Chapitre 1 : Analyse factorielle

L'analyse factorielle permet de fournir des représentations synthétiques de grands tableaux de données sous forme de visualisations graphiques. Plus généralement, le but de l'analyse factorielle est de réduire les dimensions d'un tableau de données, en visualisant les liens entre les individus et entre les variables dans des espaces de dimensions plus faibles. Selon la nature des données et leur codage, on distingue différentes méthodes d'analyse factorielle :
- l' Analyse en Composantes Principales (ACP),
- l'Analyse Factorielle des Correspondances (AFC),
- l'Analyse des Correspondances Multiples (ACM).

1 Analyse en Composantes Principales

1.1 Objectifs

On dispose d'un tableau rectangulaire, appelé tableau individus×variables, dont les colonnes représentent des variables quantitatives $Y_1, \cdots, Y_p$, $p \geq 1$, et les lignes représentent des individus sur lesquels ces variables sont mesurées. Pour analyser un tel tableau, l'ACP se fixe comme objectifs de :

- Visualiser les ressemblances entre les individus afin de construire une typologie des individus ;
- Visualiser les corrélations entre les variables afin de construire une typologie des variables ;
- Etablir les liens entre ces deux typologies
- Construire des variables synthétiques appelées *composantes principales*, qui sont des combinaisons linéaires des variables initiales $Y_1, \cdots, Y_p$ centrées et sont non corrélées entre elles.

Deux approches se distinguent pour réaliser une ACP : l'approche géométrique de Pearson(1901) basée sur le critère d'inertie et l'approche de Hotelling (1933) basée sur le critère de corrélation ou de variance.

1.2 Fondements de la méthode

Nous présentons ici l'approche basée sur le critère d'inertie. On supposera les données centrées et réduites ; ce qui correspond à l'ACP normée. Si les variables sont seulement centrées, mais non réduites, on parlera d'ACP non-normée.
On considère le tableau X suivant des données centrées et réduites

$$X = \begin{bmatrix} & \vdots & \\ \cdots & x_{ij} & \cdots \\ & \vdots & \end{bmatrix}, \qquad \text{avec} \quad x_{ij} = \frac{y_{ij} - \bar{Y}_j}{\sigma_j}$$

où pour $j = 1, \cdots, p$ $\quad \bar{Y}_j$ est la moyenne de la variable Y_j et $\sigma_j^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_{ij} - \bar{Y})^2$ est sa variance

et pour $i = 1, \cdots, n$ $\qquad x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^p.$

Notion d'inertie. Soit $x_1, \cdots, x_n$ un ensemble de points appartenant à un espace métrique et munis respectivement des masses (poids) $p_1, \cdots, p_n$ et ω un point de cet espace ; on appelle inertie par rapport à ω la quantité

$$I_\omega = \sum_{i=1}^{n} d^2(\omega, x_i).$$

Nuage des individus et Nuage des variables

Le tableau de données X est composé de n individus et p variables quantitatives. Chaque individu correspond à une ligne i du tableau et peut être représenté par un point $x_i \in \mathbb{R}^p$. On munit tous les individus d'un même poids égal à $\frac{1}{n}$. L'ensemble des points x_i forme le nuage des individus, noté

$$N_I = \left\{ \left(x_i, \frac{1}{n} \right) : i = 1, \cdots, n \right\}.$$

Ce nuage est centré sur l'origine O de $\mathbb{R}^p$. La distance au carré entre deux points-individus x_i et $x_{i'}$ est définie par la distance euclidienne usuelle

$$d^2(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2.$$

De même chaque variable correspond à une colonne j du tableau X, et peut être représentée par un vecteur de $x^j = \begin{pmatrix} x_1^j \\ \vdots \\ x_n^j \end{pmatrix} \in \mathbb{R}^n$. On donne la même importance à toutes les variables en les munissant d'un même poids égal à 1, par exemple. L'ensemble des extrémités de ces vecteurs, notées aussi x^j, forme le nuage des variables

$$N_V = \left\{ (x^j, 1) : j = 1, \cdots, p \right\}.$$

La distance au carré entre deux vecteurs-variables x^j et $x^{j'}$ est définie par la distance euclidienne usuelle, à un coefficient $1/n$ près,

$$d^2(x^j, x^{j'}) = \frac{1}{n} \sum_{i=1}^{n} (x_i^j - x_i^{j'})^2.$$

Avec cette distance, les points représentant les variables sont toutes sur la sphère de rayon 1 et que le cosinus de l'angle formé par deux vecteurs x^j et $x^{j'}$ est égal au coefficient de corrélation des variables correspondantes Y_j et $Y_{j'}$, i.e.

$$\cos(Ox^j, Ox^{j'}) = \mathrm{corr}(Y_j, Y_{j'}).$$

Définition 1 *On appelle inertie totale du nuage N_I par rapport à l'origine O la quantité*

$$I_O = \frac{1}{n} \sum_{i=1}^{n} d^2(O, x_i).$$

L'inertie totale du nuage N_V se définit de la même manière. On peut montrer que les deux nuages ont la même inertie totale, qui est égale à p (nombre de variables) du fait que les données sont centrées et réduites.

