Inversion d'un tableau de Burt

F.-G. Carpentier

1 Introduction

L'enseignement des statistiques en général et de l'ACM en particulier à un public de non spécialistes (psychologie, sociologie, etc.) nécessite de faire travailler les étudiants sur des exemples pertinents et relevant de leur discipline.

De tels exemples ne manquent pas. Il suffit de consulter les publications des chercheurs de la discipline concernée pour en trouver. Malheureusement, ces publications indiquent toujours les résultats des traitements statistiques effectués, mais n'indiquent pratiquement jamais les données observées. Il est donc particulièrement tentant d'essayer de générer des données fictives mais "possibles", autrement dit des données conduisant à des résultats identiques ou proches de ceux publiés.

En ce qui concerne plus particulièrement l'ACM, les exemples trouvés dans les manuels sont généralement donnés sous la forme d'un tableau de Burt, et les résultats figurant dans les publications permettent au mieux de retrouver ce tableau. Il est pourtant essentiel que l'étudiant puisse confronter les résultats de l'ACM à l'ensemble des réponses faites par les sujets pour mieux appréhender la méthode, et c'est bien cette démarche qui sera la sienne, s'il utilise l'ACM sur des données qu'il aura recueillies.

Le problème abordé dans cette note est donc le suivant : dans une situation relevant de l'ACM, comment retrouver un ensemble "vraisemblable" de réponses individuelles à partir d'un tableau de Burt.

2 Notations et contexte

2.1 Notations

On a recueilli sur une population d'effectif N les valeurs prises par Q variables nominales (les "questions"). La variable q possède K_q modalités.

Soit $K = \sum_{q=1}^{Q} K_q$ le nombre total de modalités.

Chaque réponse individuelle peut être codée sous forme disjonctive à l'aide de la matrice $X = (x_j) \in \mathcal{M}_{1,K}$ où $x_j = 1$ si x_j correspond à une modalité de réponse choisie par le sujet, et $x_j = 0$ sinon.

On peut aussi, éventuellement, introduire les notations D_q et F_q, indices respectifs de la première et de la dernière modalités de la question q. On a ainsi : $D_1 = 1, F_1 = K_1, D_2 = K_1 + 1, F_2 = K_1 + K_2, \ldots, F_Q = K$. La condition précédente s'écrit alors : $\forall q \quad \sum_{D_q}^{F_q} = 1.$

Les réponses possibles sont appelées "patrons de réponses". Le nombre de tels patrons est donné par : $\text{NP} = \prod_{q=1}^{Q} K_q$. L'ensemble des patrons pourra aussi être noté : $\{P_p / p \in [1..\text{NP}]\}$ avec une indexation convenable de l'ensemble des patrons.

Le tableau disjonctif complet (TDC) des réponses est une matrice $\text{TDC} = (x_{ij}) \in \mathcal{M}_{N,K}$ où la ligne i (notée X_i) est la réponse du sujet i.

Le tableau disjonctif des patrons, TDP est obtenu en additionnant les lignes identiques du TDC : les coefficients non nuls d'une ligne indiquent alors le nombre de sujets ayant choisi le patron de réponse correspondant.

Le tableau de Burt est la matrice symétrique $B = (b_{ij}) \in \mathcal{M}_{K,K}$ telle que b_{ij} soit égal au nombre de réponses comportant à la fois la modalité i et la modalité j. Plus précisément :

- Si $i = j$, b_{ii} est l'effectif de la modalité i ;
- Si i et j désignent deux modalités distinctes d'une même variable, alors $b_{ij} = 0$;
- Si i et j désignent des modalités provenant de deux variables différentes, alors b_{ij} est le nombre de sujets ayant choisi conjointement l'une et l'autre de ces modalités.

