MATHÉMATIQUES ET SCIENCES HUMAINES

CLAUDE FLAMENT

L'analyse booléenne de questionnaires

Mathématiques et sciences humaines, tome 12 (1965), p. 3-10
<http://www.numdam.org/item?id=MSH_1965__12__3_0>

Claude FLAMENT *

Cet article fait appel à un certain nombre de notions sur les Algèbres de Boole finies et les Simplexes. Pour les définitions des termes qui ne sont pas familiers au lecteur, celui-ci pourra consulter :

- *Simplexe, Algèbre simpliciale : P. ROSENSTIEHL, M.S.H. N° 9;*
- *Algèbres de Boole, Algèbres de Boole libres: M. EYTAN, M.S.H. N° 10, pp. 54-44.*
- *Morphismes, idéaux, quotients: Marc BARBUT, M.S.H. N° 12 (le présent numéro), p*

N.D.L.R.

L'ANALYSE BOOLEENNE DE QUESTIONNAIRES

Dans le but d'organiser et d'interpréter un protocole de passation de questionnaire (d'attitudes, d'opinions, ...), on recourt souvent à l'analyse hiérarchique qui tend à faire apparaître une échelle de Guttman (cf. MATALON, 1965): on sait du reste que cette méthode ne s'applique pas qu'aux questionnaires, comme l'a montré, par exemple, V. ELISSEEF (1965) en étudiant les bronzes chinois archaïques.

Cependant, il est exceptionnel, qu'on trouve, dans des données d'observation, une échelle de Guttman parfaite, ou même valablement approximative. D'où, la recherche d'analyses plus générales que celle de Guttman. Coombs et Kao (cf. COOMBS, 1964, chap. 12) ont proposé une généralisation multidimentionnelle de l'analyse hiérarchique, que MATALON (1965, chap. 7) a algébrisée, et que nous avons encore généralisée par ce qu'on peut appeler les "modèles de fermeture" (FLAMENT, 1965). Ces travaux utilisent l'algèbre simpliciale sur les parties de l'ensemble des questions (si chaque question ne comporte que deux réponses, exclusives et exhaustives); ce qui correspond à l'une des présentations possibles des échelles de Guttman (considérées comme chaînes maximales d'inclusion sur le simplexe). Mais il existe au moins une autre présentation des échelles de Guttman, et donc sans doute, une autre voie de généralisation. Il s'agit d'une présentation utilisant la notion d'implication, qui nous conduira à l'analyse booléenne de questionnaire.

*

* *

* Laboratoire de Psychologie Sociale, Aix en Provence.

Soit un questionnaire dont chaque question X admet deux réponses, x et x', exclusives et exhaustives (un individu interrogé donne une réponse et une seule à chaque question). Soit R l'ensemble des patrons P possibles, ou suites de réponses qu'un individu peut donner; par exemple, si on a 3 questions A, B, C, (a b c') est un patron P de R.

A partir d'un protocole de passation, nous définissons l'ensemble R_1 des patrons représentés, c'est-à-dire, donnés chacun par au moins un membre de la population interrogée; R_0, complémentaire de R_1 par rapport à R, est l'ensemble des patrons non représentés.

Si, pour notre questionnaire (A, B, C), nous obtenons: $R_1 = \{P_1, P_2, P_3, P_4\}$, avec :

P_1 : a b c
P_2 : a' b c
P_3 : a' b' c
P_4 : a' b' c',

nous disons que R_1 constitue une échelle de Guttman. On sait alors (cf. MATALON, 1965, p. 42) que, si on "croise" les questions deux à deux, on obtient toujours des tableaux à 4 cases, dont l'une est vide, comme dans cet exemple:

		A	
		a	a'
B	b	P_1	P_2
	b'		P_3, P_4

On voit alors que si un individu donne la réponse a, cela implique qu'il donne aussi la réponse b (s'il donnait b', ayant donné a, la case (a b') ne serait pas vide). Ce que nous symboliserons par :

$$a \longrightarrow b \ ;$$

(on remarque que $(a \rightarrow b)$ est équivalent à $(b' \rightarrow a')$).

