

Rappel sur le rapport de corrélation
et
Exemple d'analyse des correspondances multiples

A.B. Dufour

Dans cette séance, nous présentons le rapport de corrélation afin de mieux appréhender les objectifs des analyses des correspondances et des méthodes inter et intra tableaux.

Table des matières

1	Rapport de corrélation	2
1.1	La notion de variation	2
1.2	Le rapport de corrélation	3
1.3	Remarque : la variation intra-groupe	3
1.4	Représentation Graphique	4
1.5	Exercice, extrait de Dodge [2003]	5
2	Analyse des correspondances multiples	5
2.1	Rappel	5
2.2	Présentation des données	5
2.3	Quelques questions autour de ces données	7
	Références	11

1 Rapport de corrélation

Pour étudier le relation entre une variable qualitative et une variable quantitative, on décompose la variation totale en variation intergroupe et en variation intragroupe. Pour mesurer l'intensité de la relation, on peut calculer un paramètre appelé rapport de corrélation.

1.1 La notion de variation

La variance d'une variable quantitative est, par définition, la moyenne des carrés des écarts à la moyenne. On définit :

- la variation totale c'est-à-dire la somme des carrés des écarts à la moyenne :

$$vartot = \sum_{i=1}^n (x_i - \bar{x})^2$$

Sous \mathbb{R} , on l'écrit :

```
vartot <- function(x) {
  res <- sum((x - mean(x))^2)
  return(res)
}
```

Exemple. On considère la variable quantitative X , note obtenue par 15 étudiants.

```
notes <- c(13, 11, 10, 11, 12, 5, 8, 7, 2, 4, 16, 17, 13, 16, 15)
vartot(notes)
[1] 301.3333
```

- La variation intergroupe

Reprenons la variable `note` précédente. Les 15 étudiants sont répartis dans $p = 3$ groupes : (1) ceux qui ont suivi la moitié des cours, (2) ceux qui ne sont jamais venus, (3) ceux qui ont suivi tous les cours.

Maintenant, l'objectif est de savoir si la note est liée au choix des étudiants de participer ou non aux cours. Supposons que tous les étudiants aient la même note, qu'ils participent beaucoup, moyennement, ou pas du tout au cours. Alors cette valeur commune serait égale à la moyenne calculée sur l'échantillon global. Pour évaluer l'erreur réalisée si on considère que les étudiants ont la même note, on calcule le carré des écarts entre les valeurs mesurées et la moyenne globale, c'est-à-dire la variation totale vue ci-dessus.

Si on considère que la note dépend du choix des étudiants de participer ou non aux cours, alors la valeur est la moyenne du groupe d'appartenance. Pour évaluer l'erreur réalisée si on considère que la note des étudiants est liée au suivi des cours, on va calculer le carré des écarts entre la moyenne du groupe et la moyenne globale.

$$varinter = \sum_{k=1}^p n_k (\bar{x}_k - \bar{x})^2$$

où \bar{x}_k désigne la note des étudiants du groupe k et n_k , le nombre d'étudiants appartenant à ce même groupe.

Sous \mathbb{R} , on écrit :

```
varinter <- fonction(x, gpe) {
  moyennes <- tapply(x, gpe, mean)
  effectifs <- tapply(x, gpe, length)
  res <- (sum(effectifs * (moyennes - mean(x))^2))
  return(res)
}
```

Exemple. On considère que les étudiants sont classés ici par groupe de 5.

```
suivi <- as.factor(rep(c("1", "2", "3"), rep(5, 3)))
varinter(notes, suivi)
[1] 264.1333
```

1.2 Le rapport de corrélation

Pour étudier la relation entre une variable qualitative et une variable quantitative, on calcule le rapport de corrélation noté η^2 :

$$\eta^2 = \frac{\sum_{k=1}^p n_k (\bar{x}_k - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Si le rapport est proche de 0, les deux variables ne sont pas liées.

