MATHÉMATIQUES ET SCIENCES HUMAINES

JEAN-MARC BERNARD
SÉBASTIEN POITRENAUD

L'analyse implicative bayésienne multivariée d'un questionnaire binaire : quasi-implications et treillis de Galois simplifié

Mathématiques et sciences humaines, tome 147 (1999), p. 25-46
<http://www.numdam.org/item?id=MSH_1999__147__25_0>

© Centre d'analyse et de mathématiques sociales de l'EHESS, 1999, tous droits réservés.

L'accès aux archives de la revue « Mathématiques et sciences humaines » (http://msh.revues.org/) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Math. Inf. Sci. hum., (37e année, n° 147, *1999*, pp. 25-46)

L'ANALYSE IMPLICATIVE BAYÉSIENNE MULTIVARIÉE D'UN QUESTIONNAIRE BINAIRE : QUASI-IMPLICATIONS ET TREILLIS DE GALOIS SIMPLIFIÉ *

Jean-Marc BERNARD[1] et Sébastien POITRENAUD[1]

RÉSUMÉ — *Nous proposons une nouvelle méthode pour simplifier le treillis de Galois associé à un questionnaire binaire (n individus classés selon q questions binaires), méthode basée sur l'affaiblissement des implications portées par le treillis en quasi-implications. Au niveau descriptif, la méthode proposée fait intervenir un nouvel indice pour la mesure des quasi-implications (l'"indice implicatif multivarié") qui satisfait certaines conditions d'invariance par équivalence logique. Au niveau inductif, l'incertitude sur les fréquences vraies des profils est exprimée par un "modèle Dirichlet imprécis". Ce modèle répond aux difficultés des modèles bayésiens usuels fondés sur une distribution de Dirichlet unique, notamment pour le cas où n est petit devant 2^q. Un aspect important de la méthode est que les résumés descriptif et inductif qu'elle fournit constituent des treillis de Galois, versions simplifiées du treillis initial.*

MOTS CLÉS — *Données binaires, Méthodes booléennes, Indice implicatif multivarié, Inférence bayésienne, Probabilités imprécises, Modèle Dirichlet imprécis.*

SUMMARY — Multivariate Bayesian implicative analysis for a binary questionnaire : quasi-implications and simplified Galois lattice

We propose a new method for simplifying the Galois lattice associated to a binary questionnaire (n units classified according to q binary questions). The method consists in weakening the implications borne by the lattice into quasi-implications. At the descriptive level, the method involves a new measure for quasi-implications (the "multivariate implicative index") which satisfies some requirements of invariance by logical equivalence. At the inductive level, uncertainty about the patterns' true frequencies is expressed by an imprecise-Dirichlet model. This model is shown to have several advantages over the usual non-informative Bayesian approach based on a single Dirichlet prior, especially for the case where n is small in comparison to 2^q. An important feature of the method is that it provides two implicative summaries, descriptive and inductive, which both constitute simplified versions of the initial Galois lattice.

KEY WORDS — *Binary data, Boolean methods, Multivariate implicative index, Bayesian inference, Imprecise probabilities, Imprecise Dirichlet model.*

* Une version préliminaire de cet article a été présentée au congrès de la Société Francophone de Classification en Septembre 1998 à Montpellier, France. Nous remercions deux referees anonymes de ce congrès pour leurs remarques et suggestions qui nous ont aidé à en améliorer le contenu.

[1]Laboratoire Cognition et Activités Finalisées, CNRS ESA 7021, Université Paris 8, 2 rue de la Liberté, 93526 Saint-Denis Cedex, France, e-mail : berj@univ-paris8.fr, poitrenaud@univ-paris8.fr.

1. INTRODUCTION

Notre objectif dans cet article est de proposer une nouvelle méthode pour l'analyse d'un questionnaire binaire, c'est-à-dire de données sous la forme de n "individus" ayant répondu à un ensemble de q "questions" binaires, dont les modalités correspondent à la présence ou absence de q traits. Parmi les méthodes existantes, beaucoup adoptent une vision symétrique, et souvent globale, du problème de l'association entre les questions. Notre but est au contraire de pouvoir parvenir, lorsque les données l'autorisent, à des conclusions dissymétriques telles que "la réponse a à la question A implique, en général, la réponse b à la question B" (on dira "a quasi-implique b"). L'article présent est ainsi dans la ligne des travaux sur les *échelles de Guttman* (Guttman, 1944 ; Loevinger, 1948) et, plus généralement, des méthodes qui mettent en avant les structures booléennes ou ordinales sur les modalités des questions, comme l'*analyse booléenne de questionnaires* (Degenne, 1972 ; Flament, 1966 et 1976) et la théorie des *treillis de Galois* (Barbut, Monjardet, 1970 ; Duquenne, 1987 ; Ganter, 1995 ; Wille, 1982).

Nombre de travaux dans ce domaine ont visé au développement d'une analyse qualitative de données exhaustives, uniquement fondée sur la présence ou absence des profils de réponses possibles. Cette approche conduit à caractériser la structure des associations entre modalités par un ensemble de relations implicatives (Duquenne, 1987 ; Guigues & Duquenne, 1986). Mais lorsque le nombre de questions q est grand, la structure implicative peut devenir fort complexe. Il devient alors important de tenter d'en fournir un résumé simplifié qui puisse s'exprimer en termes de quasi-implications au lieu d'implications strictes. De plus, lorsque les individus constituent un échantillon d'une population plus vaste, le problème de l'inférence vient s'ajouter au précédent : Le résumé trouvé au niveau descriptif peut-il être étendu, généralisé, à la population entière ? Ce besoin de simplifier (niveau descriptif) et de généraliser (niveau inductif) nécessite la prise en compte d'aspects quantitatifs.

