Analyse Factorielle des Correspondances - Analyse des Correspondances Multiples

STID - 2A

Maxime FRANCOISE

2020 – 2021

Plan

- **Analyse Factorielle des Correspondances (AFC)**
  - Tableau de contingence et distance du χ^2
  - Visualisation, compression, débruitage
  - Utilisation de l'outil logiciel
- **Analyse des Correspondances Multipes (ACM)**

Analyse Factorielle des Correspondances

Données et Problème

- **Données**
  - Tableau de contingence entre 2 variables qualitatives
  - Plus généralement tableau de nombres non-négatifs

- **Problème**
  - Visualiser les **correspondances** entre les modalités d'une même variable
  - **Représentation simultanée** des modalités des 2 variables pour analyser les **liens** entre les 2 variables

Exemple

pr	pathcats			
	2-5 cm	<= 2 cm	> 5 cm	Total
Negative	121	248	6	375
Positive	105	336	3	444
Unknown	57	242	3	302
Total	283	826	12	1121

Distribution des effectifs croisés selon les modalités du statut des récepteurs Progestérone et de la taille de la tumeur pour $N = 1121$ femmes atteintes du cancer du sein.

$\Longrightarrow n = 3$ modalités pour la variable "statut des récepteurs Progestérone", $p = 3$ modalités pour la variable "taille de la tumeur".

Pourquoi pas une ACP ?

Tableau à $n \times p$ modalités $\longrightarrow$ pourquoi ne pas effectuer directement une ACP dessus ?

Parce-que :

- la distance euclidienne entre 2 modalités n'a pas de sens,
- les lignes et les colonnes du tableau jouent des rôles symétriques $\longrightarrow$ on doit pouvoir représenter les modalités indifféremment selon les lignes ou les colonnes,
- on cherche plutôt à représenter les **distributions** des modalités sur la population.

$\Longrightarrow$ On s'intéresse plus aux **distributions conditionnelles** de chaque modalité : 2 modalités proches auront des distributions conditionnelles comparables.

Notations

- Tableau de contingence à n lignes et p colonnes.
- $f_{i,j}$ = fréquence des individus prenant simultanément les modalités x_i de la variable ligne et y_j de la variable colonne (distribution jointe).
- **$f_{i\bullet}$ (resp. $f_{\bullet j}$)** = fréquence des individus prenant la modalité x_i (resp. y_j) de la variable ligne (resp. colonne)

$$f_{i\bullet} = \sum_{j=1}^{p} f_{i,j}, \quad f_{\bullet j} = \sum_{i=1}^{n} f_{i,j} \quad \text{(distributions marginales)}$$

- **Profils-lignes (resp. profils-colonnes)** = tableau des fréquences conditionnelles aux modalités de la variable ligne (resp. colonne)

$$f_{j/i} = \frac{f_{i,j}}{f_{i\bullet}}, \qquad f_{i/j} = \frac{f_{i,j}}{f_{\bullet j}} \qquad \text{(distributions conditionnelles)}$$

- **Profil moyen** : profil $f_I = (f_{\bullet j})_{1\leq j\leq p}$ (resp. $f_J = (f_{i\bullet})_{1\leq i\leq n}$) de la distribution marginale en colonnes (resp. en lignes).

Fréquences

$X \backslash Y$	y_1	y_2	$\dots$	y_j	$\dots$	y_p	Profil moyen f_J
x_1	f_{11}	f_{12}	$\dots$	f_{1j}	$\dots$	f_{1p}	$f_{1\bullet}$
x_2	f_{21}	f_{22}	$\dots$	f_{2j}	$\dots$	f_{2p}	$f_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	f_{i1}	f_{i2}	$\dots$	f_{ij}	$\dots$	f_{ip}	$f_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	f_{n1}	f_{n2}	$\dots$	f_{nj}	$\dots$	f_{np}	$f_{n\bullet}$
Profil moyen f_I	$f_{\bullet 1}$	$f_{\bullet 2}$	$\dots$	$f_{\bullet j}$	$\dots$	$f_{\bullet p}$	1

Profils-lignes

$X \backslash Y$	y_1	y_2	$\dots$	y_j	$\dots$	y_p	**Total**
x_1	$f_{1/1}$	$f_{2/1}$	$\dots$	$f_{j/1}$	$\dots$	$f_{p/1}$	1
x_2	$f_{1/2}$	$f_{2/2}$	$\dots$	$f_{j/2}$	$\dots$	$f_{p/2}$	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_i	$f_{1/i}$	$f_{2/i}$	$\dots$	$f_{j/i}$	$\dots$	$f_{p/i}$	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_n	$f_{1/n}$	$f_{2/n}$	$\dots$	$f_{j/n}$	$\dots$	$f_{p/n}$	1
Profil moyen f_I	$f_{\bullet 1}$	$f_{\bullet 2}$	$\dots$	$f_{\bullet j}$	$\dots$	$f_{\bullet p}$	1

