Chapitre 3: Analyse Factorielle des Correspondances

Mohamed Essaied Hamrita

ISMAI, Université Kairouan. Tunisie
mhamrita@gmail.com
http://hamrita.e-monsite.com/

Février 2014

Plan du chapitre

Motivation

On suppose donnée l'observation de deux variables, X et Y, qualitatives sur n individus. Le but de l'AFC est de résumer les dépendances entre les diverses modalités de X et Y afin de donner une vue résumée des données.

Motivation

On suppose donnée l'observation de deux variables, X et Y, qualitatives sur n individus. Le but de l'AFC est de résumer les dépendances entre les diverses modalités de X et Y afin de donner une vue résumée des données.
Généralement, ce type de données est représenté au travers d'une table de contingence, T:

	y_1	$\cdots$	y_j	$\cdots$	y_p	Total
x_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1p}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	$\ldots$	n_{kj}	$\ldots$	n_{kp}	$n_{k\cdot}$
Total	$n_{\cdot 1}$	$\ldots$	$n_{\cdot j}$	$\ldots$	$n_{\cdot p}$	$n_{\cdot\cdot}$

Exemple

Considérons l'exemple où X désigne la couleur des cheveux et Y celle des yeux, de la base de `HairEyeColor`.

Exemple

Considérons l'exemple où X désigne la couleur des cheveux et Y celle des yeux, de la base de `HairEyeColor`.

```
data(HairEyeColor)
(ex1 <- HairEyeColor[, , Sex = "Female"])
```

```
       Eye
Hair    Brown Blue Hazel Green
  Black    36    9     5     2
  Brown    66   34    29    14
  Red      16    7     7     7
  Blond     4   64     5     8
```

Exemple

Considérons l'exemple où X désigne la couleur des cheveux et Y celle des yeux, de la base de `HairEyeColor`.

```
data(HairEyeColor)
(ex1 <- HairEyeColor[, , Sex = "Female"])
```

```
       Eye
Hair    Brown Blue Hazel Green
  Black    36    9     5     2
  Brown    66   34    29    14
  Red      16    7     7     7
  Blond     4   64     5     8
```

Question: Existe-t-il une relation entre la couleur des cheveux et celle des yeux chez les femmes?

Notations

Definition (Effectifs marginaux)

On définit les effectifs marginaux par:

$$n_{i\cdot} = \sum_{j=1}^{p} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{k} n_{ij}.$$

```
# Effectifs marginaux
(ni. <- apply(ex1, 1, sum))  # ou ni.<-rowSums(ex1)

Black Brown   Red Blond
   52   143    37    81

(n.j <- apply(ex1, 2, sum))  # ou n.j<-colSums(ex1)

Brown  Blue Hazel Green
  122   114    46    31
```

Notations

Definition (Tableau des fréquences)

On définit le tableau des fréquences F par $F = f_{ij}$ où $f_{ij} = n_{ij}/n$ et n est la somme des n_{ij} (i.e. $n = \sum_i \sum_j n_{ij}$).

```
# Tableau des fréquences
(fij <- ex1/sum(ex1))
```

```
       Eye
Hair        Brown     Blue    Hazel     Green
  Black 0.115016 0.028754 0.015974 0.0063898
  Brown 0.210863 0.108626 0.092652 0.0447284
  Red   0.051118 0.022364 0.022364 0.0223642
  Blond 0.012780 0.204473 0.015974 0.0255591
```

Notations

Definition (Fréquences marginales)

On définit les fréquences marginales par:

$$f_{i\cdot} = \sum_{j=1}^{p} f_{ij}, \quad f_{\cdot j} = \sum_{i=1}^{k} f_{ij}.$$

```
# Fréquences marginales
(fi. <- apply(fij, 1, sum))

  Black   Brown     Red   Blond
0.16613 0.45687 0.11821 0.25879

(f.j <- apply(fij, 2, sum))

   Brown     Blue    Hazel    Green
0.389776 0.364217 0.146965 0.099042
```

Notations

Definition (Profils lignes)

