PROGRAMME POUR LE CALCUL DE COEFFICIENTS D'ASSOCIATION ENTRE VARIABLES RELATIONNELLES

Mohamed OUALI ALLAH

C.R.E.C. Saint-Cyr Coëtquidan 56381 GUER CEDEX
Tel : 02 97 73 50 43 Fax : 02 97 73 50 83
E-mail : ouali@mailhostesm-stcyr.terre.defense.gouv.fr

I.R.I.S.A. Campus Universitaire de Beaulieu 35042 RENNES CEDEX
Tel : 02 99 84 74 79 Fax : 02 99 84 71 71 E-mail : ouali@irisa.fr

Le programme (code Fortran77, exécutable et données) est sur le WEB à l'adresse de la revue. Pour toute information veuillez contacter l'auteur.

Introduction

L'objet du programme, nommé AVARE, est l'élaboration de la matrice des coefficients d'association entre variables qualitatives de différents types.

Dans notre approche, toutes les variables qualitatives sont considérées comme des variables préordonnances : l'ensemble des couples de modalités de chaque descripteur est muni d'une relation de préordre total. Celle-ci est quantifiée par la notion de rang moyen. Cette préordonnance peut être fournie directement par le praticien, sinon elle sera établie par le programme à partir d'un graphe valué.

Chaque variable engendre par conséquent une relation binaire sur l'ensemble des objets. Il s'agit alors de définir une mesure de similarités entre ces variables, basée sur le produit scalaire des valuations induites par les relations qu'elles engendrent.

Les similarités entre ces variables sont mesurées par un coefficient d'association, dont la normalisation par rapport à une hypothèse d'indépendance, est de nature combinatoire et statistique.

Plan de l'article :

1. Système de codage

A chaque variable qualitative est associé un préordre total sur l'ensemble des couples de ses modalités. Ce préordre appelé préordonnance, peut être préalablement établie par l'expert, pour concrétiser ainsi les particularités de telle ou telle variable. Dans le cas contraire, la préordonnance est construite à partir d'un graphe valué qui caractérise la structure associée à l'ensemble des modalités de chaque variable.

1.1. Préordonnance associée à un graphe valué

1.1.1 Graphe valué

Chaque variable qualitative ω est représentée par le graphe valué :

$$G_\omega = \langle M_\omega, \Gamma_\omega, f_\omega \rangle$$

Où :

— $M_\omega = \{1, 2, \ldots, m_\omega\}$ est l'ensemble d'indexation des m_ω modalités de ω,

— Γ_ω est un sous-ensemble de $M_\omega \times M_\omega$,

— $f_\omega : \Gamma_\omega \to \mathbb{R}$ est une fonction de valuation.

On distingue notamment deux cas :

- Si M_ω est sans structure ordinale, on a : $\Gamma_\omega = P_2(M_\omega)$ (ensemble des paires de modalités de ω). Dans ce cas, la description de deux objets consiste simplement à les « réunir » ou à les « séparer », d'où :

$$f_\omega : P_2(M_\omega) \to \mathbb{R}$$

$$\{k, l\} \to \delta_{kl} = \begin{cases} 1 & \text{si } k = l \\ 0 & \text{sinon} \end{cases}$$

Cas particulier : Si M_ω a une structure booléenne, les deux modalités *vrai* (*présence*) et *faux* (*absence*) ne sont pas équivalentes. La *présence* du caractère chez deux objets est considérée comme plus significative que leur *absence*. Ce choix peut être inversé par l'expert qui décide laquelle des deux modalités jouera le rôle de *présence* :

$$f_\omega(\{k, l\}) = \begin{cases} 2 & \text{si } k = l = \textit{présence} \\ 1 & \text{si } k = l = \textit{absence} \\ 0 & \text{si } k \neq l \end{cases}$$

- Si M_ω est structuré ordinalement, on a : $\Gamma_\omega = M_\omega \times M_\omega$ (ensemble des couples de modalités de ω). Les objets dans ce cas, ne sont pas seulement séparés ou réunis par la variable, mais en cas de séparation on doit tenir compte de l'amplitude et du sens de celle-ci, d'où :

$$f_\omega : M_\omega \times M_\omega \to \mathbb{R}$$

$$(k, l) \to k - l$$

Cas particulier : dans le cas d'une comparaison avec une variable sans structure ordinale on a : $\Gamma_\omega = P_2(M_\omega)$, on considère donc la restriction de la fonction ci-dessus à $P_2(M_\omega)$:

$$f_\omega : P_2(M_\omega) \rightarrow \mathbb{R}$$
$$\{k, l\} \rightarrow |k - l|$$