Exemple : Profils-lignes

Pourcentage
Pct en ligne

Table de pr par pathcats			
	pathcats		
pr	2-5 cm	<= 2 cm	> 5 cm
Negative	10.79 32.27	22.12 66.13	0.54 1.60
Positive	9.37 23.65	29.97 75.68	0.27 0.68
Unknown	5.08 18.87	21.59 80.13	0.27 0.99
Total	283 25.25	826 73.68	12 1.07

Profils-colonnes

$X \backslash Y$	y_1	y_2	$\dots$	y_j	$\dots$	y_p	Profil moyen f_J
x_1	$f_{1/1}$	$f_{1/2}$	$\dots$	$f_{1/j}$	$\dots$	$f_{1/p}$	$f_{1\bullet}$
x_2	$f_{2/1}$	$f_{2/2}$	$\dots$	$f_{2/j}$	$\dots$	$f_{2/p}$	$f_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	$f_{i/1}$	$f_{i/2}$	$\dots$	$f_{i/j}$	$\dots$	$f_{i/p}$	$f_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	$f_{n/1}$	$f_{n/2}$	$\dots$	$f_{n/j}$	$\dots$	$f_{n/p}$	$f_{n\bullet}$
Total	1	1	$\dots$	1	$\dots$	1	1

Exemple : Profils-colonnes

Pourcentage
Pct en col.

Table de pr par pathcats				
	pathcats			
pr	2-5 cm	<= 2 cm	> 5 cm	Total
Negative	10.79 42.76	22.12 30.02	0.54 50.00	33.45
Positive	9.37 37.10	29.97 40.68	0.27 25.00	39.61
Unknown	5.08 20.14	21.59 29.30	0.27 25.00	26.94

Transformations du tableau de contingence (LPM [1])

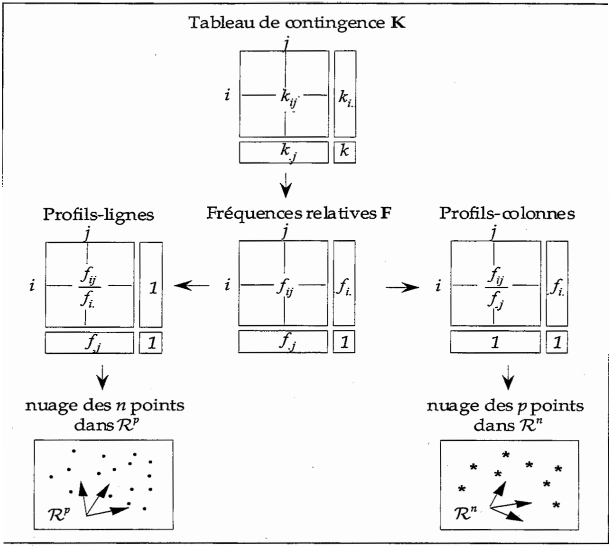


Figure 4.2 – 1. Transformations du tableau de contingence

Distance du χ^2

- **Distance euclidienne pondérée** entre les n **points constitués des profils-lignes** (resp. entre les p **points constitués des profils-colonnes**).
- **Profils-lignes** :

$$d^2(x_i, x_{i'}) = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{i,j}}{f_{i.}} - \frac{f_{i',j}}{f_{i'.}} \right)^2 = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(f_{j/i} - f_{j/i'} \right)^2$$

- **Profils-colonnes** :

$$d^2(y_j, y_{j'}) = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{i,j}}{f_{.j}} - \frac{f_{i,j'}}{f_{.j'}} \right)^2 = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(f_{i/j} - f_{i/j'} \right)^2$$

- **Adaptée** à la comparaison des distributions conditionnelles : 2 modalités ayant les **mêmes distributions conditionnelles** sont à **distance nulle**.

AFC : Principe général ($p \leq n$) [2]

- **Transformation** du tableau de contingence afin de récupérer les profils-lignes et colonnes.

- **Pondération** de chaque modalité i par sa fréquence $f_{i.}$.

- **ACP** sur le tableau des profils-lignes **avec la distance du χ^2** :
  - Maximisation sur chaque axe de la distance de chaque modalité x_i de la variable ligne au profil-ligne moyen f_I.
  - Association de chaque ligne i à un point M_i barycentre des p facteurs pondéré par les fréquences conditionnelles $(f_{i,j}/f_{.j})_{1 \leq j \leq p}$.

- **Représentation** des modalités dans les plans factoriels centrés sur le profil moyen f_I.

- **Choix des axes** pour l'analyse des correspondances de la même manière que pour l'ACP.

2. si $n \leq p$, on travaille plutôt sur le tableau des profils-colonnes

Résultats mathématiques

On montre que :

- la matrice $p \times p$ Σ diagonalisée dans l'ACP avec la distance du χ^2 a pour terme général

$$\Sigma_{j,j'} = \sum_{i=1}^{n} \frac{f_{i,j} f_{i,j'}}{f_{i.} f_{.j'}}$$

- Σ admet $p-1$ **valeurs propres** réelles positives $\lambda_1 \geqslant \ldots \geqslant \lambda_{p-1}$ **différentes de** 1 et, pour tout $j = 1, \ldots, p-1$, $0 \leq \lambda_j \leq 1$.
- **L'inertie totale** du nuage I_{tot} est **proportionnelle** à la **statistique** T **du test du** χ^2 **d'indépendance** : soit N le nombre d'individus,

$$I_{tot} = \sum_{i=1}^{n} f_{i.} d^2(x_i, f_I) = \sum_{j=1}^{p} f_{.j} d^2(y_j, f_J) = \frac{T}{N}$$

- **Relation quasi-barycentrique** : les p modalités en colonnes peuvent être représentées sur les **mêmes axes factoriels** que les n modalités en lignes.

