MAXIMISER L'ASSOCIATION PAR AGRÉGATION DANS UN TABLEAU CROISÉ

Gilbert RITSCHARD, Djamel A. ZIGHED,
Nicolas NICOLOYANNIS[1]

RÉSUMÉ — L'intensité de l'association entre la variable ligne et la variable colonne d'un tableau croisé varie avec le regroupement de catégories. Dans plusieurs contextes, comme la discrétisation simultanée de deux variables, il importe de déterminer le niveau de regroupement qui maximise l'association. Cet article étudie comment se comportent les principales mesures d'association suite à une agrégation de lignes ou colonnes. Une heuristique est ensuite proposée pour déterminer le regroupement qui (quasi-)maximise le degré d'association.

MOTS-CLÉS — table de contingence, agrégation, association, discrétisation

SUMMARY — *Maximizing Association by Aggregation in a Crosstable*: The strength of association between the row and column variables in a cross table varies with the level of aggregation of each variable. In many settings like the simultaneous discretization of two variables, it is useful to determine the aggregation level that maximizes the association. This paper deals with the behavior of the main association measures with respect to the aggregation of rows and columns and proposes an heuristic algorithm to (quasi-)maximize the association through aggregation.

KEYWORDS — crosstable, aggregation, association, discretization

1. INTRODUCTION

L'étude de l'association entre variables catégorielles repose en général sur une analyse du tableau croisé des variables concernées. Ainsi, on commence par tester l'indépendance en examinant les statistiques du khi-2 de Pearson ou du rapport de vraisemblance associées au tableau, puis, pour saisir l'intensité de l'association on se réfère à des mesures d'associations telles que v de Cramer, τ ou γ de Goodman et

[1]Adresser toute correspondance à G. Ritschard, Département d'économétrie, Université de Genève, bd du Pont-d'Arve 40, CH-1211 Genève 4. Travail réalisé en hiver 2000 pendant que G. Ritschard, professeur de statistique pour sciences sociales à l'Université de Genève, était invité par le laboratoire ERIC de l'Université Lyon 2. Les professeurs D. A. Zighed et N. Nicoloyannis sont membres d'ERIC, laboratoire de recherche dont le premier est directeur. E-mail : *gilbert.ritschard@themes.unige.ch, {zighed,nicolas.nycoloyannis}@univ-lyon2.fr* .

Kruskal, coefficient d'incertitude de Theil, τ_b de Kruskal, d de Somers, qui sont elles mêmes calculées sur la base des informations fournies par la table de contingence.

La question abordée dans cet article est celle de l'effet de l'agrégation de lignes ou colonnes d'un tableau croisé sur les mesures d'association. Il est bien connu par exemple que le regroupement de colonnes ou lignes qui ont la même distribution conditionnelle n'affecte pas les statistiques du khi-2 de Pearson et du rapport de vraisemblance (voir par exemple [1] ou [6], p. 450) mais renforce en général le degré d'association comme l'illustre en particulier les simulations présentées dans [7]. Notre propos est ici d'étudier ce lien entre fusion de lignes ou colonnes et degré d'association dans le but de déterminer le regroupement optimal, à savoir celui qui donne lieu à la plus forte association entre les variables ligne et colonne.

La maximisation du degré d'association trouve sa motivation dans plusieurs domaines. Par exemple, l'analyse de données collectées par questionnaires nécessite en général, pour des raisons d'effectifs notamment, un regroupement a posteriori des items en catégories. Lorsqu'il s'agit, comme c'est souvent le cas en sciences sociales, d'étudier l'association entre variables, il est utile de comprendre les effets du regroupement sur l'association et de choisir le cas échéant celui qui rend le mieux compte du lien. Une seconde motivation a trait à la discrétisation qui est en particulier un problème majeur en apprentissage. Les solutions optimales par rapport à la discrimination recherchée que propose la littérature et répertoriées par exemple dans [9] procèdent individuellement pour chaque variable. Dans une optique prédictive, des solutions bidimensionnelles, où l'on discrétise conjointement deux variables, devraient s'avérer plus efficaces. Breiman et al. [4] ont étudié le cas de la dichotomisation conjointe de deux variables. Notre propos est de généraliser ce cas à un nombre quelconque de catégories.

Hormis les cas triviaux avec un nombre initial réduit, trois par exemple, de catégories, rechercher la solution optimale par recombinaison systématique ne peut être envisagé à cause du nombre exponentiel de cas à considérer. Il s'agit alors de formuler une procédure qui puisse être exploitée algorithmiquement pour trouver les regroupements conjoints des catégories lignes et colonnes qui (quasi-)maximisent un critère donné d'association. Notons qu'il conviendra ici de distinguer les cas des variables nominales de celui des variables ordinales pour lesquelles seuls des regroupements de catégories adjacentes n'ont de sens.

L'optimisation considérée n'est pas sans rappeler la question des partitions des lignes et des colonnes qui maximisent le khi-2 de Pearson, abordée par Benzécri [3], et traitée en particulier par des techniques de classification automatique par Celeux et al. [5]. Ces auteurs se placent cependant dans un contexte où, contrairement à celui retenu ici, seules des partitions avec un nombre de classes fixé a priori sont pertinentes.

