Analyse Factorielle des Correspondances

MSPES-ENSAI

Introduction

Introduction

Idée générale L'analyse factorielle des correspondances (AFC) est une méthode factorielle pour l'exploration statistique d'une **table de contingence** définie par **deux** variables **qualitatives**.

Objectif Le but est d'étudier la liaison entre ces deux variables qualitatives (notamment quelles associations de modalités sont sur-représentées).

Notations: On considère deux variables qualitatives A et B ayant respectivement n et p modalités. Ces variables sont observées simultanément sur k individus affectés de poids identiques $1/n$. On peut donc construire le tableau de contingence suivant

$$K = \left[k_{ij}\right]_{\substack{i=1,\dots,n \\ j=1,\dots,p}}$$

où k_{ij} est le nombre d'observations prenant le niveau i pour la variable A et le niveau j pour la variable B.

Introduction

$$
K = \begin{array}{c|c|c|c|c||c}
 & & & & & \text{eff. marg.} \\
k_{11} & \dots & k_{1j} & \dots & k_{1p} & k_{1\bullet} \\
\vdots & & \vdots & & \vdots & \vdots \\
k_{i1} & \dots & k_{ij} & \dots & k_{ip} & k_{i\bullet} \\
\vdots & & \vdots & & \vdots & \vdots \\
k_{n1} & \dots & k_{nj} & \dots & k_{np} & k_{n\bullet} \\
\hline\hline
\text{eff. marg.} \quad k_{\bullet 1} & \dots & k_{\bullet j} & \dots & k_{\bullet p} & k
\end{array}
$$

Table 1: Table de contingence K

On note

- $k_{i\bullet} = \sum_{j=1}^{p} k_{ij}$ l'effectif marginal de la ligne i.
- $k_{\bullet j} = \sum_{i=1}^{n} k_{ij}$ l'effectif marginal de la colonne j.
- $k = \sum_{i=1}^{n} \sum_{j=1}^{n} k_{ij}$ l'effectif total.

Test du χ^2

Test du χ^2

On dit que deux variables A et B sont *non liées* relativement à K si et seulement si

$$\forall (i,j) \in \{1,\ldots,n\} \times \{1,\ldots,p\} : \; k_{ij} = \frac{k_{i\bullet} k_{\bullet j}}{k}.$$

Cette notion est cohérente avec celle d'indépendance en probabilité. En effet,

$$X \perp Y \Leftrightarrow \underbrace{\mathbb{P}(A = i \cap B = j)}_{\text{estimée par } \frac{k_{ij}}{k}} = \underbrace{\mathbb{P}(A = i)}_{\text{estimée par } \frac{k_{i\bullet}}{k}} \times \underbrace{\mathbb{P}(B = j)}_{\text{estimée par } \frac{k_{\bullet j}}{k}} , \forall (i,j)$$

Test du χ^2

On souhaite étudier la liaison entre A et B à partir de nos observations.

La représentation graphique des profils-lignes ou des profils-colonnes, au moyen de diagrammes en barres parallèles, ainsi que le calcul de coefficients de liaison (Cramer) donnent une première idée de la variation conjointe des deux variables (*cf.* cours de stat desc).

Le test du χ^2 permet de plus de s'assurer du caractère significatif de cette liaison.

Test du χ^2

On test l'indépendance entre deux variables qualitatives A et B par le test du χ^2 construit de la manière suivante:

- l'hypothèse nulle est H_0: A et B sont indépendantes,
- l'hypothèse alternative est H_1: A et B ne sont pas indépendantes.

La statistique de test est alors

$$T = \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(k_{ij} - \frac{k_{i\bullet} k_{\bullet j}}{k})^2}{\frac{k_{i\bullet} k_{\bullet j}}{k}}.$$

Pour des grandes valeurs de k, et si H_0 est vraie,

$$T \sim \chi^2_{(n-1)(p-1)}.$$

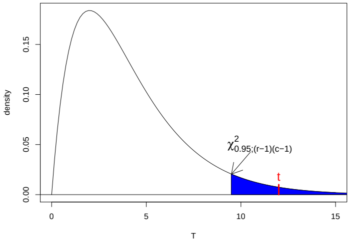
Test du χ^2

On rejette donc H_0 et on conclut au caractère significatif de la liaison entre A et B si T dépasse une valeur particulière (valeur ayant une probabilité faible et fixée a priori α d'être dépassée par une loi du χ^2 à $(n-1)(p-1)$ degrés de liberté).

