Analyse Factorielle des Correspondances (AFC)

L3{*MA*, *Info_SI*, *Info_Reseaux*} - UFR S.A.T

Pr. Ousmane THIARE

othiare@ugb.edu.sn
[www.ousmanethiare.com]

16 avril 2020

Analyse Factorielle des Correspondances (AFC)

Analyse Factorielle des Correspondances (AFC)

Introduction

- L'analyse factorielle des correspondances (AFC), ou analyse des correspondances simples, est une méthode exploratoire d'analyse des tableaux de contingence. Elle a été développée par J.-P. Benzecri durant la période 1970-1990.

Introduction

- L'analyse factorielle des correspondances (AFC), ou analyse des correspondances simples, est une méthode exploratoire d'analyse des tableaux de contingence. Elle a été développée par J.-P. Benzecri durant la période 1970-1990.
- L'AFC considérée comme une ACP particulière dotée de la métrique du $\mathcal{X}^2$ (Khi-2) qui ne dépend que du profil des colonnes du tableau. L'analyse permet, dans le plan des deux premiers axes factoriels, une représentation simultanée des ressemblances entre les colonnes ou les lignes du tableau et de la proximité entre lignes et colonnes.

Introduction

- L'analyse factorielle des correspondances (AFC), ou analyse des correspondances simples, est une méthode exploratoire d'analyse des tableaux de contingence. Elle a été développée par J.-P. Benzecri durant la période 1970-1990.
- L'AFC considérée comme une ACP particulière dotée de la métrique du $\mathcal{X}^2$ (Khi-2) qui ne dépend que du profil des colonnes du tableau. L'analyse permet, dans le plan des deux premiers axes factoriels, une représentation simultanée des ressemblances entre les colonnes ou les lignes du tableau et de la proximité entre lignes et colonnes.
- Etudier sur N individus les "liaisons" entre deux variables X et Y. Chaque variable détermine deux partitions de l'ensemble des individus selon les **modalités**.

- On note souvent I l'ensemble des modalités de la variable X et J celui des modalités de Y.

- On note souvent I l'ensemble des modalités de la variable X et J celui des modalités de Y.
- Le cardinal de I est noté n et celui de J est noté m. Pour chercher les liaisons entre X et Y nous allons croiser les deux partitions pour obtenir un **tableau de contingence** indexé par $I \times J$ (on définit un ordre sur I et J, qui peut être éventuellement arbitraire, afin de pouvoir construire ce tableau).

- On note souvent I l'ensemble des modalités de la variable X et J celui des modalités de Y.
- Le cardinal de I est noté n et celui de J est noté m. Pour chercher les liaisons entre X et Y nous allons croiser les deux partitions pour obtenir un **tableau de contingence** indexé par $I \times J$ (on définit un ordre sur I et J, qui peut être éventuellement arbitraire, afin de pouvoir construire ce tableau).
- Dans la case associée à la i-ème ligne et à la j-ème colonne on écrit l'effectif des individus ayant la i-ème modalité pour la variable X et la j-ème modalité pour la variable Y, celui-ci est noté k_{ij}.

Introduction (suite)

Tableau de contingence complété par ses marges

X \ Y	...	j-ème colonne	...	marge
...	...	...	...	...
i-ème ligne	...	k_{ij}	...	$k_{i.}$
...	...	...	...	...
marge	...	$k_{.j}$	...	N

On pose :

$$k_{.j} = \sum_{i=1}^{n} k_{ij} \text{ et } k_{i.} = \sum_{j=1}^{m} k_{ij}$$

Introduction (suite)

$k_{.j}$ est appelé l'**effectif marginal de la j-ème modalité** de Y,
$k_{i.}$ est appelé l'**effectif marginal de la i-ème modalité** de X.
Les éléments du tableau de contingence divisés par l'effectif total N constituent le **tableau des fréquences** où l'on note f_{ij} l'élément générique. Ce tableau permet de définir deux "marges" : une colonne indexée par i d'élément générique $f_{i.} = \sum_{j \in J} f_{ij}$ et une ligne indexée par j d'élément générique $f_{.j} = \sum_{i \in I} f_{ij}$, ce sont les **fréquences marginales**.

Introduction (suite et fin)

La fréquence $f_{i.}$ (respectivement $f_{.j}$) peut être interprétée comme le **poids de la i-ème modalité** de X, on peut noter celui-ci $p_{i.}$ (respectivement le poids $p_{.j}$ de la j-ème modalité de Y).

Variables qualitatives

Soit $\mathcal{X}$ une variable qualitative. On dispose d'un échantillon de n individus sur lesquels la variable est mesurée.

Modalités (ou catégories) les valeurs que peut prendre une variable qualitative ; si la variable a r modalités (valeurs possibles), on note $x_i, 1 \leq i \leq r$, ces modalités.

Effectif le nombre d 'occurrence de la modalité x_i dans l'échantillon ; on le note n_i, et on a $\sum_{i=1}^{r} n_i = n$.