Si le rapport est proche de 1, les variables sont liées.

Sous \mathbb{R} , on écrit :

```
eta2 <- fonction(x, gpe) {
  res <- varinter(x, gpe)/vartot(x)
  return(res)
}
```

Exemple. Que peut-on dire finalement des notes et du suivi des cours par les étudiants ?

```
tapply(notes, suivi, mean)
  1     2     3
11.4  5.2 15.4
tapply(notes, suivi, sd)
  1     2     3
1.140175 2.387467 1.516575
eta2(notes, suivi)
[1] 0.8765487
```

Oui, nous le reconnaissons, l'exemple est un peu démagogique.

1.3 Remarque : la variation intra-groupe

D'une manière générale, la variation totale est la somme de la variation inter-groupes et de la variation intragroupe. Cette dernière est la somme pondérée des variances calculées à l'intérieur de chaque groupe.

$$varintra = \sum_{k=1}^p n_k s_k^2 = \sum_{k=1}^p (n_k - 1) \widehat{\sigma}_k^2$$

où s_k^2 est la variance descriptive au sein du groupe k , $\widehat{\sigma_k^2}$ est la variance estimée de la population k . Elle s'obtient facilement par différence entre la variation totale et la variation inter-groupes. On a alors la relation suivante, fondamentale en statistique :

$$\text{Variation Totale} = \text{Variation Inter-groupes} + \text{Variation intragroupe}$$

Exemple. Calculs de la variation intra-groupe

```

vartot(notes) - varinter(notes, suivi)
[1] 37.2
effectifs <- tapply(notes, suivi, length)
varest <- tapply(notes, suivi, var)
sum((effectifs - 1) * varest)
[1] 37.2

```

1.4 Représentation Graphique

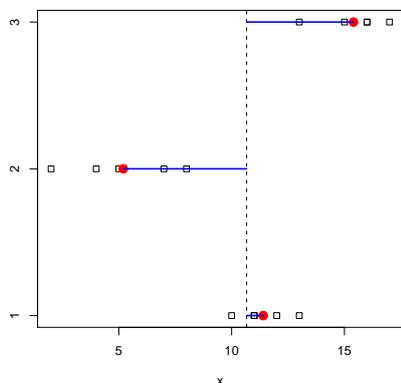
Afin de bien visualiser la relation entre une variable quantitative et une variable qualitative, on construit la représentation suivante.

- * Les groupes sont représentés en vertical, la variable quantitative en horizontal.
- * Un carré blanc représente un individu.
- * Les points rouges représentent les moyennes dans chaque groupe.
- * La ligne en pointillé représente la moyenne de l'ensemble des individus.
- * Les traits bleus représentent les écarts entre les moyennes des groupes et la moyenne de l'ensemble soit une visualisation de la variation intergroupe.

```

graphnf <- function(x, gpe) {
  stripchart(x ~ gpe)
  points(tapply(x, gpe, mean), 1:length(levels(gpe)), col = "red",
        pch = 19, cex = 1.5)
  abline(v = mean(x), lty = 2)
  moyennes <- tapply(x, gpe, mean)
  traitnf <- function(n) segments(moyennes[n], n, mean(x), n,
                                col = "blue", lwd = 2)
  sapply(1:length(levels(gpe)), traitnf)
}
graphnf(notes, suivi)

```



1.5 Exercice, extrait de Dodge [2003]

On étudie le lien entre la durée de chômage (exprimée en semaines) et trois catégories socio-professionnelles (cadres - CA, ouvriers qualifiés - OQ et ouvriers non qualifiés - ONQ).

