Data Mining

Analyse Factorielle des Correspondances

W. Toussile
wilson.toussile@gmail.com

[1]Département MSP
École Nationale Supérieure Polytechnique

06/01/2020

Section 1

Introduction

Introduction

- Proposée par J.P. Benzécri, l'AFC (binaire) permet d'étudier la liaison (correspondance) entre deux variables qualitatives.
- Elle repose sur l'analyse du table de contingence associé à deux variables qualitatives X et Y sur un ensemble de n individus, et certains tableaux binaires (dits d'indicatrices).
- L'AFC peut être vue comme une ACP associée à la métrique du chi2.
- C'est une méthode très utilisée en analyse des données textuelles.

Position du pb

- Données: $\{(x_i, y_i)\}_{i=1}^{n}$
- $(x_i, y_i) \in X(\Omega) \times Y(\Omega)$, observations d'un couple (X, Y) de variables qualitatives avec les modalités $X(\Omega) = \{a_1, \cdots, a_k, \cdots, a_K\}$ et $Y(\Omega) = \{b_1, \cdots, b_l, \cdots, b_L\}$ resp.

Problème

Représentater les liens (correspondances) entre les modalités de X et celles de Y en dimension réduite, en général 2, en perdant le moins d'information possible.

Section 2

Concepts

Table de contingence

$X \mid Y$	$\cdots$	b_l^Y	$\cdots$	Total
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k^X	$\cdots$	$n_{k,l}$	$\cdots$	$n_{k,+} = \sum_l n_{k,l}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Total	$\vdots$	$n_{+,l} = \sum_k n_{k,l}$	$\vdots$	$n = \sum_{k,l} n_{k,l}$

- On note $N = (n_{k,l})_{k,l} \in \mathbb{N}^{K \times L}$

Table des fréquences

$X \mid Y$	$\cdots$	b_l^Y	$\cdots$	Total
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k^X	$\cdots$	$f_{k,l} = \frac{n_{k,l}}{n}$	$\cdots$	$f_{k,+} = \frac{n_{k,+}}{n} = \sum_l f_{k,l}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Total	$\vdots$	$f_{+,l} = \frac{n_{+,l}}{n} = \sum_k f_{k,l}$	$\vdots$	1

- On note $F = (f_{k,l})_{k,l} \in [0,1]^{K \times L}$

Profils

Profils lignes

- Ligne k: $L_k = \left(\frac{n_{k,l}}{n_{k,+}} \right)_l \in \mathbb{R}^L$
- Matrice des profils lignes: $L = \left(\frac{n_{k,l}}{n_{k,+}} \right)_{k,l} \in \mathbb{R}^{K \times L}$
- Profil ligne moyen: $\left(\sum_k f_{k,+} \frac{f_{k,l}}{f_{k,+}} \right)_l = (f_{+,l})_l \in \mathbb{R}^L$

Exo

On pose

$$D_{K,.} = diag\left((f_{k,+})_k \right).$$

Montrer que $L = D_{K,.}^{-1} F$

Profils

Profils colonnes

- Colonne l: $C_l = \left(\frac{n_{k,l}}{n_{+,l}} \right)_l \in \mathbb{R}^K$
- Matrice des profils colonnes: $C = \left(\frac{n_{k,l}}{n_{+,l}} \right)_{k,l} \in \mathbb{R}^{K \times L}$
- Profil colonne moyen: $\left(\sum_l f_{+,l} \frac{f_{k,l}}{f_{+,l}} \right)_k = (f_{k,+})_k \in \mathbb{R}^K$

Exo

On pose

$$D_{\cdot,L} = diag\left((f_{+,l})_l\right).$$

Montrer que $C = D_{\cdot,L}^{-1} {}^t F$

Distance du Khi 2 entre profils

Profils lignes

$$d^2\left(L_k, L_{k'}\right) := \sum_l \frac{1}{f_{+,l}} \left(\frac{f_{k,l}}{f_{k,+}} - \frac{f_{k',l}}{f_{k',+}} \right)^2$$

