

Initiation à l'analyse factorielle des correspondances

A.B. Dufour & M. Royer & J.R. Lobry

Dans cette fiche, on étudie l'Analyse Factorielle des Correspondances. Cette technique statistique permet de réduire le nombre de variables, afin d'obtenir une représentation graphique des tableaux de contingence. Elle vise à y rassembler la quasi-totalité de l'information initiale, en s'attachant aux correspondances entre les caractères.

Table des matières

1 Exemple introductif	2
1.1 Les données	2
1.2 Définition d'un score <i>a priori</i>	3
1.3 Notion de score optimum	4
1.4 Représentations graphiques	4
2 Table de Contingence	5
2.1 Tableau des données	5
2.2 Tableaux des profils lignes et colonnes	8
2.3 Lien avec le test du χ^2 d'indépendance	9
3 Compréhension des résultats d'une AFC	10
3.1 Le tableau analysé	10
3.2 Les pondérations	10
3.3 La matrice diagonalisée	11
3.4 Les coordonnées des lignes	11
3.5 Les coordonnées des colonnes	12
3.6 Lien avec le Khi-Deux	12
3.7 Carte Factorielle sur les axes 1 et 2	13
4 Aides à l'interprétation	13
4.1 Décomposition de l'inertie totale	13
4.2 Contributions absolues des lignes (resp. des colonnes)	14
4.3 Contributions relatives des lignes (resp. des colonnes)	15
4.4 Contributions relatives cumulées	15

5 Application 16

Références 16

1 Exemple introductif

1.1 Les données

L'exemple porte sur la couleur des yeux et la couleur des cheveux de 592 étudiants. Les données ont été collectées dans le cadre d'un projet de classe par les étudiants d'un cours de statistique élémentaire à l'Université de Delaware [3].

```
snee74 <- read.table("http://pbil.univ-lyon1.fr/R/donnees/snee74.txt", header=TRUE)
names(snee74)
[1] "cheveux" "yeux" "sexe"
head(snee74)
  cheveux yeux sexe
1   Noir Marron Male
2   Blond Bleu Femelle
3   Noir Bleu Male
4 Marron Marron Femelle
5   Roux Marron Male
6 Marron Bleu Male
```

La couleur des cheveux est définie par 4 modalités : blond, marron, noir et roux.

```
cheveux <- snee74$cheveux
summary(cheveux)
Blond Marron Noir Roux
 127   286  108   71
```

La couleur des yeux est définie par 4 modalités : bleu, marron, noisette et vert.

```
yeux <- snee74$yeux
summary(yeux)
  Bleu Marron Noisette Vert
  215   220    93    64
```

Le lien entre les deux couleurs s'obtient à l'aide d'un tableau croisé qui ventile la population entre les modalités de ces deux variables qualitatives. C'est une table de contingence.

```
(table(yeux,cheveux) -> couleurs)
      cheveux
yeux  Blond Marron Noir Roux
Bleu   94   84  20  17
Marron 7  119  68  26
Noisette 10  54  15  14
Vert   16   29   5  14
```

Par commodité, on transforme cet objet en un `data.frame` :

```
(dfcouleurs <- data.frame(unclass(couleurs)))
      Blond Marron Noir Roux
Bleu   94   84  20  17
Marron 7  119  68  26
Noisette 10  54  15  14
Vert   16   29   5  14
```

1.2 Définition d'un score *a priori*

On va affecter *a priori* un score à chacune des colonnes (*couleur des cheveux*), par exemple (1,-1,-1,1), qui opère une opposition entre cheveux foncés (Marron, Noir) et clairs (Blond, Roux).

```
scorecheveux <- c(1,-1,-1,1)
names(scorecheveux) <- colnames(couleurs)
scorecheveux
Blond Marron Noir Roux
  1     -1     -1     1
```

Pour chaque ligne de la table de contingence (*couleur des yeux*), une fréquence observée correspond à chaque couleur de cheveux. Ainsi, pour la modalité yeux Bleu on obtient :

```
dfcouleurs <- data.frame(unclass(couleurs))
dfcouleurs["Bleu",]/sum(dfcouleurs["Bleu",])
      Blond Marron Noir Roux
Bleu 0.4372093 0.3906977 0.09302326 0.07906977
```

Il est alors possible de calculer le score moyen pour la modalité yeux Bleu :

```
dfcouleurs["Bleu",]/sum(dfcouleurs["Bleu",]) -> yeux.bleu
yeux.bleu*scorecheveux
      Blond Marron Noir Roux
Bleu 0.4372093 -0.3906977 -0.09302326 0.07906977
sum(yeux.bleu*scorecheveux)
[1] 0.03255814
```

Ce score moyen positif montre que les individus aux yeux Bleu ont des cheveux plutôt clairs.