Fréquence c'est la grandeur $f_i = n_i/n$; la somme des fréquences sur les modalités est 1. On utilise souvent le pourcentage $100 f_i$.

Variables qualitatives (suite et fin)

Représentation on peut utiliser un tableau avec r lignes de la forme

$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i
$\vdots$	$\vdots$	$\vdots$

Tableau de contingence

Soient $\mathcal{X}$ et $\mathcal{Y}$ deux variables qualitatives à r et s modalités respectivement décrivant un ensemble de n individus.

Définition le tableau de contingence est une matrice à r lignes et s colonnes renfermant les effectifs n_{ij} d'individus tels que $\mathcal{X} = x_i$ et $\mathcal{Y} = y_i$

$$\mathbf{N} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1s} \\ n_{21} & n_{22} & & \\ & \ddots & & \\ \vdots & & n_{ij} & \vdots \\ & & \ddots & \\ n_{r1} & & & n_{rs} \end{pmatrix}$$

La constitution de ce tableau est ce que les praticiens des enquêtes appellent un « tri croisé ».

Marges et profils

Marge en ligne C'est la somme $n_{i.} = \sum_{j=1}^{s} n_{ij}$, c'est-à-dire l'effectif total de la modalité x_i de $\mathcal{X}$.

Marge en colonne C'est la somme $n_{.j} = \sum_{i=1}^{r} n_{ij}$, c'est-à-dire l'effectif total de la modalité y_i de $\mathcal{Y}$.

Deux lectures possibles selon la variable que l'on privilégie, on peut définir :

- le tableau des profils-lignes $n_{ij}/n_{i.}$, qui représente la fréquence de la modalité y_j conditionnellement à $\mathcal{X} = x_i$; la somme de chaque ligne est ramenée à 100%.
- le tableau des profils-colonnes $n_{ij}/n_{.j}$, qui représente la fréquence de la modalité x_i conditionnellement à $\mathcal{Y} = y_i$; la somme de chaque colonne est ramenée à 100%.

Propriétés des profils

Moyenne la moyenne des profils-lignes (avec poids correspondant aux effectifs marginaux des lignes) est le profil marginal des colonnes :

$$\sum_{i=1}^{r} \frac{n_{i.}}{n} \times \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$$

et de même pour les colonnes $\sum_{j=1}^{s} \frac{n_{.j}}{n} \times \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$

Propriétés des profils (suite et fin)

Indépendance empirique lorsque tous les profils lignes sont identiques, il y a indépendance entre $\mathcal{X}$ et $\mathcal{Y}$, puisque la connaissance de $\mathcal{X}$ ne change pas la répartition de $\mathcal{Y}$.
On a pour tout j

$$\frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = \cdots = \frac{n_{rj}}{n_{r.}} = \frac{n_{1j} + \cdots + n_{rj}}{n_{1.} + \cdots + n_{r.}} = \frac{n_{.j}}{n}$$

et donc $n_{ij} = \dfrac{n_{i.}n_{.j}}{n}$

Le $\mathcal{X}^2$ d'écart à l'indépendance

Définition c'est la grandeur suivante, aussi notée $\mathcal{X}^2$ ou X^2.

$$d^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} = n\left[\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right]$$

Si les variables sont indépendantes, $d^2 = 0$.

Borne supérieure Comme $n_{ij} \leq n_{i.}$, on a

$$\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i.}n_{.j}} \leq \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}}{n.j} = \sum_{j=1}^{s} \frac{\sum_{i=1}^{r} n_{ij}}{n_{.j}} = \sum_{j=1}^{s} \frac{n_{.j}}{n.j} = s$$

et donc $d^2 \leq n(s-1)$.On fait de même pour r et

$$\varphi^2 = \frac{d^2}{n} \leq min(s-1, r-1)$$

Le $\mathcal{X}^2$ d'écart à l'indépendance (suite et fin)

Dépendance fonctionnelle si $\varphi^2 = s - 1$, alors pour chaque i soit $n_{ij} = ni.$, soit $n_{ij} = 0$: il existe une unique case non nulle par ligne. $\mathcal{Y}$ est fonctionnellement liée à $\mathcal{X}$.
Dépendance inverse cette relation ne signifie pas que $\mathcal{X}$ est fonctionnellement liée à $\mathcal{Y}$, sauf si r=s. On peut alors représenter le tableau comme une matrice diagonale.
Contribution au $\mathcal{X}^2$ c'est le terme

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

qui permet de mettre en évidence les associations significatives entre catégories de deux variables.

Caractère significatif du $\mathcal{X}^2$

Problème à partir de quelle valeur de d^2 doit-on considérer que les variables $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes ?.
Méthode on suppose que $\mathcal{X}$ et $\mathcal{Y}$ sont issus de tirages de deux variables aléatoires indépendantes. On peut alors montrer que d^2 est une réalisation d'une variable aléatoire D^2 qui suit une loi $\mathcal{X}^2_{(r-1)(s-1)}$.