Les données sont résumées dans les commandes ci-dessous.

```
semaine <- 2:14
effCA <- c(5, 3, 8, 7, 2, 1, rep(0, 7))
effOQ <- c(1, 2, 2, 5, 5, 13, 10, 3, 5, 1, 2, 1, 0)
effONQ <- c(2, 4, 4, 7, 6, 22, 21, 13, 13, 6, 7, 3, 1)
donCA <- rep(semaine, effCA)
donOQ <- rep(semaine, effOQ)
donONQ <- rep(semaine, effONQ)
chomage <- c(donCA, donOQ, donONQ)
profession <- factor(rep(c("CA", "OQ", "NOQ"), c(length(donCA),
length(donOQ), length(donONQ))))
```

1. Donner la moyenne et la variance de la durée du chômage, toutes catégories confondues.
2. Donner les moyennes et les variances par catégorie socio-professionnelle.
3. Existe-t-il une différence de durée du chômage entre catégories socio-professionnelles ?

2 Analyse des correspondances multiples

2.1 Rappel

En analyse des correspondances multiples, on recherche une combinaison linéaire \mathbf{y} des p variables qualitatives \mathbf{q}^j maximisant la somme des rapports de corrélations :

$$\sum_{j=1}^p \eta^2(\mathbf{y}, \mathbf{q}^j)$$

2.2 Présentation des données

Les données proviennent d'une enquête réalisée dans des supermarchés angevins et parisiens entre 1996 et 1998 dans le but de connaître l'avis de consommateurs quant aux produits biologiques et aux produits diététiques. Elles nous sont proposées par Gilles Hunault de l'université d'Angers et se trouvent originalement à l'adresse <http://www.info.univ-angers.fr/~gh/Datasets/pbio.txt> avec une copie sur le site pédagogique <http://pbil.univ-lyon1.fr/R/donnees/pbio.txt>.

419 individus ont répondu aux questions suivantes :

CONNAITRE Connaissez-vous les produits biologiques ?

0 non réponse

1 oui

2 non

DIFF Y a-t-il une différence entre produit biologique et produit diététique ?

0 non réponse

1 oui

- 2 non
- CONSOM Avez-vous déjà consommé des produits biologiques ?
- 1 non jamais
 - 2 oui une seule fois
 - 3 oui rarement
 - 4 oui de temps en temps
 - 5 oui plusieurs fois par mois
 - 6 oui plusieurs fois par semaine
 - 7 ne se prononce pas
- MARQUE Parmi les marques suivantes, laquelle connaissez-vous ?
- 0 non réponse
 - 1 bio vivre
 - 2 bjorg
 - 3 carrefour bio
 - 4 la vie
 - 5 vrai
 - 6 prosain
 - 7 favrichon
- CONSVIE Avez-vous déjà consommé des produits 'la vie' ?
- 0 non réponse
 - 1 oui une fois
 - 2 oui occasionnellement
 - 3 oui régulièrement
 - 4 non jamais
- SEXE Sexe de la personne
- 1 homme
 - 2 femme
- AGE Classe d'âge
- 1 moins de 25 ans
 - 2 entre 25 et 35 ans
 - 3 entre 35 et 45 ans
 - 4 entre 45 et 55 ans
 - 5 entre 55 et 65 ans
 - 6 plus de 65 ans
- ETATCIVIL Etat Civil
- 0 autre
 - 1 marié
 - 2 célibataire
 - 3 divorcé
 - 4 en concubinage
 - 5 veuf
- NBENF Nombre d'enfants
- 1 sans enfant
 - 2 1 enfant
 - 3 2 enfants
 - 4 3 enfants
 - 5 plus de 3 enfants
- SITPROF Situation Professionnelle
- 1 agriculteur
 - 2 artisan

- 3 cadre supérieur
 - 4 cadre moyen
 - 5 employé
 - 6 ouvrier
 - 7 retraité
 - 8 autre
 - 9 non réponse
- REVENU Classe de revenus mensuels
- 0 non réponse
 - 1 moins de 5 kF
 - 2 entre 5 et 10 kF
 - 3 entre 10 et 15 kF
 - 4 entre 15 et 20 kF
 - 5 plus de 20 kF
 - 6 ne se prononce pas