2.2 Quelques relations remarquables

2.2.1 Liens entre le tableau de Burt et les autres variables introduites

Somme des coefficients du tableau de Burt :

$$\sum_i \sum_j b_{ij} = Q^2 N$$

Somme sur une ligne, sur une colonne : pour tout i_0 et tout j_0

$$\sum_i b_{ij_0} = \sum_j b_{i_0 j} = QN$$

Somme des coefficients diagonaux : pour toute question j

$$\sum_{i=D_j}^{F_j} b_{ii} = N$$

Somme des coefficients sur une ligne ou une colonne d'un des tableaux de contingence : pour $i \in \{D_q..F_q\}$ et $q' \neq q$

$$\sum_{j=D_{q'}}^{F_{q'}} b_{ij} = b_{ii}$$

Etant donné une matrice M, on désigne par M' la transposée de M.

Le tableau de Burt et le TDC sont liés par les relations suivantes :

$$B = \text{TDC}' \ \text{TDC}$$

$$B = \sum_{i=1}^{N} X_i' X_i$$

où la somme est étendue à l'ensemble des sujets et où X_i désigne la réponse du sujet i ;

$$B = \sum_{p=1}^{\mathrm{NP}} n_p X_p' X_p$$

où la somme est étendue à l'ensemble des patrons de réponses, et où n_p désigne l'effectif du patron X_p.

En revanche, le produit TDP′TDP vaut $\sum n_p^2 X_p' X_p$, qui est en général différent du tableau de Burt.

2.2.2 Taux de liaison dans le tableau de Burt

Le tableau $T = (t_{ij})$ des taux de liaison entre individus lignes et individus colonnes est défini classiquement par :

$$t_{ij} = N \frac{b_{ij}}{b_{i\cdot} b_{\cdot j}} - 1$$

où $b_{i\cdot}$ et $b_{\cdot j}$ désignent respectivement la somme des coefficients de la ligne i et celle de la colonne j.

Pour un tableau de Burt, on vérifie que :

- Si i et j correspondent à deux questions différentes q et q', t_{ij} est le taux de liaison "classique" entre les deux modalités, dans le tableau de contingence des variables q et q' ;
- Si i et j correspondent à la même variable, avec $i \neq j$, alors $t_{ij} = -1$;
- Si $i = j$, $t_{ii} = \dfrac{N}{b_{ii}} - 1$; autrement dit, t_{ii} est positif, d'autant plus grand que la modalité est d'effectif plus faible.

Par ailleurs, dans le cas où une ligne ou une colonne du tableau B est de somme nulle, on convient d'attribuer la valeur -1 à tous les taux de liaison de cette ligne ou cette colonne.

3 Position du problème

Le tableau de Burt est entièrement déterminé par la donnée du TDC, ou du TDP. Mais, des TDC différents peuvent conduire au même tableau de Burt.

Par exemple, supposons que l'on ait 3 variables comportant chacune deux modalités, avec l'indexation suivante de l'ensemble des patrons :

X_1	1	0	1	0	1	0
X_2	1	0	1	0	0	1
X_3	1	0	0	1	1	0
X_4	1	0	0	1	0	1
X_5	0	1	1	0	1	0
X_6	0	1	1	0	0	1
X_7	0	1	0	1	1	0
X_8	0	1	0	1	0	1

On constate que l'on peut ajouter :

$$X_1 - X_2 - X_3 + X_4 - X_5 + X_6 + X_7 - X_8$$

à un ensemble de réponses sans modifier pour autant le tableau de Burt, ou encore que $X_1 + X_4 + X_6 + X_7$ et $X_2 + X_3 + X_5 + X_8$ correspondent au même tableau de Burt.

On se pose donc le problème suivant :

> Etant donné un tableau de Burt B, trouver un tableau disjonctif complet S tel que $S'S = B$.

ou encore :

> Trouver des coefficients n_p tels que : $\sum n_p X'_p X_p = B$.

Ce problème est en apparence simple au niveau théorique : il s'agit de résoudre un système d'équations linéaires, comportant K^2 équations et NP inconnues. Mais, ce système comporte de nombreuses équations non principales et de nombreuses inconnues non principales, et par ailleurs, on exige que la solution trouvée soit formée de nombres entiers positifs ou nuls.

La première idée est de résoudre le problème pas à pas, en déterminant un patron X_p tel que $X'_p X_p$ puisse être retranché de B en préservant la positivité des coefficients de B puis en appliquant la méthode sur le nouveau tableau de Burt obtenu. Mais on se retrouve rapidement avec un tableau de Burt non nul, mais dont aucun patron ne peut plus être retiré.

