Une analyse des correspondances multiples topologique

Rafik Abdesselam

Laboratoire COACTIS-ISH, UFR de Sciences Economiques et de Gestion
Université de Lyon, Lumière Lyon 2,
16, quai Claude Bernard 69365 Lyon cedex 07
rafik.abdesselam@univ-lyon2.fr

Résumé. L'objectif de ce papier est de proposer une méthode topologique d'analyse des données qui consiste à explorer, analyser et représenter les associations entre plusieurs variables qualitatives dans un contexte d'analyse des correspondances multiples. Les mesures de similarité jouent un rôle important dans de nombreux domaines de l'analyse des données. Les résultats de toute opération de structuration, de classification ou de classement d'objets dépendent fortement de la mesure de proximité choisie. Basées sur la notion de graphes de voisinage, certaines de ces mesures de proximité sont plus ou moins équivalentes. La notion d'équivalence topologique entre deux mesures est définie et statistiquement testée selon leur degré de description des associations entre les modalités de ces variables qualitatives. Un exemple sur données réelles illustre cette méthode.

Mots-clés. Tableau de Burt, mesure de proximité, graphe de voisinage, matrice d'adjacence, équivalence topologique, associations, représentations graphiques.

Abstract. The objective of this paper is to propose a topological method of data analysis that consists in exploring, analyzing and representing the associations between several qualitative variables in a context of multiple correspondence analysis. Similarity measures play an important role in many areas of data analysis. The results of any operation of structuring, clustering or classifying objects depend strongly on the proximity measure chosen. Based on the notion of neighborhood graphs, some of these proximity measures are more or less equivalents. The notion of topological equivalence between two measures is defined and statistically tested according to their description degree of the associations between the modalities of these qualitative variables. An example on real data illustrates this method.

Keywords. Burt table, proximity measure, neighborhood graph, adjacency matrix, topological equivalence, associations, graphical representations.

1 Introduction

Le choix d'une mesure de proximité est un problème important en analyse des données topologique. La comparaison d'objets, de situations ou d'idées est une tâche essentielle

pour évaluer une situation, classer des préférences ou structurer un ensemble d'éléments. Pour ce faire, nous utilisons des mesures de proximité pour mettre en évidence les similarités ou les dissimilarités entre objets. Nous savons pertinemment que le résultat dépend de la mesure utilisée. Laquelle est alors la plus utile ? Sont-elles équivalentes ? Comment identifier celle qui est la plus appropriée pour résumer la liaison entre plusieurs variables qualitatives ? Selon la mesure choisie, les résultats de cette problématique d'analyse des correspondances multiples topologique changent.

Nous nous intéressons ici à l'analyse topologique des associations entre les modalités de plusieurs variables qualitatives. Une approche dans le cas de deux variables qualitatives a t proposée (Abdesselam (2018)), elle est basée sur la notion d'équivalence topologique de mesures de proximité (Zighed *et al.* (2012)). Nous avons considéré et comparé 22 mesures de proximité pour des données binaires (Warrens (2008)).

Soit $\{x^k\ ;\ k = 1, \cdots, p\}$, un ensemble de $p > 2$ variables qualitatives à m_k modalités chacune, décrivant un ensemble de $n = \sum_{k=1}^{p} n_k$ individus-objets avec un nombre total de modalités $m = \Sigma_{k=1}^{p} m_k$. L'objectif ici est de décrire les éventuels liens topologiques entre toutes les modalités de ces variables. On utilise les notations suivantes :
- $X_k = X_{(n,m_k)}$ le tableau disjonctif associé aux m_k variables indicatrices de la variable x^k à n lignes-objets et m_k colonnes-modalités, avec $\Sigma_{k=1}^{m_k} x_i^k = 1,\ \forall_i$ et $\Sigma_{i=1}^{n} x_i^k = n_k$,
- $X_{(n,m)} = [\, X_1 | X_2 | \cdots | X_p \,]$ désigne le tableau disjonctif complet, juxtaposition des p tableaux X_k, n lignes-objets et $m = \sum_{k=1}^{p} m_k$ colonnes-modalités, avec $\Sigma_{k=1}^{m_k} x_i^k = p,\ \forall_i$ et $\Sigma_{i=1}^{n} \Sigma_{k=1}^{m_k} x_i^k = np$,
- $\mathcal{B}_{(m,m)} = {}^t X\ X$ est le tableau symétrique de Burt associé au tableau disjonctif complet X, juxtaposition de tableaux de contingence,
- $W_{(m,m)} = diag[\mathcal{B}]$ est la matrice diagonale des effectifs des m modalités,
- $U = 1_m\ {}^t 1_m$ est la $m \times m$ matrice dont tous les éléments sont égaux à 1, I_m la matrice identité d'ordre m, 1_m et 1_n désignent respectivement le vecteur d'ordre m et d'ordre n de composantes toutes égales à 1.