Ce score moyen peut être calculé pour toutes les couleurs de yeux.

```
freqyeux <- apply(dfcouleurs, 1, fonction(x) x/sum(x))
freqyeux
      Bleu Marron Noisette Vert
Blond 0.43720930 0.03181818 0.1075269 0.250000
Marron 0.39069767 0.54090909 0.5806452 0.453125
Noir 0.09302326 0.30909091 0.1612903 0.078125
Roux 0.07906977 0.11818182 0.1505376 0.218750
t(freqyeux)
      Bleu Marron Noir Roux
Bleu 0.43720930 0.3906977 0.09302326 0.07906977
Marron 0.03181818 0.5409091 0.30909091 0.11818182
Noisette 0.10752688 0.5806452 0.16129032 0.15053763
Vert 0.25000000 0.4531250 0.07812500 0.21875000
scoreyeux <- apply(t(freqyeux), 1, fonction(x) sum(x*scorecheveux))
scoreyeux
      Bleu Marron Noisette Vert
0.03255814 -0.70000000 -0.48387097 -0.06250000
```

Pour les yeux marrons, on obtient un score moyen égal à -0.7 qui est négatif et indique donc que les cheveux foncés dominent dans cette sous-population.

On pourrait assez bien séparer les 4 couleurs des yeux sur la base du score proposé pour la couleur des cheveux. Cependant, deux questions se posent :

- Existe-t-il un score des cheveux qui permet de discriminer encore mieux la couleur des yeux ?
- Lorsqu'on connaît moins bien le sujet, (ici, l'opposition *clair/foncé* est naturelle), comment définir un score qui permette de mieux comprendre la structure du tableau de données ?

1.3 Notion de score optimum

L'Analyse Factorielle des Correspondances (AFC) est la méthode permettant de définir pour une table de contingence un score sur les colonnes tel que les scores moyens des lignes (obtenus en utilisant les fréquences des tableaux de profils) soient les plus séparés possibles, au sens de la variance de ces scores moyens. Et inversement.

Cette méthode choisit comme score optimal pour les colonnes (cheveux) les valeurs :

```
library(ade4)
library(adegraphics)
ac <- dudi.coa(dfcouleurs, scannf=F, nf=3)
rownames(ac$c1)
[1] "Blond" "Marron" "Noir" "Roux"
ac$c1[,1]
[1] 1.8282287 -0.3244635 -1.1042772 -0.2834725
```

On vérifie que les valeurs extrêmes sont obtenues pour les modalités **Blond** et **Noir**, ce qui reflète que la structure majeure de ce jeu de données est l'opposition *clair/foncé*.

Exercice. Retrouver le score moyen des lignes (couleur des yeux) à partir des scores optimaux de la couleur des cheveux obtenus par l'AFC.

Réponse :

```
      Bleu      Marron      Noisette      Vert
0.5474139 -0.4921577 -0.2125969  0.1617534
```

Il est important de noter que si on cherche d'abord les scores optimaux pour le critère *couleur des yeux* par la méthode AFC (coordonnées des lignes, `li` sous `ade4`), on obtient le même résultat :

```
[1] "Bleu"      "Marron"     "Noisette"   "Vert"
[1] 0.5474139 -0.4921577 -0.2125969  0.1617534
```

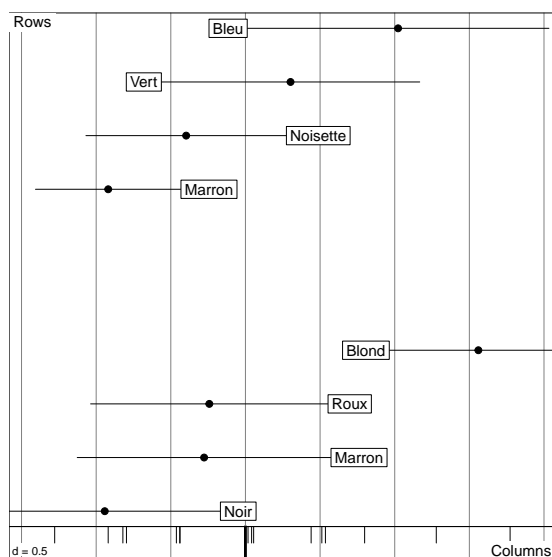
Le raisonnement que l'on vient de tenir peut se reproduire dans la recherche de score moyen des couleurs de cheveux à partir des scores optimaux de la couleur des yeux.

```
rownames(ac$co)
[1] "Blond" "Marron" "Noir" "Roux"
ac$co[,1]
[1] 0.8353478 -0.1482527 -0.5045624 -0.1295233
```