Définition Loi du khi-deux à p degrés de libertés $\mathcal{X}^2_p$ est la loi de la variable $\sum_{i=1}^{p} = U_i^2$, où les U_i sont des variables gaussiennes réduites indépendantes.

Le test du $\mathcal{X}^2$ on se fixe un risque d'erreur α (0.01 ou 0.05 en général) et on calcule la valeur d_c^2 telle que $P(\mathcal{X}^2_{(r-1)(s-1)} > d_c^2) = \alpha$. Si $d^2 > d_c^2$ on considère que l'événement est très improbable et que donc que l'hypothèse originale d'indépendance doit être rejetée. On trouvera en général ces valeurs dans une table précalculée.

Analyse des correspondances de deux variables : les données

Effectifs on a un tableau de contingence N à m_1 lignes et m_2 colonnes résultant du croisement de deux variables qualitatives $\mathcal{X}_1$ et $\mathcal{X}_2$ à m_1 et m_2 modalités respectivement. On note D_1 et D_2 les matrices diagonales des effectifs marginaux.

$$\mathbf{D}_1 = \begin{pmatrix} n_{1.} & & 0 \\ & n_{2.} & \\ & & \ddots \\ 0 & & n_{m_1.} \end{pmatrix} \quad \mathbf{D}_2 = \begin{pmatrix} n_{.1} & & 0 \\ & n_{.2} & \\ & & \ddots \\ 0 & & n_{.m_2} \end{pmatrix}$$

Profils le tableau des profils des lignes $n_{ij}/n_{i.}$ est donné par $D_1^{-1}N$ et celui des profils des colonnes $n_{ij}/n_{.j}$ par ND_2^{-1}.

Représentation géométrique des profils

Nuage de points les profils-lignes forment un nuage de m_1 points de $\mathbb{R}^{m_2}$. Chaque point est affecté d'un poids égal à sa fréquence marginale $n_{i.}/n$, et la matrice des poids est donc $\frac{1}{n}D_1$.
Centre de gravité c'est le profil marginal car

$$g_l = \frac{1}{n}(D_1^{-1}N)^t D_1 1_{m_1} = (\frac{n_{.1}}{n}, \cdots \frac{n_{.m_2}}{n})^t$$

Profils-colonne les lignes du tableau $D_1^{-1}N^t$ forment un nuage de m_2 points de $\mathbb{R}^{m_1}$, avec matrice des poids $\frac{1}{n}D_2$ et centre de gravité

$$g_c = (\frac{n_{1.}}{n}, \cdots \frac{n_{m_1.}}{n})^t$$

Comment étudier ces données

Cas indépendant en cas d'indépendance empirique, on aura

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n} \text{ et } \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$$

et les deux nuages sont alors réduits à leurs centres de gravité respectifs.

Dimension des nuages comme les profils somment à 1, les m_1 profils-lignes sont situés dans le sous-espace W_1 de dimension $m_2 - 1$ défini par $\sum_{j=1}^{m_2} x_j = 1$ et $x_j \geq 0$.

ACP l'étude de la forme des nuages au moyen de l'Analyse en Composantes Principales permettra de rendre compte de la structure des écarts à l'indépendance.

La métrique du $\mathcal{X}^2$

Profils-lignes la distance entre deux profils-lignes i et i' est

$$d^2_{\mathcal{X}^2}(i,i') = \sum_{j=1}^{m_2} \frac{n}{n_{.j}}\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2$$

ce qui revient à utiliser la métrique diagonale nD_2^{-1}.
Inertie l'inertie totale du nuage des profils-lignes par rapport à g_l est

$$I_{g_l} = \sum_{i=1}^{m_1} \frac{n_{i.}}{n} d^2_{\mathcal{X}^2}(i,g_l) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{n_{i.}}{n_{.j}}\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{.j}}{n}\right)^2$$

$$= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{1}{n_{i.}n_{.j}}\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2 = \varphi^2$$

Cette inertie mesure donc l'écart à l'indépendance.

Pourquoi la métrique du $\mathcal{X}^2$?

Pondération la pondération $n/n_{.j}$ permet de donner des importances comparables aux différentes "variables".
Equivalence distributionnelle si deux colonnes j et j' de N ont le même profil, il est logique de les regrouper en une seule d'effectif $n_{ij} + n_{ij'}$; on a alors $n_{ij}/n_{.j} = n_{ij'}/n_{.j'}$

$$\frac{n}{n_{.j}}\left(\frac{n_{ij}}{n} - \frac{n_{i'j}}{n_{i'.}}\right)^2 + \frac{n}{n_{.j'}}\left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}}\right)^2$$

$$= \frac{n}{n_{.j} + n_{.j'}}\left(\frac{n_{ij} + n_{ij'}}{n_{i.}} - \frac{n_{i'j} + n_{i'j'}}{n_{i'.}}\right)^2$$

La distance entre les profils-ligne est donc inchangée.