L'introduction d'éléments quantitatifs a été envisagée de façon descriptive en filtrant le protocole selon le critère de fréquence/rareté des profils (*e.g.* Flament, 1976 ; Duquenne, 1996) mais, comme le soulignent Hildebrand *et al.* (1977) et Bernard & Charron (1996a), ce critère ne peut, à lui-seul, servir de base pour une définition satisfaisante des quasi-implications : il peut conduire à conclure que a quasi-implique b alors que les questions A et B sont en fait statistiquement indépendantes. Une approche alternative, proposée par Luxenburger (1991), fait appel à la notion d'"implications partielles" définies sur la base de fréquences conditionnelles, mais, elle non plus, ne répond pas à la critique précédente ; de plus, les quasi-implications auxquelles elle conduit ne vérifient ni la transitivité, ni l'invariance par équivalence logique.

Hildebrand *et al.* (1977) ont étudié en détail le cas des questionnaires bivariés ($q = 2$), pour lesquels ils proposent l'indice descriptif "*Del*" comme mesure de l'adéquation des données observées à un modèle logique d'intérêt ; pour le cas de questions binaires, cet indice se réduit à l'indice de Loevinger (1948). Pour les inférences relatives à l'indice *Del* ces auteurs ont recours à une approche fréquentiste qui fait appel à des arguments asymptotiques et qui, de ce fait, se heurte à de sérieuses difficultés pour les petits échantillons ainsi que pour les données extrêmes. L'approche bayésienne qu'ont proposé Bernard & Charron (1996a, 1996b) pour le cas de données bivariées, l'*analyse implicative bayésienne*, permet de lever ces difficultés.

Hildebrand *et al.* (1977, ch. 7) proposent une généralisation de leur méthode au cas de données multivariées qui se révèle insatisfaisante : d'une part, elle ne remplit

pas certaines conditions minimales d'invariance par équivalence logique, d'autre part, elle souffre des mêmes limitations au niveau inférentiel que celles rencontrées pour les données bivariées.

Pour répondre à ces difficultés, nous proposons ici une extension de l'analyse implicative bayésienne de Bernard & Charron (1996a) au cas de données binaires multivariées. Au niveau descriptif, notre méthode conduit à un résumé descriptif du protocole en terme de quasi-implications dont la définition fait intervenir un nouvel indice d, l'*indice implicatif multivarié*, qui généralise l'indice *Del*. La "quasi-logique" qui en découle constitue une généralisation de la logique usuelle tout en en conservant certaines propriétés essentielles.

Au niveau inductif, nous recourons à une approche bayésienne non-informative, fondée sur le concept de probabilités imprécises, pour l'évaluation du degré de généralisabilité d'un tel résumé. Le *modèle Dirichlet imprécis* ("imprecise Dirichlet model" en anglais) que nous utilisons se revèle plus satisfaisant que les modèles bayésiens non-informatifs usuels qui reposent sur une distribution initiale unique. Cet avantage est manifeste lorsque q est grand et qu'ainsi nombre des 2^q profils possibles sont non observés par construction. On trouvera une présentation complète des modèles à probabilités imprécises dans Walley (1991) et du modèle Dirichlet imprécis en particulier dans Walley (1996), pour l'inférence paramétrique, et dans Walley & Bernard (1998) pour l'inférence prédictive.

Un aspect important de la méthode que nous proposons est que chacun des deux résumés, descriptif et inductif, qu'elle fournit peut être représenté par un treillis de Galois qui constitue une version simplifiée du treillis de Galois observé.

Cet article est organisé comme suit. Dans la section 2, nous introduisons les notations relatives à un questionnaire binaire. Les sections 3 et 4 présentent le treillis de Galois et les implications associés à un questionnaire binaire. L'aspect descriptif de notre méthode est introduit à la section 5 avec la définition des quasi-implications. La section 6 aborde le niveau inductif avec la présentation et les difficultés des méthodes bayésiennes non-informatives usuelles, difficultés auxquelles répond le modèle Dirichlet imprécis présenté à la section 7. Enfin, nous présentons une application de notre méthode sur un exemple réel à la section 8 avant de conclure (section 9).

2. QUESTIONNAIRE BINAIRE MULTIVARIÉ

2.1. PROTOCOLE OBSERVÉ

Soit $I = \{i_1, \cdots, i_n\}$ un ensemble de n *individus*, que, selon le contexte, on pourra aussi appeler des "unités" ou des "objets". Chaque individu peut ou non présenter certains *traits* parmi l'ensemble $F = \{a, b, c, \cdots\}$ de cardinal q. Un tel ensemble de données, que nous appellons *protocole* (ou *protocole observé*), s'exprime d'abord comme une *relation binaire*, $\mathcal{R}$, sur $I \times F$, définie par

$$i\mathcal{R}f \quad ssi \quad \text{l'individu } i \text{ possède le trait } f, \text{ avec } i \in I,\ f \in F. \tag{1}$$

A chaque trait a, on associe la *question binaire* $A = \{a, a'\}$ dont les modalités représentent respectivement la présence (a) ou l'absence (a') du trait. Le protocole

peut ainsi également être décrit comme un questionnaire binaire relativement à n individus répondant chacun aux q questions binaires, A, B, C, *etc.*.