Exemple : Résultats

Décomposition de l'inertie et du Khi 2					
Valeur singulière	Inertie principale	Khi 2	Pourcentage	Pourcent. cumulé	19 38 57 76 95 ----+----+----+----+----+---
0.12845	0.01650	18.4947	96.74	96.74	*************************
0.02357	0.00056	0.6229	3.26	100.00	*
Total	0.01705	19.1176	100.00		
Degrés de liberté = 4					

- **Valeur singulière** : racine carrée de la valeur propre correspondante.
- **2 axes factoriels au maximum** $= \min(n-1, p-1)$.
- La ligne "Total" donne la valeur de la **statistique T du test du χ^2** d'indépendance entre les 2 variables.
- Le premier axe factoriel concentre la plus grosse part de l'inertie.

Exemple : Résultats (2)

Coordonnées des lignes	Dim1	Dim2
Negative	0.1726	0.0101
Positive	-0.0472	-0.0278
Unknown	-0.1448	0.0283

Contributions partielles à l'inertie des points des lignes	Dim1	Dim2
Negative	0.6039	0.0616
Positive	0.0536	0.5503
Unknown	0.3425	0.3881

Carré des cosinus pour les points des lignes	Dim1	Dim2
Negative	0.9966	0.0034
Positive	0.7430	0.2570
Unknown	0.9632	0.0368

- **Coordonnées** des modalités lignes sur chaque axe.
- **Contribution** des modalités lignes à chaque axe : axe expliqué par les modalités dont la contribution est $\geqslant 1/n = 0,33$.
- **Le carré des cosinus** représente la position relative des modalités par rapport aux axes : plus il est proche de 1, et plus la modalité est bien représentée sur l'axe.
- $\Longrightarrow$ l'axe 1 est construit sur les modalités opposées "Negative" et "Unknown", bien représentées par cet axe.

Exemple : Résultats (3)

Coordonnées des colonnes		
	Dim1	Dim2
2-5 cm	0.2109	-0.0121
<= 2 cm	-0.0766	0.0010
> 5 cm	0.2979	0.2199

Contributions partielles à l'inertie des points des colonnes		
	Dim1	Dim2
2-5 cm	0.6805	0.0671
<= 2 cm	0.2619	0.0012
> 5 cm	0.0576	0.9317

Carré des cosinus pour les points des colonnes		
	Dim1	Dim2
2-5 cm	0.9967	0.0033
<= 2 cm	0.9998	0.0002
> 5 cm	0.6473	0.3527

- Axe expliqué par les modalités dont la contribution est $\geqslant 1/p = 0,33$
- L'axe 1 est construit sur la modalité "2-5 cm".
- Les modalités opposées "2-5 cm" et "$<=$ 2 cm" sont bien représentées sur l'axe 1.

Représentation simultanée (LPM [3])

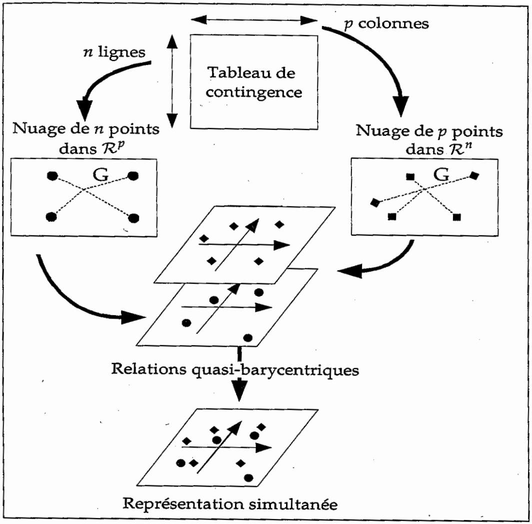


Figure 4.2 – 3. Schéma de la représentation simultanée

Interprétation géométrique (LPM [4])