Dans la section 2, nous illustrons avec un exemple numérique la variété des effets d'une agrégation de catégories sur les statistiques du khi-2 et les mesures d'association. Le cadre formel et les notations sont précisées en 3. La section 4 précise le degré de complexité de la recherche de la solution optimale et introduit le principe de l'heuristique. La section 6 propose une étude analytique de l'effet

du regroupement de deux catégories lignes ou colonnes sur un choix de mesures d'association nominales et ordinales. Il s'agit d'une part d'obtenir dans la mesure du possible des expressions simples de la variation des critères d'association, et d'autre part de mieux comprendre la sensibilité des mesures d'association, en particulier dans le cas de l'équivalence distributionnelle. Enfin, nous concluons à la section 7 avec quelques remarques notamment du point de vue des perspectives de développement.

2. EXEMPLE ET RÉSULTATS INTUITIFS

Considérons le tableau croisé suivant entre une variable ligne x et une variable colonne y

$M =$

$x\backslash y$	A	B	C	D
a	10	10	1	1
b	10	10	1	1
c	1	1	10	10
d	1	1	10	10

Intuitivement, un regroupement des deux premières $\{A, B\}$ ou des deux dernières colonnes $\{C, D\}$, qui sont identiques, devraient renforcer le degré d'association. Il en est de même pour un regroupement des lignes $\{a, b\}$ ou $\{c, d\}$ Par contre, un regroupement des catégories B et C par exemple, diminue le contraste entre les distributions colonnes et devrait donc se traduire par une réduction de l'association. Pour illustrer ces aspects on considère alors d'une part les tableaux

$M_y^+ =$

$x\backslash y$	A	B	$\{C, D\}$
a	10	10	2
b	10	10	2
c	1	1	20
d	1	1	20

$M_{xy}^+ =$

$x\backslash y$	A	B	$\{C, D\}$
a	10	10	2
b	10	10	2
$\{c, d\}$	2	2	40

et d'autre part les tableaux

$M_y^- =$

$x\backslash y$	A	$\{B, C\}$	D
a	10	11	1
b	10	11	1
c	1	11	10
d	1	11	10

$M_{xy}^- =$

$x\backslash y$	A	$\{B, C\}$	D
a	10	11	1
$\{b, c\}$	11	22	11
d	1	11	10

Le tableau 1 résume les valeurs d'un choix de mesures d'association pour le tableau M et les quatre regroupements considérés. On observe que si, conformément au principe de l'équivalence distributionnelle, les mesures du khi-2 restent les mêmes pour les tableaux M, M_y^+ et M_{xy}^+, les mesures d'association augmentent par contre comme pressenti avec l'agrégation de lignes et colonnes semblables. Les valeurs pour les regroupements M_y^- et M_{xy}^- font apparaître que l'agrégation de colonnes ou lignes

	M	M_y^+	M_{xy}^+	M_y^-	M_{xy}^-
Lignes	4	4	3	4	3
Colonnes	4	3	3	3	3
Degrés liberté	9	6	4	6	4
Khi-2 Pearson	58.91	58.91	58.91	29.45	14.73
Rapport vraisemblance	68.38	68.38	68.38	34.19	17.09
t Tschuprow	0.47	0.52	0.58	0.37	0.29
v Cramer	0.47	0.58	0.58	0.41	0.29
$\tau_{y\leftarrow x}$ Goodman-Kruskal	0.22	0.40	0.40	0.13	0.07
$\tau_{x\leftarrow y}$ Goodman-Kruskal	0.22	0.22	0.40	0.11	0.07
$u_{y\leftarrow x}$ Coefficient incertitude	0.28	0.37	0.37	0.19	0.09
$u_{x\leftarrow y}$ Coefficient incertitude	0.28	0.28	0.37	0.14	0.09
γ Goodman-Kruskal	0.68	0.77	0.80	0.63	0.57
τ_b Kendall	0.55	0.60	0.65	0.45	0.37
$d_{y\leftarrow x}$ Somers	0.55	0.55	0.65	0.41	0.37
$d_{x\leftarrow y}$ Somers	0.55	0.65	0.65	0.49	0.37

Tableau 1 – Mesures d'association selon le regroupement

distribuées de façon très différentes conduit à une réduction aussi bien des statistiques du khi-2 que des mesures d'association.

On a ainsi comme premières indications que l'association

1. s'*amplifie* par le regroupement de catégories d'une variable dont les effectifs se distribuent de façon similaire selon l'autre variable ;
2. s'*atténue* par le regroupement de catégories d'une variable dont les effectifs se distribuent de façon très différentes selon l'autre variable.

3. CADRE FORMEL ET NOTATIONS

Soit deux variables x et y prenant respectivement ℓ et c états différents. Le croisement des variables donne lieu à une table de contingence $T_{\ell\times c}$ à ℓ lignes et c colonnes. On note n_{ij} l'effectif de la cellule de la table se trouvant à l'intersection de la i-ème ligne et de la j-ème colonne. Les totaux des lignes et des colonnes sont représentés en remplaçant l'indice de sommation par un point : $n_{i\cdot} = \sum_j n_{ij}, n_{\cdot j} = \sum_i n_{ij}$. Enfin, on note n l'effectif total de la table : $n = \sum_i n_{i\cdot} = \sum_j n_{\cdot j}$. Ceci concerne évidemment les cas observés. Le cas échéant, on se référera aux probabilités p_{ij}, $p_{i\cdot}$ et $p_{\cdot j}$ qu'un individu choisi au hasard dans la population de référence soit respectivement dans la case (i, j), la ligne i ou la colonne j.

On considère les critères d'association θ_{xy} entre x et y en tant que fonction des éléments de la table de contingence $\theta_{xy} = \theta(T_{\ell\times c})$.