Si on adopte la vision d'une relation binaire, les individus et les traits jouent un rôle symétrique. Dans un contexte statistique, en revanche, l'ensemble I constitue typiquement un échantillon d'une population plus vaste dont il n'importe pas de distinguer les éléments. Parallèlement, la question d'intérêt portera alors essentiellement dans ce cas sur l'étude des associations entre traits qui, eux, doivent être distingués. Avec cette vision dissymétrique, le protocole peut être décrit comme un *protocole pondéré* composé de *profils* $p \in P$, avec $P = A \times B \times C \cdots$, pondéré chacun par l'effectif n_p des individus présentant le profil p.[2] Le nombre de profils possibles est $K = |P| = 2^q$. (Dans tout le début de cet article, nous considérons qu'aucun profil n'est *a priori* impossible ; le cas de contraintes structurelles connues *a priori* sera envisagé à la section 7.3.)

2.2. PROFILS PARTIELS ET PROJECTIONS D'UN PROTOCOLE

L'ensemble des *profils de base* est défini en référence à l'ensemble des q questions initiales, $Q = \{A, B, C, \ldots\}$. Mais le protocole de base peut être *projeté* sur un sous-ensemble non trivial quelconque Q' de Q ; chaque profil de base est alors projeté en un *profil partiel*, dont le poids associé s'obtient évidemmment par sommation des poids initiaux pertinents.

3. TREILLIS DE GALOIS ASSOCIÉ À UN QUESTIONNAIRE BINAIRE

A la relation binaire $\mathcal{R}$ sur $I \times F$, on associe un *treillis de Galois* (Barbut, Monjardet, 1970). Rappelons brièvement les éléments clefs de cette construction.

Pour chaque partie $J \subseteq I$, et, dualement, pour chaque partie $G \subseteq F$, on définit

$$\begin{aligned} J^{\uparrow} &= \{f \in F \,/\, \forall i \in J,\ i\mathcal{R}f\}, \quad \text{et} \\ G^{\downarrow} &= \{i \in I \,/\, \forall f \in G,\ i\mathcal{R}f\}. \end{aligned}$$

L'ensemble $J^{\uparrow}$, appelé l'*intension* (ou *compréhension*) de J, est l'ensemble des traits partagés par tous les éléments de J, et $G^{\downarrow}$, appelé l'*extension* de G, est l'ensemble des individus possédant tous les attributs de G.

Les deux applications $J \mapsto J^{\uparrow}$ et $G \mapsto G^{\downarrow}$ satisfont

$$\begin{aligned} J_1 \subseteq J_2 &\implies {J_2}^{\uparrow} \subseteq {J_1}^{\uparrow}, \quad \text{avec } J_1 \subseteq I,\ J_2 \subseteq I, \\ G_1 \subseteq G_2 &\implies {G_2}^{\downarrow} \subseteq {G_1}^{\downarrow}, \quad \text{avec } G_1 \subseteq F,\ G_2 \subseteq F, \end{aligned}$$

et constituent ainsi une *correspondance de Galois* entre l'ensemble des parties de I et celui des parties de F.

Un *pavé maximal* (ou *rectangle maximal*) de la relation $\mathcal{R}$ est un couple $(J \subseteq I, G \subseteq F)$ satisfaisant la *propriété de fermeture*, c'est-à-dire tel que

$$J^{\uparrow} = G \quad \text{et} \quad G^{\downarrow} = J.$$

[2]Comme Degenne (1972) ou Flament (1976), on pourrait aussi dire "*patron de réponse*" à la place de "profil".

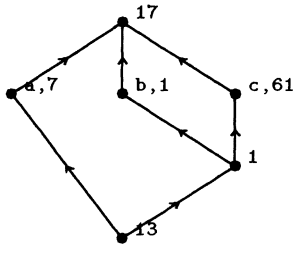
Dans le langage de la "Formal Concept Analysis" de Wille (1982) et Ganter (1995), un tel couple (J, G) est appelé un *concept* (d'extension J et d'intension G) dans le contexte de la relation $\mathcal{R} \subset I \times F$. L'ensemble de tous les concepts est partiellement et dualement ordonné par l'inclusion entre parties de I et l'inclusion entre parties de F. Cet ordre partiel induit une structure de treillis sur l'ensemble des concepts, d'où l'appelation "treillis de Galois".

Nous renvoyons le lecteur à Duquenne (1987), Guénoche & Van Mechelen (1991), et Wille (1982), pour une présentation détaillée du treillis de Galois et de ses propriétés, à Bordat (1986) et Guénoche (1990) pour sa construction, et enfin aux logiciels GLAD de Duquenne (1992a), PROCOPE de Poitrenaud (1995, 1998) ainsi qu'à Wille (1989) pour son implémentation informatique. Un accent particulier a été mis sur le développement de représentations graphiques du treillis de Galois (cumulé/décumulé, pondéré/non pondéré) permettant d'en faciliter l'interprétation (voir *e.g.* Duquenne, 1992b). Lorsque, notamment, on introduit la dissymétrie entre individus et traits, une représentation typique est celle dans laquelle chaque concept (G, J) est représenté par son intension G et par le cardinal de son extension $|J|$.