4 Algorithme d'inversion du tableau

4.1 Première étape de l'algorithme

Etant donné un patron de réponse X , on peut calculer les taux de liaison des différentes combinaisons des modalités qui le composent.

Soit T le tableau des taux de liaison du tableau de Burt courant, dans lequel on a mis à 0 les coefficients t_{ii} de la diagonale. Introduisons le tableau $B(X) = X'X$ (tableau de Burt associé au patron X). Les taux de liaison recherchés sont les coefficients non nuls du tableau $B(X) \otimes T$, où $\otimes$ désigne la multiplication terme à terme des deux tableaux.

Par ailleurs, le patron X représente une réponse susceptible d'avoir été fournie si et seulement si $B(X) \otimes T$ ne comporte pas de coefficient égal à -1.

On peut penser qu'un patron a d'autant plus de chances d'avoir été présent dans les données initiales que les taux de liaison entre les modalités qui le compose sont élevés. D'où l'idée de calculer le score d'un patron réponse à l'aide de la formule :

$$Sc(X) = \text{Minnn}\{B(X) \otimes T\}$$

où Minnn désigne le minimum (positif ou négatif) des valeurs *non nulles* de l'ensemble considéré.

La première partie de l'algorithme consiste donc à effectuer les opérations suivantes :

1. Le tableau de Burt courant est le tableau de Burt initial
2. Tant que le nombre de patrons retirés est strictement inférieur au nombre de sujets

(a) Calcul des taux de liaison du tableau de Burt courant

(b) Calcul du maximum M des scores des patrons de réponses par rapport au tableau de Burt courant et de l'indice p du patron correspondant

(c) Si $M = -1$, on exécute la la deuxième étape (décrite ci-dessous)

(d) Sinon (donc si $M > -1$), on place le patron X_p dans l'ensemble des solutions, on calcule le nouveau tableau de Burt courant : $B \leftarrow B - B(X_p)$.

3. Fin de Tant que

4.2 Deuxième étape de l'algorithme

A ce stade de la résolution du problème, l'idée est de modifier l'ensemble des solutions, de la façon la plus mineure possible, de façon à pouvoir retirer du tableau de Burt des patrons correspondant aux réponses restantes. En particulier, on cherche un algorithme permettant de modifier l'ensemble des solutions, sans faire diminuer le nombre de réponses traitées.

Un patron peut être vu comme un ensemble de $\frac{Q(Q-1)}{2}$ liens entre les modalités qui le composent.

Par exemple, avec 3 questions à 2 modalités chacune, le patron $\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$ est l'ensemble des trois liens :

$$Q_1M_1 \leftrightarrow Q_2M_3$$
$$Q_2M_3 \leftrightarrow Q_3M_6$$
$$Q_1M_1 \leftrightarrow Q_3M_6$$

A chacun de ces liens correspond un coefficient dans le tableau de Burt. Un patron peut être retiré si tous ses liens correspondent à des coefficients positifs. A l'inverse, si un patron ne peut pas être retiré, il existe au moins un lien $Q_aM_i \leftrightarrow Q_bM_j$ pour lequel $b_{ij} = 0$.

Le retrait d'un patron supplémentaire est alors réalisé selon le schéma suivant :

- On identifie un patron X ("patron candidat") qui pourrait être retiré du tableau B pourvu que l'on rétablisse le lien $Q_aM_i \leftrightarrow Q_bM_j$
- On recherche alors dans les solutions déjà trouvées deux patrons Y_1 et Y_2 tels que :
  - Y_1 et Y_2 sont identiques sur les questions autres que Q_a et Q_b
  - Y_1 contient le lien $Q_aM_i \leftrightarrow Q_bM_j$
  - Y_2 contient un lien $Q_aM_{i'} \leftrightarrow Q_bM_{j'}$, avec $i \neq i'$ et $j \neq j'$
  - Les coefficients $b_{i'j}$ et $b_{ij'}$ du tableau de Burt courant sont non nuls (strictement positifs)

Si de tels patrons X, Y_1 et Y_2 sont trouvés :