Soit P_x une partition des états de la variable ligne x et P_y une partition des valeurs de y. Chaque couple (P_x, P_y) de partitions donne lieu a une table de contingence T

différente. Le problème général envisagé est alors la recherche du couple de partitions qui maximise la mesure d'association

$$\max_{P_x,P_y} \ \theta\big(T(P_x,P_y)\big) \ . \tag{1}$$

Pour des variables ordinales, et donc en particulier pour les variables mesurables de type intervalle ou ratio, seules les partitions obtenues par regroupement de catégories adjacentes sont pertinentes. Dans ce cas, on considère le problème restreint

$$\begin{cases} \max_{P_x,P_y} \ \theta\big(T(P_x,P_y)\big) \\ \text{s.c.} \quad P_x \in \mathcal{A}_x \text{ et } P_y \in \mathcal{A}_y \ . \end{cases} \tag{2}$$

où $\mathcal{A}_x$ et $\mathcal{A}_y$ désignent respectivement l'ensemble des partitions par regroupements adjacents des catégories de x et de y. En désignant par $\mathcal{P}_x$ et $\mathcal{P}_y$ les ensembles non restreints de partitions, on a, pour $c, \ell > 2$, $\mathcal{A}_x \subset \mathcal{P}_x$ et $\mathcal{A}_y \subset \mathcal{P}_y$. Notons encore que, l'association entre variables ordinales pouvant être négative, le critère $\theta\big(T(P_x,P_y)\big)$ à maximiser est dans ce cas la valeur absolue d'une mesure d'association.

4. STRATÉGIE GÉNÉRALE

Le propos de cette section est de préciser le degré de complexité de la recherche de la solution optimale par exploration et d'esquisser l'heuristique qui sera exploitée. La complexité d'une exploration complète empêche sa généralisation à un nombre quelconque de lignes et de colonnes et justifie en effet le recours à une heuristique.

4.1. Complexité de la solution optimale

La détermination de la solution optimale nécessite l'exploration, de toutes les regroupements possibles de lignes ou de colonnes, c'est-à-dire l'ensemble des couples (P_x, P_y). Le nombre de situations à examiner est donné par le produit du nombre de regroupements possibles des lignes par celui des colonnes, soit $\#\mathcal{P}_x \#\mathcal{P}_y$.

Pour le cas de variables nominales, le nombre de regroupements correspond au nombre $B(c) = \#\mathcal{P}$ de partitions de l'ensemble des c catégories de la variable. Il peut être obtenu récursivement par la formule de Bell [2].

$$B(c) = \sum_{0 \le k \le c-1} \binom{c-1}{k} B(k)$$

avec $B(0) = 1$. Pour $c = \ell$ le nombre $B(c)B(\ell)$ de configurations possibles à explorer est ainsi par exemple respectivement $25, 225, 2704$ et $41'209$ pour $c = 3, 4, 5, 6$ et passe à plus de 13 milliard pour $c = \ell = 10$.

Pour des variables ordinales, et donc en particulier dans les problèmes de discrétisation, seuls les regroupements de catégories adjacentes sont pertinentes. Le nombre

de cas à explorer s'en trouve dès lors réduit. Le nombre $G(c) = \#\mathcal{A}$ de groupements différents de c catégories est

$$G(c) = \sum_{k=0}^{c-1} \binom{c-1}{k} = 2^{(c-1)}$$

ce qui donne respectivement $G(c)G(\ell) = 16, 64, 256, 1'024$ configurations à explorer pour un tableau carré avec $c = \ell = 3, 4, 5, 6$ et plus d'un million pour $c = \ell = 10$.

4.2. Heuristique pas à pas

Compte tenu des limites évidentes de l'exploration exhaustive de toutes les configurations, nous proposons une approche procédant par regroupements successifs de deux catégories. Cette façon de faire ne conduit évidemment pas nécessairement à la solution optimale, mais de façon générale seulement à une solution quasi-optimale.

La stratégie proposée est une stratégie itérative. A chaque étape, on recherche parmi les regroupements de deux catégories ligne ou de deux catégories colonne celui qui maximise le critère d'association $\theta(T)$ retenu. Formellement, en notant (P_x^k, P_y^k) la partition obtenue à l'étape k, il s'agit à chaque étape k de chercher la solution (P_x^k, P_y^k) du programme

$$\begin{cases} \max\limits_{P_x, P_y} \ \theta\big(T(P_x, P_y)\big) \\ \text{s.c.} \quad P_x = P_x^{(k-1)} \quad \text{et} \quad P_y \in \mathcal{P}_y^{(k-1)} \\ \qquad \text{ou} \\ \qquad P_x \in \mathcal{P}_x^{(k-1)} \quad \text{et} \quad P_y = P_y^{(k-1)} \end{cases} \tag{3}$$

où $\mathcal{P}_x^{(k-1)}$ désigne l'ensemble des partitions sur la variable x obtenues par un regroupement de deux classes de la partition $P_x^{(k-1)}$.

Pour des variables ordinales, il convient de remplacer $\mathcal{P}_x^{(k-1)}$ et $\mathcal{P}_y^{(k-1)}$ par les ensembles $\mathcal{A}_x^{(k-1)}$ et $\mathcal{A}_y^{(k-1)}$ de partitions obtenues par le regroupement de deux éléments adjacents.