Ce genre d'analyse conduit à exprimer l'échelle de Guttman que nous avons observée par le schéma d'implication:

$$a \rightarrow b \rightarrow c \ ,$$

ou, de façon équivalente :

$$c' \rightarrow b' \rightarrow a' .$$

Dans ces schémas, il est inutile de porter $(a \rightarrow c)$ ou $(c' \rightarrow a')$, puisque l'implication est une relation transitive.

Supposons maintenant que nous ayons 4 questions $\underline{A}$, $\underline{B}$, $\underline{C}$, $\underline{D}$, et que les patrons représentés constituent l'ensemble $\underline{R}_1$ suivant :

$\underline{P}_1$:	$\underline{a}$	$\underline{b}$	$\underline{c}$	$\underline{d}$
$\underline{P}_2$:	$\underline{a}'$	$\underline{b}$	$\underline{c}$	$\underline{d}$
$\underline{P}_3$:	$\underline{a}'$	$\underline{b}$	$\underline{c}'$	$\underline{d}$
$\underline{P}_4$:	$\underline{a}'$	$\underline{b}'$	$\underline{c}$	$\underline{d}$
$\underline{P}_5$:	$\underline{a}'$	$\underline{b}'$	$\underline{c}'$	$\underline{d}$
$\underline{P}_6$:	$\underline{a}'$	$\underline{b}'$	$\underline{c}'$	$\underline{d}'$

Un tel protocole ne peut pas se résumer parfaitement par une échelle de Guttman - mais rien ne nous empêche de considérer les 6 "tableaux croisés" $\underline{A} \times \underline{B}$, $\underline{A} \times \underline{C}$, etc..., et d'en déduire un schéma d'implication; on trouve le schéma suivant :

$$\begin{tikzcd} & \underline{b} \arrow[dr] & \\ \underline{a} \arrow[ur] \arrow[dr] & & \underline{d} \; ; \\ & \underline{c} \arrow[ur] & \end{tikzcd}$$

le tableau $(\underline{B} \times \underline{C})$ ne comporte aucune case vide, d'où l'absence de flèche entre $\underline{b}$ et $\underline{c}$.

On peut considérer ce schéma d'implication comme résumant le protocole d'observation, mais il faut remarquer que <u>ce résumé n'est pas univoque</u>; en effet, si nous supprimons le patron $\underline{P}_2$, ou le patron $\underline{P}_5$, ou les deux à la fois, on constate qu'on obtient toujours le même schéma d'implication. D'où, deux problèmes :

1) Comment trouver la classe des protocoles qui admettent un même schéma d'implication;

2) Comment généraliser cette analyse d'implication de telle sorte qu'il y ait correspondance biunivoque entre l'ensemble des protocoles et l'ensemble des schémas.

Quelque soit l'intérêt du premier problème, nous ne traiterons ici que du second. L'idée est de considérer, non seulement le croisement des questions deux à deux, mais aussi trois à trois, quatre à quatre, etc.., et de trouver des implications non seulement entre deux réponses, mais aussi entre deux groupes de réponses. Ainsi, on montre que, si dans l'exemple précédent on supprime $\underline{P}_2$, le protocole admet pour résumé univoque le schéma suivant :

$$\begin{tikzcd} & & \underline{b} \arrow[dr] & \\ (\underline{b} \text{ et } \underline{c}) \arrow[r] & \underline{a} \arrow[ur] \arrow[dr] & & \underline{d} \; . \\ & & \underline{c} \arrow[ur] & \end{tikzcd}$$

Il semble commode de présenter cette généralisation dans le cadre de l'algèbre de Boole libre qu'on peut construire sur le questionnaire étudié.

*

* *

Soit donc un questionnaire comportant n questions X_i ($i = 1, 2, \ldots, n$): chaque question X_i admet deux réponses, x_i, x'_i de telle sorte que tout individu interrogé donne soit x_i, soit x'_i.

Pour chaque question, choisissons arbitrairement une réponse; pour des commodités d'écriture, choisissons les x_i, et considérons l'ensemble $Q = \{x_i \mid i = 1, 2, \ldots, n\}$ comme l'ensemble des generateurs d'une algèbre de Boole libre Q, correspondant au questionnaire étudié.