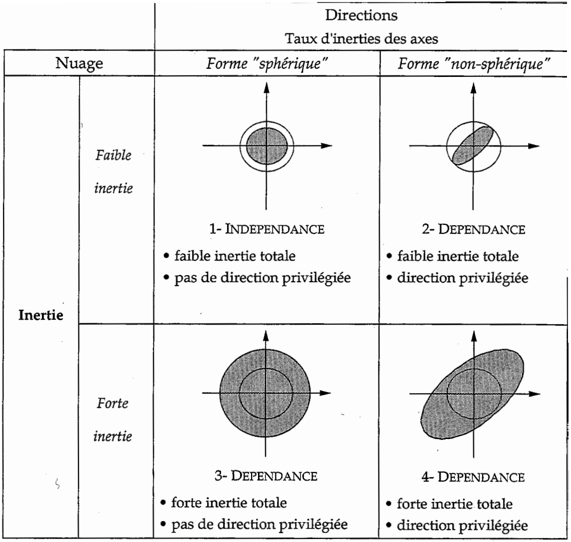
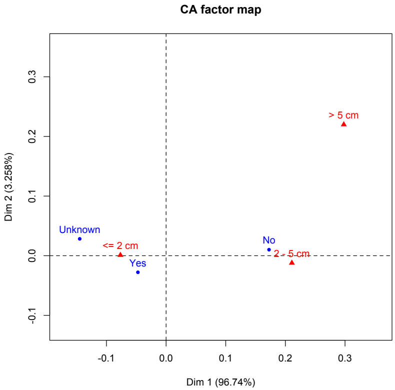


Figure 4.3 – 1. Indépendance et dépendances

Exemple : Représentation dans le premier plan factoriel



Axe 2 très peu représentatif ⇒ forte dépendance. Axe 1 porté par les modalités ("Negative" ; "2-5 cm"), et "Unknown". La modalité ">5 cm" est atypique : elle pourrait être traitée en modalité illustrative.

Interprétation des résultats

- **Analyse simultanée**
  - des contributions,
  - des cosinus carrés.

- Comme dans le cadre de l'ACP, une modalité ligne (resp. colonne) n'intervient dans l'interprétation d'un axe factoriel **que si** elle est **bien représentée** sur l'axe, i. e. son cosinus carré est proche de 1.

- Si deux modalités **bien représentées** sur un plan factoriel sont **proches**, leurs **distributions** sont **comparables**.

  $\Longrightarrow$ les individus prenant ces modalités se comportent de manière comparable.

Modalités atypiques

- Modalités ayant de **fortes contributions**, mais relativement **mal représentées** sur les axes factoriels.
- Effet souvent dû à la distance du χ^2, qui a tendance à sur-représenter les modalités de faible effectif.
- **Que faire ?**
  - les **éliminer** de l'analyse,
  - **uniquement si ce sont des modalités de faible effectif** (**apurement**),
    - les **regrouper** avec des modalités comparables,
    - les **ventiler** sur les autres modalités : attribuer de manière aléatoire une autre modalité aux individus concernés.

Effet Guttman

- Si les $p-1$ valeurs propres différentes de 1 ont **toutes** une valeur **proche de 1**, on parle d'**effet Guttman**.
- Chaque **modalité ligne** correspond alors exactement à une **modalité colonne**.

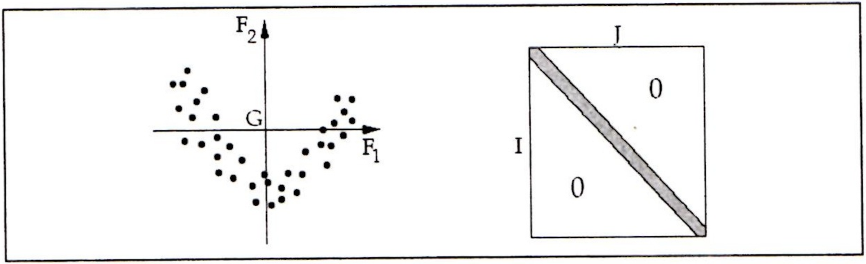


Figure 4.3 – 4. Effet Guttman et structure possible du tableau

Choix du nombre de facteurs

- **Choix du nombre de facteurs** résumant convenablement l'information contenue dans le tableau de contingence.
- **Méthodes** identiques à celles utilisées pour l'ACP.
- **Ebouli des valeurs propres** : garder les facteurs pour lesquels les valeurs propres se trouvent avant le point d'inflexion.
- **Règle de Kaiser** :
  - Calcul de la moyenne des $p-1$ valeurs propres différentes de 1

$$\overline{\lambda} = \frac{\sum_{j=1}^{p-1} \lambda_j}{p-1}$$

  - Sélection des facteurs pour lesquels $\lambda_j \geqslant \overline{\lambda}$.

Exemple : Choix du nombre de facteurs

- 2 valeurs propres : $\lambda_1 = 0,016$, $\lambda_2 = 0,0006$.
  $\Longrightarrow$ l'ébouli n'apporte rien dans ce cas.
- Moyenne des valeurs propres : $\overline{\lambda} = 0,009$
  $\Longrightarrow$ on ne retient qu'un seul axe factoriel.
- Le premier axe factoriel explique plus de 95% de l'inertie du nuage de points-modalités.
- Les modalités sont réparties sur le premier axe factoriel (hors modalité atypique ">5 cm").
  $\Longrightarrow$ Les deux variables sont fortement dépendantes.

AFC sous R : fonction CA

Avec le package FactoMineR.