En partant de $T^0 = T_{\ell \times c}$ le tableau initial associé aux catégories les plus fines des variables x et y, l'heuristique proposée consiste alors à rechercher successivement les tableaux $T^k, k = 1, 2, \ldots$ définis par les partitions solution de (3). La procédure se poursuit tant que $\theta(T^k) \geq \theta(T^{(k-1)})$ et est arrêtée dès que cette condition n'est plus vérifiée. En d'autre terme on procède successivement au regroupement de deux catégories qui maximise l'accroissement du critère d'association jusqu'à ce que seul un accroissement négatif puisse être obtenu.

Le *regroupement quasi-optimal* est le couple (P_x^k, P_y^k) solution de (3) à l'étape k où

$$\theta\big(T^{(k+1)}\big) - \theta\big(T^k\big) < 0 \quad \text{et} \quad \theta\big(T^k\big) - \theta\big(T^{(k-1)}\big) \geq 0 \ .$$

Par convention, on fixe le critère d'association $\theta(T)$ à zéro pour toute table ayant une seule ligne ou colonne. L'algorithme conduit ainsi à une table 1×1 ne contenant qu'une valeur si seulement si toutes les lignes, et donc toutes les colonnes sont identiquement distribuées.

5. LES CRITÈRES D'ASSOCIATION

Nous rappelons ici les formules des critères d'association considérés. Pour plus de détails, voir par exemple [7].

Statistiques du khi-2

Pearson	$X^2 = \sum_{i=1}^{\ell}\sum_{j=1}^{c} \frac{(n\, n_{ij} - n_{i\cdot} n_{\cdot j})^2}{(n n_{i\cdot} n_{\cdot j})}$
Rapport vraisemblance	$G^2 = 2\sum_i \sum_j n_{ij} \log(\frac{n\, n_{ij}}{n_{i\cdot} n_{\cdot j}})$

Mesures d'association basées sur le khi-2 de Pearson

	théorique	empirique
Phi	$\phi = \sqrt{\sum_i \sum_j \frac{(p_{ij} - p_{i\cdot} p_{\cdot j})^2}{p_{i\cdot} p_{\cdot j}}}$	$\hat{\phi} = \sqrt{\frac{X^2}{n}}$
Contingence	$c_c = \sqrt{\frac{\phi^2}{1+\phi^2}}$	$\hat{c}_c = \sqrt{\frac{X^2}{n+X^2}}$
Tschuprow	$t = \sqrt{\frac{\phi^2}{\sqrt{(\ell-1)(c-1)}}}$	$\hat{t} = \sqrt{\frac{X^2}{n\sqrt{(\ell-1)(c-1)}}}$
Cramer	$v = \sqrt{\frac{\phi^2}{\min\{\ell, c\} - 1}}$	$\hat{v} = \sqrt{\frac{X^2}{n(\min\{\ell, c\} - 1)}}$

Mesures PRE nominales

τ de Goodman-Kruskal

$$\tau_{y \leftarrow x} = \frac{\sum_i \sum_j \frac{p_{ij}^2}{p_{i\cdot}} - \sum_j p_{\cdot j}^2}{1 - \sum_j p_{\cdot j}^2} \qquad \hat{\tau}_{y \leftarrow x} = \frac{n \sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot}} - \sum_j n_{\cdot j}^2}{n^2 - \sum_j n_{\cdot j}^2}$$

Coefficient d'incertitude u de Theil

$$u_{y \leftarrow x} = \frac{\sum_i \sum_j p_{ij} \log_2 \left(\frac{p_{i\cdot} p_{\cdot j}}{p_{ij}}\right)}{\sum_j p_{\cdot j} \log_2 p_{\cdot j}} \qquad \hat{u}_{y \leftarrow x} = \frac{n \log_2 n - \sum_i \sum_j n_{ij} \log_2 \left(\frac{n_{i\cdot} n_{\cdot j}}{n_{ij}}\right)}{n \log_2 n - \sum_j n_{\cdot j} \log_2 n_{\cdot j}}$$

On note π^c, π^d, π_x, p_y les probabilités respectivement d'une pair $\big\{(x_i, y_i), (x_j, y_j)\big\}$ avec un ordre concordant, i.e. $x_i > x_j$ et $y_i > y_j$, avec un ordre discordant, i.e. $x_i > x_j$ and $y_i < y_j$, avec égalité sur x seulement et avec égalité sur y seulement. De même, on note m^c, m^d, m_x and m_y le nombre de pairs d'observations respectivement concordantes, discordantes, avec égalité sur x seulement et avec égalité sur y seulement.

γ de Goodman-Kruskal

$$\gamma = \frac{\pi^c - \pi^d}{\pi^c + \pi^d} \qquad \hat{\gamma} = \frac{m^c - m^d}{m^c + m^d}$$

d de Somers

$$d_{y \leftarrow x} = \frac{\pi^c - \pi^d}{\pi^c + \pi^d + \pi_y} \qquad \hat{d}_{y \leftarrow x} = \frac{m^c - m^d}{m^c + m^d + m_y}$$

τ_b de Kendall

$$\tau_b = \frac{\pi^c - \pi^d}{\sqrt{(\pi^c + \pi^d + \pi_x)(\pi^c + \pi^d + \pi_y)}} \qquad \hat{\tau}_b = \frac{m^c - m^d}{\sqrt{(m^c + m^d + m_x)(m^c + m^d + m_y)}}$$

τ_c de Kendall

$$\tau_c = \pi^c - \pi^d \left(\frac{\min\{\ell, c\}}{\min\{\ell, c\} - 1}\right) \qquad \hat{\tau}_c = \frac{m^c - m^d}{m_{\text{tot}}} \left(\frac{\min\{\ell, c\}}{\min\{\ell, c\} - 1}\right)$$

6. REGROUPEMENT DE DEUX CATÉGORIES

Afin d'étudier formellement l'effet d'un regroupement de deux catégories, on explicite analytiquement cet effet sur un choix de critères d'association. On traite tout d'abord le cas des mesures d'associations fondées sur le khi-2 de Pearson, puis les mesures nominales de type PRE et finalement les mesures pour variables ordinales. Pour les mesures symétriques, on considère le regroupement des colonnes j et k. Pour les mesures asymétriques (mesures PRE et d de Somers), on retient y comme variable dépendante et on examine également l'effet d'un regroupement des catégories i et s de la variable indépendante.