En pratique, on choisit souvent un risque de 1ère espèce $\alpha = 0.05$ (risque de rejeter H_0 à tort).

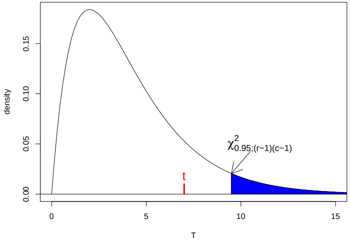
On compare la statistique observée t (*i.e.*, la réalisation de T) au quantile $1-\alpha$ du $\chi^2_{(n-1)(p-1)}$.

Test du χ^2



- La statistique observée t est dans la zone de rejet (t supérieur au quantile 0.95 du $\chi^2_{(n-1)(p-1)}$).
- Au risque de première espèce $\alpha = 0.05$, on rejette H_0.
- Lien significatif entre A et B.

Test du χ^2



- La statistique observée t n'est pas dans la zone de rejet (t inférieur au quantile 0.95 du $\chi^2_{(n-1)(p-1)}$).
- Au risque de première espèce $\alpha = 0.05$, on ne peut pas rejeter H_0.
- On ne conclut pas à un lien significatif entre A et B. L'AFC a peu d'intérêt dans ce cas.

Test du χ^2

La majorité des logiciels retourne la *p-valeur* d'un test.

Ici la *p-valeur* associée à la statistique de test observée t est

$$\text{p-valeur} = \mathbb{P}(\chi^2_{(n-1)(p-1)} > t)$$

- Si p-valeur $> \alpha$ alors t n'est pas dans la zone de rejet (*i.e,* $t < \chi^2_{1-\alpha;(n-1)(p-1)}$). Au risque de première espèce $\alpha = 0.05$, on ne peut donc pas rejeter H_0.
- Si p-valeur $< \alpha$ alors t est dans la zone de rejet (*i.e,* $t > \chi^2_{1-\alpha;(n-1)(p-1)}$). Au risque de première espèce $\alpha = 0.05$, on peut donc rejeter H_0. Au plus p-valeur $<< \alpha$, au plus la probabilité que A et B soient liées est forte.

Profils-lignes et profils-colonnes

Fréquences relatives

On considère le tableau des fréquences relatives

$$F = \frac{1}{k} K = \left[\frac{k_{ij}}{k} \right]_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}} = \left[f_{ij} \right]_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}}$$

et on note $f_{i\bullet} = \sum_{j=1}^{n} \frac{k_{ij}}{k}$ et $f_{\bullet j} = \sum_{i=1}^{p} \frac{k_{ij}}{k}$.

Profils-lignes

Le tableau des profils-lignes X est le tableau des fréquences conditionnelles de la modalité j de B sachant la modalité i de A:

$$X = \left[\frac{f_{ij}}{f_{i\bullet}}\right]_{\substack{i=1,\ldots,n\\ j=1,\ldots,p}}.$$

Ainsi avec les notations du cours d'ACP

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_i^\top \\ \vdots \\ x_n^\top \end{bmatrix} \text{ avec } x_i = \begin{bmatrix} f_{i1}/f_{i\bullet} \\ \vdots \\ f_{ij}/f_{i\bullet} \\ \vdots \\ f_{ip}/f_{i\bullet} \end{bmatrix} \in \mathbb{R}^p.$$

Profils-lignes

On observe n profils-lignes $x_i = \left[f_{i1}/f_{i\bullet}, \ldots, f_{ij}/f_{i\bullet}, \ldots, f_{ip}/f_{i\bullet}\right]^\top \in \mathbb{R}^p$.

Les profils-lignes sont définis sur un simplex (*i.e.*, coordonnées du vecteur positives et sommes des éléments du vecteur égale à 1).