Les matrices de dissimilarité associées aux mesures de proximité sont calculées à partir du tableau de Burt $\mathcal{B}$. Le temps de calcul est ainsi considérablement réduit.

$A_{(m,m)} = (a_{kl}) = \mathcal{B}$, dont l'élément, $a_{kl} = |x^k \cap x^l| = \sum_{i=1}^{n} x_i^k x_i^l$ correspond au nombre d'attributs communs aux deux points x^k et x^l,

$B_{(m,m)} = (b_{kl}) = {}^t X\ (1_n\ {}^t 1_m - X) = {}^t X\ 1_n\ {}^t 1_m - {}^t X\ X = W\ 1_m\ {}^t 1_m - A = W\ U - A$

dont l'élément, $b_{lk} = |X^l - X^k| = |X^l \cap \overline{X^k}| = \sum_{i=1}^{n} x_i^l (1 - x_i^k)$ est le nombre d'attributs présents dans x^l mais pas dans x^k,

$C_{(m,m)} = (c_{kl}) = {}^t(1_n\ {}^t 1_m\ -\ X)\, X = 1_m\ {}^t 1_n\ X\ -\ {}^t X\ X = UW - A$

dont l'élément, $c_{kl} = |X^k - X^l| = |Z^k \cap \overline{X^l}| = \sum_{i=1}^{n} x_i^k (1 - x_i^l)$ est le nombre d'attributs présents dans x^k mais pas dans x^l.

$D_{(m,m)} = (d_{kl}) = {}^t(1_n\ {}^t 1_m\ -\ X)\,(1_n\ {}^t 1_m\ -\ X) = nU - (A + B + C)$

dont l'élément, $d_{kl} = |\overline{x^k} \cap \overline{x^l}| = \sum_{i=1}^{n} (1 - x_i^k)(1 - x_i^l)$ est le nombre d'attributs qui ne sont présents ni dans x^k ni dans x^l.

$X^l = \{i/x_i^l = 1\}$ et $X^k = \{i/x_i^k = 1\}$ étant les ensembles d'attributs présents respectivement dans les données du point-modalité x^l et x^k et $|.|$ désigne le cardinal de l'ensemble. Les quatre quantités sont liées : $\forall k = 1, p\,;\, \forall l = 1, p\ a_{kl} + b_{kl} + c_{kl} + d_{kl} = n$.

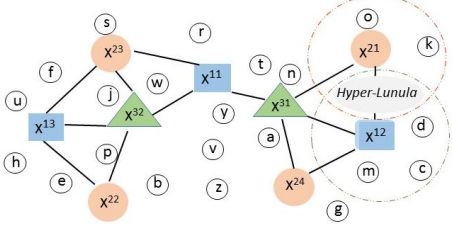
2 Equivalence topologique

L'équivalence topologique repose sur la notion de graphe topologique que l'on désigne également par graphe de voisinage. Deux mesures de proximité sont équivalentes si les graphes topologiques induits sur l'ensemble des objets restent identiques. Mesurer la ressemblance entre mesures de proximité revient à comparer leurs graphes de voisinage.