L'étude se limite aux versions empiriques des mesures d'association. Les résultats s'étendent cependant aisément aux mesures d'association théoriques exprimées en termes de probabilités.

6.1. Variation des mesures fondées sur le khi-2 de Pearson

La variation du X^2 suite à l'agrégation des colonnes j et k est,

$$\begin{aligned}\Delta_y X^2 &= \frac{1}{n}\sum_{i=1}^{\ell}\left(\frac{\big(n(n_{ij}+n_{ik})-n_{i\cdot}(n_{\cdot j}+n_{\cdot k})\big)^2}{n_{i\cdot}(n_{\cdot j}+n_{\cdot k})} - \right.\\ &\qquad \left. - \frac{(nn_{ij}-n_{i\cdot}n_{\cdot j})^2}{n_{i\cdot}n_{\cdot j}} - \frac{(nn_{ik}-n_{i\cdot}n_{\cdot k})^2}{n_{i\cdot}n_{\cdot k}}\right) \qquad (4)\end{aligned}$$

En développant et en simplifiant le terme sous le signe de sommation cette variation s'écrit

$$\Delta_y X^2 = \frac{-n}{n_{\cdot j}n_{\cdot k}(n_{\cdot j}+n_{\cdot k})}\sum_i \frac{(n_{\cdot j}n_{ik}-n_{\cdot k}n_{ij})^2}{n_{i\cdot}} \quad . \qquad (5)$$

Clairement, cette quantité est non positive. Un regroupement de catégories ne peut donc en aucun cas augmenter le X^2. Au mieux, la variation est nulle. Ceci se produit lorsqu'on a l'équivalence distributionnelle des deux colonnes :

$$\frac{n_{ij}}{n_{\cdot j}} = \frac{n_{ik}}{n_{\cdot k}} \quad \Leftrightarrow \quad (n_{\cdot j}n_{ik}-n_{\cdot k}n_{ij})^2 = 0$$

La réduction du X^2 est d'autant plus importante, que l'écart entre les distributions est grand.

Le coefficient de contingence ϕ est une fonction croissante de X^2 et sa variation est donc également non positive.,

Le t de Tschuprow et le v de Cramer dépendent également du nombre de catégories. Une réduction de c ou ℓ peut donc compenser la réduction de x^2 et conduire à un accroissement de la valeur de ces deux mesures d'association.

Dans le cas de l'agrégation de deux colonnes lorsque $c \leq \ell$

$$\Delta_y v > 0 \quad \Leftrightarrow \quad \frac{X^2+\Delta_y X^2}{X^2} > \frac{c-2}{c-1}$$

Pour $c = 3$ par exemple, on a un accroissement du v de Cramer tant que $-\Delta_y X^2$ reste inférieur à $X^2/2$.

Notons que le regroupement de deux catégories n'affecte le dénominateur dans l'expression du v de Cramer que si elle porte sur la variables qui a le moins de catégories. Le v de Cramer ne peut donc augmenter que dans ce cas. La mesure n'est donc pas appropriée pour l'heuristique décrite précédemment.

6.2. Variation des mesures de type PRE

Les mesures de type PRE sont asymétriques par construction. Il convient alors de distinguer le cas du regroupement sur la variable dépendante (y, colonne dans notre cas) et indépendante (x, ligne).

Notons $S_y^{\tau_{yx}} = \sum_j n_{.j}^2$ et $S_{yx}^{\tau_{yx}} = \sum_i S_{yi}$, où les indices x et y indiquent respectivement que la quantité S est sensible à un regroupement sur la variable x (ligne) ou y (colonne). On a

$$\tau_{yx} = \frac{S_{yx}^{\tau_{yx}} - S_y^{\tau_{yx}}}{n^2 - S_y^{\tau_{yx}}}$$

d'où il apparaît clairement que la variation $\Delta_x \tau_{yx}$ suite à un regroupement sur la variable indépendante x peut être analysée par le biais de la seule variation $\Delta_x S_{yx}^{\tau_{yx}}$, tandis que pour la variation $\Delta_y \tau_{yx}$ suite à un regroupement sur la variable dépendante y, on doit prendre en compte les variations de $S_y^{\tau_{yx}}$ et $S_{yx}^{\tau_{yx}}$.