1.4 Représentations graphiques

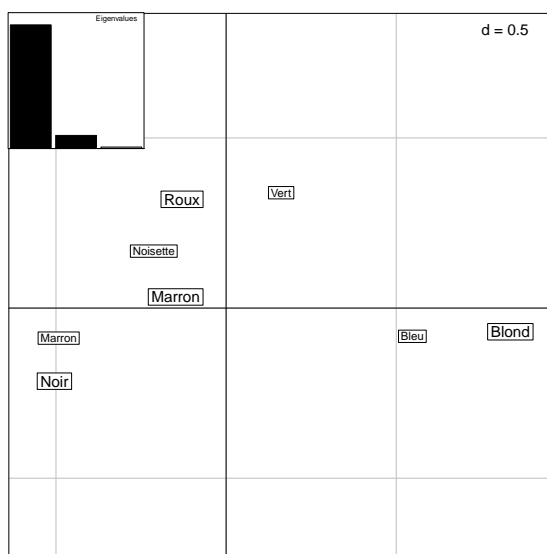
On peut alors donner une représentation graphique des valeurs obtenues pour les scores des lignes (resp. des colonnes) pour le premier score optimal des colonnes (resp. des lignes).

```
score(ac)
```



Et pour finir, on donne une représentation graphique des résultats obtenus sur les deux premiers scores optimaux.

`scatter(ac)`



2 Table de Contingence

2.1 Tableau des données

La table engendrée par le croisement de deux variables qualitatives s'appelle une *table de contingence* observée. Il est important de rappeler que :

- i) tout individu présente une modalité et une seule de chaque variable ;

- ii) chaque modalité doit avoir été observée au moins une fois, sinon elle est supprimée.

Les données sont extraites d'une étude de J. Blondel et H Farre en 1988 [1]. On s'intéresse à l'abondance d'oiseaux - toutes espèces confondues - dans 4 régions européennes, sur des sites où les stades de végétation varient. Les deux variables retenues pour l'analyse sont :

1. la région : Pologne, Bourgogne, provence ou Corse ;
2. le stade de végétation : du milieu le plus ouvert S1 au milieu le plus fermé S6.

Les données sont stockées dans `oiseaux.txt`.

```
oiseaux <- read.table("http://pbil.univ-lyon1.fr/R/donnees/oiseaux.txt",h=T)
oiseaux
  Pologne Bourgogne Provence Corse
S1      33       507       77  278
S2     157       512      164  288
S3     751       444      194  406
S4     621       221      357  494
S5     776       375      347  512
S6     854       536      580  598
taboiseaux <- as.table(as.matrix(oiseaux))
```

Les informations de base sont le nombre total d'observations (n), le nombre de modalités pour la variable 'région' (I) et le nombre de modalités pour la variable 'stade de végétation' (J).

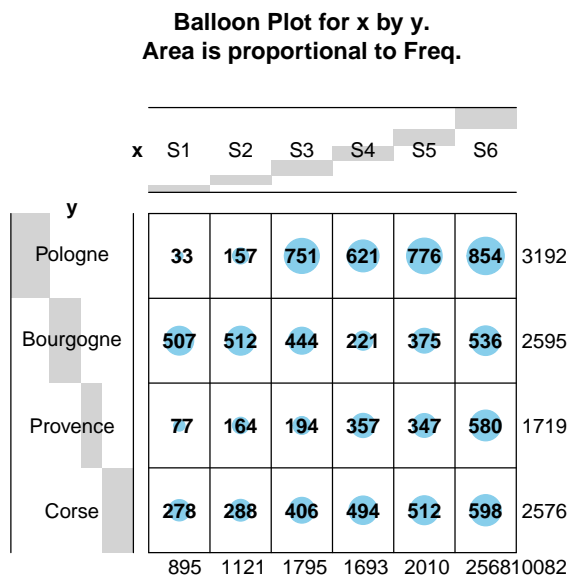
```
(n <- sum(oiseaux))
[1] 10082
(I <- nrow(oiseaux))
[1] 6
(J <- ncol(oiseaux))
[1] 4
```

On peut construire le tableau des fréquences relatives où chaque terme est de la forme $f_{ij} = \frac{n_{ij}}{n}$.

```
frequoiseaux <- oiseaux/n
round(frequoiseaux,digits=4)
  Pologne Bourgogne Provence Corse
S1 0.0033 0.0503 0.0076 0.0276
S2 0.0156 0.0508 0.0163 0.0286
S3 0.0745 0.0440 0.0192 0.0403
S4 0.0616 0.0219 0.0354 0.0490
S5 0.0770 0.0372 0.0344 0.0508
S6 0.0847 0.0532 0.0575 0.0593
```