Autres propriétés de la métrique du du $\mathcal{X}^2$

Propriétés de g_l le vecteur Og_l est orthogonal à W_1 au sens de la métrique du $\mathcal{X}^2$ car

$$\langle g_l x, Og_l \rangle = (x - g_l)^t n D_2^{-1} g_l = (x - g_l)^t 1_{m_2} = 0$$

et la norme de g_l est $\|g_l\|^2_{\mathcal{X}^2} = g_l^t n D_2^{-1} g_l = g_l^t 1_{m_2} = 1$. Tous les vecteurs centrés du nuage sont donc orthogonaux à g_l.

Profils-colonnes on définit la distance entre deux profils-colonnes j et j' comme

$$d^2_{\mathcal{X}^2}(j, j') = \sum_{i=1}^{m_1} \frac{n}{n_{i.}} \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}}\right)^2$$

ce qui correspond à une métrique de matrice nD_1^{-1}. Ses propriétés sont similaires à celles sur les profils-lignes.

ACP des deux nuages de profils

Il y a deux possibilités qui sont en dualité exacte

Profils-lignes

- tableau de données $X = D_1^{-1}N$;
- métrique $M = nD_2^{-1}$;
- poids $D = \frac{D}{n}$.

Profils-colonnes

- tableau de données $X = D_2^{-1}N^t$;
- métrique $M = nD_1^{-1}$;
- poids $D = \frac{D}{n}$.

Autres données

- Centre de gravité $g = X^t D1$
- Matrice de variance-covariance

$$V = X^t DX - gg^t = (X - 1g^t)^t D(X - 1g^t)$$

g est un facteur principal

Pourquoi ? g est vecteur propre de VM associé à la valeur propre 0 car, comme g est $\mathcal{X}^2$-orthogonal à W.

$$VMg = (X - 1g^t)^t D(X - 1g^t)Mg$$

et on a donc $X^t DXMg = VMg + gg^t = 0 + g\|g\|_{\mathcal{X}^2} = g$.
Autres axes les autres valeurs et vecteurs propres de VM et $X^t DXM$ sont identiques car, pour tout vecteur $u \perp g$.

$$X^t DXMu = VMu + gg^t Mu = VMu + g\langle g, u\rangle_{\mathcal{X}^2} = VMu$$

Centrage il est inutile de centrer les tableaux de profil ; on effectue une ACP non centrée et on élimine la valeur propre 1 associée à l'axe principal g et au facteur principal Mg=1.

Calcul de l'ACP

On fait d'abord le calcul pour les profils-lignes.

Facteurs principaux ils sont vecteurs propres de

$$MX^t DX = (nD_2^{-1})(D_1^{-1}N)^t \frac{D_1}{n}(D_1^{-1}N) = D_2^{-1}N^t D_1^{-1}N$$

On a donc pour chaque axe principal k

$$D_2^{-1}N^t D_1^{-1}Nu_k = \lambda_k u_k$$

Composantes principales la composition principale associée au facteur u_k est $a_k = Xu_k = D_1^{-1}Nu_k$; elle est vecteur propre de la matrice $D_1^{-1}ND_2^{-1}N^t$ car

$$D_1^{-1}ND_2^{-1}N^t a_k = D_1^{-1}ND_2^{-1}N^t D_1^{-1}Nu_k = \lambda_k D_1^{-1}Nu_k = \lambda_k a_k$$

Profils-colonnes on échange les indices 1 et 2 et on transpose N.

Comparaison lignes-colonnes

	ACP profils-lignes	*ACP profils-colonnes*
Facteurs principaux	Vecteurs propres de $\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}$	Vecteurs propres de $\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'$
Composantes principales	Vecteurs propres de $\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'$ normalisés par $\mathbf{a}_k'\frac{\mathbf{D}_1}{n}\mathbf{a}_k = \lambda_k$	Vecteurs propres de $\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}$ normalisés par $\mathbf{b}_k'\frac{\mathbf{D}_2}{n}\mathbf{b}_k = \lambda_k$

Comparaison les deux analyses conduisent aux mêmes valeurs propres et les facteurs principaux de l'une sont les composantes principales de l'autre (à un facteur près).

Interprétation des résultats

Coordonnées des points les coordonnées des profils-lignes et profils-colonnes s'obtiennent en cherchant les vecteurs propres des produits des tableaux de profils. Ce sont les grandeurs principales à obtenir.
Projection des nuages il est possible de projeter les deux nuages de points sur la même représentation. On justifiera plus tard le sens de cette représentation et son interprétation.
Cercle des corrélations il n'a aucun intérêt ici, puisque les véritables variables sont qualitatives.
(non) effet de taille comme les composantes variables sont centrées ($\sum_{i=1}^{m_1} n_{i.} a_{ki} = \sum_{j=1}^{m_2} n_{.j} b_{kj}$), on sait que les cordonnées des a_k et b_k ne peuvent être toutes de même signe ; il n'y a donc jamais d'effet de "taille".