La première colonne CODE correspond à l'identifiant associé à la personne interrogée.

```
pbio <- read.table("http://pbil.univ-lyon1.fr/R/donnees/pbio.txt",
  h = T, row.names = 1)
```

2.3 Quelques questions autour de ces données

1. Quelle est la dimension de ce data frame ?
2. Ecrire le résumé statistique du data frame `pbio`. Que constate-t-on ? Modifier-le pour le rendre conforme à la réalité des données.
3. Ecrire le nouveau résumé statistique. Donner le nombre d'enquêtés connaissant la marque `carrefour bio`.
4. On note que certains enquêtés n'ont pas répondu aux questions posées mais que la non réponse n'obéit pas toujours au même codage. On modifie le data frame (1) en remplaçant les modalités 'non réponse' codées par 0 (sauf dans un cas par 7) par des 'NA' et (2) en ne conservant qu'un data frame des données complètes.

```
int <- read.table("http://pbil.univ-lyon1.fr/R/donnees/pbio.txt",
  h = T, row.names = 1)
temp <- which(int == 0, arr.ind = TRUE)
for (i in 1:100) int[temp[i, 1], temp[i, 2]] <- NA
for (i in 1:419) if (int[i, 3] == 7) int[i, 3] <- NA
for (j in 1:11) int[, j] <- factor(int[, j])
pbio.cc <- int[complete.cases(int), ]
summary(pbio.cc)
```

CONNAITRE	DIFF	CONSOM	MARQUE	CONSVIE	SEXE	AGE	ETATCIVIL	NBENF
1:305	1:251	1:76	1: 1	1: 9	1: 96	1:46	1:168	1:176
2: 9	2: 63	2:12	2:135	2: 47	2:218	2:93	2: 89	2: 59
		3:70	3: 23	3: 16		3:51	3: 16	3: 53
		4:94	4: 91	4:242		4:77	4: 33	4: 16
		5:20	5: 46			5:24	5: 8	5: 10
		6:42	6: 5			6:23		
			7: 13					

SITPROF	REVENU
8 :94	1:18
5 :87	2:79
4 :64	3:64
7 :31	4:49
3 :25	5:83
2 : 9	6:21
(Other): 4	

On constate que, après avoir enlevé les données manquantes, la modalité `agriculteur` de la variable `SITPROF` vaut 0.

```
summary(pbio.cc$SITPROF)
 1  2  3  4  5  6  7  8
 0  9 25 64 87  4 31 94

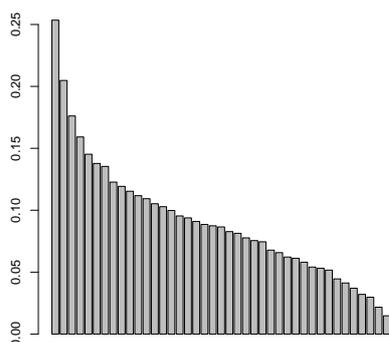
levels(pbio.cc$SITPROF)
[1] "1" "2" "3" "4" "5" "6" "7" "8"
```

Il faut donc redéfinir les modalités de cette variable.

```
pbio.cc$SITPROF <- factor(pbio.cc$SITPROF)
levels(pbio.cc$SITPROF)
[1] "2" "3" "4" "5" "6" "7" "8"
```

- Commenter les résultats de l'analyse des correspondances multiples réalisée sur l'ensemble des variables du tableau.

```
library(ade4)
acmtot <- dudi.acm(pbio.cc, scannf = F)
barplot(acmtot$eig)
```