Soit l'ensemble $E = \{x^{11}, \ldots, x^{1m_1}, \ldots, x^{k1}, \ldots, x^{km_k}, \ldots, x^{p1}, x^{pm_p}\}$ à $|E| = m$ modalités dans $\{0,1\}^n$, associé aux p variables qualitatives. On peut à l'aide d'une mesure de proximité u, définir une relation de voisinage V_u qui sera une relation binaire sur $E \times E$.

Pour une mesure de proximité donnée u, on peut construire un graphe de voisinage sur l'ensemble des objets-modalités où les sommets sont les modalités et les arêtes sont définies par une relation de voisinage. Il existe de nombreuses définitions pour construire cette relation binaire de voisinage, par exemple, l'Arbre de Longueur Minimale, le Graphe de Gabriel, ou encore le Graphe des Voisins Relatifs (GVR)(Toussaint (1980)), dont les couples de points voisins (x^{kr}, x^{ls}) vérifient la propriété GVR suivante :

$$\begin{cases} V_u(x^{kr}, x^{ls}) = 1 & si\ \ u(x^{kr}, x^{ls}) \leq \max[u(x^{kr}, x^{qt}), u(x^{qt}, x^{ls})]; \\ & \forall x^{kr},\ x^{ls},\ x^{qt} \in E,\ x^{qt} \neq x^{kr}\ \ et\ \ x^{qt} \neq x^{ls} \\ V_u(x^{kr}, x^{ls}) = 0 & sinon \end{cases}$$



V_u	x^{11}	x^{12}	x^{13}	x^{21}	x^{22}	x^{23}	x^{24}	x^{31}	x^{32}
x^{11}	1	0	0	0	0	1	0	1	1
x^{12}	0	1	0	1	0	0	1	1	0
x^{13}	0	0	1	0	1	1	0	0	1
x^{21}	0	1	0	1	0	0	0	1	0
x^{22}	0	0	1	0	1	0	0	0	1
x^{23}	1	0	1	0	0	1	0	0	1
x^{24}	0	1	0	0	0	0	1	1	0
x^{31}	1	1	0	1	0	0	1	1	0
x^{32}	1	0	1	0	1	1	0	0	1

Figure 1: GVR de neuf modalités de trois variables - Matrice d'adjacence associée

Pour toute mesure de proximité donnée u, on peut lui associer une matrice dite d'adjacence V_u binaire et symétrique d'ordre $m = \Sigma_{k=1}^{p} m_k$. La Figure 1 illustre un ensemble de n objets-individus autour de $m = 9$ modalités associées à trois variables qualitatives x^1, x^2 et x^3 respectivement à trois, quatre et deux modalités. Par exemple, $V_u(x^{12}, x^{21}) = 1$ signifie sur le plan géométrique que l'hyper-Lunule (intersection des deux hypersphères centrées sur les deux points-modalités x^{12} et x^{21}) est vide.

- **Comparaison et sélection de mesures de proximité**

Pour mesurer l'équivalence topologique entre deux mesures de proximité u_i et u_j, nous proposons de tester si les matrices d'adjacence associées V_{u_i} et V_{u_j} sont différentes ou pas. L'indice d'équivalence topologique entre deux matrices d'adjacence est mesuré par la propriété de concordance suivante :

$$S(V_{u_i}, V_{u_j}) = \frac{\sum_{k=1}^{p}\sum_{r=1}^{m_k}\sum_{l=1}^{p}\sum_{s=1}^{m_l}\delta_{kr\,ls}(x^{kr},x^{ls})}{r^2} \quad \text{avec} \quad \delta_{kr\,ls}(x^{kr},x^{ls}) = \begin{cases} 1 & \text{si } \mathrm{V}_{u_i}(x^{kr},x^{ls}) = V_{u_j}(x^{kr},x^{ls}) \\ 0 & \text{ailleurs.} \end{cases}$$

La mesure de similarité $S(V_{u_i}, V_{u_j}) = 1$ signifie que les deux matrices d'adjacence sont identiques et par conséquent, la structure topologique induite par les deux mesures est la même. Dans ce cas, on parle d'équivalence topologique parfaite entre les deux mesures de proximité. La valeur $S(V_{u_i}, V_{u_j}) = 0$ signifie que la topologie a totalement changé.