Contributions à l'inertie

Contribution des profils-lignes On sait que $\lambda_k = \sum_{i=1}^{m_1} \frac{n_{i.}}{n}(a_{ki})^2$, où a_{ki} est la coordonnée du profil-ligne i sur la kième composante principale de l'ACP sur les profils-lignes. On définit donc la contribution de i à l'axe principal k comme

$$\frac{n_{i.}}{n} \cdot \frac{(a_{ki})^2}{\lambda_k}$$

On considérera les catégories ayant les influences les plus importantes (typiquement $> n_{i.}/n$) comme constitutives des axes ; on regardera aussi le signe de la coordonnée.

Contribution des profils-colonnes pour les mêmes raisons, la contribution du profil-colonne k est

$$\frac{n_{.j}}{n} \cdot \frac{(b_{kj})^2}{\lambda_k}$$

Formules de transition

But on cherche une relation entre les vecteurs a_k et b_k pour éviter de faire deux diagonalisations de matrices. Par exemple, si $m_1 < m_2$, on diagonalisera la matrice $D_1^{-1} N D_2^{-1} N^t$.

Formules un calcul simple donne les formules suivantes

$$b_k = \frac{1}{\sqrt{\lambda_k}} D_2^{-1} N^t a_k, \text{ soit } b_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{m_1} \frac{n_{ij}}{n_{.j}} a_{ki};$$

$$a_k = \frac{1}{\sqrt{\lambda_k}} D_1^{-1} N b_k, \text{ soit } a_{ki} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{m_2} \frac{n_{ij}}{n_{i.}} b_{kj}.$$

Méthode comme a_k est (à une normalisation près) le facteur principal associé à b_k, on sait que $b_k = \alpha D_2^{-1} N^t a_k$. Pour déterminer α, il suffit d'écrire que $b_k^t \frac{D_2}{n} b_k = \lambda_k$.

Décomposition de l'inertie

φ^2 et valeurs propres on sait que l'inertie totale (et donc la somme des valeurs propres) est égale à φ^2. Comme il y a au plus $min(m_1 - 1, m_2 - 1)$ valeurs propres, on obtient si $m_1 < m_2$

$$\varphi^2 = \sum_{k=1}^{m_1-1} \lambda_k.$$

Choix du nombre de valeurs propres c'est un problème plus difficile

- la règle de Kaiser ($\lambda_k > 1$) ne s'applique plus ;
- la règle du coude reste valide, mais est un peu subjective ;
- on peut par contre s'aider de la part d'inertie expliquée.

Analyse des correspondances multiples

But on veut étendre l'AFC au cas de $p \geq 2$ variables $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_p$ à $m_1, m_2, \cdots, m_p$ modalités. Ceci est particulièrement utile pour l'exploration d'enquêtes où les questions sont à réponses multiples.

Problème l'analyse des correspondances utilise une table de contingence qui est difficilement généralisable au cas p>2.

Méthode on cherche un moyen différent de calculer l'AFC pour p = 2 et on vérifie que les résultats sont comparables. Si on a de la chance, on pourra étendre cette nouvelle version pour p>2 .

Les données

Données brutes chaque individu est décrit par les numéros pour chacune des p variables des modalités qu'il possède. Il n'est pas possible de faire des calculs sur ce tableau, où les valeurs sont arbitraires.

Tableau disjonctif on remplace la jème colonne par m_j colonnes d'indicatrices : on met un zéro dans chaque colonne, sauf celle correspondant à la valeur x_i^j de l'individu i qui recoit 1.

Exemple on a trois variables (avec respectivement 3, 2 et 2 modalités) mesurées sur 4 individus. Les tableaux (ci-dessous à gauche) sont équivalents aux tableaux disjonctifs à droite.

$$\begin{pmatrix} 1 \\ 3 \\ 2 \\ 3 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 1 \\ 2 \end{pmatrix} \qquad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Tableau disjonctif et tableau de contingence

Tableau disjonctif à la variable $\mathcal{X}_j$ on associe le tableau disjonctif X_j à n lignes et m_j colonnes.

Tableau de contingence on vérifie facilement que le tableau de contingence des variables $\mathcal{X}_j$ et $\mathcal{X}_k$ est donné par

$$N_{jk} = X_j^t X_k$$

Effectifs marginaux la matrice diagonale des effectifs marginaux de la variable $\mathcal{X}_j$ est donnée par

$$D_j = X_j^t X_j$$

Exemple

$$N_{21} = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 0 \end{pmatrix}$$

$$D_2 = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

Tableau disjonctif joint

Définition c'est la matrice juxtaposée $X = (X_1|X_2|\cdots|X_p)$ qui possède n lignes et $m_1 + \cdots + m_p$ colonnes. On va la considérer comme un tableau de contingence et lui appliquer une analyse de correspondances.
Les lignes la somme des éléments de chaque ligne de X est égale à p. Le tableau des profils-lignes est donc $\frac{1}{p}X$
Les colonnes la somme des éléments de chaque colonne de X est égale à l'effectif marginal de la modalité correspondante. le tableau des profils-colonnes est donc XD^{-1}, où D est la matrice diagonale par blocs

$$\mathrm{D} = \begin{pmatrix} \mathrm{D}_1 & & 0 \\ & \ddots & \\ 0 & & \mathrm{D}_p \end{pmatrix}$$