Réaliser l'ACP revient à ajuster (projeter) le nuage des individus d'une part, et le nuage des variables d'autre part, sur des systèmes d'axes orthogonaux deux à deux de sorte que l'inertie des nuages projetés sur ces axes soit maximale. L'inertie totale I_O sera ainsi décomposée suivant ces deux suites d'axes orthogonaux.

Ajustement ou analyse du nuage des individus

Soit u_1 un vecteur unitaire de $\mathbb{R}^p$ engendrant un axe Δ_1 sur lequel la projection du nuage des individus N_I maximise l'inertie des points projetés par rapport à O. On peut montrer que cette inertie projetée est égale à la plus grande valeur propre, notée λ_1, de la matrice $V = X'X$ ou " $'$ " désigne la transposée et que u_1 est un vecteur propre de V. Δ_1 est appelée premier axe factoriel.

On choisit un deuxième vecteur unitaire u_2 orthogonal à u_1 et engendrant un axe Δ_2 sur lequel la projection du nuage des individus N_I maximise l'inertie projetée. Celle-ci est égale à la deuxième plus grande valeur propre, notée λ_2, de la matrice V. Δ_2 est appelé deuxième axe factoriel. Le plan (Δ_1, Δ_2) est appelé premier plan factoriel.

Ainsi, de proche en proche, on construit l'axe de rang α, Δ_α, qui maximise l'inertie du nuage projeté des individus, et qui est orthogonal aux axes déjà construits. Δ_α sera engendré par un vecteur propre unitaire de

V, associé à la α^{ieme} valeur propre, λ_α, de V.
Le nombre maximal d'axes qu'on peut construire est égal à $r = rang(X)$.

Facteur prinicipal. On appelle facteur principal de rang α du nuage des individus le vecteur c_α de $\mathbb{R}^n$, dont la i^{eme} composante $c_\alpha(i)$ est la coordonnée de l'individu i sur l'axe factoriel Δ_α (engendré par le vecteur unitaire $u_\alpha \in \mathbb{R}^p$).
On appelle α^{ieme} composante principale la variable F_α qui, à chaque individu i, associe sa coordonnée sur l'axe Δ_α notée $F_\alpha(i) = c_\alpha(i)$.
Les plans factoriels fournissent des représentations planes approchées du nuage initial N_I, sur lesquelles on peut visualiser les ressemblances entre les individus. L'analyse se restreint souvent sur le premier plan factoriels, mais d'autres plans factoriels peuvent aussi être considérées s'ils présentent un pourcentage d'inertie assez important.

Ajustement ou analyse du nuage des variables
Pour visualiser les corrélations entre les variables, l'ACP applique la même démarche que précédemment au nuage des variables N_V. C'est à dire on projette successivement le nuage des variables N_V sur un système d'axes orthogonaux deux à deux de sorte que l'inertie projetée soit maximale. On obtient ainsi une suite d'axes factoriels orthogonaux deux à deux : $D_1, D_2, \cdots, D_r$, avec $r = rang(X)$. Chaque axe D_α est engendré par un vecteur unitaire $v_\alpha \in \mathbb{R}^n$, qui est vecteur propre de la matrice $W = XX'$, associé à un valeur propre μ_α. Ces axes sont associés à de nouvelles variables synthétiques, appelées composantes principales et qui résument au mieux les variables initiales.
Le facteur principal de rang α du nuage N_V est défini par le vecteur d_α de $\mathbb{R}^p$ dont les composantes $d_\alpha(j), j = 1, \cdots, p$ sont les coordonnées des p variables sur l'axe factoriel D_α.
On appelle α^{ieme} composante principale du nuage N_V la variable G_α qui, à chaque variable j, associe sa coordonnée sur l'axe D_α, notée $G_\alpha(j) = d_\alpha(j)$.
Dualité entre les deux analyses
Les nuages N_I et N_V sont des représentations d'un même tableau de données à travers ses lignes et à travers ses colonnes. Il existe donc des relations de dualité entre ces deux nuages. Les inerties projetées des deux nuages sur les axes factoriels de même rang α sont égales ; i.e. $\lambda_\alpha = \mu_\alpha$ valeur λ_α. On a les relations de transition suivantes qui permettent de passer d'un nuage à l'autre :

$$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} c_\alpha \qquad \text{et} \qquad u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} d_\alpha.$$

Cela signifie que les vecteurs propres de l'un des nuages peuvent être obtenus à partir des facteurs principaux de l'autre nuage, à un facteur $\frac{1}{\sqrt{\lambda_\alpha}}$ près. Ces relations permettent d'interpréter les deux nuages en même temps.