On construit la matrice d'adjacence notée V_{u*}, qui correspond au mieux au tableau de Burt $\mathcal{B}$. Pour cela, on examine les similitudes entre les modalités à partir de l'écart entre chaque profil-modalité et son profil moyen, c'est-à-dire l'écart jusqu'à l'indépendance. Cette matrice d'adjacence dite de référence est définie comme suit:

$$\begin{cases} V_{u*}(x^{kr},x^{ls}) = 1 & si \;\; \frac{\mathcal{B}_{kr\,ls}}{\mathcal{B}_{kr\,..}} \geq \frac{\mathcal{B}_{kr\,..}}{np^2} \quad \forall k,l = 1,p\;;\; r = 1,m_k \; et \; s = 1,m_l \\ V_{u*}(x^{kr},x^{ls}) = 0 & sinon \end{cases}$$

$\mathcal{B}_{kr\,ls} = \Sigma_{i=1}^{n} x_i^{kr} x_i^{ls}$, est l'élément de la matrice de Burt qui correspond au nombre d'individus possédant la modalité r de la variable k et la modalité s de la variable l,

$\mathcal{B}_{kr\,..} = \Sigma_{l=1}^{p}\Sigma_{s=1}^{m_s} b_{kr\,ls}$, $\frac{\mathcal{B}_{kr\,ls}}{\mathcal{B}_{kr\,..}}$ et $\frac{\mathcal{B}_{kr\,..}}{np^2}$ désignent respectivement la marge ligne, le profil-ligne et le profil moyen de la modalité r de la variable k.

Cette matrice d'adjacence binaire et symétrique V_{u*} est associée à une mesure de proximité de référence, inconnue, notée u_*.

Pour visualiser les mesures de proximité, on peut par exemple appliquer une Classification Ascendante Hiérarchique (CAH) sur les facteurs significatifs de l'Analyse en Composantes Principales (ACP) du tableau des dissimilarités $[D]_{ij} = 1 - S(V_{u_i}, V_{u_j})_{i,j=1,22}$. De plus, pour déterminer la classe de mesures de proximité la plus proche de la mesure de référence u_*, cette dernière sera considérée comme élément illustratif dans les analyses, en projetant *a posteriori* le vecteur de dissimilarité $[D]_{*i} = 1 - S(V_{u*}, V_{u_i})_{i=1,22}$.

- **Comparaison statistique de l'équivalence topologique**

Soient V_{u_i} et V_{u_j} les matrices d'adjacence associées à deux mesures de proximité u_i et u_j. Pour comparer le degré d'équivalence topologique entre deux mesures, nous testons si les matrices d'adjacence associées sont statistiquement différentes ou pas, en utilisant le test non paramétrique de Kappa (Cohen, (1960)). Ces matrices binaires et symétriques d'ordre m, sont dépliées selon deux vecteurs de composantes appariées, formées des $\frac{m(m-1)}{2}$ valeurs supérieures (ou inférieures) de la diagonale.

Le degré d'équivalence topologique entre les deux mesures u_i et u_j est évalué et testé à partir du coefficient de concordance de Kappa, calculé sur le tableau 2×2 de contingence formé par les deux vecteurs :

$$\widehat{\kappa}(V_{u_i}, V_{u_j}) = \frac{P_o - P_e}{1 - P_e}$$

avec $\begin{cases} \mathrm{P}_o = \frac{2}{r(r+1)} \sum_{k=0}^{1} n_{kk} & \text{la proportion de concordance observée,} \\ \mathrm{P}_e = \frac{4}{r^2(r+1)2} \sum_{k=0}^{1} n_{k.} n_{.k} & \text{la proportion de concordance attendue sous l'hypothèse d'indépendance.} \end{cases}$

Il y a une parfaite indépendance d'accord ou de concordance sous l'hypothèse nulle $H_0 : \kappa = 0$. La concordance est d'autant plus élevée que sa valeur tend vers $+1$, parfaite ou maximale si $\kappa = 1$ et une parfaite discordance lorsque $\kappa = -1$.