Cas p=2

On applique l'analyse des correspondances de X ; on cherche les composantes principales de l'ACP en colonnes. Elles sont vecteurs propres de

$$(XD^{-1})^t \frac{1}{2} X = \frac{1}{2} D^{-1} X^t X$$

Or on a

$$X^t X = \begin{bmatrix} X_1^t \\ X_2^t \end{bmatrix} \begin{bmatrix} X_1 & X_2 \end{bmatrix} = \begin{bmatrix} X_1^t X_1 & X_1^t X_2 \\ X_2^t X_1 & X_2^t X_2 \end{bmatrix} = \begin{bmatrix} D_1 & N \\ N^t & D_2 \end{bmatrix}$$

Finalement, les composantes principales sont valeurs propres de

$$\frac{1}{2} \begin{bmatrix} D_1^{-1} & 0 \\ 0 & D_2 \end{bmatrix} \begin{bmatrix} D_1 & N \\ N^t & D_2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} I_{m_1} & D_1^{-1} N \\ D_2^{-1} N^t & I_{m_2} \end{bmatrix}$$

Résolution des équations

On note a (resp. b) les m_1 premières (resp. m_2 dernières) coordonnées de la composante principale recherchée et μ la valeur propre correspondante :

$$\begin{bmatrix} I_{m_1} & D_1^{-1}N \\ D_2^{-1}N^t & I_{m_2} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = 2\mu \begin{bmatrix} a \\ b \end{bmatrix}$$

On obtient les équations

$$\begin{cases} D_1^{-1}Nb &= (2\mu - 1)a \\ D_2^{-1}N^t a &= (2\mu - 1)b \end{cases}$$

et donc on retrouve les coordonnées des lignes et des colonnes de N dans l'AFC classique (avec $\lambda = (2\mu - 1)^2$) :

$$\begin{cases} D_2^{-1}N^t D_1^{-1}Nb &= (2\mu - 1)^2 b \\ D_2^{-1}ND_2^{-1}N^t a &= (2\mu - 1)^2 a \end{cases}$$

Le nombre de valeurs propres

Problème on a a priori $m_1 + m_2 - 1$ valeurs propres non nulles, ce qui est plus important que dans le cas classique. En particulier pour chaque λ, on a deux μ possibles

$$\begin{cases} \mu = \frac{1+\sqrt{\lambda}}{2} \text{ associée à } \begin{bmatrix} a \\ b \end{bmatrix} \\ \mu = \frac{1-\sqrt{\lambda}}{2} \text{ associée à } \begin{bmatrix} a \\ -b \end{bmatrix} \end{cases}$$

On ne garde donc que les valeurs $\mu > 1$. On peut montrer qu'il y en a $min(m_1 - 1, m_2 - 1)$.

Interprétation L'interprétation de la part d'inertie expliquée par les valeurs propres est maintenant très différente. En particulier les valeurs propres qui étaient très séparées dans l'AFC de N le sont beaucoup moins dans celle de X.

Le nombre de valeurs propres

But on cherche à faire une représentation des $m_1, m_2, \cdots, m_p$ comme points d'un espace de faible dimension.

Méthode on fait une AFC sur le tableau disjonctif joint $X = (X_1|X_2|\cdots|X_p)$, qui possède n lignes et $m_1 + m_2 + \cdots + m_p$ colonnes.

Le tableau de Burt c'est le tableau $B = X^t X$, qui est un super-tableau de contingence des variables $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_p$

$$\mathbf{B} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 & \cdots & \mathbf{X}_1'\mathbf{X}_p \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 & & \\ \vdots & & \ddots & \vdots \\ \mathbf{X}_p'\mathbf{X}_1 & \cdots & & \mathbf{X}_p'\mathbf{X}_p \end{bmatrix}$$

où on rappelle que $X_i^t X_i = D_i$. Le tableau de Burt est donc formé de tableaux de contingence et de matrices d'effectifs marginaux.