Considérons l'exemple d'un questionnaire comprenant $q = 3$ questions binaires, A, B et C, représenté sous la forme d'un protocole pondéré $(p, n_p)_{p \in P}$ (voir Tableau 1). On y a également figuré le treillis de Galois décumulé (tout noeud hérite des traits situés au-dessus de lui et des individus situés en-dessous) et pondéré (pour les individus).

Tableau 1: *Exemple 1. Protocole pondéré comprenant* $q = 3$ *questions binaires, A, B and C, et portant sur* $n = 100$ *individus (données fictives), avec le treillis de Galois décumulé et pondéré associé.*

n_p	b		b'	
	c	c'	c	c'
a	13	0	0	7
a'	1	1	61	17



Un aspect important des méthodes basées sur le treillis de Galois est qu'elles fournissent une description exhaustive des associations entre traits à tous les niveaux possibles de complexité : binaire (faisant intervenir deux questions), ternaire (faisant intervenir trois questions), *etc.*. En contrepartie, et même pour un nombre modéré de questions, le treillis de Galois peut se révéler fort complexe. De plus, comme le soulignent aussi Guénoche & Van Mechelen (1991), le treillis peut être extrèmement sensible à des modifications mineures de l'ensemble des individus. Ces deux faits motivent la recherche de simplifications du treillis de Galois. Il s'agit d'en obtenir un résumé simplifié qui parvienne à capturer l'essentiel des associations entre traits tout en présentant un caractère suffisamment stable, *i.e.* généralisable. En ce qui nous concerne, la simplification se fondera sur l'équivalence entre structure du treillis et liste d'implications et portera sur l'affaiblissement de ces implications ; la recherche de généralisabilité, quant à elle, sera abordée plus loin par le développement d'une méthode d'inférence bayésienne appropriée.

4. IMPLICATIONS ASSOCIÉES AU TREILLIS DE GALOIS

Si on s'attache seulement aux dépendances strictes entre modalités des questions, il suffit de ne retenir du protocole que le caractère présent ($n_p > 0$) ou absent

($n_p = 0$) de chaque profil p. A ce niveau qualitatif, le protocole peut être caractérisé par une liste d'implications entre modalités (Flament, 1976) qui peut s'exprimer sous une forme minimale (Guigues, Duquenne, 1986). Comme le souligne Duquenne (1987), cette vision implicative est en dualité avec le treillis de Galois : plus le treillis est complexe et moins la structure implicative est riche, moins il l'est et plus cette structure est riche. Avant d'aborder cet aspect, il est nécessaire de préciser certaines notations et identités relatives aux expressions logiques qui nous permettront d'énoncer ces implications.

4.1. EXPRESSIONS LOGIQUES

Du point de vue logique, les traits $a, b, c, \ldots$ sont assimilés à des propositions élémentaires qui peuvent être vraies ou fausses. La concaténation (*e.g.* dans ab) désigne le "et logique", le symbole "prime" (*e.g.* dans a') la négation, le symbole "$\Longrightarrow$" l'implication logique. Nous aurons également recours aux symboles "$\wedge$" et "$\vee$" pour le "et" et le "ou" logiques respectivement, ainsi qu'au symbole "$\emptyset$" pour désigner une proposition toujours fausse.

En notant r, s et t trois propositions quelconques, on vérifiera aisément les trois identités logiques,

$$r \Longrightarrow \emptyset \equiv r', \tag{2}$$

$$(r \wedge s) \Longrightarrow t \equiv r \Longrightarrow (t \vee s'), \quad \text{et} \tag{3}$$

$$(rs \Longrightarrow \emptyset) \wedge (rs' \Longrightarrow \emptyset) \equiv r \Longrightarrow \emptyset, \tag{4}$$

où le symbole "$\equiv$" indique l'équivalence logique de deux propositions.

4.2. IMPLICATIONS PORTÉES PAR UN PROTOCOLE

Implications élémentaires. L'absence d'un profil de base $p \in P$ dans le protocole s'exprime simplement par l'énoncé p' (p est faux), soit, en utilisant l'identité (2), par

$$p \Longrightarrow \emptyset.$$

Nous appelons *implication élémentaire* une telle implication, le qualificatif "élémentaire" indiquant qu'elle concerne *un seul profil de base.*

Dans l'exemple du Tableau 1, le profil $ab'c$ est absent, soit $(ab'c)'$ ou encore $ab'c \Longrightarrow \emptyset$. Par l'identité (3), cette absence peut encore s'exprimer par $ab' \Longrightarrow c'$, $ac \Longrightarrow b$ ou bien $b'c \Longrightarrow a'$. De ces diverses écritures équivalentes, l'expression $ac \Longrightarrow b$ ("la conjonction de a et c implique b") est la plus simple pour l'interprétation. Cependant nous utiliserons la forme $ab'c \Longrightarrow \emptyset$, et plus généralement $r \Longrightarrow \emptyset$, de façon privilégiée; cette forme canonique présente en effet l'avantage d'identifier directement les profils de base ou partiels absents qui jouent un rôle central dans la méthode que nous proposons.

Structure implicative du protocole. L'absence simultanée de plusieurs profils de base, $p_1, p_2, p_3, \cdots$, dans le protocole observé s'exprime par la conjonction

$$(p_1)' \wedge (p_2)' \wedge (p_3)' \wedge \cdots,$$

ou encore, en utilisant (2), par la conjonction d'implications élémentaires

$$(p_1 \Longrightarrow \emptyset) \wedge (p_2 \Longrightarrow \emptyset) \wedge (p_3 \Longrightarrow \emptyset) \wedge \cdots$$

Ainsi, la vision qualitative du protocole, qui n'en retient que l'absence ou la présence des profils de base, se traduit par une conjonction d'implications élémentaires.