Les variations à considérer sont

$$\Delta_y S_y^{\tau_{yx}} = (n_{.j} + n_{.k})^2 - n_{.j}^2 - n_{.k}^2 \tag{6}$$

$$\Delta_y S_{yx}^{\tau_{yx}} = n \sum_i \frac{(n_{ij} + n_{ik})^2 - n_{ij}^2 - n_{ik}^2}{n_{i.}} \tag{7}$$

$$\Delta_x S_{yx}^{\tau_{yx}} = n \sum_j \Big(\frac{(n_{ij} + n_{sj})^2}{n_{i.} + n_{s.}} - \frac{n_{ij}^2}{n_{i.}} - \frac{n_{sj}^2}{n_{s.}} \Big) \tag{8}$$

De même, pour le coefficient d'incertitude de Theil, on a, en notant $S_y^{u_{yx}} = \sum_j n_{.j} \log_2 n_{.j}$ et $S_{yx}^{u_{yx}} = \sum_j \sum_i n_{ij} \log_2 \left(n_{i.} n_{.j} / n_{ij} \right)$

$$u_{yx} = \frac{n \log_2 n - S_{yx}^{u_{yx}}}{n \log_2 n - S_y^{u_{yx}}}$$

Pour la variation $\Delta_x u_{yx}$ il suffit d'étudier la variation $\Delta_x S_{yx}^{u_{yx}}$ de la double somme du numérateur, tandis que pour la variation suite à un regroupement sur y, on doit prendre en compte les variations de $S_y^{u_{yx}}$ et $S_{yx}^{u_{yx}}$.

Les variations à considérer sont

$$\Delta_y S_y^{u_{yx}} = (n_{.j} + n_{.k}) \log_2 (n_{.j} + n_{.k}) - n_{.j} \log_2 n_{.j} - n_{.k} \log_2 n_{.k} \tag{9}$$

$$\Delta_y S_{yx}^{u_{yx}} = \sum_i \Bigg((n_{ij} + n_{ik}) \ \log_2 \Big(\frac{(n_{.j} + n_{.k}) n_{i.}}{n_{ij} + n_{ik}} \Big) - n_{ij} \ \log_2 \Big(\frac{n_{i.} n_{.j}}{n_{ij}} \Big) - n_{ik} \ \log_2 \Big(\frac{n_{i.} n_{.k}}{n_{ik}} \Big) \Bigg) \tag{10}$$

$$\Delta_x S_{yx}^{u_{yx}} = \sum_j \Bigg((n_{ij} + n_{sj}) \ \log_2 \Big(\frac{n_{.j} (n_{i.} + n_{s.})}{n_{ij} + n_{sj}} \Big) - n_{ij} \ \log_2 \Big(\frac{n_{i.} n_{.j}}{n_{ij}} \Big) - n_{sj} \ \log_2 \Big(\frac{n_{s.} n_{.j}}{n_{sj}} \Big) \Bigg) \tag{11}$$

6.2.1. *Regroupement sur la variable indépendante x.*

Les quantités S_y étant insensibles à un regroupement sur x, les variations $\Delta_x S_y$ sont nulles. Les variations de τ_{yx} et u_{yx} suite au regroupement de deux lignes (catégories de x) sont alors

$$\Delta_x \tau_{yx} = \frac{\Delta_x S_{yx}^{\tau_{yx}}}{n^2 - S_y^{\tau_{yx}}} \tag{12}$$

$$\Delta_x u_{yx} = \frac{-\Delta_x S_{yx}^{u_{yx}}}{n \log_2 n - S_y^{u_{yx}}} \tag{13}$$

Les dénominateurs sont dans les deux expressions ci-dessus des quantités positives indépendantes du regroupement. Le maximum de la variation de la mesure correspond alors au maximum de la variation du numérateur, soit de $\Delta_x S_{yx}$.

On peut vérifier que tant $\Delta_x \tau_{yx}$ que le $\Delta_x u_{yx}$ sont non positifs. Ces variations sont nulles lorsque les deux lignes agrégées sont colinéaires, soit lorsque $n_{sj}/n_{s\cdot} = n_{ij}/n_{i\cdot}$.

La variation $\Delta_x S_{yx}^{\tau_{yx}}$ par exemple peut s'écrire sous la forme

$$\Delta_x S_{yx}^{\tau_{yx}} = n \sum_j \frac{-(n_{ij} n_{s\cdot} - n_{sj} n_{i\cdot})^2}{n_{i\cdot} n_{s\cdot} (n_{i\cdot} + n_{s\cdot})}$$

qui est clairement non positive. Le regroupement de catégories de la variable indépendante ne peut pas alors accroître la valeur de la mesure d'association τ_{yx}. Ceci est en fait assez intuitif : le regroupement ne peut pas accroître le contenu prédictif de la variable indépendante.

On établit un résultat similaire pour le coefficient d'incertitude u_{yx}. En effet, $\Delta_x S_{yx}^{u_{yx}}$ s'écrit

$$\Delta_x S_{yx}^{u_{yx}} = n \sum_j \left(n_{ij} \log_2 \left(\frac{n_{ij}}{n_{i\cdot}} \right) + n_{sj} \log_2 \left(\frac{n_{sj}}{n_{s\cdot}} \right) - (n_{ij} + n_{sj}) \log_2 \left(\frac{n_{ij} + n_{sj}}{n_{i\cdot} + n_{s\cdot}} \right) \right) \tag{14}$$

Le terme sous le signe de sommation est non négatif. En effet, il est de la forme

$$f(a,b,c,d) = a \log_2 \left(\frac{a}{b} \right) + c \log_2 \left(\frac{c}{d} \right) - (a+c) \log_2 \left(\frac{a+c}{b+d} \right)$$

avec $0 \leq a \leq b$ et $0 \leq c \leq d$. La fonction f atteint son minimum en $a/b = c/d = (a+c)/(b+d)$ où l'on a $f(a,b,c,d) = 0$ et ne peut donc pas être négative. Il en est donc de même de (14) ce qui implique une variation (13) de u_{yx} non positive.