- Création du tableau de contingence dans le cas de données brutes :
```
contingence = data.frame(matrix(table(var1,var2),
ncol = length(unique(var2))))
rownames(contingence) = levels(var1)
colnames(contingence) = levels(var2)
```

- AFC sur le tableau de contingence (où var2 a le plus petit nombre de modalités) :
```
var12.afc = CA(contingence,
ncp = length(unique(var2))-1,
graph = T)
```
ncp : donne le nombre d'axes factoriels à garder (défaut = 5).
graph : produit la représentation des modalités dans le premier plan factoriel.

Sorties de la fonction `CA`

- `var12.afc$eig` : tableau des valeurs propres et des pourcentages d'inertie expliquée pour chaque facteur.
- `var12.afc$col` : liste des coordonnées, contributions et cosinus carrés des modalités colonnes sur chaque facteur.
- `var12.afc$row` : liste des coordonnées, contributions et cosinus carrés des modalités lignes sur chaque facteur.

Exemple sous R : sorties de la fonction CA

```
> cancer.afc$eig
       eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.0164984197               96.74199                          96.74199
dim 2 0.0005556225                3.25801                         100.00000
>
```

```
> cancer.afc$row$coord
              Dim 1       Dim 2
No       0.17257769  0.01011458
Yes     -0.04724481 -0.02778541
Unknown -0.14483424  0.02829057
```

```
> cancer.afc$row$contrib
           Dim 1    Dim 2
No      60.38828  6.15945
Yes      5.35850 55.03401
Unknown 34.25322 38.80654
```

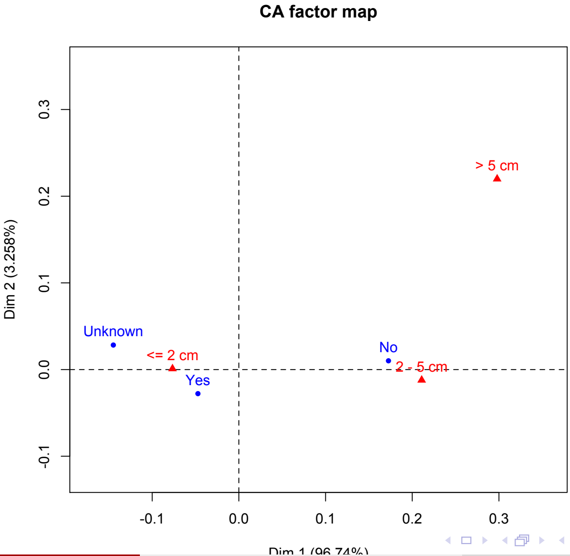
```
> cancer.afc$row$cos2
            Dim 1       Dim 2
No      0.9965768 0.003423241
Yes     0.7430083 0.256991680
Unknown 0.9632482 0.036751844
```

```
> cancer.afc$col$coord
               Dim 1        Dim 2
<= 2 cm   -0.0765800  0.000967516
2 - 5 cm   0.2108828 -0.012148617
> 5 cm     0.2979387  0.219907541
```

```
> cancer.afc$col$contrib
              Dim 1      Dim 2
<= 2 cm   26.191650  0.1241398
2 - 5 cm  68.048820  6.7058637
> 5 cm     5.759531 93.1699965
```

```
> cancer.afc$col$cos2
              Dim 1        Dim 2
<= 2 cm   0.9998404 0.0001595939
2 - 5 cm  0.9966923 0.0033077498
> 5 cm    0.6473386 0.3526614063
```

Exemple sous R : sorties de la fonction CA (2)



AFC sous SAS : procédure `corresp`

```
/* AFC sur une table de contingence */
proc corresp data = cancer;
tables pr,pathcats;
run;
```

Options utiles :

- **short** : n'affiche que les tableaux des valeurs singulières et des coordonnées.
- **NOPRINT** : n'affiche pas les sorties.
- **Attention ! !** : ne pas oublier la virgule entre les noms des variables. Sinon `corresp` effectue une ACM.

Sorties (utiles) de la procédure corresp

Décomposition de l'inertie et du Khi 2					
Valeur singulière	Inertie principale	Khi 2	Pourcentage	Pourcent. cumulé	19 38 57 76 95 ----+----+----+----+----+---
0.12845	0.01650	18.4947	96.74	96.74	************************
0.02357	0.00056	0.6229	3.26	100.00	*
Total	0.01705	19.1176	100.00		
Degrés de liberté = 4					

Sorties (utiles) de la procédure `corresp` (2)

Coordonnées des lignes		
	Dim1	Dim2
Negative	0.1726	0.0101
Positive	-0.0472	-0.0278
Unknown	-0.1448	0.0283

Contributions partielles à l'inertie des points des lignes		
	Dim1	Dim2
Negative	0.6039	0.0616
Positive	0.0536	0.5503
Unknown	0.3425	0.3881

Carré des cosinus pour les points des lignes		
	Dim1	Dim2
Negative	0.9966	0.0034
Positive	0.7430	0.2570
Unknown	0.9632	0.0368

Sorties (utiles) de la procédure corresp (3)

Coordonnées des colonnes		
	Dim1	Dim2
2-5 cm	0.2109	-0.0121
<= 2 cm	-0.0766	0.0010
> 5 cm	0.2979	0.2199