Dans l'exemple du Tableau 1, les deux profils abc' et $ab'c$ sont absents. La structure implicative du protocole s'exprime ainsi par "$(abc' \Longrightarrow \emptyset) \wedge (ab'c \Longrightarrow \emptyset)$", ou bien de façon équivalente, en utilisant (3), par

$$(ab \Longrightarrow c) \quad \wedge \quad (ac \Longrightarrow b).$$

Implications binaires. L'utilisation récursive de (4) permet le cas échéant (mais pas dans notre premier exemple) de condenser cette liste d'implications élémentaires en une liste plus compacte d'implications, dont chacune exprime l'absence simultanée de plusieurs profils de base. La forme la plus compacte est celle des *implications binaires*, c'est-à-dire des implications dont l'expression ne fait plus intervenir que *deux* questions, comme par exemple $ab' \Longrightarrow \emptyset$ ou $a \Longrightarrow b$, qui présentent un intérêt particulier pour l'interprétation des données.[3]

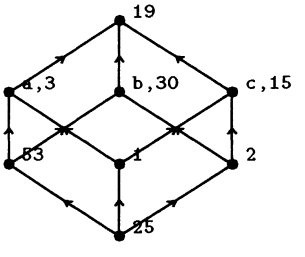
Les implications élémentaires et binaires constituent deux extrèmes parmi l'ensemble des implications possibles. Les implications élémentaires sont minimales dans la mesure où elles identifient un profil de base absent unique, mais leur écriture fait intervenir l'ensemble des q questions. A l'autre extrème, les implications binaires sont minimales dans le sens où leur écriture ne nécessite que deux questions, mais, en contrepartie, chacune identifie $2^{(q-2)}$ profils de base absents. Bien entendu, les deux notions coïncident lorsque $q = 2$.

5. ANALYSE DESCRIPTIVE : QUASI-IMPLICATIONS

Considérons un second exemple simple avec à nouveau $q = 3$ questions (voir Tableau 2). Dans celui-ci, tous les 2^q profils possibles sont observés et ainsi rien ne peut être conclu en termes d'implications strictes (le treillis associé est booléen). Cependant, si les poids des profils $ab'c$, $ab'c'$ et $a'bc$ valaient tous 0 (au lieu de 1, 3 et 2 respectivement), la structure implicative du protocole pourrait s'exprimer par "$(ab'c \Longrightarrow \emptyset) \wedge (ab'c' \Longrightarrow \emptyset) \wedge (a'bc \Longrightarrow \emptyset)$", ou, de façon plus compacte par "$(a \Longrightarrow b) \wedge (bc \Longrightarrow a)$", en utilisant (3) et (4). Notre objectif est précisément d'étudier dans quelle mesure cette dernière structure implicative — suggérée par une intuition simple, peut-être un peu trop — peut être considérée comme une approximation acceptable de ce protocole.

Tableau 2: Exemple 2. Protocole pondéré portant sur $q = 3$ questions binaires, A, B et C, et $n = 148$ individus (données fictives), et treillis de Galois associé.

	b		b'	
	c	c'	c	c'
a	25	53	1	3
a'	2	30	15	19



[3]Nous supposons ici qu'aucun trait n'est soit toujours présent, soit toujours absent, et donc que des réductions ultimes telles que $a \Longrightarrow \emptyset$ ou $a' \Longrightarrow \emptyset$ ne peuvent se produire.

Pour répondre à cet objectif, nous définissons d'abord ici la notion de quasi-implication (*q-implication* en bref) sur une base purement descriptive dans le sens précis de Rouanet *et al.* (1990, pp. 2–4), c'est-à-dire en considérant uniquement les fréquences relatives $\boldsymbol{f} = (f_p)_{p \in P}$ avec $f_p = n_p/n$ des profils et non l'effectif total n du protocole.

5.1. DEUX QUESTIONS BINAIRES ($q = 2$)

Bernard & Charron (1996a) ont fondé l'analyse implicative d'un tableau de contingence 2×2 ($q = 2$) sur l'indice "*Del*", introduit dans ce contexte par Hildebrand *et al.* (1977) et défini pour tout profil $p = ij$, $i \in \{a, a'\}$ et $j \in \{b, b'\}$, par

$$d_{ij \Longrightarrow \emptyset} = 1 - \frac{f_{ij}}{f_i f_j}, \tag{5}$$

où f_i et f_j sont les fréquences marginales de i et j respectivement. Cet indice apparaît également dans la littérature sous le nom d'"indice d'homogénéité de Loevinger" (Loevinger, 1948).[4]

L'indice $d_{ij \Longrightarrow \emptyset}$ mesure l'écart du tableau 2×2 observé à l'indépendance locale entre i et j (qui se traduit par $f_{ij} = f_i f_j$) dans la direction du *modèle logique* $ij \Longrightarrow \emptyset$, d'où notre notation. Cet indice varie dans l'intervalle $]-\infty, 1]$; en particulier, il vaut 0 en cas d'indépendance locale entre i et j et 1 lorsque $ij \Longrightarrow \emptyset$. Ainsi, c'est lorsqu'il est positif que cet indice peut être interprété comme un degré de q-implication. Pour une valeur-seuil donnée $d_{quasi} > 0$, une *q-implication de degré* d_{quasi} a été définie par

$$ij \longrightarrow \emptyset \quad ssi \quad d_{ij \Longrightarrow \emptyset} \geq d_{quasi}. \tag{6}$$