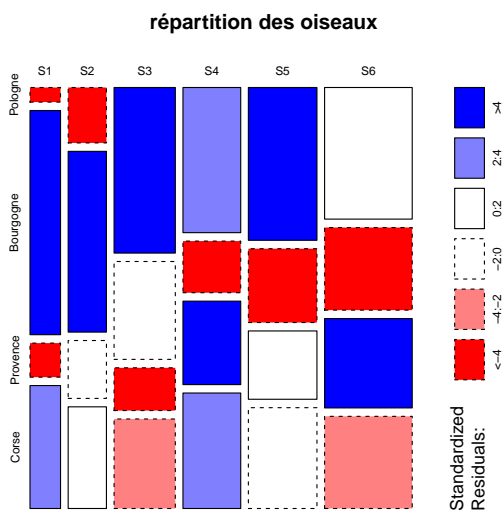
On peut obtenir différentes représentations graphiques de la table de contingence. Le principe est d'utiliser des symboles dont la surface est proportionnelle aux effectifs :

```
library(gplots)
library(caTools)
library(bitops)
balloonplot(taboiseaux)
```



La fonction `mosaicplot` permet de mettre en évidence les liens les plus importants :

```
mosaicplot(oiseaux, shade=TRUE,main="répartition des oiseaux")
```



2.2 Tableaux des profils lignes et colonnes

On calcule maintenant les fréquences conditionnelles. Pour ce faire, on note V_1 et V_2 , les deux variables qualitatives étudiées.

Profils lignes

Les fréquences conditionnelles associées aux profils lignes sont notées $f_{j|i}$ et définies par

$$f_{j|i} = P(V_2 = j | V_1 = i) = \frac{P(V_2 = j \cap V_1 = i)}{P(V_1 = i)}$$

$$f_{j|i} = \frac{\frac{n_{ij}}{n}}{\frac{n_{i\bullet}}{n}} = \frac{n_{ij}}{n_{i\bullet}}$$

```
(profLignes <- prop.table(taboiseaux,1))
      Pologne Bourgogne Provence Corse
S1 0.03687151 0.56648045 0.08603352 0.31061453
S2 0.14005352 0.45673506 0.14629795 0.25691347
S3 0.41838440 0.24735376 0.10807799 0.22618384
S4 0.36680449 0.13053751 0.21086828 0.29178972
S5 0.38606965 0.18656716 0.17263682 0.25472637
S6 0.33255452 0.20872274 0.22585670 0.23286604
```

On vérifie que les sommes en lignes sont toutes égales à 1.

```
rowSums(profLignes)
S1 S2 S3 S4 S5 S6
1 1 1 1 1 1 1
```

Profils colonnes

Les fréquences conditionnelles associées aux profils colonnes sont notées $f_{i|j}$ et définies par

$$f_{i|j} = P(V_1 = i | V_2 = j) = \frac{P(V_1 = i \cap V_2 = j)}{P(V_2 = j)}$$

$$f_{i|j} = \frac{\frac{n_{ij}}{n}}{\frac{n_{\bullet j}}{n}} = \frac{n_{ij}}{n_{\bullet j}}$$

```
(profColonnes <- prop.table(taboiseaux,2))
      Pologne Bourgogne Provence Corse
S1 0.01033835 0.19537572 0.04479348 0.10791925
S2 0.04918546 0.19730250 0.09540430 0.11180124
S3 0.23527569 0.17109827 0.11285631 0.15760870
S4 0.19454887 0.08516378 0.20767888 0.19177019
S5 0.24310777 0.14450867 0.20186155 0.19875776
S6 0.26754386 0.20655106 0.33740547 0.23214286
```

On vérifie également que les sommes en colonnes sont toutes égales à 1.

```
colSums(profColonnes)
      Pologne Bourgogne Provence Corse
1 1 1 1 1 1
```


2.3 Lien avec le test du χ^2 d'indépendance

Le test du Khi-Deux d'indépendance entre deux variables est caractérisé par les deux hypothèses :

- H_0 : les deux variables sont indépendantes
- H_1 : les deux variables sont liées.

Sous l'hypothèse nulle H_0 , $P(V_2 = j \cap V_1 = i) = P(V_1 = i) \times P(V_2 = j)$.

Ainsi, sous H_0 , la fréquence théorique est égale à $\frac{n_{i\bullet}}{n} \times \frac{n_{\bullet j}}{n}$.