- **Représentation graphique de l'association topologique**

Afin d'analyser et de visualiser graphiquement les éventuels liens entre les m modalités des p variables qualitatives, on propose d'effectuer l'ACMT qui revient à effectuer l'ACP du triplet $\{V_{u*}\ ;\ M\ ;\ D_m\}$ où V_{u*} est la matrice d'adjacence associée à la mesure de proximité $u*$, la mesure la plus adaptée aux données considérées, $M = I_m$ est la matrice identité d'ordre m et $D_m = \frac{W}{np}$ la matrice diagonale pondérée des poids des modalités. Cette analyse peut être effectuée à partir de n'importe quelle matrice d'adjacence V_u associée à chacune des 22 mesures de proximité u considérées.

- **Exemple d'application**

Pour illustrer l'ACMT, on a considéré les données d'une étude sur l'entrepreneuriat féminin réalisée à Dakar en 2014. Les données de la Table 1 ont été collectées auprès de 153 femmes entrepreneures de la région de Dakar, l'objectif ici est de donner une description topologique de leur signalétique.

Variables / Modalités	Age			Situation matrimoniale				Nombre d'enfants			Niveau d'études		
Moins de 25 ans	22	0	0	18	2	1	1	13	3	6	3	1	18
entre 25 et 50 ans	0	80	0	16	9	21	34	14	11	55	58	5	17
Plus de 50 ans	0	0	51	3	8	24	16	8	35	8	30	10	11
Célibataire	18	16	3	37	0	0	0	20	3	14	9	1	27
Divorcée	2	9	8	0	19	0	0	3	10	6	13	5	1
Mariée monogame	1	21	24	0	0	46	0	7	21	18	26	5	15
Mariée Polygame	1	34	16	0	0	0	51	5	15	31	43	5	3
Sans enfant	13	14	8	20	3	7	5	35	0	0	11	5	19
De 1 à 3 enfants	3	11	35	3	10	21	15	0	49	0	27	9	13
Plus de 3 enfants	6	55	8	14	6	18	31	0	0	69	53	2	14
Analphabète-Primaire	3	58	30	9	13	26	43	11	27	53	91	0	0
Secondaire	1	5	10	1	5	5	5	5	9	2	0	16	0
Supérieur	18	17	11	27	1	15	3	19	13	14	0	0	46

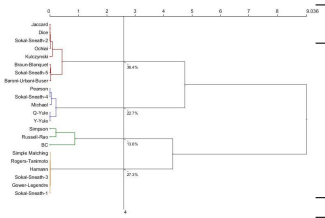
Table 1: Tableau de Burt - Entrepreneuriat féminin à Dakar, Sénégal

Les principaux résultats de l'approche topologique proposée sont présentés dans les tableaux et graphiques suivants. Ils permettent de visualiser les mesures de proximité proches les unes des autres selon les données considérées et de représenter dans un contexte topologique les associations entre les $m = 13$ modalités des $p = 4$ variables qualitatives signalétiques des entrepreneures.

On a établi l'ensemble des équivalences topologiques $S(V_{u_i}, V_{u_j})$ et $S(V_{u_i}, V_{u*})$ entre les 22 mesures de proximité considérées et avec la mesure de référence u_*. Tous les tests statistiques de Kappa sur les équivalences topologiques sont significatifs avec un

risque d'erreur $\alpha \leq 5\%$. Les similarités par paire diffèrent quelque peu, certaines sont plus proches que d'autres et certaines sont identiques en parfaite équivalence topologique $S(V_{u_i}, V_{u_j}) = 1$, avec une concordance parfaite $\widehat{\kappa}(V_{u_i}, V_{u_j}) = 1$.