En associant à chaque x_i sa fréquence relative $f_{i\bullet}$ comme pondération, on obtient le nuage pesant de $\mathbb{R}^p$ suivant

$$N_L = \{\{x_i, f_{i\bullet}\}; i = 1, \ldots, n\}.$$

Le centre de gravité du nuage N_L est

$$g = \left[f_{\bullet 1}, \ldots, f_{\bullet j}, \ldots, f_{\bullet p}\right]^\top.$$

Profils-colonnes

Le tableau des profils-colonnes $Y^\top$ est le tableau des fréquences conditionnelles de la modalité i de A sachant la modalité j de B:

$$Y = \begin{bmatrix} y_1^\top \\ \vdots \\ y_j^\top \\ \vdots \\ y_p^\top \end{bmatrix} \text{ avec } y_j = \begin{bmatrix} f_{1j}/f_{\bullet j} \\ \vdots \\ f_{ij}/f_{\bullet j} \\ \vdots \\ f_{nj}/f_{\bullet j} \end{bmatrix} \in \mathbb{R}^n.$$

En associant à chaque y_j sa fréquence relative $f_{\bullet j}$ comme pondération, on obtient le nuage pesant de $\mathbb{R}^n$ suivant

$$N_C = \{\{y_j, f_{\bullet j}\}; j = 1, \ldots, p\}.$$

Le centre de gravité du nuage N_C est

$$h = \left[f_{1\bullet}, \ldots, f_{i\bullet}, \ldots, f_{n\bullet}\right]^\top.$$

Inertie

En négligeant l'aléatoire...

$$
\begin{aligned}
A \text{ et } B \text{ indpt } &\Leftrightarrow \forall (i,j),\ f_{ij} = f_{i\bullet} f_{\bullet j} \\
&\Leftrightarrow \forall (i,j),\ \frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j} \\
&\Leftrightarrow \forall i,\ x_i = g \\
&\Leftrightarrow \forall (i,j),\ \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet} \\
&\Leftrightarrow \forall j,\ y_j = h
\end{aligned}
$$

Inertie

L'inertie comme mesure d'écart à l'indépendance:

- sur le tableau des profils-lignes X

$$I = \sum_{i=1}^{n} f_{i\bullet} d_{M_L}(x_i, g)^2.$$

- sur le tableau des profils-lignes Y

$$J = \sum_{j=1}^{p} f_{\bullet j} d_{M_C}(y_j, h)^2.$$

L'étude de la liaison entre A et B se fait par l'étude des inerties I et J à travers une ACP particulière.

Dans le cas de l'indépendance statistique, les tableaux des *profils-lignes* et de *profils-colonnes* sont alors réduits à un point en leurs centres de gravité respectifs.

L'étude de la forme de ces nuages au moyen d'une ACP permettra donc de rendre compte de la structure des *écarts à l'indépendance.* Il faut donc choisir une métrique pour chacun de ces espaces.

Remarques:

- Toute structure d'ordre existant éventuellement sur les modalités de A ou de B est ignorée par l'AFC.
- Tout individu présente une modalité et une seule de chaque variable.
- Chaque modalité doit avoir été observée au moins une fois, sinon elle est supprimée.

ACP de N_L

ACP de N_L

Objectif: l'ACP va produire des axes orthogonaux de plus grande inertie (donc ici de plus grande dépendance).

Données: on considère le nuage pesant

$$N_L = \{\{x_i, f_{i\bullet}\}; i = 1, \ldots, n\} \text{ où } x_i = \frac{1}{f_{i\bullet}} \left[f_{i1}, \ldots, f_{ij}, \ldots, f_{ip}\right]^\top \in \mathbb{R}^p.$$

Ce nuage est défini par la matrice de données X et la matrice des poids $D_{f_{i\bullet}} = \text{diag}(f_{1\bullet}, \ldots, f_{n\bullet}) \in \mathbb{R}^{n \times n}$.