On appelle profils lignes les fréquences conditionnelles:

$$PL_{ij} = f_{j|i} = \frac{f_{ij}}{f_{i.}} = D_1 F$$

où D_1 =diag$(1/f_{1.}, \ldots, 1/f_{k.})$.

```
(plij <- fij/fi.)  # ou plij<-diag(1/fi.)%*%fij

       Eye
Hair        Brown    Blue     Hazel    Green
  Black 0.692308 0.17308 0.096154 0.038462
  Brown 0.461538 0.23776 0.202797 0.097902
  Red   0.432432 0.18919 0.189189 0.189189
  Blond 0.049383 0.79012 0.061728 0.098765
```

Notations

Definition

On appelle profil ligne moyen le vecteur:

$$\overline{PL} = \sum_i f_i.PL_{ij}$$

```
(plbar <- t(fi.) %*% plij)

      Eye
         Brown    Blue   Hazel    Green
  [1,] 0.38978 0.36422 0.14696 0.099042
```

Notations

Definition (Profiles colonnes)

De même on définit les profiles colonnes et le profile colonne moyen par:

$$PC_{ij} = f_{j|i} = \frac{f_{ij}}{f_{\cdot j}} = D_2 F^t, \quad \overline{PC} = \sum_j f_{\cdot j} PC_{ij}$$

où $D_2 = \text{diag}(1/f_{\cdot 1}, \ldots, 1/f_{\cdot p})$.

```
(pcij <- t(t(fij)/f.j))  # ou pcij<-diag(1/f.j)%*%t(fij)

       Eye
Hair        Brown     Blue   Hazel    Green
  Black 0.295082 0.078947 0.10870 0.064516
  Brown 0.540984 0.298246 0.63043 0.451613
  Red   0.131148 0.061404 0.15217 0.225806
  Blond 0.032787 0.561404 0.10870 0.258065

t(pcbar <- pcij %*% f.j)

      Hair
         Black   Brown     Red   Blond
  [1,] 0.16613 0.45687 0.11821 0.25879
```

Hypothèse d'indépendance

Definition (Distance de χ^2)

On appelle distance de chi-deux entre X et Y la quantité:

$$\begin{aligned}\chi^2 &= n\varphi = n\sum_{i,j}\frac{(f_{ij}-f_{i.}\times f_{.j})^2}{f_{i.}\times f_{.j}}\\ &= n\sum_{i,j}\frac{(n_{ij}-n_{i.}\times n_{.j})^2}{n_{i.}\times n_{.j}} = n\left(\sum_{i,j}\frac{n_{ij}^2}{n_{i.}\times n_{.j}}-1\right)\end{aligned}$$

Hypothèse d'indépendance

Definition (Distance de χ^2)

On appelle distance de chi-deux entre X et Y la quantité:

$$\begin{aligned}\chi^2 &= n\varphi = n\sum_{i,j}\frac{(f_{ij}-f_{i\cdot}\times f_{\cdot j})^2}{f_{i\cdot}\times f_{\cdot j}} \\ &= n\sum_{i,j}\frac{(n_{ij}-n_{i\cdot}\times n_{\cdot j})^2}{n_{i\cdot}\times n_{\cdot j}} = n\left(\sum_{i,j}\frac{n_{ij}^2}{n_{i\cdot}\times n_{\cdot j}}-1\right)\end{aligned}$$

Cette grandeur est souvent utilisée comme test d'indépendance. En effet sous l'hypothèse nulle d'indépendance, χ^2 suit la loi chi-deux à $(k-1)\times(p-1)$ degrés de liberté .

Hypothèse d'indépendance

```
(phi <- sum((fij - fi. %o% f.j)^2/fi. %o% f.j))

[1] 0.34078

(chi2 <- phi * sum(ni.))

[1] 106.66

ddl <- (nrow(plij) - 1) * (ncol(plij) - 1)
qchisq(0.95, ddl)

[1] 16.919
```

Ici, au seuil $\alpha = 5\%$, on accepte l'hypothèse nulle d'indépendance car $\chi^2_{\text{tab}} > \chi^2_{\text{ob}}$.