6.2.2. *Regroupement sur la variable dépendante y.*

Pour un regroupement sur la variable dépendante (colonne), on doit également tenir compte des changements dans les totaux $n_{\cdot j}$ des colonnes concernées qui impliquent

des variations dans les dénominateurs des formules définissant τ_{yx} et u_{yx}. Les variations de ces mesures sont dans ce cas

$$\Delta_y \tau_{yx} = \frac{(n^2 - S_y^{\tau_{yx}})\Delta_y S_{yx}^{\tau_{yx}} - (n^2 - S_{yx}^{\tau_{yx}})\Delta_y S_y^{\tau_{yx}}}{(n^2 - S_y^{\tau_{yx}})^2 - (n^2 - S_y^{\tau_{yx}})\Delta_y S_y^{\tau_{yx}}} \tag{15}$$

$$\Delta_y u_{yx} = \frac{(n\log_2 n - S_{yx}^{u_{yx}})\Delta_y S_y^{u_{yx}} - (n\log_2 n - S_y^{u_{yx}})\Delta_y S_{yx}^{u_{yx}}}{(n\log_2 n - S_y^{u_{yx}})^2 - (n\log_2 n - S_y^{u_{yx}})\Delta_y S_y^{u_{yx}}} \tag{16}$$

Les dénominateurs des deux expressions sont positifs. Le signe de la variation est donc déterminé pour chacune des deux mesures par celui du numérateur.

En ce qui concerne τ_{yx}, il ressort clairement des formules (6) et (7) que $\Delta_y S_{yx}^{\tau_{yx}}$ et $\Delta_y S_y^{\tau_{yx}}$ sont tous les deux positifs. On peut donc avoir aussi bien des variations positives ou que négatives

$$\Delta_y \tau_{yx} \geq 0 \quad \Leftrightarrow \quad \frac{\Delta_y S_{yx}^{\tau_{yx}}}{\Delta_y S_y^{\tau_{yx}}} \geq \frac{(n^2 - S_{yx}^{\tau_{yx}})}{(n^2 - S_y^{\tau_{yx}})} \tag{17}$$

$$\Delta_y \tau_{yx} \leq 0 \quad \Leftrightarrow \quad \frac{\Delta_y S_{yx}^{\tau_{yx}}}{\Delta_y S_y^{\tau_{yx}}} \leq \frac{(n^2 - S_{yx}^{\tau_{yx}})}{(n^2 - S_y^{\tau_{yx}})} \tag{18}$$

Pour u_{xy}, il ressort également des formules (9) et (10) que $\Delta_y S_{yx}^{u_{yx}}$ et $\Delta_y S_y^{u_{yx}}$ sont tous deux positifs. La variation peut donc là aussi être positive ou négative

$$\Delta_y u_{yx} \geq 0 \quad \Leftrightarrow \quad \frac{\Delta_y S_y^{u_{yx}}}{\Delta_y S_{yx}^{u_{yx}}} \geq \frac{(n^2 - S_y^{u_{yx}})}{(n^2 - S_{yx}^{u_{yx}})} \tag{19}$$

$$\Delta_y u_{yx} \leq 0 \quad \Leftrightarrow \quad \frac{\Delta_y S_y^{u_{yx}}}{\Delta_y S_{yx}^{u_{yx}}} \leq \frac{(n^2 - S_y^{u_{yx}})}{(n^2 - S_{yx}^{u_{yx}})} \tag{20}$$

6.3. Mesures ordinales

On considère ici les mesures pour variables ordinales fondées sur les notions de paires concordantes et discordantes, deux observations étant dites concordantes si on a la même relation d'ordre sur les variables x et y, et discordantes si la relation d'ordre est inversée. L'association entre variables catégorielles pouvant être négative, il convient de maximiser la valeur absolue des mesures. Par ailleurs, on se limite à l'étude de l'effet du regroupement de catégories adjacentes.

Les mesures examinées sont le γ de Goodman & Kruskal, le τ_b de Kendall, le τ_c de Kendall et Stuart et les d_y et d_x asymétriques de Somers.

Il s'agit donc d'étudier les quantités

$$
\begin{aligned}
m^c &= \sum_{i=1}^{\ell-1}\sum_{j=1}^{c-1}\Big(n_{ij}\sum_{i'>i}\sum_{j'>j}n_{i'j'}\Big)\\
m^d &= \sum_{i=1}^{\ell-1}\sum_{j=2}^{c}\Big(n_{ij}\sum_{i'>i}\sum_{j'<j}n_{i'j'}\Big)\\
m_x &= \sum_{i=1}^{\ell}\sum_{j=1}^{c-1}\Big(n_{ij}\sum_{j'>j}n_{ij'}\Big)\\
m_y &= \sum_{j=1}^{c}\sum_{i=1}^{\ell-1}\Big(n_{ij}\sum_{i'>i}n_{i'j}\Big)\\
m_{xy} &= \sum_{i=1}^{c}\sum_{j=1}^{c}\frac{n_{ij}(n_{ij}-1)}{2}\\
m_{tot} &= m^c + m^d + m_x + m_y + m_{xy} = \frac{n\,(n-1)}{2}
\end{aligned}
$$

dont les variations $\Delta_y m$ suite à l'agrégation des colonnes k et $k+1$ sont

$$\Delta_y m^c = -\sum_{i=1}^{\ell-1} n_{ik}\sum_{i'>i} n_{i'(k+1)} \tag{21}$$

$$\Delta_y m^d = -\sum_{i=1}^{\ell-1} n_{i(k+1)}\sum_{i'>i} n_{i'k} \tag{22}$$

$$\Delta_y m_x = -\sum_{i=1}^{\ell} n_{ik} n_{i(k+1)} \tag{23}$$

$$\Delta_y m_y = \sum_{i=1}^{\ell-1}\Big(n_{ik}\sum_{i'>i} n_{i'(k+1)} + n_{i(k+1)}\sum_{i'>i} n_{i'k}\Big) \tag{24}$$

$$\Delta_y m_{xy} = \sum_{i=1}^{\ell} n_{ik} n_{i(k+1)} \tag{25}$$

Le nombre total $m_{\rm tot}$ de paires restant évidemment inchangé.