On en déduit la table des **effectifs théoriques** (qui serait observée sous H_0), en conservant les effectifs marginaux observés.

$$\frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

```
reschi <- chisq.test(oiseaux)
reschi$expected
  Pologne Bourgogne Provence Corse
S1 283.3604 230.3635 152.5992 228.6768
S2 354.9129 288.5335 191.1326 286.4209
S3 568.3039 462.0140 306.0509 458.6312
S4 536.0103 435.7603 288.6597 432.5697
S5 636.3737 517.3527 342.7088 513.5648
S6 813.0387 660.9760 437.8488 656.1365
```

La statistique du test est la suivante :

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n})^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}}$$

Elle tend vers une loi du χ^2 à $(I - 1) \times (J - 1)$ degrés de liberté. Le résultat au test du Khi-Deux est

```
reschi
      Pearson's Chi-squared test
data:  oiseaux
X-squared = 1286.4, df = 15, p-value < 2.2e-16
```

Comme la p -value est très faible, on rejette l'hypothèse nulle. Les variables sont liées. Il est alors intéressant d'explorer la structure de cette relation.

Définition

On appelle **lien** entre la modalité i de la variable V_1 et la modalité j de la variable V_2 la quantité :

$$\frac{1}{n} \frac{(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n})^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}}$$

Les couples de modalités (i, j) qui correspondent aux liens les plus importants sont les plus responsables de la dépendance entre la variable V_1 et la variable V_2 .

Conclusion. Que la liaison entre les 2 variables soit statistiquement significative ou non, on peut explorer la structure du tableau plus en détail. Lorsque les variables présentent de nombreuses modalités, il est difficile d'extraire une information pertinente si on se contente d'observer le tableau de données. La technique de l'Analyse Factorielle des Correspondances (AFC) est là pour pallier cette déficience.

3 Compréhension des résultats d'une AFC

Les résultats de l'AFC de la table de contingence permettant d'étudier l'abondance globale des 79 espèces entre les régions et les stades de végétations.

```
afc <- dudi.coa(oiseaux, scannf=F, nf=3)
names(afc)
[1] "tab" "cw" "lw" "eig" "rank" "nf" "c1" "l1" "co" "l1" "call"
[12] "N"
```

3.1 Le tableau analysé

Le tableau analysé est :

```
afc$tab
      Pologne  Bourgogne  Provence  Corse
S1 -0.88354056  1.20086931 -0.49541015  0.215689302
S2 -0.55763796  0.77449050 -0.14195700  0.005513046
S3  0.32147604 -0.03899013 -0.36611848 -0.114757176
S4  0.15855979 -0.49284040  0.23675044  0.142012415
S5  0.21940922 -0.27515601  0.01252145 -0.003046877
S6  0.05038053 -0.18907797  0.32465807 -0.088604250
```

C'est le lien entre les effectifs théoriques et les effectifs observés.

```
(oiseaux-reschi$expected)/reschi$expected
      Pologne  Bourgogne  Provence  Corse
S1 -0.88354056  1.20086931 -0.49541015  0.215689302
S2 -0.55763796  0.77449050 -0.14195700  0.005513046
S3  0.32147604 -0.03899013 -0.36611848 -0.114757176
S4  0.15855979 -0.49284040  0.23675044  0.142012415
S5  0.21940922 -0.27515601  0.01252145 -0.003046877
S6  0.05038053 -0.18907797  0.32465807 -0.088604250
```

3.2 Les pondérations

Les pondérations des lignes et des colonnes sont les fréquences marginales de la table de contingence observée.

```
afc$cw
      Pologne  Bourgogne  Provence  Corse
0.3166038  0.2573894  0.1705019  0.2555049
apply(oiseaux,2,function(x) sum(x)/n)
      Pologne  Bourgogne  Provence  Corse
0.3166038  0.2573894  0.1705019  0.2555049

afc$lw
      S1      S2      S3      S4      S5      S6
0.08877207  0.11118826  0.17804007  0.16792303  0.19936521  0.25471137
apply(oiseaux,1,function(x) sum(x)/n)
      S1      S2      S3      S4      S5      S6
0.08877207  0.11118826  0.17804007  0.16792303  0.19936521  0.25471137
```

3.3 La matrice diagonalisée

La matrice diagonalisée est **H**.

```
matZ <- as.matrix(afc$stab)
DI <- diag(afc$lw)
DrJ <- diag(sqrt(afc$cw))
matH <- DrJ %*% t(matZ) %*% DI %*% matZ %*% DrJ
```

