Visualisation multidimensionnelle de données qualitatives : Analyse Factorielle des Correspondances & Analyse des Correspondances Multiples.

Dhafer Malouche

http://www.essai.rnu.tn
& http://essai.academia.edu/DhaferMalouche
dhafer.malouche@me.com

Mars-Avril 2016

Plan

ACP généralisée

Les éléments de l'ACP

- $w_1, \ldots, w_n$ un échantillon de n individus. $x^1, \ldots, x^d$ d variables telles que
$$x^j(w_i) = x_i^j$$
- $p_1, \ldots, p_n$ les poids des individus. On pose alors
$$D = \begin{pmatrix} p_1 & \ldots & 0 \\ \vdots & \ddots & \\ 0 & \ldots & p_n \end{pmatrix}$$
- On suppose $\mathbb{R}^d$ est muni d'une métrique M qui est une matrice $d \times d$ définie positive.
- $X = (x_i^j)$ est le tableau centré des données : L'origine $\underline{0} = (0, \ldots, 0) \in \mathbb{R}^d$ est le centre de gravité du nuage des individus $\mathcal{N}_I = \{ (x_i, p_i),\ i = 1 \ldots n \}$.

Inertie d'un nuage de points

- L'inertie du nuage des points $\mathcal{N}_I$ est définie par

$$I(\mathcal{N}_I) = \sum_{i=1}^{n} p_i \left[\text{dist}_M(0, x_i)\right]^2$$

- On a $\text{dist}(0, x_i)^2 = ||x_i||_M^2 = (x_i)' M x_i$
- Soit $a \in \mathbb{R}^d$ de $M-$norme 1, $||a|| = 1$, Le nuage obtenu par projection orthogonal sur $< a >$ est

$$\mathcal{N}_I(a) = \{ (x_i(a), p_i) = (< x_i, a > \ a, p_i), \ i = 1 \ldots n \}$$

- L'inertie de $\mathcal{N}_I(a)$ sera

$$I(\mathcal{N}_I)(a)) = \sum_{i=1}^{n} p_i \left[\text{dist}_M(0, x_i(a))\right]^2 = a' MVMa$$

L'objectif de l'ACP

Objectif : Chercher un vecteur unitaire $a \in \mathbb{R}^d$ tel que la projection orthogonale de $\mathcal{N}_I$ sur $\langle a \rangle$ soit d'inertie maximale.

Problème : Résoudre

$$\max_{a \in \mathbb{R}^d,\, a'Ma=1} a'MVMa \qquad (1)$$

Solution : Si les valeurs propres de VM sont $\lambda_1 > \lambda_2 > \ldots > \lambda_d$ alors a^1 vecteur propre $M-$unitaire associé à λ_1 est solution de (1), a^2 vecteur propre $M-$unitaire associé à la valeur propre λ_2 est solution de (1) $M-$othogonale à a^1...

Les résultats de l'ACP

- **Les composonates principales** $C^1, \ldots, C^d$: vecteurs contenants les coordonnées des individus sur les axes $\langle a^1 \rangle, \ldots, \langle a^d \rangle$:

$$\forall\ k = 1 \ldots d\ C^k = XMa$$

- **Les axes principaux** $a^1, \ldots, a^d$: les axes de projection des individus $\langle a^1 \rangle, \ldots, \langle a^d \rangle$
- **Les facteurs principaux** $u^1, \ldots, u^d$: les cœfficients de la combinaison linéaire dans l'écriture des composantes principales en fonction des variables :

$$\forall\ k = 1 \ldots d\ C^k = \sum_{j=1}^{d} u_j^k x^j \iff C^k = Xu^k$$

$u^1, \ldots, u^k$ sont les vecteurs propres $M^{-1}-$orthonormés de la matrice MV.

Produit Scalaire des Matrices

- $\mathcal{M}(n, d)$ est l'ensemble des matrices réelles centrées à n lignes et d colonnes
- D et M deux matrices diagonales d'éléments diagonaux positifs et de dimensions resp. $n \times n$ et $d \times d$.
- Pour tout $X = (x_i^j)$ et $Y = (y_i^j)$ dans $\mathcal{M}(n, p)$, on a

$$\langle X, Y \rangle = \mathrm{tr}(X'DYM) = \sum_{i,j} p_i m_j x_i^j y_i^j \qquad (2)$$

- Si X est une matrice centrée alors

$$\|X\|^2 = \sum_{i=1}^{n} p_j \left(\sum_{j=1}^{d} m_j (x_i^j)^2 \right) = \sum_{i=1}^{n} p_j d_M^2(0, x_i)$$

o'u x_i est la $i-$ème ligne de X (un point de $\mathbb{R}^d$).