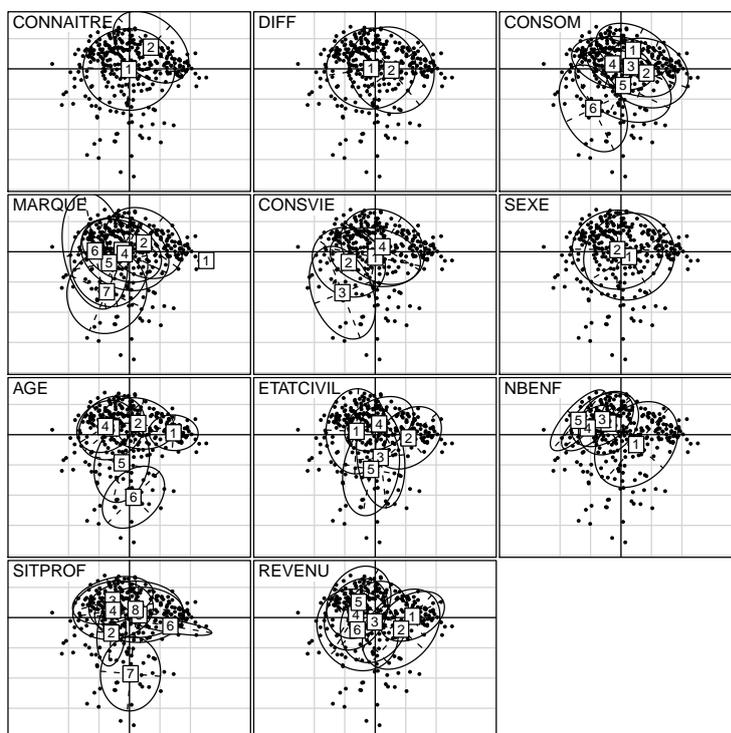


On note que le nombre important des valeurs propres (liées on le rappelle non aux variables mais aux modalités de ces variables) ne permet pas d'énoncer un critère de sélection du nombre de facteurs à conserver. On conserve 4 valeurs propres mais on ne détaillera dans la présentation que les deux premiers. A charge au lecteur de regarder les facteurs 3 et 4.

```
head(inertia.dudi(acmtot)$TOT)
  inertia      cum      ratio
1 0.2534726 0.2534726 0.06800484
2 0.2047920 0.4582646 0.12294905
3 0.1761199 0.6343845 0.17020072
4 0.1591724 0.7935569 0.21290550
5 0.1452219 0.9387788 0.25186747
6 0.1377689 1.0765476 0.28882985
```

En gardant les quatre premiers facteurs, on ne conserve que 21.29% de l'inertie totale. Mais ce pourcentage est relativement courant dans ce genre d'analyse.