Conséquence. La coordonnée d'une variable Y_j sur l'axe factoriel Δ_α est égale au coefficient de corrélation entre cette variable et la composante principale F_α du nuage des individus, i.e.

$$G_\alpha(j) = \text{corr}(Y_j, F_\alpha).$$

1.3 Aides à l'interprétation

Pour pouvoir interpréter les proximités de deux points (ou éléments) d'un nuage sur un plan factoriel, il est nécessaire de mesurer la qualité de représentation globale du nuage, mais aussi la qualité de représentation de chaque point. **Qualité de représentation d'un nuage par un axe**
Elle est mesurée par le rapport :

$$\tau_\alpha = \frac{\text{Inertie projetée du nuage sur l'axe } \Delta_\alpha}{\text{Inertie totale du nuage}} = \frac{\lambda_\alpha}{I_O}.$$

τ_α est la part d'inertie expliquée par l'axe Δ_α, encore appelée pourcentage d'inertie. Le pourcentage d'inertie expliqué par le plan factoriel $(\Delta_\alpha, \Delta_\beta)$ est la somme des pourcentages d'inertie de ces deux axes

$$\tau_\alpha + \tau_\beta = \frac{\lambda_\alpha + \lambda_\beta}{I_O}.$$

Qualité de représentation d'un élément par un axe
Elle est mesurée par le cosinus carré de l'angle formé par cet élément (x_i) et l'axe (Δ_α). On note

$$cos^2_\alpha(i) = \frac{F^2_\alpha(i)}{||x_i||^2} = cos^2(Ox_i, u_\alpha).$$

$cos^2_\alpha(i)$ représente la part d'inertie du point x_i expliquée par l'axe Δ_α.
La qualité de représentation d'un élément sur un plan factoriel $(\Delta_\alpha, \Delta_\beta)$ est la somme des cosinus carrés des angles formés par cet élément les deux axes Δ_α et Δ_β. On note

$$QLT_2(i) = cos^2_\alpha(i) + cos^2_\beta(i).$$

Contribution d'un élément à l'inertie d'un axe
Elle est mesurée par le rapport

$$CTR_\alpha(i) = \frac{\text{Inertie projetée de l'élément i sur l'axe } \Delta_\alpha}{\text{Inertie totale projetée sur l'axe } \Delta_\alpha} = \frac{p_i F_\alpha(i)}{\lambda_\alpha}.$$

Cet indicateur mesure le rôle joué par un élément sur le positionnement d'un axe. On s'intéressera en priorité aux éléments à forte contribution dans les interprétations.
Point supplémentaire ou élément illustratif
C'est un point n'appartenant pas au nuage analysé, mais que l'on peut représenter sur les plans factoriels pour enrichir l'interprétation. Il ne contribue pas au positionnement des axes.

Application. Données sur le budget de consommation des ménages

2 Analyse Factorielle des Correspondances (AFC)

2.1 Objectifs

L' AFC s'applique sur un tableau de contingence. Son objectif fondamental est d'analyser la liaison entre deux variables qualitatives ou nominales ; c'est à dire d'étudier l'écart à l'indépendance entre ces deux variables représentées par un tableau de contingence. Pour cela, elle cherche à :

- comparer les profils lignes entre eux ;
- comparer les profils colonnes entre eux ;
- mettre en évidence des associations entre les modalités lignes et les modalités colonnes.

2.2 Fondements de la méthode

On observe deux variables qualitatives ou nominales X et Y dans une population composée de k individus. Le tableau de contingence est représenté par le tableau des fréquences conjointes données par

$$f_{ij} = k_{ij}/k, \quad \text{où} \quad k_{ij} \text{ est l'effectif conjoint de la modalité ligne } i \text{ et de la modalité colonne } j.$$

$f_{i.} = \sum_{j=1}^{p} f_{ij}$ est la fréquence marginale de la modalité ligne i ;
$f_{.j} = \sum_{i=1}^{n} f_{ij}$ est la fréquence marginale de la modalité colonne j.

f_{11}	...	f_{1j}	...	f_{1p}	$f_{1.}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
f_{i1}	...	f_{ij}	...	f_{ip}	$f_{i.}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
f_{n1}	...	f_{nj}	...	f_{np}	$f_{n.}$
$f_{.1}$	...	$f_{.j}$	...	$f_{.p}$	1

Soit $I = \{1, \cdots, i \cdots, n\}$ l'ensemble des lignes et $J = \{1, \cdots, j, \cdots, p\}$ l'ensemble des colonnes.