- On retire de l'ensemble des solutions les patrons Y_1 et Y_2
- On ajoute au tableau de Burt courant les tableaux $Y_1'Y_1$ et $Y_2'Y_2$
- On peut alors retirer du tableau de Burt courant les trois patrons suivants :
  - Le patron Z_1, comportant le lien $Q_aM_{i'} \leftrightarrow Q_bM_j$, identique à Y_1 et Y_2 sur les autres questions
  - Le patron Z_2, comportant le lien $Q_aM_i \leftrightarrow Q_bM_{j'}$, identique à Y_1 et Y_2 sur les autres questions
  - Le patron candidat X

L'état du tableau de Burt à chaque étape est résumé ci-dessous.

Etat inital

		Q_b	
		j	j'
Q_a	i	0	$b_{ij'}$
	i'	$b_{i'j}$	$b_{i'j'}$

Ajout de Y_1 et Y_2

		Q_b	
		j	j'
Q_a	i	1	$b_{ij'}$
	i'	$b_{i'j}$	$b_{i'j'}+1$

Retrait de Z_1 et Z_2

		Q_b	
		j	j'
Q_a	i	1	$b_{ij'}-1$
	i'	$b_{i'j}-1$	$b_{i'j'}+1$

Retrait de X

		Q_b	
		j	j'
Q_a	i	0	$b_{ij'}-1$
	i'	$b_{i'j}-1$	$b_{i'j'}+1$

L'algorithme permettant de réaliser dans cette deuxième étape dans la boucle "tant que" de la première étape est alors le suivant :

On dispose d'une liste de "patrons exclus", initialisée à $\emptyset$.

1. On sélectionne, en dehors de la liste de patrons exclus, le patron X dont le nombre de "liens brisés" est minimal, c'est-à-dire le patron qui, pour pouvoir être retiré, exigera le moins grand nombre de liens rétablis, et si possible un seul lien
2. On recherche Y_1 et Y_2 comme décrit ci-dessus
3. Si la recherche précédente aboutit, on opère la substitution $-Y_1 - Y_2 + Z_1 + Z_2$ et, si de plus, le patron X avait un seul "lien brisé", on retire X.
4. Si la recherche précédente n'aboutit pas, on place X dans la liste des patrons exclus
5. Fin de la deuxième étape

5 Exemple de mise en oeuvre

L'algorithme précédent a été implémenté dans le programme R disponible sur ce site. Un exemple de démonstration est fourni avec ce programme. L'exemple, adapté de [Crucianu M., Asselin de Beauville J-P., Boné R., Méthodes factorielles pour l'analyse de données, Hermès 2004], comporte 5 questions, 22 modalités, et 383 sujets.

Le programme retire 378 des 383 patrons avant de rentrer dans la deuxième étape, ce qui montre la pertinence de l'utilisation des taux de liaison pour construire le critère de choix du patron à retirer. La deuxième étape extrait ensuite 4 des 5 patrons restants, le 5è et dernier, quant à lui, est directement extrait du tableau de Burt.

Le temps d'exécution est de l'ordre de 30 mn sur un MacBook G3 à 400 Mhz, et de 10 mn sur un ordinateur équipé d'un processeur Intel à 2 Ghz.

6 Conclusion et discussion

La méthode semble donner des résultats pertinents sur les exemples traités. Certes, certaines réponses hautement improbables peuvent se retrouver dans les patrons choisis, mais avec un effectif très faible. Cela est dû essentiellement aux propriétés du tableau de Burt (cf section 3).

La méthode converge-t-elle toujours vers une solution ? La réponse à cette question est difficile, car nous n'avons pas de critère montrant de façon absolue qu'un tableau est un tableau de Burt, c'est-à-dire a été obtenu à partir d'un tableau disjonctif complet. Ainsi, une non-convergence de la méthode peut simplement être due à une erreur dans les données d'entrée. Il serait d'ailleurs souhaitable de faire vérifier par le programme que le tableau de Burt fourni en entrée vérifie les critères simples de cohérence (égalité des effectifs dans tous les tableaux de contingence juxtaposés, notamment).