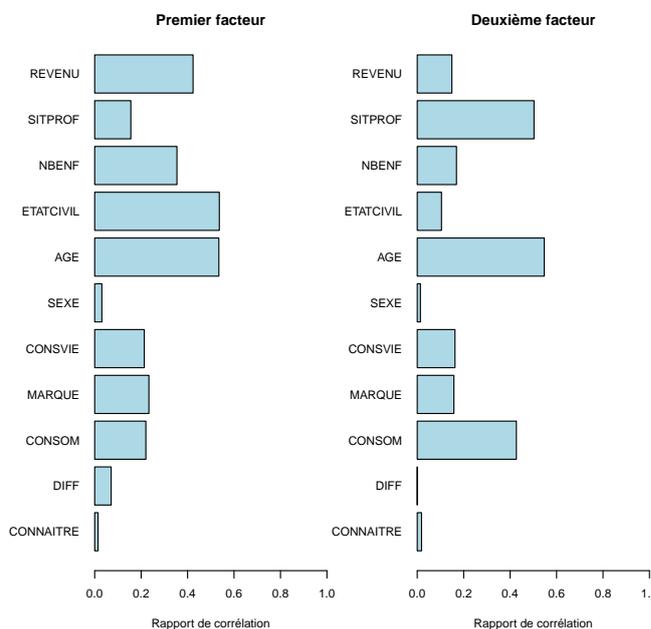
```
scatter(acmtot)
```



Dans cette représentation graphique, le même plan factoriel est répété autant de fois qu'il y a de variables qualitatives. Sur chaque plan, il y a 419 points correspondant aux individus enquêtés. Pour faciliter l'interprétation, on représente, variable par variable, la modalité prise par chaque individu et une ellipse résumant la dispersion des points. On voit par exemple que pour la variable CONSO, il y a opposition entre ceux qui consomment des produits biologiques plusieurs fois par semaine [6] et tous les autres, de ceux qui ne consomment jamais [1] à ceux qui consomment plusieurs fois par mois [5].

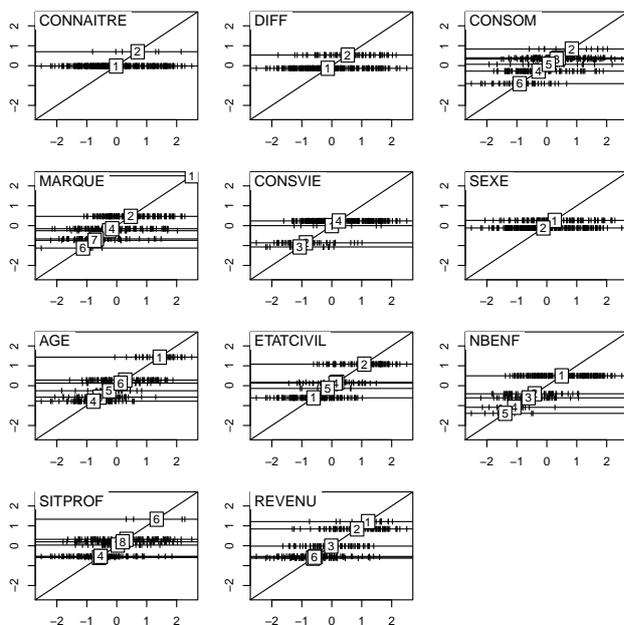
L'objectif de l'ACM étant d'obtenir des scores numériques des individus maximisant la somme des rapports de corrélation entre ces scores et les variables qualitatives, il est intéressant de les représenter.

```
par(mfrow = c(1, 2), mar = c(5, 6, 2, 0), cex = 0.7)
barplot(acmtot$cr[, 1], horiz = TRUE, xlim = c(0, 1), names.arg = colnames(pbio),
       las = 1, main = "Premier facteur", col = "lightblue", xlab = "Rapport de corrélation")
barplot(acmtot$cr[, 2], horiz = TRUE, xlim = c(0, 1), names.arg = colnames(pbio),
       las = 1, main = "Deuxième facteur", col = "lightblue", xlab = "Rapport de corrélation")
```



La fonction `score()` permet de visualiser les variables qualitatives avec un facteur. Pour chaque variable, les individus sont positionnés sur l'axe des abscisses par leur score sur l'axe factoriel considéré et, sur l'axe des ordonnées par le score de la modalité qu'ils portent. Le score d'une modalité est la moyenne des scores des individus portant cette modalité, ce qui est mis en évidence par la première bissectrice.

```
score(acmtot, xax = 1)
```



En étudiant le résultat de `score(acmtot, xax=2)`, faire le lien avec les

plans factoriels.

6. On considère les cinq premières variables comme des variables actives et les suivantes comme des variables illustratives.
 - (a) Réaliser une analyse des correspondances multiples sur les individus conservés dans le tableau et les 5 variables actives.
 - (b) Réaliser une analyse des correspondances multiples sur les individus conservés dans le tableau et les 6 variables illustratives.
 - (c) Calculer le coefficient de corrélation entre le premier facteur de l'analyse sur les variables actives et le premier facteur de l'analyse sur les variables illustratives. Commenter.

Références

Y. Dodge. *Premiers pas en statistique*. Springer-Verlag, 2003.