Exo

Montrer que la distance du Khi 2 entre profils lignes est définie par la métrique $D_{\cdot,L}^{-1}$

Distance du Khi 2 entre profils

Profils Colonnes

$$d^2\left(C_l, C_{l'}\right) := \sum_k \frac{1}{f_{k,+}} \left(\frac{f_{k,l}}{f_{+,l}} - \frac{f_{k,l'}}{f_{+,l'}} \right)^2$$

Exo

Montrer que la distance du Khi 2 entre profils colonnes est définie par la métrique $D_{K,\cdot}^{-1}$

Section 3

Test d'indépendance du Khi 2

Test d'indépendance du Khi 2

Les hypothèses

$$\begin{aligned} \mathcal{H}_0 &: \text{Les deux variables sont indépendantes} \\ \mathcal{H}_1 &: \text{Les deux variables ne sont pas indépendantes} \end{aligned}$$

- Cette question est globale par rapport à l'objectif de l'AFC
- L'AFC permet d'examiner plus finement les liens (correspondances) entre les modalités des deux variables.
- Sous l'hypothèse $\mathcal{H}_0$, la fréqeunce attendue de la modalité (a_k, b_l) est

$$\widehat{f}_{k,l} = f_{k,+} f_{+,l}$$

Test d'indépendance du Khi 2

Stat. de test et décision

$$\mathbb{X}^2 = n \sum_k \sum_l \frac{\left(f_{k,l} - \widehat{f}_{k,l}\right)^2}{\widehat{f}_{k,l}} \xrightarrow{\mathcal{H}_0} \mathcal{X}^2_{(K-1)(L-1)}$$

Au seuil $\alpha \in]0,1[$ (en général $\alpha = 5\%$), on rejette $\mathcal{H}_0$ si et seulement si

$$\mathbb{X}^2_{obs} > q_{1-\alpha}\left(\mathcal{X}^2_{(K-1)(L-1)}\right)$$

Section 4

AFC (binaire)

AFC (binaire)

Note

- Reppelons que l'objectif est de représenter les profils lignes et colonnes dans un espace de dimension réduite (en général 2), de sorte à conserver au mieux les distances entre profils lignes et profils colonnes.
- La proximité entre une modalité a_k de X et une modalité b_l de Y représente alors le lien (la correspondance) positive entre a_k et b_l
- L'AFC est une double ACP: ACP sur les profils lignes et ACP sur les profiles colonnes
- Notons:
  - $\overline{l} = (f_{+,l})_l \in \mathbb{R}^L$ le profil ligne moyen
  - $\overline{c} = (f_{k,+})_k \in \mathbb{R}^K$ le profil colonne moyen

AFC (binaire)

Données pour les ACP

ACP	AFC lignes	AFC colonnes
Données centées	$L_c = D_{K,\cdot}^{-1} F - 1_K {}^t\bar{l}$	$C_c = D_{\cdot,L}^{-1\,t} F - 1_L {}^t\bar{c}$
Poids	$D_{K,\cdot}^{-1}$	$D_{\cdot,L}^{-1}$
Métrique	$D_{\cdot,L}^{-1}$	$D_{K,\cdot}^{-1}$
C.P.	$L_c D_{\cdot,L}^{-1} U$	$C_c D_{K,\cdot}^{-1} V$

Exemple

On s'interesse à la couleur des yeux et celle des cheveux de $n = 592$ femmes. Les données sont résumées dans le tableau suivant:

eux | Cheveux	Chatains	Roux	Blonds
Marrons	119	26	7
Noisette	54	14	10
Verts	29	14	16
Bleus	84	17	94

- Existe-t-il un lien entre couleur des yeux et couleur des cheveux?
- Si oui, quelles sont les correspondances?