Contributions partielles à l'inertie des points des colonnes		
	Dim1	Dim2
2-5 cm	0.6805	0.0671
<= 2 cm	0.2619	0.0012
> 5 cm	0.0576	0.9317

Carré des cosinus pour les points des colonnes		
	Dim1	Dim2
2-5 cm	0.9967	0.0033
<= 2 cm	0.9998	0.0002
> 5 cm	0.6473	0.3527

Analyse des Correspondances Multipes (ACM)

Données et Problème

- **Données**
  - Tableau X de n individus et $s > 2$ variables **qualitatives**.
  - Correspond à la donnée de plusieurs tableaux de contingences observés sur les mêmes individus.
- **Problème**
  - Etablir les **correspondances** entre les modalités d'une même variable.
  - Visualiser les liens entre plusieurs variables à l'aide d'une **représentation simultanée**.

Exemple

VIEWTABLE: Work.Cancer

	lnpos	hisgrad	er	pr	status	pathcats	ln_yesno
1	1	Unknown	Unknown	Unknown	Censored	<= 2 cm	Yes
2	0	1	Positive	Positive	Censored	<= 2 cm	No
3	0	1	Unknown	Unknown	Censored	<= 2 cm	No
4	0	1	Unknown	Unknown	Censored	<= 2 cm	No
5	0	2	Unknown	Unknown	Censored	<= 2 cm	No
6	0	Unknown	Unknown	Unknown	Censored	<= 2 cm	No
7	0	1	Unknown	Unknown	Censored	<= 2 cm	No
8	0	3	Unknown	Unknown	Censored	<= 2 cm	No
9	0	2	Unknown	Unknown	Censored	<= 2 cm	No
10	0	1	Positive	Positive	Censored	<= 2 cm	No
11	0	2	Unknown	Unknown	Censored	<= 2 cm	No
12	0	3	Unknown	Unknown	Censored	<= 2 cm	No
13	0	2	Unknown	Unknown	Censored	<= 2 cm	No
14	0	3	Negative	Negative	Censored	<= 2 cm	No
15	0	2	Unknown	Unknown	Censored	<= 2 cm	No
16	0	2	Unknown	Unknown	Censored	<= 2 cm	No
17	0	2	Unknown	Unknown	Censored	<= 2 cm	No
18	0	3	Negative	Negative	Censored	<= 2 cm	No
19	0	3	Positive	Negative	Censored	<= 2 cm	No
20	0	2	Unknown	Unknown	Censored	<= 2 cm	No
21	0	2	Negative	Negative	Censored	<= 2 cm	No
22	0	3	Negative	Negative	Censored	<= 2 cm	No
23	0	Unknown	Unknown	Unknown	Censored	<= 2 cm	No
24	0	2	Unknown	Unknown	Censored	<= 2 cm	No
25	1	3	Positive	Positive	Censored	<= 2 cm	Yes
26	0	2	Unknown	Unknown	Censored	<= 2 cm	No
27	0	2	Unknown	Unknown	Censored	<= 2 cm	No
28	0	2	Unknown	Unknown	Censored	<= 2 cm	No
29	0	2	Positive	Negative	Censored	<= 2 cm	No
30	0	1	Unknown	Unknown	Censored	<= 2 cm	No
31	0	2	Unknown	Unknown	Censored	<= 2 cm	No
32	3	3	Unknown	Unknown	Censored	<= 2 cm	Yes
33	0	Unknown	Positive	Positive	Censored	<= 2 cm	No

$s = 7$ variables nominales pour $n = 1121$ femmes atteintes du cancer du sein.

Fondements et principes

- Lorsque $s = 2$, il est équivalent d'effectuer une AFC sur le tableau de contingence à p_1 lignes et p_2 colonnes ou sur le tableau **binaire** à n lignes et $p_1 + p_2$ colonnes correspondant.
- **Généralisation immédiate** au cas de $s > 2$ variables nominales.
- $\Longrightarrow$ **Analyse des correspondances** du **tableau disjonctif complet** ou du **tableau de Burt** Z issu du tableau initial X :
  - Transformations de Z en profils-lignes et profils-colonnes,
  - Pondération des points par leurs profils marginaux,
  - ACP avec la distance du χ^2 sur le tableau des profils ayant le plus de lignes.

Tableau disjonctif complet (LPM[5])

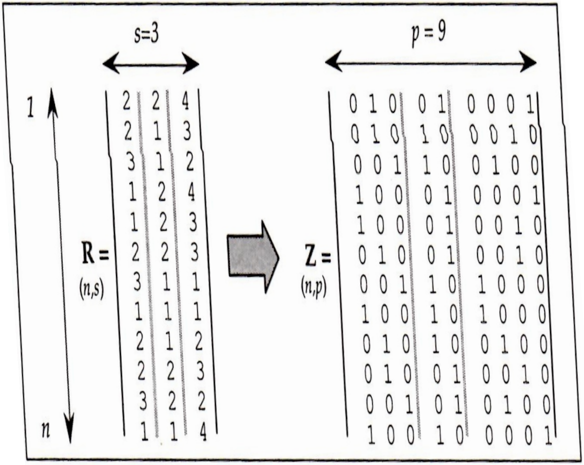


Figure 5.1 - .2. Construction du tableau disjonctif complet Z

Tableau de Burt (LPM[6])

