment peuvent –elles être regroupées ?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **marque** | **puissanc** | long | haut | poids | pollu |
| Golf | 75 | 421 | 149 | 1217 | 143 |
| CitrC4 | 138 | 426 | 146 | 1381 | 142 |
| Pg607 | 204 | 491 | 145 | 1723 | 223 |
| Velsatis | 150 | 486 | 158 | 1735 | 188 |
| CitrC2 | 61 | 367 | 147 | 932 | 141 |
| Chrys3 | 340 | 502 | 148 | 1835 | 291 |
| AudiA3 | 102 | 421 | 143 | 1205 | 168 |
| OutLand | 202 | 455 | 167 | 1595 | 237 |
| PtCruiser | 223 | 429 | 154 | 1595 | 235 |
| SantaF | 125 | 450 | 173 | 1757 | 197 |

Créer cette base sur R avant d’effectuer l’analyse.

Cas 3 :

* Visualiser le jeu de données *decathlon2 ( factoextra).*
* Considérer les individus actifs (1 :23) et comme variables actives (1 :10).
* Présenter les composantes principales
* Etablir le graphe des éboulis des valeurs propres
* Donner le plan de corrélations des variables
* Donner le plan factoriel des individus
* Le plan biplot

Cas 4 : (équation des axes factoriels, facteurs discriminants)

L’objectif de l’analyse est de produire à partir d’un ensemble d’individus répartis dans des groupes décrits par plusieurs descripteurs, de nouvelles variables, les composantes principales, les axes factoriels, on parle ici de variables discriminantes, qui séparent au mieux les groupes.

Nous travaillons ici sur un fichier de données. On a récence 34 crus de vin de Bordeaux répartis en 3 groupes. Les descripteurs correspondent à des variables météorologiques

(température, ensoleillement, chaleur, pluie)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Temper ( F)** | **Ensol** | **Chal** | **Pluie** | **Qualité** |
| 3064 | 1201 | 10 | 361 | medium |
| 3000 | 1053 | 11 | 338 | bad |
| 3155 | 1133 | 19 | 393 | medium |
| 3085 | 970 | 4 | 467 | bad |
| 3245 | 1258 | 36 | 294 | good |
| 3267 | 1386 | 35 | 225 | good |
| 3080 | 966 | 13 | 417 | bad |
| 2974 | 1189 | 12 | 488 | bad |
| 3038 | 1103 | 14 | 677 | bad |
| 3318 | 1310 | 29 | 427 | medium |
| 3317 | 1362 | 25 | 326 | good |
| 3182 | 1171 | 28 | 326 | bad |
| 2998 | 1102 | 9 | 349 | bad |
| 3221 | 1424 | 21 | 382 | good |
| 3019 | 1230 | 16 | 275 | medium |
| 3022 | 1285 | 9 | 303 | medium |
| 3094 | 1329 | 11 | 339 | medium |
| 3009 | 1210 | 15 | 536 | bad |
| 3227 | 1331 | 21 | 414 | medium |
| 3308 | 1366 | 24 | 282 | good |
| 3212 | 1289 | 17 | 302 | medium |
| 3361 | 1444 | 25 | 253 | good |
| 3061 | 1175 | 12 | 261 | medium |
| 3478 | 1317 | 42 | 259 | good |
| 3126 | 1248 | 11 | 315 | medium |
| 3458 | 1508 | 43 | 286 | good |
| 3252 | 1361 | 26 | 346 | medium |
| 3052 | 1186 | 14 | 443 | bad |
| 3270 | 1399 | 24 | 306 | good |
| 3198 | 1259 | 20 | 367 | good |
| 2904 | 1164 | 6 | 311 | bad |
| 3247 | 1277 | 19 | 375 | good |
| 3083 | 1195 | 5 | 441 | bad |
| 3043 | 1208 | 14 | 371 | bad |

1. Identifier les descripteurs qui apportent la « bonne qualité ».

Indication : Package FactoMineR, fonction MCA

**Cas** 5 ( Rcmdr)

**6** Anova à 2 facteurs ( Importance haute)

|  |  |  |  |
| --- | --- | --- | --- |
| Rendement | Méthode | Type de champ | |
| 20 | 1 | 1 |  |
| 7 | 1 | 1 |  |
| 39 | 1 | 2 |  |
| 17 | 1 | 2 |  |
| 34 | 1 | 3 |  |
| 13 | 1 | 3 |  |
| 13 | 1 | 4 |  |
| 5 | 1 | 4 |  |
| 35 | 2 | 1 |  |
| 52 | 2 | 1 |  |
| 30 | 2 | 2 |  |
| 28 | 2 | 2 |  |
| 58 | 2 | 3 |  |
| 73 | 2 | 3 |  |
| 64 | 2 | 4 |  |
| 62 | 3 | 1 |  |
| 44 | 3 | 1 |  |
| 82 | 3 | 2 |  |
| 81 | 3 | 2 |  |
| 69 | 3 | 3 |  |
| 84 | 3 | 3 |  |

Les données correspondent à une expérience où 3 méthodes de culture ont été testées sur quatre types de champs (mêmes sols, mais des expositions différentes).

Le rendement de la culture a été mesuré après la moisson.

***Remarque :***

La 3ième méthode n'a pu être testée ici sur le 4ième type de champ (manque de graines), et qu'une expérience n'a pu être menée à bout (méthode 2, type de champ 4) à cause d'un orage de grêle. Nous parlons d’une situation d’anova déséquilibrée.

L’étude de l'interaction entre les méthodes et les types de champ est intéressante pour l'expérimentateur.

D’où l’anova à deux facteurs

Nous cherchons à déterminer s'il existe une influence significative de la méthode, du type de champ, et éventuellement de leur interaction.

Indication et commande :

1. Réaliser d’abord une anova à un facteur. ( analyse de normalité et d’homoscedasticité de la colonne quantitative)
2. Anova à 2 facteurs