Les critères de choix de d_{quasi} sont discutés dans Bernard & Charron (1996a). Par la suite, nous utilisons fréquemment la valeur-seuil $d_{quasi} = 0.50$ qui correspond au cas d'un profil deux fois moins représenté qu'en cas d'indépendance locale. Cette valeur-seuil particulière n'est utilisée qu'à titre indicatif et on aura souvent intérêt, dans l'étude des quasi-implications, à utiliser une grille de valeurs-seuils telle que, par exemple, [0; 0.20; 0.40; 0.60; 0.80; 1].

Considérons, par exemple, le protocole projeté sur $Q' = \{A, B\}$ du protocole de base du Tableau 2. Pour chaque profil p parmi ab, ab', $a'b$ et $a'b'$, l'indice $d_{p \Longrightarrow \emptyset}$ prend comme valeurs -0.28, 0.81, 0.35 et -1.01 respectivement. Pour la valeur-seuil $d_{quasi} = 0.50$ par exemple, on trouve une seule q-implication élémentaire, $ab' \longrightarrow \emptyset$, *i.e.* $a \longrightarrow b$.

Soulignons qu'avec cette définition, il n'est possible de conclure à une quasi-implication que lorsque les données s'écartent de l'indépendance, et que l'écart qui est pris en compte est un écart *dans une direction donnée*, ce qui autorise des conclusions dissymétriques : les énoncés $a \longrightarrow b$ et $b \longrightarrow a$ ne sont pas traités de façon équivalente, du fait que les modèles logiques stricts associés identifient ab' et $a'b$ respectivement comme "case d'erreur" associée, pour reprendre le terme de Hildebrand *et al.* (1977). Enfin, dans cette construction, deux énoncés logiques qui identifient les mêmes cases d'erreur sont traités de manière identique ; ceci assure l'invariance des q-implications par équivalence logique. Nous renvoyons le lecteur à Bernard & Charron (1996a) pour d'autres propriétés de l'indice "*Del*", et notamment ses liens avec le coefficient de contingence ϕ^2.

[4]Dans Bernard & Charron (1996a), c'est l'indice de Loevinger, avec sa notation usuelle H, qui est mis en avant. Nous privilégions ici une présentation plus proche des travaux de Hildebrand *et al.* (1977) ainsi qu'une notation (5) plus conforme à celle de l'indice "*Del*" proposé par ces auteurs et déjà utilisée dans Bernard & Charron (1996b).

5.2. GÉNÉRALISATION À PLUSIEURS QUESTIONS BINAIRES ($q > 2$)

5.2.1. *Premières suggestions et leurs difficultés*

Il y a plusieurs façons d'envisager *a priori* la généralisation de la notion de q-implication à plus de deux questions. Une première direction est de se centrer sur les implications binaires, c'est-à-dire de ne considérer du protocole que les paires de questions, comme cela est fait dans Bernard & Charron (1996a). Un premier inconvénient de cette approche est que, par définition, elle "oublie" les possibles dépendances ternaires et d'ordre supérieur. Plus problématique encore, elle peut conduire à un résumé non-transitif du protocole. Considérons à nouveau notre premier exemple, le protocole donné au Tableau 1, et les trois protocoles projetés qu'on peut en dériver (voir Tableau 3). A partir de ceux-ci, on obtient respectivement

$$d_{ab'\Longrightarrow\emptyset} = 0.59, \quad d_{bc'\Longrightarrow\emptyset} = 0.73, \quad \text{et} \quad d_{ac'\Longrightarrow\emptyset} = -0.40.$$

Pour $d_{quasi} = 0.50$, on est ainsi conduit aux conclusions que $a \longrightarrow b$, $b \longrightarrow c$, mais non que $a \longrightarrow c$. On a même en fait une répulsion entre les modalités a et c.

Tableau 3: *Exemple 1. Protocoles projetés sur* $\{A,B\}$, $\{B,C\}$ *et* $\{A,C\}$.

	b	b'
a	13	7
a'	2	78

	c	c'
b	14	1
b'	61	24

	c	c'
a	13	7
a'	62	18

Une autre direction consiste à se centrer sur les implications élémentaires, mais, même dans ce cadre plus restreint, plusieurs possibilités de généralisation peuvent encore être envisagées. Deux voies possibles, explorées par Hildebrand *et al.* (1977, ch. 7), ont en commun de chercher à se ramener au cas des implications binaires, mais aussi, comme ces auteurs le notent, de conduire à une non-invariance par équivalence logique. Revenons à notre second exemple, le protocole donné au Tableau 2, et considérons, par exemple, le problème de la définition d'une mesure du degré de la quasi-implication élémentaire $ab'c \longrightarrow \emptyset$.

Une première idée est de décomposer cet énoncé en faisant intervenir la modalité composée ab' d'une part et la modalité c d'autre part. Avec ce *regroupement dissymétrique*, on est ramené au cas $q = 2$ et, par (5), on obtient $d_{(ab')c\Longrightarrow\emptyset} = 0.14$. Une difficulté de cette approche est qu'elle dépend du choix de regroupement effectué : la décomposition en ac et b' conduit à $d_{(ac)b'\Longrightarrow\emptyset} = 0.85$, et la décomposition en $b'c$ et a donne $d_{(b'c)a\Longrightarrow\emptyset} = 0.89$, et ceci, malgré que les expressions logiques qui interviennent dans ces trois écritures soient logiquement équivalentes.