Métrique: on choisit la métrique du χ^2 notée M_L où

$$M_L = D_{1/f_{\bullet j}} = \text{diag}(1/f_{\bullet 1}, \ldots, f_{\bullet p}) \in \mathbb{R}^{p \times p}.$$

ACP de N_L

Métrique: on choisit la métrique du χ^2 notée M_L où

$$M_L = D_{1/f_{\bullet j}} = \mathrm{diag}(1/f_{\bullet 1}, \ldots, f_{\bullet p}) \in \mathbb{R}^{p \times p}.$$

Pour calculer la distance entre deux profils-lignes x_i et $x_{i'}$, on utilise la formule suivante

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

La métrique du χ^2 introduit l'inverse des fréquences marginales des modalités de B comme pondérations des écarts entre éléments de deux profils relatifs à A.

ACP de N_L

Principe d'équivalence distributionnelle Le choix de la métrique du χ^2 permet de garantir une certaine invariance vis-à-vis du choix des modalités de A et B.

Condition d'invariance Il faut proportionnalité entre les modalités avant et après regroupement.

Le terme de métrique du χ^2 vient du fait que le nuage N_L a pour inertie totale la quantité mesurant l'écart à l'indépendance:

$$I = \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{\left(k_{ij} - \frac{k_{i\bullet} k_{\bullet j}}{k}\right)^2}{\frac{k_{i\bullet} k_{\bullet j}}{k}}.$$

ACP de N_L

Le cours sur l'ACP a été fait pour un nuage centré.

Le centre de gravité du nuage N_L est g.

Notations

- $\tilde{X}$ matrice centrée en colonne

$$\tilde{X} = X - \mathbf{1}_n g^\top = \begin{bmatrix} (x_1 - g)^\top \\ \vdots \\ (x_i - g)^\top \\ \vdots \\ (x_n - g)^\top \end{bmatrix}.$$

- $\tilde{V}$ matrice d'inertie calculée en $\tilde{X}$

$$\tilde{V} = \tilde{X}^\top D_{f_{i\bullet}} \tilde{X}.$$

- V matrice d'inertie calculée en X

$$V = X^\top D_{f_{i\bullet}} X.$$

ACP de N_L

1. g est vecteur propre de $\tilde{V}M_L$ associé à la valeur propre 0.
2. g est vecteur propre de VM_L associé à la valeur propre 1.

Les autres vecteurs propres et valeurs propres de $\tilde{V}M_L$ et VM_L sont égaux.

En pratique, en AFC, on ne centre pas X. On écarte juste le vecteur propre g de valeur propre 1.

On a potentiellement $p-1$ valeurs propres d'intérêt

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{p-1}$$

avec vecteurs propres

$$\mu_1, \ldots, \mu_{p-1}.$$

ACP de N_L

Autres résultats issus de l'ACP

- On a la relation suivante pour l'inertie

$$I = \lambda_1 + \ldots + \lambda_{p-1}.$$

- % d'inertie expliquée = % de dépendance expliquée.
- Composantes principales

$$\mathbb{R} \supset C_{\bullet k} = X D_{1/f_{\bullet j}} \mu_k, \quad k = 1, \ldots, p-1.$$

- Qualité de représentation de l'individu i sur l'axe k

$$Co2(i,k) = \frac{C_{ik}^2}{\|x_i\|^2_{D_{1/f_{\bullet j}}}}.$$

- Contribution de l'individu i à l'axe k

$$CTR(i,k) = \frac{f_{i\bullet} C_{ik}^2}{\lambda_k}.$$

- Les vecteurs $\frac{C_{\bullet k}}{\sqrt{\lambda_k}}$ sont $D_{f_{i\bullet}}$ normés et orthogonaux 2 à 2.

ACP de N_L

En résumé:

- L'AFC est une ACP particulière.
- Cette ACP ne se fait pas sur les données de départ (ici le tableau de contingence) mais sur le tableau de profils-lignes.
- Cette ACP considère une métrique particulière qui permet de lier l'inertie du nuage à la statistique du χ^2 (qui mesure la dépendance entre 2 variables qualitatives).
- On récupère tous les indicateurs de l'ACP.

ACP de N_L

Questions:

- Dans le cours de l'ACP, la notion d'individus et de variables était simple. Ici, ce n'est plus le cas lorsqu'on considère le tableau de contingence.
- Peut-on projeter les variables (ici les modalités de B) sur le même espace que les observations (et non pas sur un "cercle des corrélations")?
- Que se passe-t-il si on considère la transposée du tableau de contingence initial?