Les fonctions de valuation ainsi définies, induisent un préordre total $\prec_\omega$ sur Γ_ω :

$$\forall (x, y) \in \Gamma_\omega \times \Gamma_\omega \qquad x \prec_\omega y \Leftrightarrow f_\omega(x) \leq f_\omega(y)$$

1.1.2 Matrices de rangs

La préordonnance définie ci-dessus est ensuite quantifiée, en affectant un rang à chacun de ses éléments (constitué des deux modalités k et l). En présence d'ex æquo, on attribue aux éléments de la classe d'ex æquo, la moyenne arithmétique des rangs qu'ils auraient eu s'ils étaient totalement ordonnés. Ces rangs —dits *moyens*— présentent l'avantage d'avoir une somme constante, quel que soit le préordre choisi.

On procède enfin, à un "centrage-réduction" de cette table de rangs pour aboutir à des rangs —notés r_{kl}^{ω} — dans l'intervalle [–1, +1].

Les matrices de rangs $R_\omega = \left(r_{kl}^{\omega}\right)$ ainsi obtenues, sont antisymétriques ou symétriques à diagonale nulle. Elles caractérisent chaque variable, et tiennent compte de la nature de la variable à laquelle elle est comparée. On montre dans [OUA91b] que :

- Si la variable ω est sans structure ordinale, on a :

$$r_{kl}^{\omega} = \begin{cases} 0 & \text{si } k = l \\ -1 & \text{sinon} \end{cases}$$

Cas particulier : dans le cas booléen on a :

$$r_{kl}^{\omega} = \begin{cases} 1 & \text{si } k = l = présence \\ 0 & \text{si } k = l = absence \\ -1 & \text{si } k \neq l \end{cases}$$

- Si la variable ω est structurée ordinalement, on a :

$$r_{kl}^{\omega} = \frac{(k-l)(2m_{\omega} - |k-l|)}{m_{\omega}^2 + 1}$$

Cas particulier : dans le cas d'une comparaison avec une variable sans structure ordinale on a :

$$r_{kl}^{\omega} = \frac{|k-l|(|k-l| - 2m_{\omega})}{m_{\omega}^2 + 1}$$

1.2. Préordonnance fournie

Il s'agit des variables préordonnances au sens classique du terme, c'est à dire que le praticien fournit un préordre total sur l'ensemble des paires de modalités de chaque variable.

A chaque paire de modalités est attribué son rang moyen. On effectue là aussi le centrage-réduction de la table de rangs pour aboutir à des demi-matrices à valeurs dans l'intervalle [–1, +1].

Ces demi-matrices seront ensuite complétées par :

- symétrie, lorsque la variable ou celle à laquelle elle sera comparée est sans structure ordinale,

- antisymétrie, lorsque la variable et celle à laquelle elle sera comparée sont structurées ordinalement

2. Mesure de similarités

2.1 Un critère général

Chaque variable est représentée par une relation binaire sur l'ensemble d'objets Ω. Ainsi une variable relationnelle ω définit une matrice dite de codage (ou de pondération) sur l'ensemble $\Omega \times \Omega$:

$$C_\omega = \left(c_{ij}^{\omega}\right)_{(i,j)\in I\times I}$$

Où $I = \{1, \ldots, i, \ldots, n\}$ est l'ensemble d'indexation de Ω.

On considère alors — relativement à la comparaison de deux variables ω et ϖ — un critère d'association classique, produit scalaire (hors diagonale) des valuations engendrées par les deux variables sur l'ensemble des couples d'objets :

$$s(\omega,\varpi) = \sum_{i\neq j} c_{ij}^{\omega} c_{ij}^{\varpi}$$

L'étude et la normalisation de cet indice ont fait l'objet de nombreux travaux, et ce, dans différents contextes. Le nôtre s'inscrit dans la méthode de la vraisemblance du lien [LER81], où une hypothèse d'indépendance, à caractère permutationnel, consiste à associer à l'indice s — dit brut — deux variables aléatoires duales :

$$S_1 = s(\omega, \varpi^*) = \sum_{i \neq j} c_{ij}^{\omega} c_{\tau(i)\tau(j)}^{\varpi}$$

$$S_2 = s(\omega^*, \varpi) = \sum_{i \neq j} c_{\tau(i)\tau(j)}^{\omega} c_{ij}^{\varpi}$$

Où τ est une permutation aléatoire sur *I*, muni d'une probabilité uniforme.