Une idée alternative est celle du *conditionnement* : dire que "$ab'c \Longrightarrow \emptyset$" est logiquement équivalent à dire que "$ab' \Longrightarrow \emptyset$ conditionnellement à c", ce que nous écrivons "$ab' \Longrightarrow \emptyset|c$". On est à nouveau ramené au cas $q = 2$, et par (5), on obtient $d_{ab'\Longrightarrow\emptyset|c} = 0.90$. Malheureusement, les deux autres conditionnements équivalents conduisent ici aussi à des conclusions incohérentes : $d_{ac\Longrightarrow\emptyset|b'} = 0.41$, et $d_{b'c\Longrightarrow\emptyset|a} = 0.21$.

5.2.2. *Indice implicatif multivarié*

Les tentatives précédentes de définir des q-implications sur la base des implications binaires et de l'indice défini en (5) mènent à des incohérences, d'où notre proposition

de généraliser l'approche de Bernard & Charron (1996a), rappelée à la section 5.1, aux implications élémentaires pour q quelconque. Pour cela, nous introduisons un nouvel indice, appelé *indice implicatif multivarié* et noté $d_{p\Longrightarrow\emptyset}$. Cet indice généralise l'indice défini en (5) en faisant intervenir l'ensemble des q questions du protocole de base. Pour tout profil $p = ijk\cdots$, avec $i \in \{a, a'\}$, $j \in \{b, b'\}$, $k \in \{c, c'\}$, *etc.*, cet indice est défini par

$$d_{p\Longrightarrow\emptyset} \;=\; 1 - \frac{f_p}{f_i f_j f_k \cdots}, \tag{7}$$

où f_i, f_j, f_k, *etc.* désignent les fréquences marginales des modalités i, j, k, *etc.*.

L'indice implicatif multivarié $d_{p\Longrightarrow\emptyset}$ constitue une mesure locale de l'écart du protocole par rapport à l'*indépendance complète*, définie par : $\forall p = ijk\cdots, f_p = f_i f_j f_k \cdots$ (Kendall & Stuart, 1973, p. 600). Cette mesure est locale parce l'écart mesuré est celui dans la direction spécifique du modèle logique $p \Longrightarrow \emptyset$. L'indice $d_{p\Longrightarrow\emptyset}$ vaut 0 en cas d'indépendance locale ($f_p = f_i f_j f_k \ldots$), il est positif lorsque le profil p est sous-représenté ($f_p < f_i f_j f_k \ldots$), et vaut 1 dans le cas d'une sous-représentation maximale, *i.e.* où le profil p est absent ($f_p = 0$). Par exemple, une valeur de 0.50 pour $d_{p\Longrightarrow\emptyset}$ signifie que le profil p est sous-représenté de 50% (*i.e.* deux fois moins représenté) par rapport à la situation d'indépendance complète entre A, B, C, *etc.*.

Dans les définitions (5) et (7) des indices implicatifs bivarié et multivarié, deux aspects sont essentiels : *(i)* le fait de prendre la *densité* de la mesure $(f_p)_{p\in P}$ par rapport à une mesure de référence $(\widehat{f_p})_{p\in P}$, *i.e.* un indice du type $f_p/\widehat{f_p}$; *(ii)* le fait de choisir comme mesure de référence celle qui exprime l'indépendance complète, *i.e.* $\widehat{f_p} = f_i f_j f_k \ldots$ C'est *(i)* qui assure que la combinaison d'énoncés relatifs à plusieurs profils se fait par moyennage — on le verra plus loin —, et c'est *(ii)* qui exprime la question d'intérêt, *i.e.* celle d'opposer dépendance orientée à indépendance. Par contre, la forme particulière des indices (5) et (7), qui fait précéder la densité par "1−", est secondaire et a comme motivation celle de la cohérence avec les indices usuels de liaison, pour lesquels la valeur 0 indique une absence de liaison et la valeur 1 une liaison maximale.

5.2.3. *Quasi-implications*

Pour une valeur-seuil donnée $d_{quasi} > 0$, nous étendons la notion d'implication élémentaire "$p \Longrightarrow \emptyset$" à celle de *quasi-implication élémentaire* "$p \longrightarrow \emptyset$" par

$$p \longrightarrow \emptyset \quad ssi \quad d_{p\Longrightarrow\emptyset} \geq d_{quasi}, \tag{8}$$

ce qu'on lira "le profil p quasi-implique $\emptyset$" et que nous exprimerons aussi par "le profil p est *quasi-absent* (ou *q-absent*)". Nous aurons parfois recours aux écritures "$p \xrightarrow{Q} \emptyset$" ou "$p \xrightarrow{ABC} \emptyset$" qui indiquent explicitement le niveau de projection auquel les q-implications élémentaires sont définies, la valeur-seuil d_{quasi} restant toujours implicite.