Intuitivement, un terme x_i signifie: "donner la réponse x_i"; le conjugué de x_i est sa négation: "ne pas donner x_i", c'est-à-dire: "donner la réponse x'_i; l'opération sup ($\vee$) se lira ou ($x_i \vee x_j$: "donner la réponse x_i, ou la réponse x_j, ou les deux à la fois"), et l'opération inf ($\wedge$) se lira ET ($x_i \wedge x_j$; "donner à la fois x_i et x_j"). Le pôle supérieur V peut être considéré comme signifiant la "configuration (de réponses) universelle", puisque $V = x_i \vee x'_i$, ce qui est toujours le cas, les réponses étant exhaustives; et le pôle inférieur Λ, la "configuration absurde", puisque $\Lambda = x_i \quad x'_i$, ce qui ne peut jamais se présenter concrètement, les réponses étant exclusives.

Notons que les éléments de Q qui s'écrivent à l'aide seulement des n générateurs, de l'opération inf et de la conjugaison ($/$), correspondent aux patrons de réponses constituant l'ensemble R défini précédemment. Nous appellerons donc ces termes, des patrons, et les désignerons par P. Nous appellerons sous-patrons p un élément de Q écrit à l'aide d'une partie seulement des générateurs, de l'opération inf et de la conjugaison. Ainsi, si $n = 3$, $(x_1 \wedge x'_2 \wedge x_3)$ est un patron, et $(x_1 \wedge x'_2)$ est un sous-patron.

Si, à une population donnée, nous posons le questionnaire à partir duquel nous avons construit notre algèbre Q, nous observons, comme on l'a vu, un ensemble R_1 de patrons représentés, et un ensemble R_0 de patrons non représentés. A partir de cela, il nous faut maintenant définir une fonction de représentation r sur Q :

$$Q \xrightarrow{\quad r \quad} \{0,1\};$$

si α est un élément de Q, $r(\alpha) = 1$ signifie que α est représenté; $r(\alpha) = 0$, qu'il est non représenté.

Etant donné l'ensemble R_0 des patrons non représentés, nous définissons r par les règles suivantes :

1 – $r(\Lambda) = 0$;

2 – $r(P) = 0$ si le patron P est non représenté (appartient à R_0);

3 – α et β étant des éléments de Q, $r(\alpha \vee \beta) = 0$ si $r(\alpha) = r(\beta) = 0$;

4 – si un élément α de Q n'est pas déclaré non représenté par l'application des règles précédentes, alors il est représenté: $r(\alpha) = 1$.

On peut se demander si $\underline{r}$ n'est pas un épimorphisme de $\underline{Q}$ sur l'ensemble $\{0, 1\}$ considéré comme une algèbre de Boole; il n'en est rien, comme le montrent les tableaux suivants:

$\underline{r}\ (\alpha)$	$\underline{r}\ (\beta)$	$\underline{r}\ (\alpha \vee \beta)$	$\underline{r}\ (\alpha \wedge \beta)$
0	0	0	0
0	1	1	0
1	0	1	0
1	1	1	?

$\underline{r}\ (\alpha)$	$\underline{r}\ (\alpha')$
0	1
1	?

Les valeurs de $\underline{r}$ portées dans les colonnes de droite de ces tableaux se démontrent assez facilement à partir des règles de construction de la fonction $\underline{r}$. Les points d'interrogation ne signifient pas que les valeurs de $\underline{r}\ (\alpha \wedge \beta)$ et de $\underline{r}\ (\alpha')$ sont indéterminées: elles ne peuvent se déduire des valeurs de $\underline{r}\ (\alpha)$ et $\underline{r}\ (\beta)$, mais seulement, se construire par les règles définissant $\underline{r}$; et, suivant les cas, on peut obtenir soit 1, soit 0. Ceci suffit à montrer que $\underline{r}$ n'est pas un épimorphisme, et suggère, par ailleurs, une certaine dissymétrie (parfois gênante) entre les deux opérations sup et inf, tout comme entre les deux valeurs de $\underline{r}$.