Exemple de tableau de Burt

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 2 & 2 & 0 & 0 & 2 \\ 0 & 1 & 2 & 3 & 0 & 1 & 2 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 2 & 0 \\ 0 & 0 & 2 & 2 & 0 & 0 & 2 \end{pmatrix}$$

Les coordonnées factorielles des catégories

Notation on note $a = (a_1, \cdots, a_p)^t$ le vecteur à $m_1 + m_2 + \cdots + m_p$ composantes des coordonnées factorielles des catégories sun axe

Calcul de l'AFC sur X comme la matrice des profils-lignes est $\frac{1}{p}X$ et celle des profils colonnes XD^{-1}, a est vecteur propre de

$$(XD^{-1})^t \frac{1}{p} X = \frac{1}{p} D^{-1} X^t X = \frac{1}{p} D^{-1} B$$

et donc l'équation des coordonnées des catégories est

$$\frac{1}{p} D^{-1} B a = \mu a$$

avec la convention de normalisation

$$\frac{1}{np} a^t D a = \mu$$

Formules barycentriques

Les coordonnées des individus soit z le vecteur à n composantes des coordonnées des n individus sur un axe factoriel associé à la valeur propre μ. D'après les résultats de sur l'AFC, on a

$$z = \frac{1}{\sqrt{\mu}} \frac{1}{p} Xa$$

On a donc la normalisation

$$V(z) = \frac{1}{n} z^t z = \frac{1}{\mu n p^2} a^t X^t X a = \frac{1}{\mu n p^2} a^t (p \mu D a) = \frac{1}{np} a^t D a = \mu$$

Les seuls termes non nuls dans le calcul de Xa sont les coordonnées de la catégorie de chaque variable possédée par l'individu.

Barycentre des catégories à $1/\sqrt{\mu}$ près, la coordonnée d'un individu est égale à la moyenne arithmétique simple des coordonnées des catégories auxquelles il appartient.

Formules barycentriques (suite et fin)

On a de même la seconde formule

$$a = \frac{1}{\sqrt{\mu}} D^{-1} X^t z$$

Les seuls termes non nuls de $X^t z$ sont les coordonnées des individus ayant une modalité donnée.
Barycentre des individus à $1/\sqrt{\mu}$ près, la coordonnée d'un individu est égale à la moyenne arithmétique des coordonnées des n_j individus de cette catégorie.

Propriétés des valeurs propres

Valeurs propres triviales la valeur propre 1 est associée (comme en AFC) à la composante $z_0 = (1, \cdots, 1)$ dans l'espace des individus. Les autres vecteurs propres lui sont orthogonaux, et donc de moyenne nulle.
Autres valeurs propres si $n > \sum_{i=1}^{p} m_i$, le rang de X est $\sum_{i=1}^{p} m_i - p + 1$ et le nombre de valeurs propres non trivialement égales à 0 ou 1 est $q = \sum_{i=1}^{p} m_i - p$.
Somme la somme des valeurs propres non triviales est donc

$$\sum_{k=1}^{q} \mu_k = Trace(\frac{1}{p} D^{-1} B) - 1 = \frac{1}{p} \sum_{i=1}^{p} m_i - 1 = \frac{q}{p}$$

La moyenne des q valeurs propres vaut 1/p.

Barycentres et représentation

Représentation commune Les points représentatifs des catégories sont barycentres des groupes d'individus. On peut donc représenter individus et catégories dans un même plan factoriel.

Moyennes Comme z est une variable de moyenne nulle, la formule de barycentre indique que pour chaque variable $\mathcal{X}_i$, les coordonnées de ses catégories (pondérés par les effectifs) sont de moyenne nulle. Aucun centrage n'est donc nécessaire

Echelle pour que les catégories se trouvent visuellement au barycentre des individus qui les représentent on peut remplacer a par

$$a = D^{-1}X^t z = \sqrt{\mu}a$$

Variables et axes factoriels

Si n_j est l'effectif de la catégorie j et a_j sa coordonnée sur l'axe factoriel associé μ, alors

$$\frac{1}{np} \sum_{j \in \text{catégorie}} n_j (a_j)^2 = \mu$$

Catégorie la contribution de la catégorie j à l'axe factoriel est

$$\frac{1}{\mu} \frac{n_j}{np} (a_j)^2,$$

intéressante si elle est supérieure au poids n_j/np.
Variable la contribution totale de la variable $\mathcal{X}_i$ à l'axe factoriel est

$$\frac{1}{\mu} \frac{1}{np} \sum_{j \text{ modalité de} \mathcal{X}_i} n_j (a_j)^2$$

Individus et axes factoriels

La normalisation de z est $\sum_{i=1}^{n}(z_i)^2 = n\mu$, où z_i est la cordonnées de l'individu i sur l'axe factoriel associé à la valeur propre μ.

Contribution d'un individu elle est égale pour l'individu i à

$$\frac{1}{n\mu}(z_i)^2$$

Cette contribution est jugée en la comparant au poids 1/n.

Qualité de la représentation pour le sous-espace formé par les l premiers axes, c'est le cosinus carré habituel

$$\frac{\sum_{k=1}^{l}(z_i^{(k)})^2}{\sum_{k=1}^{q}(z_i^{(k)})^2}$$

où l'exposant (k) correspond aux différents axes factoriels.