Le rang de la matrice analysée est donné par $\min(I - 1, J - 1)$ soit

```
min(I-1,J-1)
[1] 3
afc$rank
[1] 3
```

Les valeurs propres et les vecteurs propres issus de cette diagonalisation sont :

```
eigen(matH)
eigen() decomposition
$values
[1] 1.109603e-01 1.358960e-02 3.041339e-03 -1.754281e-17
$vectors
      [,1]      [,2]      [,3]      [,4]
[1,] 0.59399474 0.5453407 -0.1821261 -0.5626756
[2,] -0.77126117 0.2088254 -0.3227364 -0.5073356
[3,] 0.21901448 -0.7942466 -0.3882049 -0.4129187
[4,] -0.06602364 -0.1678326 0.8437821 -0.5054749
eigen(matH) -> reseigen
```

On retrouve bien les valeurs propres de l'analyse.

```
reseigen$values
[1] 1.109603e-01 1.358960e-02 3.041339e-03 -1.754281e-17
afc$eig
[1] 0.110960284 0.013589597 0.003041339
```

3.4 Les coordonnées des lignes

Les coordonnées des lignes dites axes principaux s'obtiennent par $\mathbf{ZD}_j^{1/2}\mathbf{U}$. Elles sont centrées, de variances λ et de covariances nulles.

```
matZ %*% DrJ %*% reseigen$vectors[,1:3]
      [,1]      [,2]      [,3]
S1 -0.81718935 0.00028756 0.06532488
S2 -0.50244823 -0.04296969 -0.04455945
S3 0.09342198 0.22432151 -0.01681756
S4 0.26250867 -0.09325185 0.08706607
S5 0.18223172 0.03432637 0.01926157
S6 0.12314001 -0.10353041 -0.06403649
afc$li
      Axis1      Axis2      Axis3
S1 -0.81718935 0.00028756 0.06532488
S2 -0.50244823 -0.04296969 -0.04455945
S3 0.09342198 0.22432151 -0.01681756
S4 0.26250867 -0.09325185 0.08706607
S5 0.18223172 0.03432637 0.01926157
S6 0.12314001 -0.10353041 -0.06403649
sum(afc$li$Axis1*afc$lw)
[1] 5.551115e-17
sum(afc$li$Axis2*afc$lw)
[1] 5.170289e-18
sum(afc$li$Axis1*afc$li$Axis1*afc$lw)
```

```
[1] 0.1109603
afc$eig[1]
[1] 0.1109603
sum(afc$li$Axis2*afc$li$Axis2*afc$lw)
[1] 0.0135896
afc$eig[2]
[1] 0.0135896
sum(afc$li$Axis1*afc$li$Axis2*afc$lw)
[1] -5.791334e-17
```

3.5 Les coordonnées des colonnes

Les coordonnées des colonnes dites composantes principales s'obtiennent par $D_J^{-1/2} U \Lambda^{1/2}$. Elles sont centrées, de variances λ et de covariances nulles.

```
diag(1/sqrt(afc$cw)) %% reseigen$vector[,1:3] %% diag(sqrt(afc$eig))
      [,1]      [,2]      [,3]
[1,] 0.35164811 0.11298300 -0.01785034
[2,] -0.50639559 0.04798342 -0.03508205
[3,] 0.17668188 -0.22423018 -0.05184763
[4,] -0.04350946 -0.03870615 0.09205835
afc$co
      Comp1      Comp2      Comp3
Pologne 0.35164811 0.11298300 -0.01785034
Bourgogne -0.50639559 0.04798342 -0.03508205
Provence 0.17668188 -0.22423018 -0.05184763
Corse -0.04350946 -0.03870615 0.09205835
sum(afc$co$Comp1*afc$cw)
[1] 3.122502e-17
sum(afc$co$Comp2*afc$cw)
[1] 1.214306e-17
sum(afc$co$Comp1*afc$co$Comp1*afc$cw)
[1] 0.1109603
afc$eig[1]
[1] 0.1109603
sum(afc$co$Comp2*afc$co$Comp2*afc$cw)
[1] 0.0135896
afc$eig[2]
[1] 0.0135896
sum(afc$co$Comp1*afc$co$Comp2*afc$cw)
[1] -1.084202e-17
```