Quoiqu'il en soit, nous pouvons considérer l'ensemble $\underline{Q}_0$ des éléments non représentés de $\underline{Q}$:

$$\alpha \in \underline{Q}_0 \Longleftrightarrow \underline{r}\ (\alpha) = 0.$$

On voit alors que: $\underline{Q}_0$ est un idéal de $\underline{Q}$.

En effet, un idéal $\underline{I}$ de $\underline{Q}$ est une partie de $\underline{Q}$ telle que:

- $\Lambda \in \underline{I}$;
- si $\alpha \in \underline{I}$, pour tout $\beta \in \underline{Q}$, on a: $(\alpha \wedge \beta) \in \underline{I}$;
- si α et $\beta \in \underline{I}$, alors $(\alpha \vee \beta) \in \underline{I}$.

A partir de la définition de $\underline{Q}_0$, des règles de construction de la fonction $\underline{r}$ et des tableaux donnés ci-dessus, on vérifie aisèment que $\underline{Q}_0$ a bien ces trois propriétés.

$\underline{Q}$ étant fini, on sait qu'il existe un élément unique ε de $\underline{Q}_0$ qui absorbe, pour l'opération sup, tous les éléments de $\underline{Q}_0$; ε est le sup de tous les éléments de $\underline{Q}_0$; et la donnée de ε suffit à définir entièrement $\underline{Q}_0$.

On voit alors facilement qu'à chaque protocole d'observation (défini par $\underline{R}_1$ et $\underline{R}_0$) correspond un élément ε unique (qui se calcule comme le sup des patrons non représentés) et qu'à chaque élément de $\underline{Q}$ (autre que le pôle supérieur V) correspond un protocole unique. Nous avons une correspondance bi-univoque entre $\underline{Q}$ et l'ensemble des protocoles.

En d'autres termes, <u>ε résume entiérement, et de façon univoque, tout ce qui nous intéresse</u> dans les résultats de la passation de notre questionnaire.

*

*   *

On définit sur une algèbre de Boole, une relation d'<u>implication</u> (ou d'absorbtion):

$$(\alpha \longrightarrow \beta) \Longleftrightarrow (\alpha \wedge \beta = \alpha).$$

On démontre qu'il s'agit d'une relation d'ordre, et par ailleurs que:
- si $((\alpha \vee \beta) \longrightarrow \gamma)$, alors $(\alpha \longrightarrow \gamma)$ et $(\beta \longrightarrow \gamma)$
- si $((\alpha \wedge \beta) \longrightarrow \gamma)$, alors $(\alpha \longrightarrow \gamma \vee \beta')$ et $(\beta \longrightarrow \gamma \vee \alpha')$
- si $(\alpha = \Lambda)$, alors $(\alpha \longrightarrow \Lambda)$.

Nous allons utiliser cette relation dans une nouvelle algèbre de Boole, $\underline{Q}/\underline{Q}_0$, définie comme l'algèbre quotient de $\underline{Q}$ par l'idéal $\underline{Q}_0$. Cette algèbre est généralement non <u>libre</u>; elle est isomorphe au simplexe des parties de l'ensemble $\underline{R}_1$ des patrons représentés. On peut l'écrire en conservant les générateurs de $\underline{Q}$ et en introduisant entre eux les liaisons nécessaires. Pour ce faire, il suffit de poser $\mathcal{E}$, le <u>sup</u> de $\underline{Q}_0$ dans $\underline{Q}$, égal à Λ , le pôle inférieur de $\underline{Q}/\underline{Q}_0$.

Or, on sait que tout élément d'une algèbre de Boole peut s'écrire canoniquement sous la forme d'un <u>sup</u> d'une partie des éléments que nous avons qualifiés de <u>sous patrons</u> (ou de patrons) (n'utilisant que les générateurs, et les opérations <u>inf</u> et de conjugaison).

Donc, si $\underline{p}$ fait partie de l'écriture de $\mathcal{E}$, comme $\mathcal{E} = \Lambda$, et donc $\mathcal{E} \longrightarrow \Lambda$, on peut écrire $\underline{p} \longrightarrow \Lambda$; et si $\underline{p} = \underline{x_i} \wedge \underline{x'_j} \wedge \underline{x_k}$, on peut écrire: $\underline{x_i} \wedge \underline{x'_j} \longrightarrow \underline{x'_k}$, ou: $\underline{x_i} \wedge \underline{x_k} \longrightarrow \underline{x_j}$, ou encore: $\underline{x'_j} \wedge \underline{x_k} \longrightarrow \underline{x'_i}$.