Contribution à l'inertie totale

Soit $x_j = (x_j^j)$ le vecteur colonne de X correspondant à une catégorie j. On rappelle que l'inertie totale vaut

$$\sum_{j \in \text{catégorie}} \frac{n_j}{np} d^2(j,g) = \frac{1}{p} \sum_{i=1}^{p} m_i - 1$$

La distance du profil-colonne j au centre de gravité des profils-colonnes g=1/n est

$$d^2(j,g) = \sum_{i=1}^{n} \frac{np}{p} (\frac{x_i^j}{n_j} - \frac{1}{n})^2 = n \sum_{i=1}^{n} (\frac{x_i^j}{n_j^2} + \frac{1}{n^2} - \frac{2x_i^j}{nn_j})$$

$$= \frac{n}{n_j} - 1$$

Contribution à l'inertie totale (suite et fin)

Contribution d'une catégorie la contribution absolue de la catégorie à l'inertie est

$$\frac{n_j}{np} d^2(j,g) = \frac{1}{p}(1 - \frac{n_j}{n})$$

qui est une fonction décroissante de l'effectif. Il faut dont éviter les catégories d'effectif trop faible, qui d'ailleurs se retrouveront dans les premiers axes.

Contribution d'une variable la contribution de la variable $\mathcal{X}_i$ est

$$\sum_{j \text{ modalité de} \mathcal{X}_i} \frac{1}{p}(1 - \frac{n_j}{n}) = \frac{m_i - 1}{p}$$

Elle est d'autant plus grande que le nombre de modalité est élevé. Il faut donc éviter les disparités trop grandes entre les modalités (quand on a le choix du découpage...)

Les variables supplémentaires

Leur usage est très courant en analyse des correspondances multiples.

Variables qualitatives on les place directement sur la projection sur un plan factoriel en utilisant la formule de barycentre des individus : si on veut placer une variable supplémentaire de tableau disjonctif X_{sup} et d'effectifs marginaux D_{sup}, on calcule les coordonnées de ses modalités sur un axe principal par

$$a_{sup} = \frac{1}{\sqrt{\mu}} D_{sup}^{-1} X'_{sup} z$$

Variables quantitatives on calcule "à la main" leur corrélation avec les axes factoriels et on les place sur un cercle de corrélations. On peut aussi les découper en classes et les traiter comme des variables qualitatives.

Valeurs-test pour les variables supplémentaires

But on cherche à savoir si une variable supplémentaire est liée à un axe donné.

Idée du calcul si les n_j individus d'une catégorie étaient pris au hasard, la moyenne de leurs coordonnées serait une variable aléatoire centrée (les z sont de moyenne nulle) et de variance $\frac{\mu}{n_j}\frac{n-n_j}{n-1}$. De plus, la moyenne des coordonnées est égale à μa_j.

Valeur-test c'est la version centrée et réduite de la moyenne des coordonnées

$$a_j\sqrt{n_j}\sqrt{\frac{n-1}{n-n_j}}$$

Quand n_j est assez grand, elle est significative si elle est supérieure à 2 ou 3. On ne doit pas l'utiliser sur les variables actives.

Pratique de l'analyse des correspondances multiples

Sélection des variables on décide souvent de ne garder qu'un nombre réduit de variables actives et de garder les autres comme variables supplémentaires.

Sélection des axes

- règle courante : garder les axes tels que $\mu > 1/p$ (la moyenne des valeurs propres est 1/p).
- les axes intéressants sont ceux que l'on peut interpréter, en regardant les contributions des variables actives et les valeurs-tests associées aux variables supplémentaires.
- en pratique on se contente souvent d'interpréter le premier plan principal.

Inertie expliquée elle est moins intéressante qu'en ACP.

Points communs entre AFC et ACM

But	décrire les liaisons entre plusieurs variables qualitatives
Cas $p = 2$	les coordonnées des modalités sont les mêmes pour les deux analyses
Représentation	toutes les modalités peuvent être représentées sur le même diagramme
Contribution d'une modalité à un axe	$\text{poids} \times \dfrac{(\text{coordonnée})^2}{\text{valeur propre}}$
Qualité de la représentation d'une modalité par un sous espace	$\cos^2\theta = \dfrac{\sum_{\text{axes du sous esp.}} (\text{coord sur l'axe})^2}{\sum_{\text{tous les axes}} (\text{coord sur l'axe})^2}$

Points communs entre AFC et ACM

	AFC	ACM
Individus	non	oui
Données	tableau de contingence profils lignes/colonnes	tableau disjonctif tableau de Burt
Poids d'une modalité	$\frac{n_{i\cdot}}{n}$ (profil-ligne) $\frac{n_{\cdot j}}{n}$ (profil-colonne)	$\frac{n_j}{np}$
Nb de val. propres	$\min(m_1 - 1, m_2 - 1)$	$\sum_{i=1}^{p} m_i - p$
Axes à conserver	pas de règle Kaiser ; peut-être part d'inertie.	$\mu > \frac{1}{p}$
Variables supplémentaires	pas vraiment de sens	qualitatives et quantitatives