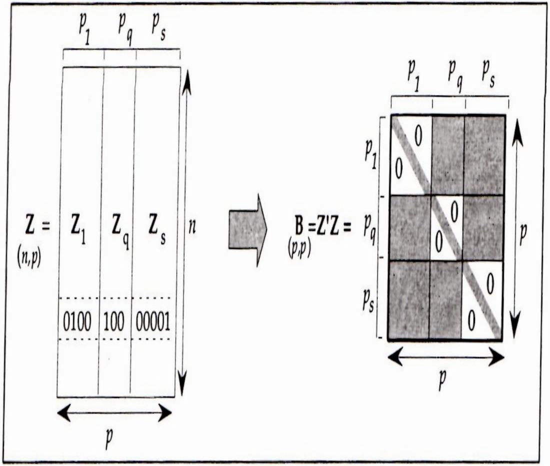


Figure 5.1 – 3. Construction du tableau des faces de l'hypercube (tableau de Burt) B à partir du tableau disjonctif complet Z

Apurement (Ventilation)

- **Apurement** : réponse au problème des modalités **de faible effectif** qui peuvent perturber l'analyse :
  - petit nuage de points très concentré et très éloigné des autres,
  - petit effectif ayant un grand poids dans l'analyse,
  - peuvent rendre instables les axes factoriels.

- L'apurement rend l'analyse plus robuste.

- Les modalités dont l'effectif est insuffisant sont ventilées aléatoirement dans les autres modalités : les individus concernés sont répartis aléatoirement dans les autres modalités.

- Les modalités ventilées sont gardées comme modalités supplémentaires dans l'analyse.

- Par défaut, une modalité est ventilée si son effectif est inférieur à **2%** de l'effectif total.

Caractéristiques de l'ACM

- **Nombre maximal** de facteurs principaux = *nombre total de modalités non-ventilées - nombre de variables* $= p' - s$.
- **Somme des valeurs propres** :

$$\sum_{j=1}^{p'-1} \lambda_j = \frac{p' - s}{s}.$$

- Les premiers axes expliquent une **faible part de l'inertie**.
- La **décroissance des valeurs propres** est moins forte que dans l'ACP ou l'AFC.
- $\Longrightarrow$ Le **nombre d'axes** à retenir pour une analyse ultérieure est plus important que pour l'ACP ou l'AFC.
- **Règle de Kaiser** : moyenne des valeurs propres =

$$\frac{1}{p'-1} \sum_{j=1}^{p'-1} \lambda_j = \frac{p' - s}{s \times (p' - 1)}.$$

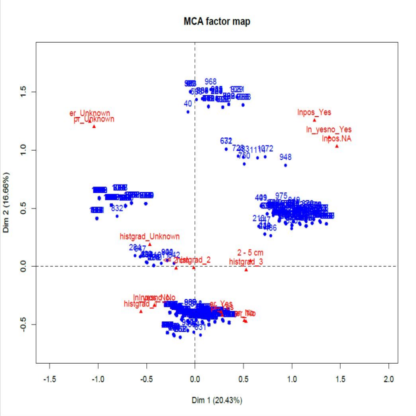
Exemple : Valeurs singulières et inertie expliquée

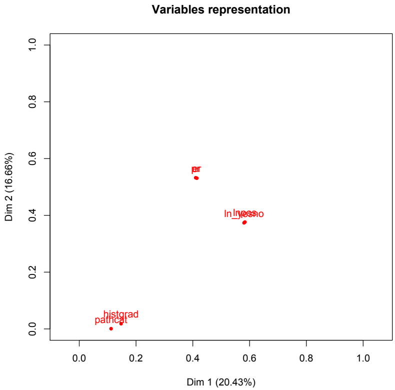
Décomposition de l'inertie et du Khi 2					
Valeur singulière	Inertie principale	Khi 2	Pourcentage	Pourcent. cumulé	6 12 18 24 30
0.40896	0.16724	4687.0	32.07	32.07	***************************
0.35161	0.12363	3464.8	23.71	55.78	********************
0.23810	0.05669	1588.8	10.87	66.66	*********
0.20857	0.04350	1219.1	8.34	75.00	*******
0.19748	0.03900	1092.9	7.48	82.48	******
0.18845	0.03551	995.3	6.81	89.29	******
0.16058	0.02578	722.6	4.94	94.23	****
0.15168	0.02301	644.8	4.41	98.65	****
0.08328	0.00694	194.4	1.33	99.98	*
0.01130	0.00013	3.6	0.02	100.00	
Total	0.52144	14613.2	100.00		
Degrés de liberté = 196					

Représentation simultanée individus/modalités

- **Relation quasi-barycentrique** : comme pour l'AFC, les n individus et p modalités sont représentés dans les mêmes plans factoriels.
- De même que pour l'AFC, les **contributions** et **cosinus carrés** des modalités permettent d'**expliquer** les axes factoriels.
- L'interprétation du nuage de points-individus est similaire à l'ACP.
- Les **variables** sont représentées dans le plan factoriel par les **centres de gravité** des modalités correspondantes.