Le regroupement de catégories de y transforme évidemment des inégalités sur cette variable en égalités. Ceci se traduit par un transfert de paires comptabilisées dans m^c et m^d vers m_y et de m_x vers m_{xy}.

Discussion

En cas d'équivalence distributionnelle ($n_{i(k+1)} = \alpha n_{ik}$) des deux colonnes k et $k+1$, la réduction de m^c et de m^d est identiques. La différence $m^c - m^d$, numérateur commun de toutes les mesures, reste donc constante tandis que la somme $m^c + m^d$

diminue. Le regroupement de colonnes identiquement distribuées induit dès lors un accroissement du γ de Goodman & Kruskal. Pour le d_y de Somers pertinent lorsque y est la variable dépendante, la diminution de $m^c + m^d$ est compensée par l'accroissement de m_y, et la mesure reste donc insensible à l'agrégation de colonnes identiquement distribuées. Pour d_x, c'est-à-dire lorsqu'on agrège des catégories de la variable indépendante, la diminution de $m^c + m^d$ est amplifiée par celle de m_x et l'on observe alors un renforcement de l'association. En ce qui concerne le τ_b, le dénominateur étant la moyenne géométrique de celui de d_y et de d_x, on a également un accroissement de l'association, mais moins marqué que pour le d_y. Enfin, le τ_c est affecté par un regroupement de catégories identiquement distribuées uniquement par l'effet sur $\min\{\ell, c\}$. Le τ_c augmente lorsque $c \leq \ell$ et reste inchangé sinon.

En dehors de l'équivalence distributionnelle, toutes les situations peuvent se présenter. Notons simplement que si dans la colonne k les les n_{ik} ont tendance à être plus importants pour les petits i que pour les grands, et dans la colonne $k+1$ les $n_{i(k+1)}$ ont tendance à être plus importants pour les grands i que pour les petits, la réduction de m^c sera plus importante que celle de m^d. Le regroupement devrait alors diminuer l'association si elle est positive et la renforcer si elle est négative.

7. CONCLUSION

Cet article est une contribution au problème de la recherche des partitions des catégories ligne et colonne qui maximisent l'association. Dans ce contexte, les résultats présentés ici, que ce soit sur la complexité de la solution ou sur la sensibilité des critères d'association à l'agrégation de catégories, ne constituent qu'une étape préliminaire. Il reste beaucoup à faire, spécialement en ce qui concerne les propriétés et la mise en œuvre de la l'algorithme décrit au point 4.2. Citons en particulier deux aspects. Premièrement, il s'agit de s'assurer empiriquement de l'efficacité de l'heuristique. Nous travaillons actuellement au développement d'une procédure de simulation pour étudier sous quelles conditions et dans quelle mesure la solution quasi-optimale fournie par l'algorithme peut s'écarter de la vraie solution optimale. Selon les résultats du paragraphe 4.1, la comparaison avec la vraie solution n'est possible que pour des tables de départs de taille raisonnable de l'ordre de six lignes par six colonnes. Deuxièmement, il importe de tenir compte de la plus grande fiabilité des grands effectifs. A cette fin, il est prévu d'introduire dans l'algorithme les estimations de Laplace $\hat{p}_{ij}^{\lambda} = \frac{n_{ij}+\lambda}{n+rc\lambda}$ des probabilités. L'utilisation de ces estimateurs permet en effet d'accroître la robustesse de la solution avec la valeur de λ.

BIBLIOGRAPHIE

[1] Jean-Paul Auray, Gérard Duru, and Djamel A. Zighed. *Analyse des données multidimensionnelles.* Editions A. Lacassagne, Lyon, 1990.

[2] E. T. Bell. The iterated exponential numbers. *Ann. Math.*, 39 :539–557, 1938.

[3] J.-P. Benzécri. *Analyse des données. Tome 2 : Analyse des correspondances.* Dunod, Paris, 1973.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification And Regression Trees.* Chapman and Hall, New York, 1993.

[5] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, and H. Ralambondrainy. *Classification automatique des données.* Informatique. Dunod, Paris, 1988.

[6] J. D. Jobson. *Applied Multivariate Data Analysis. Vol.II Categorical and Multivariate Methods.* Springer-Verlag, New York, 1992.

[7] M. Olszak and G. Ritschard. The behaviour of nominal and ordinal partial association measures. *The Statistician*, 44(2) :195–212, 1995.

[8] Ricco Rakotomalala and Djamel A. Zighed. Mesures PRE dans les graphes d'induction : une approche statistique de l'arbitrage généralité-précision. In Gilbert Ritschard, André Berchtold, François Duc, and Djamel A. Zighed, editors, *Apprentissage : des principes naturels aux méthodes artificielles*, pages 37–60. Hermes Science Publications, Paris, 1998.

[9] Djamel A. Zighed and Ricco Rakotomalala. *Graphes d'induction : apprentissage et data mining.* Hermes Science Publications, Paris, 2000.