3.6 Lien avec le Khi-Deux

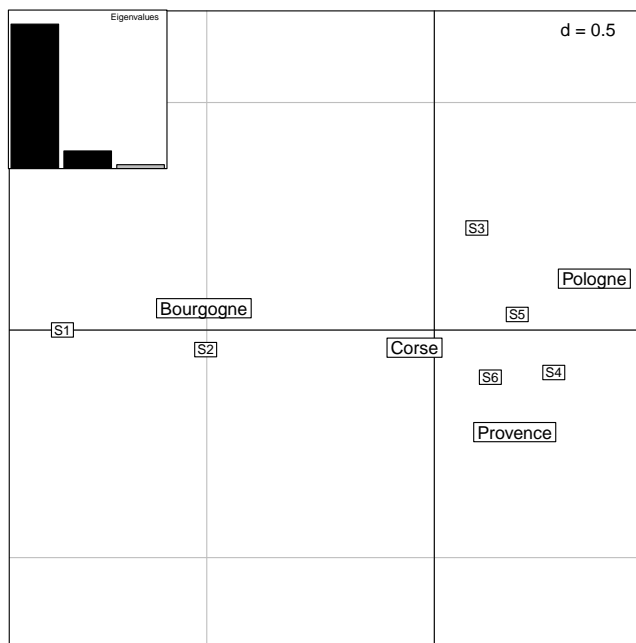
Lien entre l'inertie totale et la valeur de la statistique du Khi-Deux :

$$I_T = \frac{\chi^2}{n}$$

```
reschi$statistic
X-squared
1286.375
reschi$statistic/n
X-squared
0.1275912
sum(afc$eig)
[1] 0.1275912
```

3.7 Carte Factorielle sur les axes 1 et 2

`scatter(afc)`



Commenter.

4 Aides à l'interprétation

Les statistiques d'inertie sont importantes dans les analyses à pondérations non uniformes comme l'analyse factorielle des correspondances. Elles s'étendent à tout type d'analyse à un tableau. On les retrouve dans la fonction `inertia.dudi`. Pour les analyses à pondérations uniformes comme l'analyse en composantes principales, elles sont redondantes avec les cartes factorielles. C'est pourquoi les statistiques d'inertie sont présentées dans ce document.

```
aides <- inertia.dudi(afc, row.inertia=TRUE, col.inertia=TRUE)
names(aides)
[1] "TOT"      "row.contrib" "row.abs"    "row.rel"   "row.cum"
[6] "col.contrib" "col.abs"    "col.rel"   "col.cum"   "call"
```

4.1 Décomposition de l'inertie totale

La somme des valeurs propres est égale à l'inertie totale du nuage de points.

$$I_T = \sum_{k=1}^r \lambda_k$$

où r représente le rang de la matrice diagonalisée. La quantité λ_k/I_T est l'inertie relative du vecteur principal de rang k .

```
IT <- sum(afc$eig)
IT
[1] 0.1275912
afc$eig/IT
[1] 0.86965454 0.10650887 0.02383659
```

```
aides$TOT
      inertia      cum      cum(%)
Ax1 0.110960284 0.1109603 86.96545
Ax2 0.013589597 0.1245499 97.61634
Ax3 0.003041339 0.1275912 100.00000
```

La première colonne contient les valeurs propres λ_k de 1 à r .

La seconde colonne contient la somme des valeurs propres de 1 à K : $\sum_{k=1}^K \lambda_k$.

La dernière colonne contient l'inertie relative cumulé du nuage sur les K dimensions retenues :

$$\frac{\sum_{k=1}^K \lambda_k}{\sum_{k=1}^r \lambda_k}$$

4.2 Contributions absolues des lignes (resp. des colonnes)

L'inertie des projections sur la composante principale k se décompose en somme de contributions absolues (CA) de la variable (ou de la modalité) à la définition de k . Cela souligne les points qui contribuent le plus à l'analyse.

Soit V_1^i le vecteur associé à la modalité i de la variable V_1 comme par exemple le vecteur associé au stade de végétation S5 :

```
oiseaux[5,]
      Pologne Bourgogne Provence Corse
S5      776      375      347      512
```

La contribution du point est définie par :

$$CA_{u_k}(V_1^i) = \frac{\frac{1}{p_i} \langle V_1^i / u_k \rangle_{D_J}^2}{\lambda_k}$$

```
aides$row.abs
      Axis1(%)  Axis2(%)  Axis3(%)
S1 53.426187 5.401653e-05 12.455717
S2 25.297290 1.510695e+00 7.258949
S3 1.400388 6.592544e+01 1.655689
S4 10.428696 1.074530e+01 41.854603
S5 5.966639 1.728617e+00 2.432023
S6 3.480801 2.008989e+01 34.343018
```

que l'on retrouve facilement à l'aide des valeurs retournées par l'objet `dudi.coa`.

```
afc$li[,1]*afc$li[,1]*afc$lw/afc$eig[1]
      S1      S2      S3      S4      S5      S6
0.53426187 0.25297290 0.01400388 0.10428696 0.05966639 0.03480801
```

Notez que les résultats donnés par la fonction `inertia.dudi` sont multipliés par 1000 pour faciliter la lecture.