Si on procède ainsi pour tous les sous-patrons faisant partie de l'écriture de $\mathcal{E}$, on obtient un <u>schéma d'implication</u> qui est très exactement la solution de notre problème.

Reprenons l'un des protocoles mentionnés en introduction. Nous avions 4 questions: A, B, C, D, et $\underline{R}_1 = \{\underline{P}_1, \underline{P}_3, \underline{P}_4, \underline{P}_5, \underline{P}_6\}$; avec :

$$\begin{array}{ll}
\underline{P}_1 : & \underline{a}\ \underline{b}\ \underline{c}\ \underline{d} \\
\underline{P}_3 : & \underline{a'}\ \underline{b}\ \underline{c'}\ \underline{d} \\
\underline{P}_4 : & \underline{a'}\ \underline{b}\ \underline{c}\ \underline{d} \\
\underline{P}_5 : & \underline{a'}\ \underline{b'}\ \underline{c'}\ \underline{d} \\
\underline{P}_6 : & \underline{a'}\ \underline{b'}\ \underline{c'}\ \underline{d'};
\end{array}$$

on en déduit $\underline{R}_0$ par complémentarité, et on obtient:

$$\mathcal{E} = (\underline{a} \wedge \underline{b'}) \vee (\underline{a} \wedge \underline{c'}) \vee (\underline{b} \wedge \underline{d'}) \vee (\underline{c} \wedge \underline{d'}) \vee (\underline{a'} \wedge \underline{b} \wedge \underline{c}),$$

qui se traduit par divers schémas d'implication équivalents, dont l'un est bien celui que nous avions donné:

$$\begin{tikzcd}
 & & \underline{b} \arrow[dr] & \\
(\underline{b} \wedge \underline{c}) \arrow[r] & \underline{a} \arrow[ur] \arrow[dr] & & \underline{d} \\
 & & \underline{c} \arrow[ur] &
\end{tikzcd}$$

*

*   *

Prenons un <u>exemple réel</u>: une petite partie d'une enquête effectuée, il y a plusieurs années, par l'I.F.O.P., auprès d'un échantillon de plus de 3.000 Français. De cette enquête, nous n'avons retenu que 4 questions; voici les 4 générateurs de l'algèbre utilisée:

$\underline{g}$: être favorable à de Gaulle;

$\underline{p}$: penser que les pouvoirs du Parlement doivent être augmentés;

$\underline{n}$: penser que le niveau de vie s'est amélioré dans les derniers temps;

$\underline{s}$: être syndicaliste.

On a trouvé:

$$\mathcal{E} = (\underline{g}' \wedge \underline{p}') \vee (\underline{g} \wedge \underline{p} \wedge \underline{n}') \vee (\underline{g} \wedge \underline{p} \wedge \underline{s});$$

il a semblé que le schéma le plus significatif, à l'époque, était:

$$\begin{tikzcd}[row sep=small] (\underline{p} \wedge \underline{n}') \arrow[dr] & & \\ & \underline{g}' \arrow[r] & \underline{p}. \\ (\underline{p} \wedge \underline{s}) \arrow[ur] & & \end{tikzcd}$$

Nous laissons le lecteur interpréter psychosociologiquement ce schéma, et construire les autres schémas équivalents (mathématiquement), dont certains lui paraîtront d'un certain intérêt.

*

*   *

L'analyse booléenne de questionnaire résoud bien le problème que nous nous posions; mais pour que la méthode devienne utilisable pratiquement, d'autres problèmes doivent être résolus. Nous en mentionnerons quelques uns, sans donner les solutions plus ou moins complètes que nous leur avons trouvées.