On appelle **profil ligne** i la répartition en pourcentage, suivant les modalités colonnes, des individus présentant la modalité i. Pour i fixé, on note

$$f_j^i = \frac{f_{ij}}{f_{i.}}$$

Les $f_j^i,\ j = 1, \cdots, p$ représentent la loi de probabilité conditionnelle de Y sachant $(X = i)$. Le profil ligne i est donc le vecteur de $\mathbb{R}^p$

$$f_J^i = (f_1^i, \cdots, f_j^i, \cdots, f_p^i)'.$$

De même on appelle **profil colonne** j la répartition en pourcentage, suivant les modalités lignes, des individus présentant la modalité j. Pour tout j fixé, on note

$$f_i^j = \frac{f_{ij}}{f_{.j}}$$

Les $f_i^j,\ i = 1, \cdots, n$ représentent la loi de probabilité conditionnelle de X sachant $(Y = j)$. Le profil colonne j est donc le vecteur de $\mathbb{R}^n$

$$f_I^j = (f_1^j, \cdots, f_i^j, \cdots, f_n^j)'.$$

Ces profils lignes et colonnes sont représentés par les tableaux suivants :

Tableau des profils lignes

f_1^1	...	f_j^1	...	f_p^1	1
$\vdots$		$\vdots$		$\vdots$	$\vdots$
f_1^i	...	f_j^i	...	f_p^i	1
$\vdots$		$\vdots$		$\vdots$	$\vdots$
f_1^n	...	f_j^n	...	f_p^n	1
$f_{.1}$	...	$f_{.j}$	...	$f_{.p}$	1

Tableau des profils colonnes

f_1^1	...	f_1^j	...	f_1^p	$f_{1.}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
f_i^1	...	f_i^j	...	f_i^p	$f_{i.}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
f_n^1	...	f_n^j	...	f_n^p	$f_{n.}$
1	...	1	...	1	1

On appelle **profil ligne moyen** la distribution marginale des modalités colonnes : c'est la distribution marginale de la variable Y. Il correspond au centre de gravité de l'ensemble des profils lignes $f_J^i,\ i = 1, \cdots, n$.

On appelle **profil colonne moyen** la distribution marginale des modalités lignes : c'est la distribution marginale de la variable X. Il correspond au centre de gravité de l'ensemble des profils colonnes $f_I^j,\ j =$

$1, \ldots, p$.

Analyser la liaison entre les deux variables qualitatives X et Y revient à examiner, d'une part la proximité entre les profils lignes et la proximité entre les profils colonnes et, d'autre part la proximité de chaque profil (ligne ou colonne) par rapport à son profil moyen correspondant. Autrement dit, il s'agit d'étudier l'écart à l'indépendance des deux variables qualitatives. Pour ce faire, on considérera l'ensemble des profils lignes comme un nuage de n points dans l'espace des p colonnes du tableau de contingence, et l'ensemble des profils colonnes comme un nuage de p points dans l'espace des n lignes du tableau de contingence. On appliquera ensuite l'ACP à chacun de ces nuages pour visualiser les proximités entre les profils.
La proximité entre deux profils lignes f_J^i et $f_J^{i'}$ est mesurée par la distance du khi-deux, définie par

$$d^2(f_J^i, f_J^{i'}) = \sum_{j=1}^{p} \frac{1}{f_{.j}} (f_j^i - f_j^{i'})^2$$

et la proximité entre deux profils colonnes f_I^j et $f_I^{j'}$ est mesurée par la distance du khi-deux, définie par

$$d^2(f_I^j, f_I^{j'}) = \sum_{i=1}^{n} \frac{1}{f_{i.}} (f_i^j - f_i^{j'})^2.$$

Inertie des deux nuages
Le nuage des profils lignes est défini par

$$N_I = \{(f_J^i, f_{i.}) : i = 1, \cdots, n\}$$

Son centre de gravité g_I est le profil ligne moyen. Il sert de référence aux autres profils lignes. On a

$$g_I = \sum_{i=1}^{n} f_{i.} f_J^i \in \mathbb{R}^p.$$

Sa j^{eme} composante est donnée par :

$$\sum_{i=1}^{n} f_{i.} f_j^i = \sum_{i=1}^{n} f_{i.} \frac{f_{ij}}{f_{i.}} = f_{.j},$$

i.e. $g_I = (f_{.1}, \cdots, f_{.j}, \cdots, f_{.p})'$.

L'inertie totale du nuage N_I par rapport à son centre de gravité g_I est donnée par

$$\text{Inertie}(N_I) = \sum_{i=1}^{n} f_{i.} d^2(f_J^i, g_I) = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{f_{i.}}{f_{.j}} \left(\frac{f_{ij}}{f_{.j}} - f_{.j} \right)^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}.$$

De manière analogue, on définit le nuage des profils colonnes par

$$N_J = \{(f_I^j, f_{.j}) : j = 1, \cdots, p\}.$$

Son centre de gravité g_J est le profil colonne moyen. Il sert de référence aux autres profils colonnes. On a

$$g_J = \sum_{j=1}^{p} f_{.j} f_I^j \in \mathbb{R}^n.$$

Sa i^{eme} composante est donnée par :

$$\sum_{j=1}^{p} f_{.j} f_I^j = \sum_{j=1}^{p} f_{i.} \frac{f_{ij}}{f_{i.}} = f_{i.},$$

i.e. $g_J = (f_{1.}, \cdots, f_{i.}, \cdots, f_{n.})'$.