Ces deux variables aléatoires S_1 et S_2 ont une même distribution, et on montre [LER81] sous des conditions assez générales, que cette distribution commune —notée S— est asymptotiquement normale.

2.2 Coefficient centré-réduit

La normalisation statistique (centrage-réduction) de l'indice *s*, qui nécessite le calcul de l'espérance μ et surtout de l'écart-type σ de la variable aléatoire *S*, permet la définition du Coefficient Centré-Réduit (CCR) :

$$Q(\omega, \varpi) = \frac{s(\omega, \varpi) - \mu}{\sigma}$$

Les expressions du CCR ([LER87], [MAN67]) sont fort complexes et difficilement interprétables. Nous avons repris d'une façon plus analytique [OUA91a], les calculs de normalisation en distinguant le cas d'un codage symétrique de celui d'un codage antisymétrique.

Nous avons ainsi, abouti à une expression plus synthétique du CCR, en introduisant pour les matrices de codage, des moments factoriels centrés d'ordre 2 (où *n(n-1)* est noté $n^{[2]}$)

- covariance entre $\mathscr{C}_\omega$ et $\mathscr{C}_\varpi$:

$$\Phi_{\omega\varpi} = \frac{1}{n^{[2]}} \sum_{i \neq j} c_{ij}^{\omega} c_{ij}^{\varpi} - \left(\frac{1}{n^{[2]}} \sum_{i \neq j} c_{ij}^{\omega} \right) \left(\frac{1}{n^{[2]}} \sum_{i \neq j} c_{ij}^{\varpi} \right)$$

- variance de $\mathscr{C}_\omega$:

$$\Phi_\omega = \Phi_{\omega\omega} = \frac{1}{n^{[2]}} \sum_{i \neq j} c_{ij}^{\omega 2} - \left(\frac{1}{n^{[2]}} \sum_{i \neq j} c_{ij}^{\omega} \right)^2$$

- variance des marges de $\mathscr{C}_\omega$

$$\Psi_\omega = \frac{1}{n^{[2]}(n-1)} \sum_i \left(\sum_{j,j \neq i} c_{ij}^{\omega} \right)^2 - \left(\frac{1}{n^{[2]}} \sum_{i \neq j} c_{ij}^{\omega} \right)^2$$

L'une des formes simplifiées du CCR [OUA91b] s'écrit alors :

$$Q(\omega, \varpi) = \frac{\sqrt{n}}{2} \frac{\Phi_{\omega\varpi}}{\sqrt{\Psi_\omega \Psi_\varpi + \frac{1}{2n}\left(\Phi_\omega - 2\Psi_\omega\right)\left(\Phi_\varpi - 2\Psi_\varpi\right)}}$$

Nous montrons par ailleurs que les moments Φ et Ψ sont asymptotiquement indépendants de n. D'où, lorsque n est suffisamment grand, le deuxième terme de la variance devient négligeable devant le premier. Nous aboutissons ainsi à la *forme limite* du CCR :

$$Q(\omega, \varpi) \approx \frac{\sqrt{n}}{2} \frac{\Phi_{\omega\varpi}}{\sqrt{\Psi_\omega \Psi_\varpi}}$$

Enfin, le recours à une « réduction géométrique » du CCR, en le divisant par la moyenne géométrique des deux coefficients diagonaux correspondants ($Q(\omega,\varpi) = \frac{Q(\omega,\varpi)}{\sqrt{Q(\omega,\omega)Q(\varpi,\varpi)}}$), nous permet d'obtenir la *forme corrélative* du CCR, qui n'est autre que le coefficient de corrélation linéaire entre les matrices de codage :

$$Q(\omega, \varpi) \equiv \frac{\Phi_{\omega\varpi}}{\sqrt{\Phi_{\omega}\Phi_{\varpi}}}$$

2.2.1 Formulation relationnelle du CCR

Pour exprimer le CCR entre variables qualitatives, il faut préciser le codage qui leur est associé. Un objet i (resp. j) possède une modalité k (resp. l) de la variable ω. Dans ce cas on associe au couple d'objets (i, j) le rang moyen du couple de modalités (k, l). Soit :

$$c_{ij}^{\omega} = r_{kl}^{\omega}$$

On applique alors les expressions du CCR du paragraphe 1.2.2 pour obtenir la formulation relationnelle de la mesure de similarités.