1 - Il y a intérêt à ce que $\mathcal{E}$ soit écrit sous forme canonique <u>minimale</u>: le plus petit nombre possible de sous patrons, chacun comportant le plus petit nombre possible de générateurs: sinon, l'écriture est redondante, et aussi le schéma d'implication qui en résulte. Par exemple, si $\mathcal{E} = (\underline{a} \wedge \underline{b}') \vee (\underline{b} \wedge \underline{c}') \vee (\underline{a} \wedge \underline{c}')$, la forme minimale est $\mathcal{E} = (\underline{a} \wedge \underline{b}') \vee (\underline{b} \wedge \underline{c}')$; et effectivement, il n'est pas nécessaire (ni mathématiquement, ni psycho-sociologiquement) de considérer la relation $(\underline{a} \longrightarrow \underline{c})$, puisqu'elle n'est qu'une conséquence, par transitivité, des relations $(\underline{a} \longrightarrow \underline{b})$ et $(\underline{b} \longrightarrow \underline{c})$.

2 - Pour le calcul pratique de $\mathcal{E}$, suivant les cas, il est économique, soit de partir directement des patrons, soit de considérer les tableaux croisant les questions 2 à 2, 3 à 3, etc., successivement. Dans ce dernier cas, un algorithme peut être donné qui permet de faire un nombre minimum d'opérations matérielles, et conduit directement à la forme canonique minimale de $\mathcal{E}$.

3 - Si certaines questions du questionnaire réel comportent plus de deux réponses, l'analyse booléenne est toujours praticable, à condition de remplacer chaque question ayant plus de deux réponses par plusieurs questions n'ayant que deux réponses, comme dans l'exemple, où $\underline{A}$ (ayant 3 réponses: $\underline{a_1}$, $\underline{a_2}$, $\underline{a_3}$) est remplacé par $\underline{X}$ et $\underline{Y}$:

		X	
		x	x'
Y	y	a_1	a_2
	y'	a_3	

La case vide se traduit très naturellement dans notre langage par ($x' \longrightarrow y$); dans le schéma d'implication, on pourra traduire ($x \wedge y$) par a_1, x par (a_1 ou a_3), etc...

4 - La principale difficulté pour l'usage pratique de l'analyse booléenne de questionnaire (comme pour toute analyse algébrique de questionnaire) vient de la nécessité, très généralement reconnue, d'avoir une approximation statistique des résultats; nous avons (Flament, 1965 a) proposé une voie générale pour ce genre de problème, mais son application ici semble délicate.

5 - On l'aura remarqué, l'analyse booléenne de questionnaire est applicable à tous les protocoles d'observation possibles. Une telle méthode universelle est sans doute un bon cadre pour l'étude mathématique de la plupart des problèmes d'analyse algébrique de questionnaire, mais on peut aussi souhaiter une spécification des résultats permettant de retrouver les modèles classiques. Par exemple, on démontre qu'on a une échelle de Guttman, si et seulement si on peut réindicier les générateurs selon un certain ordre (la hiérarchie guttmanienne), de telle sorte que: $\mathcal{E} = \bigvee_{i} (x_i \wedge x'_{i+1})$, $i = 1, 2, \ldots, (n - 1)$. Il convient de rechercher ainsi les types de $\mathcal{E}$ correspondant aux divers modèles classiques, et aussi de voir si divers types de $\mathcal{E}$ particuliers ne donnent pas naissance à de nouveaux modèles (par exemple, si $\mathcal{E}$ est l'inf de divers $\mathcal{E}_i$ guttmaniens, on obtient une division de la population en classes latentes, chacune caractérisée par une échelle de Guttman différente).

*

* *

BIBLIOGRAPHIE

COOMBS, C.H., 1964, A theory of data, New York, Wiley.

ELISSEEF, V., 1965, Possibilités du scalogramme dans l'étude des bronzes chinois archaïques, Math. Sc. Hum., 11, 1-10.

FLAMENT, C., 1965, Introduction aux modèles de fermeture en analyse algébrique de questionnaire, Cah. Psychol., 8, 213-222.

FLAMENT, C., 1965 a, L'analyse algébrique de questionnaire: définition et validation, Colloque International du CNRS: Modèles et formalisation du comportement (à paraître).

MATALON, B., 1965, L'analyse hiérarchique, Paris, Gauthier, Villars et Mouton.