L'inertie totale du nuage N_J par rapport à g_J est donnée par

$$\text{Inertie}(N_J) = \sum_{j=1}^{p} f_{.j} d^2(f_I^j, g_J).$$

On peut montrer sans difficulté que les deux nuages ont la même inertie totale

$$\textbf{Inertie}(N_I) = \textbf{Inertie}(N_J).$$

On peut aussi vérifier que l'inertie totale de chaque nuage, par rapport à son centre de gravité, est égale à χ^2/k, où χ^2 est la statistique du test de Khi-deux donnée par

$$\chi^2 = k \times \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}.$$

Cette inertie peut donc être interprétée comme une mesure de l'intensité de la liaison entre les deux variables. Lorsqu'elle est nulle, tous les profils lignes sont égaux au profil ligne moyen (*idem* pour les profils colonnes) ; i.e. il y a indépendance parfaite entre les variables.

Relations de transition

Soit F_α la composante principale de rang α obtenue dans l'analyse du nuage des profils lignes et G_α celle obtenue pour l'analyse du nuage des profils colonnes. Les quantités d'inertie associées à ces deux composantes principales sont égales, et coincident avec la α^{ieme} valeur propre, λ_α, de la matrice d'inertie analysée par ACP. On a les formules suivantes :

$$F_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}} G_\alpha(j) \tag{1}$$

et

$$G_\alpha(j) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}} F_\alpha(i), \tag{2}$$

où $F_\alpha(i)$ est la coordonnée du point représentant le profil ligne i sur l'axe factoriel de rang α du nuage des profils lignes, et $G_\alpha(j)$ est la coordonnée du point représentant le profil colonne j sur l'axe factoriel du rang α du nuage des profils colonnes. Les relations **pseudo-barycentriques** (1) et (2) dites de *transition* permettent de passer des résultats d'une analyse à l'autre et donnent un sens à une représentation simultanée des lignes et des colonnes en superposant les cartes (plans) factorielles (F_1, F_2) et (G_1, G_2).

Remarques

- On peut aussi obtenir des relations barycentriques moyennant une dilatation des coordonnées $F_\alpha(i)$ et $G_\alpha(j)$ par le facteur $\sqrt{\lambda_\alpha}$ sur chaque axe factoriel. Ainsi, chaque modalité ligne i sera positionnée au barycentre des modalités colonnes j affectées des poids $\frac{f_{ij}}{f_{i.}}$ et vice-versa. Puisqu'un barycentre est plus proche des points qui ont un plus grand poids, alors une modalité ligne i sera d'autant plus attirée par une modalité colonne j que la valeur du poids $\frac{f_{ij}}{f_{i.}}$ est élevée.

- Cette double représentation barycentrique est spécifique à l'AFC du fait des rôles symétriques joués par les lignes et les colonnes.

- Sur le graphique de la représentation simultanée, la position relative de deux points représentant deux modalités d'une même variable (ligne ou colonne) s'interprète en tant que distance, alors que la position d'un point représentant une modalité d'une variable par rapport aux positions de tous les points représentant les modalités de l'autre variable s'interprète en tant que barycentre. La proximité entre deux points

représentant chacun, une modalité de l'une des deux variables peut s'interprèter comme une association, lorsque ces deux points sont situés à la périphérie du nuage. Toute association entre une modalité ligne et une modalité colonne suggérée sur le graphique doit être vérifiée sur le tableau des données initial.

- En ACP les points du nuage ont souvent le même poids, ce qui n'est pas le cas en AFC.
- En AFC l'origine des axes factoriels est le centre de gravité du nuage des profils lignes (ou colonnes), i.e. le profil moyen.
- En AFC la valeur propre associée à un axe est inférieure à 1, ($\lambda_\alpha \leq 1$). Ce qui n'est pas le cas en ACP.

Indice d'attraction/répulsion entre une modalité ligne i et une modalité colonne j :
Il est défini par le rapport

$$d_{ij} = \frac{f_{ij}}{f_{i.}f_{.j}}.$$

- Si $d_{ij} > 1$, les modalités i et j s'attirent ; cela signifie que la modalité j est sur-représentée dans la population $[X = i]$ par rapport à l'ensemble de la population, mais aussi que la modalité i est sur-représentée dans la population $[Y = j]$ par rapport à l'ensemble de la population.
- Si $d_{ij} < 1$, les modalités i et j se repousssent ; ce qui implique une sous-représentation des modalités i et j respectivement dans les populations$[Y = j]$ et $[X = i]$ par rapport à l'ensemble de la population.
- Si $d_{ij} = 1, \forall i, j$, les deux variables sont indépendantes.

3 Analyse Factorielle des Correspondances Multiples (ACM)

3.1 Objectifs

On dispose d'un tableau de données individus×variables, qui peuvent être qualitatives ou quantitatives. Pour analyser un tel tableau, l'ACM cherche à :

- mettre en évidence les associations entre les modalités des différentes variables ;
- étudier les liaisons entre les variables ;
- étudier la ressemblance entre les individus pour dégager des profils d'individus, i.e. construire une typologie des individus.