Théorème d'Eckart-Young

Théorème

Soit $X \in \mathcal{M}(n,d)$ de rang r. Pour tout $k = 1, \ldots, r$, les cœfficients de la matrice

$$\widehat{X}^k = \operatorname{argmin}_{Y \in \mathcal{M}(n,d) \text{ et } rg(Y)=k} \|X - Y\|$$

se calculent par

$$\widehat{(x_i^j)}^k = \sum_{k=1}^{r} \frac{C_i^k u_j^k}{m_j} \tag{3}$$

o'u

- $(u^1, \ldots, u^r)$ *est une famille de vecteurs* $M^{-1}-$*orthonormées vecteurs propres de la matrice* $MV = MX'DX$.
- *pour tout* $l = 1 \ldots r, \quad C^l = Xu^l$.

Les données

- Sur un échantillon de n individus on a observé 2 variables qualitatives X et Y ayant respectivement I et J modalités.
- Les modalités de X (resp. de Y) sont $\{x_1, \ldots, x_I\}$ (resp. $\{y_1, \ldots, y_J\}$)
- On construit le tableau N de taille $I \times J$ de coefficients $n_{ij} =$ nombre d'individus $X = x_i$ et $Y = y_j$.
- On considère la matrice $F = N/n = (f_{ij})$ où f_{ij} est la fréquence d'observation de $X = x_i$ et $Y = y_j$.

Les profils-lignes

- Fréquence marginales de X, pour tout $i = 1, \ldots, I$

$$f_{i+} = \sum_{j=1}^{J} f_{ij}$$

- Les profils-lignes sont les vecteurs de $\mathbb{R}^J$

$$f_i^Y = \left(\frac{f_{i1}}{f_{i+}}, \ldots, \frac{f_{iJ}}{f_{i+}} \right)'$$

- La matrice des profils-lignes est la matrice de lignes les f_i^Y :

$$F^Y = D_I^{-1} F$$

où

$$D_I = \begin{pmatrix} f_{1+} & \cdots & 0 \\ \vdots & \ddots & \\ 0 & & f_{I+} \end{pmatrix}$$

Les profils-lignes (suite)

- Nuage des profils-lignes : c'est l'ensemble des points f_i^Y de $\mathbb{R}^J$ muni chacun du poids f_{i+} :

$$N_I = \left\{ (f_i^Y, f_{i+}),\ i = 1 \dots I \right\}$$

- Dans $\mathbb{R}^J$ on considère le produit scalaire de métrique D_J^{-1} o'u

$$D_J = \begin{pmatrix} f_{+1} & \dots & 0 \\ \vdots & \ddots & \\ 0 & & f_{+J} \end{pmatrix}$$

Donc

$$\forall\, x, y \in \mathbb{R}^J,\ \langle x, y \rangle_{D_J^{-1}} = \sum_{j=1}^{J} \frac{x_j y_j}{f_{+j}}$$

Les profils-colonnes

- Fréquence marginales de Y, pour tout $j = 1, \ldots, J$

$$f_{+j} = \sum_{i=1}^{I} f_{ij}$$

- Les profils-lignes sont les vecteurs de $\mathbb{R}^I$

$$f_j^X = \left(\frac{f_{1j}}{f_{+j}}, \ldots, \frac{f_{Ij}}{f_{+j}} \right)'$$

- La matrice des profils-colonnes est la matrice de lignes les f_j^X :

$$F^X = D_J^{-1} F'$$

où

$$D_J = \begin{pmatrix} f_{+1} & \cdots & 0 \\ \vdots & \ddots & \\ 0 & & f_{+J} \end{pmatrix}$$

Les profils-colonnes (suite)