Exercice

Un échantillon aléatoire de 1367 diplômes d'université, délivrés en 1984, a donné la répartition suivante:

	Licence	Mastère	Doctorat
Masculin	534	144	22
Féminin	515	141	11

On veut tester si le sexe et le niveau de diplôme obtenu sont liés.

1) Formuler l'hypothèse nulle testée.

2) Effectuer le test approprié. Conclure.

Exercice

Un échantillon aléatoire de 1367 diplômes d'université, délivrés en 1984, a donné la répartition suivante:

	Licence	Mastère	Doctorat
Masculin	534	144	22
Féminin	515	141	11

On veut tester si le sexe et le niveau de diplôme obtenu sont liés.
1) Formuler l'hypothèse nulle testée.
2) Effectuer le test approprié. Conclure.

```
Masculin <- c(534, 144, 22)
Féminin <- c(515, 141, 11)
dd <- rbind(Masculin, Féminin)
colnames(dd) <- c("Licence", "Mastère", "Doctorat")
dd
```

```
         Licence Mastère Doctorat
Masculin     534     144       22
Féminin      515     141       11
```

Solution

```
# Tab des effectifs théoriques
ni. <- rowSums(dd)
n.j <- colSums(dd)
(effT <- ni. %o% n.j/sum(dd))

          Licence Mastère Doctorat
Masculin  537.16  145.94   16.898
Féminin   511.84  139.06   16.102

(chi2ij <- (dd - effT)^2/effT)

           Licence  Mastère Doctorat
Masculin 0.018609 0.025789   1.5402
Féminin  0.019530 0.027065   1.6164

(chi2 <- sum(chi2ij))

[1] 3.2476

(chi2Theo <- qchisq(0.05, 2, lower.tail = F))

[1] 5.9915
```

Le but de l'AFC

Les objectifs de l'analyse factorielle des correspondances (AFC) sont de

- comparer les profils-lignes entre eux,

Le but de l'AFC

Les objectifs de l'analyse factorielle des correspondances (AFC) sont de

- comparer les profils-lignes entre eux,
- comparer les profils-colonnes entre eux,

Le but de l'AFC

Les objectifs de l'analyse factorielle des correspondances (AFC) sont de

- comparer les profils-lignes entre eux,
- comparer les profils-colonnes entre eux,
- repérer les cases du tableau où les effectifs observés n_{ij} sont nettement différents des effectifs théoriques (sous l'hypothèse d'indépendance).

Le but de l'AFC

Les objectifs de l'analyse factorielle des correspondances (AFC) sont de

- comparer les profils-lignes entre eux,
- comparer les profils-colonnes entre eux,
- repérer les cases du tableau où les effectifs observés n_{ij} sont nettement différents des effectifs théoriques (sous l'hypothèse d'indépendance).

Le but de l'AFC

Les objectifs de l'analyse factorielle des correspondances (AFC) sont de

- comparer les profils-lignes entre eux,
- comparer les profils-colonnes entre eux,
- repérer les cases du tableau où les effectifs observés n_{ij} sont nettement différents des effectifs théoriques (sous l'hypothèse d'indépendance).

L'AFC est une méthode faisant apparaître les cartéristiques de la situation d'indépendance, au niveau des lignes, des colonnes, ou des cases du tableau de contingence.

Utiliser la distance de χ^2

L'idée pour comparer des profils lignes ou des profils colonnes sera d'utiliser la distance du χ^2. La distance entre deux profils lignes PL_{i_1} et PL_{i_2} sera alors:

Utiliser la distance de χ^2

L'idée pour comparer des profils lignes ou des profils colonnes sera d'utiliser la distance du χ^2. La distance entre deux profils lignes PL_{i_1} et PL_{i_2} sera alors:

$$d(PL_{i_1}, PL_{i_2}) = \sum_{j=1}^{p} \frac{n}{n_{.j}} \left(\frac{n_{i_1 j}}{n_{.j}} - \frac{n_{i_2 j}}{n_{.j}} \right)^2 = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{i_1 j}}{f_{i1.}} - \frac{f_{i_2 j}}{f_{i1.}} \right)^2$$

Ecriture matricielle

Si T désigne la matrice associée au tableau de contingence, $T = [n_{ij}]$, on note

$$D_1 = \text{diag}(1/n_1., 1/n_2., \ldots, 1/n_k.) \quad D_2 = \text{diag}(1/n.1, 1/n.2, \ldots, 1/n.p)$$

On peut alors obtenir facilment les profils lignes et colonnes, en notant que

$$PL = D_1 T, \quad PC = TD_2$$

$\Longrightarrow$ on peut alors étudier le nuage des profils lignes, dans $\mathbb{R}^k$, $o chaque point on associe un poids gal sa fr quence marginale$: $la matrice des poids est alors \frac{1}{n} D_1^{-1}$.