3.2 Fondement de la méthode

L'ACM s'applique généralement sur un tableau de données provenant des résultats d'une enquête. Il faut cependant coder les données pour pouvoir effectuer l'ACM. Les variables numériques sont transformées en variables catégorielles avec des classes.

Codage disjonctif complet : Il conduit à un tableau appelé tableau disjonctif complet (TDC) défini par la matrice X suivante. On considère m variables $Y_1, \cdots Y_m$. Chaque variable Y_j possède p_j modalités de sorte

que le nombre total de modalités est $p = \sum_{j=1}^{m} p_j$.

$X = TDC =$

	Y_1 ... Y_j ... Y_m					
	$1.....p_1$		$1..k..p_j$		$1.....p_m$	$Marge$
1						m
	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i	0100	...	x_{ijk}	...	0010	m
	$\vdots$		$\vdots$		$\vdots$	$\vdots$
n						m
$Marge$	$n_{11}....n_{1p_1}$	...	$n_{j1}..n_{jk}..n_{jp_j}$	...	$n_{m1}....n_{mp_m}$	$n \times m$

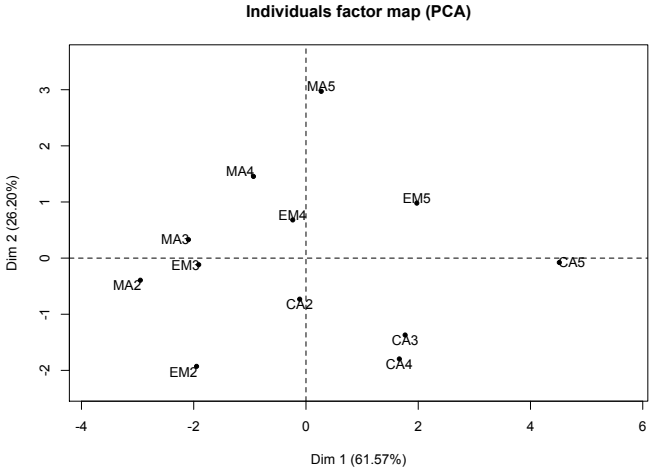
$x_{ijk} = 1$ ou 0 selon que l'individu i prend la modalité k de la variable j ou non.

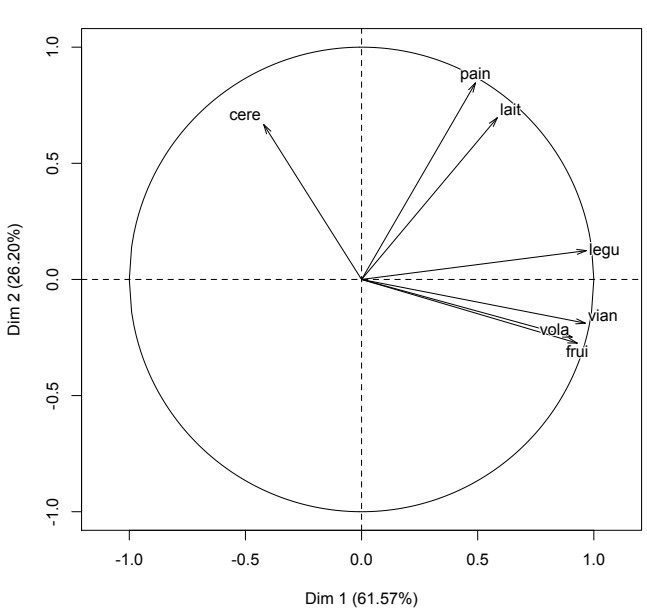
Remarque. Pour réaliser l'ACM, on peut aussi utiliser un tableau de Burt qui est défini par

$$Y = X'X.$$

Y est une matrice carrée symétrique d'ordre p, croisant toutes les modalités entre elles. L'élément y_{ij} est le nombre d'individus possédant à la fois la modalité ligne i et la modalité colonne j. Ainsi, le tableau de Burt est une juxtaposition de tableaux de contingence des variables prises deux à deux. Donc réaliser l'ACM revient à réaliser l'AFC sur le tableau de Burt ou le TDC.

Pour mettre en évidence des associations entre les modalités des différentes variables, on pourra construire, à partir du tableau disjonctif complet (TDC), un nuage de profils lignes et un nuage de profils colonnes, puis ajuster ces nuages sur des plans factoriels afin de mieux visualiser les distances entre ces modalités.
Comme en ACP et en AFC, on se servira des aides à l'interprétation : qualités de représentation, contribution, cosinus carré, qui sont définies ici pour les modalités. La contribution d'une variable Y_j est la somme des contributions de ses modalités. Sa liaison avec la composante principale F_α est mesurée par le rapport de corrélation $\eta^2(F_\alpha, Y_j)$.
Dans l'interprétation des proximités des modalités, on peut avoir deux cas de figure:
- Si deux modalités d'une même variable sont proches, cela voudrait dire que les individus possédant l'une des modalités et ceux possédant l'autre modalité sont globalement les même du point de vue des autres variables.
- Si deux modalités de deux variables différentes sont proches, cela peut signifier que ce sont globalement les même individus qui possédent l'une et l'autre modalités.