- Nuage des profils-colonnes : c'est l'ensemble des points f_j^X de $\mathbb{R}^I$ muni chacun du poids f_{+j} :

$$N_J = \left\{ (f_j^X, f_{+j}), j = 1 \dots J \right\}$$

- Dans $\mathbb{R}^I$ on considère le produit scalaire de métrique D_I^{-1} où

$$\forall\, x, y \in \mathbb{R}^I,\ \langle x, y \rangle_{D_I^{-1}} = \sum_{i=1}^{I} \frac{x_i y_i}{f_{i+}}$$

La distance du χ^2

- la distance (du χ^2) deux points profils-lignes de f_i^Y et $f_{i'}^Y$

$$d_{\chi^2}(x,y)^2 = \sum_{j=1}^{J} \frac{1}{f_{+j}} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2$$

- **Equivalence distributionnelle** : On montre que si deux colonnes de F ont les mêmes profils-colonnes : $\frac{f_{ij}}{f_{+j}} = \frac{f_{ij'}}{f_{+j'}}$ et on les regroupe en un seul effectif $f_{ij} + f_{ij'}$ les distances entre les profils-lignes ne changent pas :

$$\frac{1}{f_{+j}} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2 + \frac{1}{f_{+j'}} \left(\frac{f_{ij'}}{f_{i+}} - \frac{f_{i'j'}}{f_{i'+}} \right)^2$$

$$= \frac{1}{f_{+j} + f_{+j'}} \left(\frac{f_{ij} + f_{ij'}}{f_{i+}} - \frac{f_{ij'} + f_{i'j'}}{f_{i'+}} \right)^2$$

Centre de gravité

- Le centre de gravité du nuage N_I (resp. N_J) des profils-lignes (resp. profils-colonnes) est

$$f^Y = (f_{+1}, \ldots, f_{+J})' \text{ (resp. } f^X = (f_{1+}, \ldots, f_{I+})'$$

- En effet

$$\sum_{i=1}^{I} f_{i+} f_i^Y = \sum_{i=1}^{I} f_{i+} \begin{pmatrix} \frac{f_{i1}}{f_{i+}} \\ \vdots \\ \frac{f_{iJ}}{f_{i+}} \end{pmatrix} = f^Y$$

ACP sur les profils-lignes

- Eléments de l'ACP :

individus	profils-lignes f_i^Y
tableau des données	$F^Y = D_I^{-1} F$
métrique	D_J^{-1}
poids	D_I

- Inertie du nuage N_I par rapport au centre de gravité :

$$\begin{aligned} I_{N_I} &= \sum_{i=1}^{I} f_{i+} d^2_{\chi^2}(f_i^Y, f^Y)^2 \\ &= \sum_{i,j} f_{i+}\, f_{j+} \left(\frac{f_{ij}}{f_{i+}\, f_{j+}} - 1 \right)^2 \end{aligned} \tag{4}$$

Inertie et Indépendances

- On considère le test

$$\mathrm{H}_0 \;\; X \perp\!\!\!\perp Y \quad \text{vs} \quad \mathrm{H}_1 \;:\; X \not\perp\!\!\!\perp Y$$

Sous H_0, la statistique $T = n\, I_{N_I}$ suit approximativement une loi de $\chi^2((I-1)(J-1))$

- Si $X = x_i$ est indépendant de $Y = y_j$ pour tout i, j alors $f_{ij} = f_{i+} f_{+j}$ donc pour tout j et i

$$f^X = f_j^X, \quad \text{et} \quad f^Y = f_i^Y$$

Matrice de covariance

La matrice covariance s'exprime :

$$V = (F^Y)' D_I (F^Y) - f^Y (f^Y)'$$

Proposition

La matrice VD_J^{-1}

- *admet* 0 *comme valeur propre de vecteur propre* f^Y.
- *a les les mêmes vecteurs propres associés aux mêmes valeurs de la matrice* $(F^Y)' D_I (F^Y) D_J^{-1}$ *sauf* f^Y *est un vecteur propre de* $(F^Y)' D_I (F^Y) D_J^{-1}$ *associé à la valeur propre 1.*

Remarque. $(F^Y)' D_I (F^Y) = F' D_I^{-1} D_I D_I^{-1} F = F' D_I^{-1} F$

Petits calculs...