4.3 Contributions relatives des lignes (resp. des colonnes)

L'inertie totale se décompose en contributions à la trace des lignes (et des colonnes). Le carré de la norme de la variable ou de la modalité se décompose en contributions relatives (CR) des composantes à la représentation de la ligne i (resp. de la colonne j). Les contributions relatives sont des carrés de cosinus.

$$CR_{u_k}(V_1^i) = \frac{\langle V_1^i / u_k \rangle_{D_J}^2}{\|V_1^i\|_{D_J}^2}$$

```
aides$row.rel
      Axis1      Axis2      Axis3
S1 -99.36503  1.230396e-05  0.6349587
S2 -98.50482 -7.204425e-01 -0.7747375
S3  14.71022  8.481308e+01 -0.4767020
S4  80.89337 -1.020800e+01  8.8986367
S5  95.54255  3.390039e+00  1.0674121
S6  50.57407 -3.574912e+01 -13.6768028
```

que l'on retrouve, dans le cas du paiement semestriel :

```
(afc$li[3,]*afc$li[3,])/(sum(afc$tab[3,]*afc$tab[3,]*afc$cw))
      Axis1      Axis2      Axis3
S3 0.1471022 0.8481308 0.00476702
```

Notez que :

1. les résultats sont également multipliés par 1000,
2. les résultats obtenus sont bien sûr tous supérieurs à 0 : le signe est rajouté afin de situer la modalité sur les axes,
3. la dernière colonne contient la contribution à la trace :

```
(afc$tab[3,]*sqrt(afc$cw))*sqrt(afc$lw[3]) -> temp
sum(temp*temp)/sum(afc$eig)
[1] 0.08278964
```

4.4 Contributions relatives cumulées

Ce dernier tableau contient, pour chaque ligne V_1^i la somme des contributions relatives. Ce sont les carrés des cosinus entre un vecteur et un sous-espace de projection.

```
aides$row.cum
      Axis1  Axis1:2  Axis1:3      Axis4:3
S1 99.36503 99.36504      100 -2.220446e-14
S2 98.50482 99.22526      100  0.000000e+00
S3 14.71022 99.52330      100  1.110223e-14
S4 80.89337 91.10136      100 -4.440892e-14
S5 95.54255 98.93259      100 -2.220446e-14
S6 50.57407 86.32320      100 -2.220446e-14
```

La case `remain` est ici égale à 0 car nous avons conservé toutes les valeurs propres de l'analyse. Dans le cas contraire, cette colonne contient la somme des contributions relatives du sous-espace qui n'a pas été retenu.

5 Application

La recherche scientifique ne soulève que peu de controverse ou de résistance de la part du grand public. Dans quelques cas, rares, le débat entre science, morale et religion resurgit. C'est le cas de la recherche sur les embryons humains.

L'analyse que nous nous proposons de réaliser est adaptée de l'article *Attitudes towards Embryo research, worldviews and the moral status of the Embryo Frame* [2]. Les données ont été recueillies au près du grand public dans 9 pays européens et elles concernent le statut accordé à l'embryon.

A human embryo that is a few days old . . . (1) "is a mere cluster of cells, and it makes no sense to discuss its moral condition"; (2) "has a moral condition halfway between that of a cluster of cells and that of a human being"; (3) "is closer in its moral condition to a human being than to a mere cluster of cells"; (4) "has the same moral condition as a human being."

Les données sont rangées dans la table de contingence ci-dessous. Répondre à la question "Peut-on faire une typologie des pays?".

Pays	(1)	(2)	(3)	(4)	non réponses
Autriche	64	223	243	326	144
Danemark	373	234	112	219	62
France	227	224	166	282	100
Allemagne	88	218	259	289	146
Italie	203	157	137	373	130
Pays.Bas	207	329	154	223	88
Pologne	138	154	98	382	229
Royaume.Uni	255	168	117	236	224
Espagne	215	188	125	298	174

Pour entrer les données, utilisez :

```
res <- data.frame()
fix(res)
```

Références

- [1] J. Blondel and H. Farré. The convergent trajectories of bird communities along ecological successions in european forests. *écologia (Berlin)*, 75 :83–93, 1988.
- [2] R Pardo and F. Calvo. Attitudes towards embryo research, worldviews and the moral status of the embryo frame. *Science Communication*, 30, 1 :8–47, 2008.
- [3] R.D. Snee. Graphical display of two-way contingency tables. *The American Statistician*, 28 :9–12, 1974.