ACP de N_C

- Tableau des données: profils-colonnes

$$Y = \begin{bmatrix} y_1^\top \\ \vdots \\ y_j^\top \\ \vdots \\ y_p^\top \end{bmatrix} \text{ avec } y_j = \begin{bmatrix} f_{1j}/f_{\bullet j} \\ \vdots \\ f_{ij}/f_{\bullet j} \\ \vdots \\ f_{nj}/f_{\bullet j} \end{bmatrix} \in \mathbb{R}^n.$$

- Matrice des poids: $D_{f_{\bullet j}} = \text{diag}(f_{\bullet 1}, \ldots, f_{\bullet p})$.
- Métrique du χ^2: $M_C = D_{1/f_{i\bullet}} = \text{diag}(1/f_{1\bullet}, \ldots, 1/f_{n\bullet})$.
- Le centre de gravité du nuage N_C est $h = \left[f_{1\bullet}, \ldots, f_{i\bullet}, \ldots, f_{n\bullet}\right]^\top$.
- $\tilde{Y} = Y - \mathbf{1}_p h^\top$ est le tableau Y centré en colonnes.
- Matrice d'inertie de $\tilde{Y}$: $\tilde{W} = \tilde{Y}^\top D_{f_{\bullet j}} \tilde{Y}$.
- Matrice d'inertie de Y: $W = Y^\top D_{f_{\bullet j}} Y$.

ACP de N_C

1. h est vecteur propre trivial de valeur propre 1 pour WM_C.
2. h est vecteur propre trivial de valeur propre 0 pour $\tilde{W}M_C$.
3. les autres vecteurs propres et valeurs propres sont identiques.

En pratique, on ne centre pas Y, on écarte la valeur propre 1.

On a donc $n-1$ valeurs propres

$$\rho_1 \geq \rho_2 \geq \ldots \geq \rho_{n-1},$$

de vecteurs propres

$$\nu_1, \ldots, \nu_{n-1}.$$

ACP de N_C

Autres résultats issus de l'ACP

- On a la relation suivante pour l'inertie

$$J = \rho_1 + \ldots + \rho_{n-1}.$$

- % d'inertie expliquée = % de dépendance expliquée.
- Composantes principales

$$\mathbb{R} \supset d_{\bullet k} = Y D_{1/f_{i\bullet}} \nu_k, \quad k = 1, \ldots, n-1.$$

- Qualité de représentation de l'individu j sur l'axe k

$$Co2(j,k) = \frac{d_{jk}^2}{\|y_j\|^2_{D_{1/f_{i\bullet}}}}.$$

- Contribution de l'individu i à l'axe k

$$CTR(j,k) = \frac{f_{\bullet j} d_{jk}^2}{\nu_k}.$$

- Les vecteurs $\frac{d_{\bullet k}}{\sqrt{\rho_k}}$ sont $D_{f_{\bullet j}}$ normés et orthogonaux 2 à 2.

Liens entre les ACP

Liens entre les ACP

Soit I l'inertie de N_C obtenue avec la métrique M_C et J l'inertie de N_L obtenue avec la métrique M_L. On a

$$I = J.$$

On a $\lambda_k = \rho_k$ pour $k = 1, \ldots, \min(n-1, p-1)$. Ainsi, il y a au maximum $r = \min(n-1, p-1)$ valeurs propres non nulles.

On a la relation entre les vecteurs propres, pour $k = 1, \ldots, r$

$$\nu_k = \frac{1}{\sqrt{\lambda_k}} F D_{1/f_{\bullet j}} \mu_k \text{ et } \mu_k = \frac{1}{\sqrt{\lambda_k}} F^\top D_{1/f_{i\bullet}} \nu_k$$

On a les relations pseudo-barycentriques suivantes pour $k = 1, \ldots, r$

$$C_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{p} \frac{f_{ij}}{f_{i\bullet}} d_{jk} \text{ et } d_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{n} \frac{f_{ij}}{f_{\bullet j}} C_{ik}.$$