- $D_J^{-1} f^Y = \underline{1}_J = (1, \ldots, 1)' \in \mathbb{R}^J$, $F\underline{1}_J = f^X$, $D_J \underline{1}_J = f^Y$.
- De même $D_I^{-1} f^X = \underline{1}_I = (1, \ldots, 1)' \in \mathbb{R}^I$, $F'\underline{1}_I = f^Y$, $D_I \underline{1}_I = f^X$.
- On conclut $VD_J^{-1} f^Y = 0$ et $(F^Y)' D_I (F^Y) D_J^{-1} f^Y = 0$
- Pour calculer composantes, axes, facteurs principaux il suffit de calculer les vecteurs propres $D_J^{-1}-$orthonormés de la matrice $(F^Y)' D_I (F^Y) D_J^{-1}$.

Calcul des composantes, axes et facteurs principaux

- **Axes principaux** : $f^Y, a^2, \ldots, a^J$ sont les vecteurs propres D_J^{-1}-orthonormés de la matrice $F' D_I^{-1} F D_J^{-1}$.
- **Facteurs pricipaux** : $\underline{1}_J, u^2, \ldots, u^J$ sont les vecteurs propres D_J-orthonormés de la matrice $D_J^{-1} F' D_I^{-1} F$
- **Composantes principales** : $\underline{1}_I, C^2, \ldots, C^J$ où $C^k = D_I^{-1} F u^k = D_I^{-1} F D_J^{-1} a^k$ pour tout $k = 1 \ldots I$.
  Les C^k sont les vecteurs propres de la matrice $D_I^{-1} F D_J^{-1} F'$.
  Les C^k sont 2 à 2 D_I-orthogonales.

Contributions, Qualités

- Contributions : C_i^k étant la coordonnée du profil-ligne ${f_i}^Y$ sur le $k-$ème axe principale $\langle a_k \rangle$. Donc sa contribution à l'inertie de cet axe pour $k \geq 2$

$$CTR_k({f_i}^Y) = \frac{f_{i+}(C_i^k)^2}{\lambda_k}$$

- La Qualité de la représentation est alors

$$QLT_k({f_i}^Y) = \frac{(C_i^k)^2}{d^2_{\chi^2}({f_i}^Y, f_Y)}$$

Reconstitution des données

- Appliquant (3) à $X = D^{-1}F$, $M = D_J^{-1}$ et $D = D_I$, pour tout i, j

$$f_{ij} = f_{i+}\, f_{+j} \left(1 + \sum_{k=2}^{J} C_i^k\, u_j^k \right)$$

- Donc $\sum_{k=2}^{J} C_i^k\, u_j^k$ mesure l'écart par rapport à l'hypothèse d'indépendance.

ACP sur les profils-colonnes

- Eléments de l'ACP :

individus	profils-colonnes f_j^X
tableau des données	$F^X = D_J^{-1} F'$
métrique	D_I^{-1}
poids	D_J

- Inertie du nuage N_J par rapport au centre de gravité :

$$\begin{aligned} I_{N_J} &= \sum_{j=1}^{J} f_{j+} d^2_{\chi^2}(f_j^X, f^X)^2 \\ &= \sum_{i,j} f_{i+}\, f_{j+} \left(\frac{f_{ij}}{f_{i+}\, f_{j+}} - 1 \right)^2 \end{aligned} \tag{5}$$

Calcul des composantes, axes et facteurs principaux

- **Axes principaux** : $f^X, b^2, \ldots, b^I$ sont les vecteurs propres D_I^{-1}-orthonormés de la matrice $FD_J^{-1}F'D_I^{-1}$.
- **Facteurs principaux** : $\underline{1}_I, v^2, \ldots, v^I$ sont les vecteurs propres D_I-orthonormés de la matrice $D_I^{-1}FD_J^{-1}F'$
- **Composantes principales** : $\underline{1}_J, \widetilde{C}^2, \ldots, \widetilde{C}^I$ où $\widetilde{C}^k = D_J^{-1}\,F'v^k = D_J^{-1}\,F'D_I^{-1}b^k$ pour tout $k = 1 \ldots J$.
  Les $\widetilde{C}^k$ sont les vecteurs propres de la matrice $D_J^{-1}F'D_I^{-1}F$.
  Les $\widetilde{C}^k$ sont 2 à 2 D_J-orthogonales.

Relation entre a^k et b^k

- $b^k = \frac{1}{\sqrt{\lambda_k}} F D_J^{-1} a^k$. En effet

$$\underbrace{F D_J^{-1} F' D_I^{-1} \underbrace{F D_J^{-1} a^k}}_{=\lambda_k a^k} = \lambda_k \underbrace{F D_J^{-1} a^k}$$

De même $a^k = \frac{1}{\sqrt{\lambda_k}} F' D_I^{-1} b^k$.