Exemple : représentation dans le premier plan factoriel





Les modalités "er = Unknown" et "pr = Unknown" expliquent le plus l'axe 1, et y sont bien représentées. L'axe 2 est expliqué par les modalités "er = Positive" et "pr = Positive", et y sont relativement bien représentées.

ACM : Exemple sous R

Toujours avec le package *FactoMineR* : fonction `MCA`.

- `donneescat = na.omit(churnFR[,c(1,2,14)])`
  `churnFR.acm = MCA(donneescat, level.ventil = 0.02)`
  `level.ventil` = niveau de ventilation pour les modalités rares.
  Valeurs par défaut pour `ncp` et `graph` = celles de la fonction `PCA`.
  Sorties similaires à la fonction `PCA`.
- `churnFR.acm$eig` :

```
> churnFR.acm$eig
      eigenvalue percentage of variance cumulative percentage of variance
dim 1 4.243955e-01           4.243955e+01                        42.43955
dim 2 3.362956e-01           3.362956e+01                        76.06911
dim 3 2.393089e-01           2.393089e+01                       100.00000
```

ACM : procédure Corresp

Principe :

- ACM à partir du **tableau disjonctif complet** (garde les individus en lignes).
- Nombre maximum de facteurs principaux = *nombre total de modalités - nombre de variables*.

```
proc corresp data = churnFR outc = churnFR_acm binary dimens = 3 short;
tables cont_int cont_mv parti_c;
run;
```

- **Binary** : effectue l'ACM a partir du tableau disjonctif complet (`MCA` pour tableau de Burt).
- **Dimens** : nombre de dimensions (= facteurs principaux) demandées (défaut = 2).
- **Short** : n'affiche que la liste courte des statistiques (`all` pour tout avoir).

Sortie de la procédure `Corresp`

The CORRESP Procedure

Décomposition de l'inertie et du Khi 2					
Valeur singulière	Inertie principale	Khi 2	Pourcentage	Pourcent. cumulé	8 16 24 32 40 ----+----+----+----+----+---
0.65146	0.42440	4564.4	42.44	42.44	***************************
0.57991	0.33630	3616.9	33.63	76.07	*********************
0.48919	0.23931	2573.8	23.93	100.00	***************
Total	1.00000	10755.0	100.00		
Degrés de liberté = 17920					

Sortie de la procédure `Corresp`

Eléments de la table `OUTC` :

	TYPE	_NAME_	Quality	Mass	Inertia	Dim1	Dim2	Dim3	Contr1	Contr2	Contr3	
1	INERTIA		.	.	1	.	.	.	0.4243955137	0.3362955895	0.2393088967	
2	OBS	1	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.
3	OBS	2	1	0.00027894	0.0002790576	-0.538832452	0.8065851738	0.2439295222	0.0001908304	0.0005396226	0.0000693555	0.
4	OBS	3	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.
5	OBS	4	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.
6	OBS	5	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.
7	OBS	6	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.
8	OBS	7	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.
9	OBS	8	1	0.00027894	0.0005957352	0.9700353346	-0.437560382	1.0016401158	0.0006184651	0.0001588055	0.0011694332	0.
10	OBS	9	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.
11	OBS	10	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.
12	OBS	11	1	0.00027894	0.0002790576	-0.538832452	0.8065851738	0.2439295222	0.0001908304	0.0005396226	0.0000693555	0.
13	OBS	12	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.
14	OBS	13	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.
15	OBS	14	1	0.00027894	0.0009257816	1.0773787296	0.3910934756	-1.416060714	0.0007629162	0.0001268676	0.0023373069	0.
16	OBS	15	1	0.00027894	0.0002790576	-0.538832452	0.8065851738	0.2439295222	0.0001908304	0.0005396226	0.0000693555	0.
17	OBS	16	1	0.00027894	0.0002790576	-0.538832452	0.8065851738	0.2439295222	0.0001908304	0.0005396226	0.0000693555	0.
18	OBS	17	1	0.00027894	0.0002790576	-0.538832452	0.8065851738	0.2439295222	0.0001908304	0.0005396226	0.0000693555	0.
19	OBS	18	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.
20	OBS	19	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.
21	OBS	20	1	0.00027894	0.0000596383	-0.200777374	-0.388338465	-0.150614787	0.0000264953	0.0001250865	0.0000264416	0.

Analyse factorielle : Synthèse

- Déterminer le nombre d'axes à retenir afin de faciliter l'étude.
- Décrire les axes retenus à l'aide du cercle des corrélations (uniquement pour l'ACP) et/ou des contributions et cosinus carrés.
- Visualiser les données sur les 1 à 3 premiers plans factoriels suivant le nombre d'axes retenus.
- Détecter et traiter les individus, variables ou modalités atypiques.
- Décrire et synthétiser les résultats.