Summary(PCA)
Eigenvalues

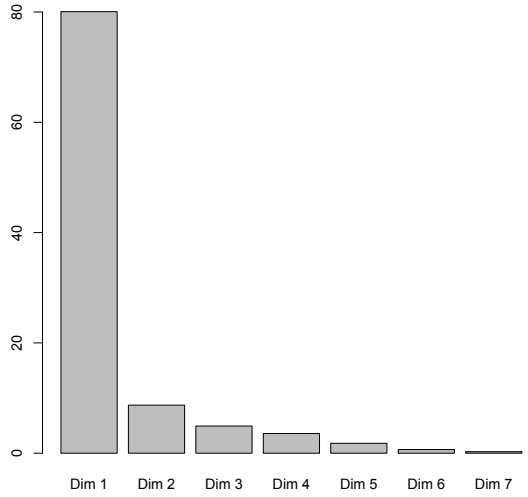
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
Variance	4.310	1.834	0.639	0.126	0.061	0.028	0.003
% of var.	61.571	26.196	9.126	1.805	0.866	0.400	0.037
Cumulative % of var.	61.571	87.766	96.892	98.697	99.563	99.963	100.000

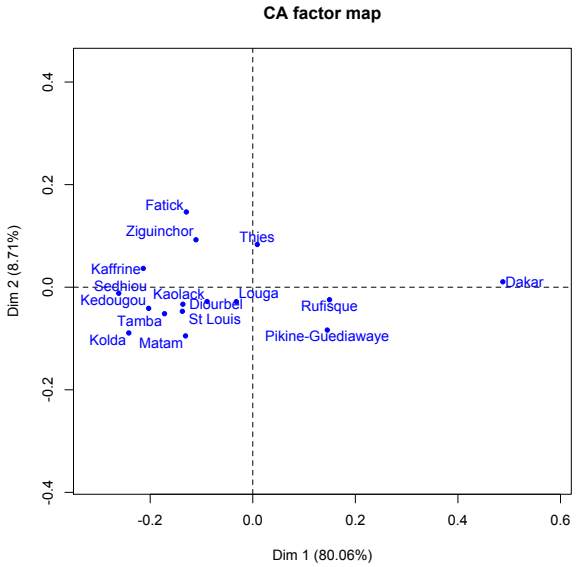
Individuals (the 10 first)

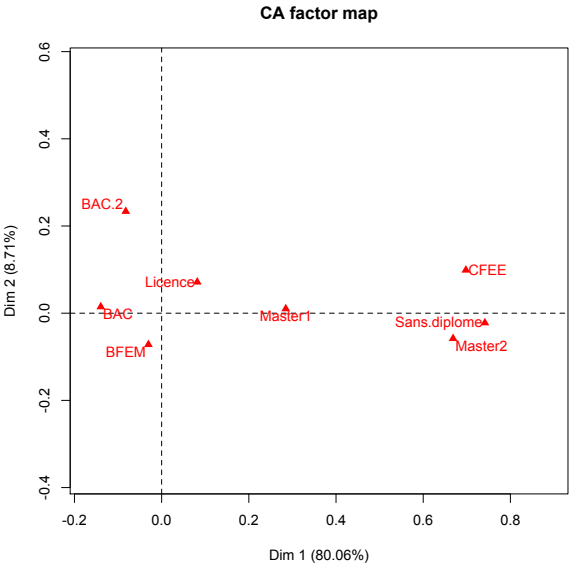
	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
MA2	3.047	-2.951	16.837	0.938	-0.394	0.705	0.017	0.452	2.661	0.022
EM2	3.039	-1.950	7.348	0.411	-1.928	16.897	0.403	-1.286	21.575	0.179
CA2	1.757	-0.112	0.024	0.004	-0.731	2.431	0.173	1.487	28.833	0.716
MA3	2.126	-2.094	8.477	0.970	0.329	0.493	0.024	-0.091	0.109	0.002
EM3	2.069	-1.911	7.062	0.853	-0.117	0.063	0.003	-0.659	5.673	0.102
CA3	2.535	1.766	6.031	0.485	-1.370	8.530	0.292	1.078	15.155	0.181
MA4	1.790	-0.936	1.693	0.273	1.455	9.623	0.661	-0.276	0.996	0.024
EM4	0.847	-0.236	0.108	0.078	0.680	2.102	0.645	0.274	0.981	0.105
CA4	2.537	1.663	5.348	0.430	-1.796	14.663	0.501	0.147	0.281	0.003
MA5	3.053	0.272	0.143	0.008	2.971	40.108	0.947	0.562	4.118	0.034