Relation entre C^k et $\widetilde{C}^k$

- $\widetilde{C}^k = \frac{1}{\sqrt{\lambda_k}} D_J^{-1} F' C^k$. En effet,

$$\widetilde{C}^k = D_J^{-1} F' D_I^{-1} b^k = \frac{1}{\sqrt{\lambda_k}} D_J^{-1} F' \underbrace{D_I^{-1} F D_J^{-1} a^k}_{=C^k}$$

De m ?me $C^k = \frac{1}{\sqrt{\lambda_k}} D_I^{-1} F \widetilde{C}^k$

- Donc pour tout i, j, on a les relations barycentriques :

$$C_i^k = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{r} \frac{f_{ij}}{f_{i+}} \widetilde{C}_j^k \quad (6)$$

$$\widetilde{C}_j^k = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{r} \frac{f_{ij}}{f_{+j}} C_i^k \quad (7)$$

Interprétation

- De la relation (6), on conclut que la représentation d'un profil-ligne f_i^Y sur le $k-$ème axe principal est le barycentre (à $\sqrt{\lambda_k}$ près) des profils-colonnes f_i^X munis des poids $\frac{f_{ij}}{f_{i+}}$:

- Dans une représentation simultanée : Un profil-ligne f_i^Y proche d'un profil-colonne f_i^X indiquerait une forte correspondance entre les deux modalités.

Elements supplémenaires

- On dispose de J_s colonnes supplémentaires qui concernent la variables Y et on veut situer ces points-colonnes par rapport aux J pofils-colonnes analysés.
- j est une colonne supplémentaire, son profil-colonne

$$(f_j^X)_s = \left(\frac{f_{1j}}{+j}, \ldots, \frac{f_{Ij}}{+j} \right)'$$

- Pour représenter $(f_j^X)_s$ on projète ce point-colonne sur les axes-principaux. Sa coordonnée sur le $k-$ème axe

$$(\widetilde{C}_j^k)_s = (f_j^X)_s' D_I^{-1} b^k$$

- On a aussi la relation barycentrique

$$(\widetilde{C}_j^k)_s = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{r} \frac{f_{ij}}{f_{+j}} C_i^k$$

Mise en œuvre des calculs

La matrice à diagonaliser est $F' D_I^{-1} F D_J^{-1}$ de dimensions $J \times J$. On pose $\widehat{A} = F' D_I^{-1} F$ qui est symétrique

Proposition

La matrice $A = D_J^{-1/2} \widehat{A} D_J^{-1/2}$ est symétrique et a les mêmes valeurs propres que $F' D_I^{-1} F D_J^{-1}$. Les vecteurs propres sont liés par

$$a = D_J^{1/2} w$$

Preuve.

$$\underbrace{F' D_I^{-1} F}_{\widehat{A}} D_J^{-1} a = a\Delta \Rightarrow D_J^{-1/2} \widehat{A} D_J^{-1/2} \underbrace{D_J^{-1/2} a}_{=w} = \underbrace{D_J^{-1/2} a}_{=w} \Delta$$

■

Données

- On observe sur n individus p variables qualitatives $x^1, \ldots, x^p$.
- Chaque variable x^j a k_j modalités.
- Chaque individu peut être muni d'un poids p_i, $i = 1, \ldots, n$.
- On transforme notre matrice données originale en une matrice $Z = (z_i^l)$ de dimensions

$$n \times \sum_{j=1}^{p} k_j$$

telle que $z_i^l = 1/0$ selon la modalité prise par w_i par rapport à une variable x^j.

L'ACP est une ACM

- Les poids des individus sont p_i.
- La matrice $X = (x_i^l)$ des l'ACP est de dimensions $n \times \sum_{j=1}^{p} k_j$

$$x_i^l = \begin{cases} -1 & \text{si } z_i^l = 0 \\ \dfrac{1}{f(k_j)} - 1 & \text{si } z_i^l = 1 \end{cases}$$

où $f(k_j)$ est la fréquence de la modalité k_j.

- Les poids sur les variables (les colonnes de X) sont $f(k_j)/p$.