2.2.2 Formulation contingentielle

Malgré leur relative compacité, ces formulations relationnelles du CCR ne se prêtent guère au calcul puisqu'elles impliquent un encombrement spatial considérable du fait de la manipulation de matrices de codage d'ordre n^2 (le nombre d'objets n étant généralement très grand).

On transfère alors, le support du codage de l'ensemble des couples d'objets à celui des couples de modalités (dont le nombre dépasse rarement quelques unités), en utilisant les tables de contingences.

Par conséquent, les moments définis dans le paragraphe 1.2.2 ne sont plus des sommations sur les $\left(c_{ij}^{\omega}\right)$, mais sur les $\left(r_{kl}^{\omega}\right)$ pondérées par les «effectifs » des tables de contingences (les expressions du CCR ainsi obtenues sont dites contingentielles[1]) :

---

[1] Les expressions précises des deux formulations relationnelle et contingentielle sont développées pour tous les cas de figure dans [OUA91b]

- covariance entre $\mathscr{C}_\omega$ et $\mathscr{C}_\varpi$:

$$\Phi_{\omega\omega} = \frac{1}{n^{[2]}} \sum_{k \neq l} n_{k\bullet} n_{l\bullet} \left(r_{kl}^{\omega}\right)^2 - \left(\frac{1}{n^{[2]}} \sum_{k \neq l} n_{k\bullet} n_{l\bullet} r_{kl}^{\omega} \right)$$

- variance de $\mathscr{C}_\omega$:

$$\Phi_{\omega\varpi} = \frac{1}{n^{[2]}} \sum_{k \neq l} \sum_{p \neq q} n_{kp} n_{lq} r_{kl}^{\omega} r_{pq}^{\varpi} - \left(\frac{1}{n^{[2]}} \sum_{k \neq l} n_{k\bullet} n_{l\bullet} r_{kl}^{\omega} \right) \left(\frac{1}{n^{[2]}} \sum_{p \neq q} n_{\bullet p} n_{\bullet q} r_{pq}^{\varpi} \right)$$

- variance des marges de $\mathscr{C}_\omega$:

$$\Psi_{\omega} = \frac{1}{n^{[2]}(n-1)} \sum_{k} n_{k\bullet} \left(\sum_{l,k \neq l} n_{l\bullet} r_{kl}^{\omega} \right)^2 - \left(\frac{1}{n^{[2]}} \sum_{k \neq l} n_{k\bullet} n_{l\bullet} r_{kl}^{\omega} \right)^2$$

Où :

- $n_{k\bullet}$ *est le nombre d'objets possédant la modalité k de la variable* ω *et la modalité p de la variable* ϖ ;

- $n_{k\bullet}$ *(resp.* $n_{\bullet p}$*) est le nombre d'objets possédant la modalité k de la variable* ω *(resp. la modalité p de la variable* ϖ*)*

Remarques :

- *Comme les matrices de rangs sont symétriques ou antisymétriques, ces sommations s'effectuent sur les matrices triangulaires* $\left(r_{kl}^{\omega}\right)_{k<l}$,

- *La deuxième partie de ces moments s'annule quand ces matrices sont antisymétriques (cas des variables à modalités ordonnées).*

Conclusion

La mesure de similarité entre variables qualitatives que nous avons développée, s'apparente donc à un coefficient de corrélation linéaire entre les matrices de codage qu'ils engendrent sur l'ensemble des objets.

D'où l'importance d'une représentation suffisamment intégrante pour assimiler toutes les caractéristiques des données et suffisamment souple pour adopter et formaliser les connaissances du domaine d'application. Dans ce sens, le codage par préordonnance peut intégrer et transcrire fidèlement l'information véhiculée par des données à modalités.

BIBLIOGRAPHIE

[LER81] I.C. Lerman Classification et analyse ordinale des données, Dunod, Paris, 1981.

[LER87] I.C. Lerman Analyse de la forme limite de coefficients statistiques d'association entre variables relationnelles, P.I. No 367 I.R.I.S.A., 1987.

[MAN67] N. Mantel Detection of disease clustering and a generalized regression approach, Cancer Research vol. 27, 1967.

[OUA91a] M. Ouali Allah Variables Relationnelles : Codage et Association, P.I. No 569 I.R.I.S.A., 1990, No 1399 I.N.R.I.A., 1991.

[OUA91b] M. Ouali Allah Analyse en préordonnances des données qualitatives. Applications aux données numériques et symboliques, Thèse de l'université de Rennes I, 1991.