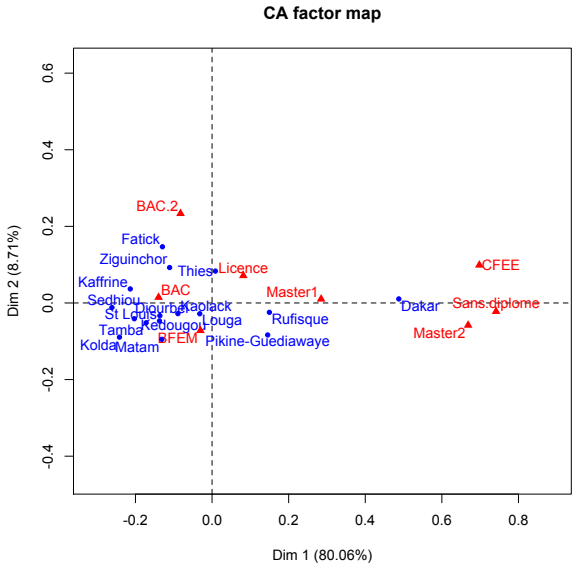
Variables

	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
pain	0.491	5.585	0.241	0.846	39.066	0.716	0.007	0.008	0.000
legu	0.968	21.722	0.936	0.124	0.838	0.015	-0.064	0.635	0.004
frui	0.929	20.019	0.863	-0.274	4.104	0.075	0.110	1.881	0.012
vian	0.964	21.543	0.928	-0.188	1.931	0.035	0.164	4.195	0.027
vola	0.907	19.081	0.822	-0.248	3.363	0.062	0.301	14.211	0.091
lait	0.585	7.931	0.342	0.696	26.448	0.485	-0.379	22.444	0.143


80
60
40
20
0
Dim 1
Dim 2
Dim 3
Dim 4
Dim 5
Dim 6
Dim 7


CA factor map
Dim 2 (8.71%)
Dim 1 (80.06%)
Fatick
Ziguinchor
Thies
Kaffrine
Sedhiou
Kedougou
Kaolack
Diourbel
Louga
Tamba
St Louis
Rufisque
Kolda
Matam
Pikine-Guediawaye
Dakar


CA factor map
Dim 2 (8.71%)
Dim 1 (80.06%)
BAC.2
Licence
BAC
BFEM
Master1
CFEE
Sans.diplome
Master2



Call:
CA(X = tab)

The chi square of independence between the two variables is equal to 6268.504 (p-value = 0).

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
Variance	0.045	0.005	0.003	0.002	0.001	0.000	0.000
% of var.	80.065	8.714	4.928	3.571	1.793	0.657	0.272
Cumulative % of var.	80.065	88.778	93.706	97.277	99.071	99.728	100.000

Rows (the 10 first)

	Iner*1000	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
Dakar	28.586	0.488	62.409	0.985	0.011	0.267	0.000	-0.049	10.408	0.010
Diourbel	1.189	-0.136	2.259	0.857	-0.033	1.230	0.051	-0.034	2.342	0.055
Fatick	2.887	-0.130	2.376	0.371	0.147	27.996	0.476	0.007	0.109	0.001
Kaffrine	1.477	-0.214	2.575	0.787	0.037	0.697	0.023	-0.070	4.494	0.085
Kaolack	0.954	-0.089	1.320	0.624	-0.028	1.187	0.061	0.004	0.037	0.001
Kedougou	0.793	-0.203	1.399	0.797	-0.041	0.530	0.033	-0.044	1.045	0.037
Kolda	3.192	-0.242	5.818	0.823	-0.089	7.283	0.112	-0.059	5.620	0.049
Louga	0.997	-0.032	0.133	0.060	-0.028	0.947	0.047	-0.092	17.662	0.492
Matam	1.040	-0.131	1.306	0.567	-0.095	6.291	0.297	-0.017	0.375	0.010
Pikine- Gdye	3.944	0.145	5.282	0.604	-0.084	16.065	0.200	0.068	19.025	0.134

Columns

	Iner*1000	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
BAC	8.873	-0.140	17.223	0.876	0.015	1.752	0.010	-0.046	30.440	0.095
BFEM	2.756	-0.030	0.681	0.111	-0.072	35.251	0.628	0.035	14.513	0.146
CFEE	7.620	0.698	15.555	0.921	0.099	2.868	0.018	0.009	0.039	0.000
Master2	2.889	0.669	4.044	0.632	-0.058	0.281	0.005	0.204	6.128	0.059
BAC.2	3.409	-0.082	0.669	0.089	0.234	49.454	0.713	0.043	2.907	0.024
Licence	2.231	0.082	1.401	0.283	0.071	9.857	0.217	0.087	26.148	0.326
Master1	6.166	0.285	12.466	0.912	0.010	0.146	0.001	0.004	0.033	0.000
Sans.diplom	22.426	0.742	47.963	0.965	-0.022	0.392	0.001	-0.118	19.793	0.025