Le dendrogramme et la table de la Figure 2 résument les principaux résultats de la partition retenue en quatre classes homogènes de mesures de proximité. La mesure de référence u_* est affectée à la classe 3 constituée des mesures de Russell & Rao, BC et Simpson. Ce sont là les mesures de proximité les plus adaptées pour l'analyse topologique de la signalétique des entrepreneures.



	Classe 1	Classe 2	Classe 3	Classe 4
Effectif	8	5	3	6
Mesure de proximité	$u_{Jaccard}$ u_{Dice} $u_{Kulczynski}$ u_{Ochiai} $u_{Sokal-Sneath-2}$ $u_{Baroni-Urbani}$ $u_{Sokal-Sneath-5}$ $u_{Braun-Blanquet}$	$u_{Pearson}$ $u_{Michael}$ u_{Q-Yule} u_{Y-Yule} $u_{Sokal-Sneath-4}$	$u_{Russell-Rao}$ $u_{Simpson}$ u_{BC}	$u_{Simple-Matching}$ $u_{Rogers-Tanimoto}$ u_{Hamann} $u_{Gower-Legendre}$ $u_{Sokal-Sneath-1}$ $u_{Sokal-Sneath-3}$
Référence	u_*			

Figure 2: Arbre hiérarchique & Affectation de la mesure de référence

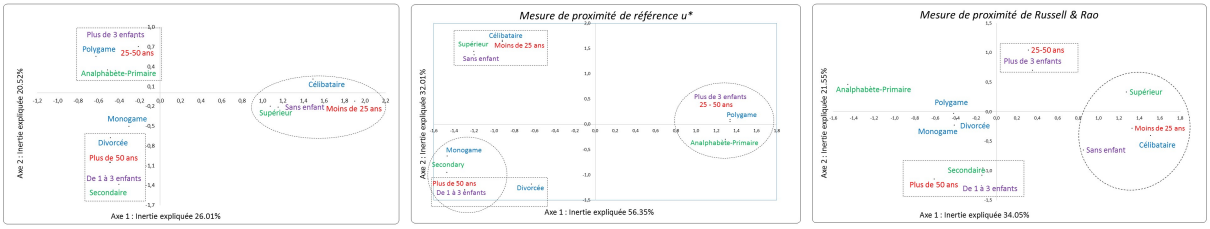


Figure 3: ACM & ACMT & ACMT Russell-Rao

La figure 3 présente à titre de comparaison, sur les premiers plans factoriels, les représentations graphiques des trois analyses des correspondances multiples, la classique (ACM) et deux topologiques (ACMT): l'analyse proposée et celle réalisée avec la mesure de proximité de Russell & Rao.

Contrairement aux deux autres méthodes qui ne décrivent que trois fortes liaisons, l'ACMT en fait ressortir quatre : deux qui s'opposent sur le premier axe factoriel (56.35%) et les deux autres sur le deuxième axe factoriel (32.01%). Les liaisons sont matérialisées par des formes géométriques.

3 Conclusion

Ce travail propose une nouvelle méthode topologique d'analyse des correspondances multiples (ACMT) qui vient enrichir les méthodes classiques d'analyse des données qualitatives.

Bibliographie

[1] Abdesselam, R. (2018), Sélection de mesures de proximité pour une analyse des correspondances topologique. *Actes des 50èmes Journées de Statistique*, Société Franaise de Statistique, SFdS-2018, Paris, Saclay, France.
[2] Cohen, J. (1960), A coefficient of agreement for nominal scales. *Educ Psychol Meas*, Vol 20, 27–46.
[3] Toussaint, G. T. (1980), The relative neighbourhood graph of a finite planar set. *In Pattern recognition*, 12, 4, 261–268.
[4] Warrens, M. J. (2008), Bounds of resemblance measures for binary (presence/absence) variables. In Journal of Classification, *Springer*, 25, 2, 195–208.
[5] Zighed, D., Abdesselam, R., and Hadgu, A. (2012), Topological comparisons of proximity measures. *16th PAKDD 2012 Conference*, Part I, LNAI 7301, Springer, 379–391.