Le centre de gravité du nuage est le profil ligne moyen

$$\overline{PL} = \frac{1}{n} (D_1 T)' D_1^{-1}$$

Une double ACP

L'AFC est, en fait, une double ACP:

Une double ACP

L'AFC est, en fait, une double ACP:

- sur les profils lignes:
  - Tableau de données: $PL = D_1 T$.
  - Matrice de poids: $\frac{1}{n} D_1^{-1}$.
  - Métrique d'écart à l'indépendance: nD_2.

Une double ACP

L'AFC est, en fait, une double ACP:

- sur les profils lignes:
  - Tableau de données: $PL = D_1 T$.
  - Matrice de poids: $\frac{1}{n} D_1^{-1}$.
  - Métrique d'écart à l'indépendance: nD_2.
- sur les profils colonnes:
  - Tableau de données: $PC = D_2 T'$.
  - Matrice de poids: $\frac{1}{n} D_2^{-1}$.
  - Métrique d'écart à l'indépendance: nD_1.

ACP sur profils lignes

- Axes principaux: ce sont les vecteurs propres, a^j, de:

$$PL'(1/nD_1^{-1})PLnD2 = T'D_1TD_2$$

La modalité j de la variable Y (ici, les "variables") est représentée par les coordonnées:

$$\sqrt{\lambda_1}a_1^j, \ \ldots, \ \sqrt{\lambda_k}a_k^j$$

- Coordonnées principales: $c^j = PL(nD_2)a^j = nD_1TD_2a^j$.

La modalité i de la variable X (ici les "individus") est représentée par les coordonnées: $c_i^1, \ \ldots \ c_i^p$.

ACP sur profils colonnes (par symétrie)

- Axes principaux: ce sont les vecteurs propres, b^j, de:

$$TD_2T'D_1$$

La modalité i de la variable X (ici, les "variables") est représentée par les coordonnées:

$$\sqrt{\lambda'_1}b^j_1, \ \ldots, \ \sqrt{\lambda'_k}b^j_k$$

- Coordonnées principales: $d^j = nD_2TD_1b^j$.

La modalité j de la variable y (ici les "individus") est représentée par les coordonnées: $d^1_j, \ \ldots \ c^p_j$.

Etude de cas

Considérons l'exemple de l'étude de la correspondance entre la catégorie socioprofessionnelle (CSP) et le type d'hébergement en vacances.
Source: M. Goguel (1967). Les vacances des Français en 1966. Etudes et conjoncture.
Le tableau de contingence est donné comme suit:

```
vac <- read.table("D://Enseignement/Proba/Analyse des données/MonCours/Data/vac
    h = T, row.names = 1, sep = "")
vac
```

	Hotel	Location	Res.Second	Parents	Amis	Camping	Sej.org
Agriculteurs	195	62	1	499	44	141	49
Patrons	700	354	229	959	185	292	119
Cadres.sup	961	471	633	1580	305	360	162
Cadre.moy	572	537	279	1689	206	748	155
Employes	441	404	166	1079	178	434	178
Ouvriers	783	1114	387	4052	497	1464	525
Autres.actifs	142	103	210	1133	132	181	46
Inactifs	741	332	327	1789	311	236	102
Total	4535	3377	2232	12780	1858	3856	1336

	Autres	Total
Agriculteurs	65	1056
Patrons	140	2978
Cadres.sup	148	4620
Cadre.moy	112	4298
Employes	92	2972
Ouvriers	387	9209
Autres.actifs	59	2006
Inactifs	102	3940
Total	1105	31079