Les q-implications non élémentaires sont définies en utilisant récursivement la même règle de composition (4) que dans la logique stricte, en partant des q-implications élémentaires. Formellement, la notion générale de q-implication $r \xrightarrow{Q} s$ est définie par,

$$r \xrightarrow{Q} s \quad ssi \quad \forall p \in P, \text{ si } (r \Longrightarrow s) \Longrightarrow (p \Longrightarrow \emptyset), \text{ alors } p \xrightarrow{Q} \emptyset, \tag{9}$$

où toutes les q-implications élémentaires $p \xrightarrow{Q} \emptyset$ sont attestées relativement au même niveau de projection Q et avec la même valeur-seuil d_{quasi}.

5.2.4. *Propriétés de cette quasi-logique*

Généralisation de la logique stricte. La logique stricte est obtenue pour la valeur-seuil $d_{quasi} = 1$. En effet $d_{p \Longrightarrow \emptyset} \geq 1$ équivaut à $d_{p \Longrightarrow \emptyset} = 1$, et donc à $p \Longrightarrow \emptyset$.

Invariance. La définition (9) des q-implications en terme de q-implications élémentaires, à un *niveau unique de projection*, garantit l'invariance des q-implications par équivalence logique. En particulier on a,

$$\text{si} \quad (r \Longrightarrow s) \Longleftrightarrow (u \Longrightarrow v), \quad \text{alors} \quad \forall d_{quasi} \geq 0, \quad (r \xrightarrow{Q} s) \Longleftrightarrow (u \xrightarrow{Q} v). \tag{10}$$

Transitivité. L'invariance par équivalence logique assure également la transitivité des q-implications. Prenons l'exemple simple du cas d'un protocole comportant $q = 3$ questions, A, B et C, pour lequel on a trouvé les deux q-implications $a \xrightarrow{ABC} b$ et $b \xrightarrow{ABC} c$ simultanément (à un degré d_{quasi} fixé). Par la définition (9), on a

$$\begin{aligned} a \xrightarrow{ABC} b \quad &\equiv \quad (ab'c \xrightarrow{ABC} \emptyset) \wedge (ab'c' \xrightarrow{ABC} \emptyset), \quad \text{et} \\ b \xrightarrow{ABC} c \quad &\equiv \quad (abc' \xrightarrow{ABC} \emptyset) \wedge (a'bc' \xrightarrow{ABC} \emptyset). \end{aligned}$$

En combinant la seconde et la troisième des quatre q-implications élémentaires ci-dessus par (9) à nouveau, on obtient aisément comme conséquence $ac' \xrightarrow{ABC} \emptyset$, soit $a \xrightarrow{ABC} c$.

Cet exemple a valeur générale. Dans notre quasi-logique, tout énoncé est, en définitive, toujours ramené à une conjonction de q-implications élémentaires, conjonction qui ne peut générer aucune contradiction. Celles-ci sont ensuite combinées entre-elles avec les mêmes règles que celles de la logique stricte, du fait de la définition (9) et aucune contradiction ne peut non plus apparaître à ce niveau.

Cohérence par projection. Une autre propriété importante des q-implications ainsi définies est la *cohérence par projection*. Supposons que, dans un questionnaire avec $Q = \{A, B, C, \cdots\}$, on trouve deux profils q-absents qui ne diffèrent que par un seul trait, $p_1 = ijk\cdots$ et $p_2 = i'jk\cdots$. Par (9), la conjonction des deux énoncés $p_1 \xrightarrow{Q} \emptyset$ et $p_2 \xrightarrow{Q} \emptyset$ s'exprime de façon simplifiée par $p \xrightarrow{Q} \emptyset$, en posant $p = jk\cdots$. Or, par la définition (8), $p_1 \xrightarrow{Q} \emptyset$ et $p_2 \xrightarrow{Q} \emptyset$ équivalent respectivement à

$$d_{p_1 \Longrightarrow \emptyset} \geq d_{quasi} \quad \text{et} \quad d_{p_2 \Longrightarrow \emptyset} \geq d_{quasi}. \tag{11}$$

De plus, on vérifie aisément que $d_{p \Longrightarrow \emptyset}$ s'exprime comme moyenne pondérée des deux indices d ci-dessus, précisément

$$d_{p \Longrightarrow \emptyset} \quad = \quad \frac{\widehat{f_1} d_{p_1 \Longrightarrow \emptyset} + \widehat{f_2} d_{p_2 \Longrightarrow \emptyset}}{\widehat{f_1} + \widehat{f_2}}, \quad \text{avec} \quad \widehat{f_1} = f_i f_j f_k \cdots \text{ et } \widehat{f_2} = f_{i'} f_j f_k \cdots . \tag{12}$$

Par conséquence, on a

$$d_{p \Longrightarrow \emptyset} \quad \geq \quad d_{quasi}.$$

Nous avons ainsi montré que

$$jk\cdots \overset{ABC\cdots}{\longrightarrow} \emptyset \implies jk\cdots \overset{BC\cdots}{\longrightarrow} \emptyset. \tag{13}$$

Si on applique récursivement ce résultat, toute q-implication vérifiée à un niveau de projection fin, Q, l'est nécessairement aussi à tout niveau plus grossier, $Q' \subset Q$. Contrairement à ce qui se passe en logique stricte, la réciproque est fausse. C'est ici que réside la principale différence entre la quasi-logique définie ici et la logique stricte : en logique stricte, le niveau de projection n'est pas pertinent et l'écriture $r \implies \emptyset$ n'a pas besoin d'autre précision, alors que dans cette quasi-logique les q-implications sont conservées par projection mais non par "raffinement" du questionnaire.

5.3. RÉSUMÉ DESCRIPTIF DU PROTOCOLE

A partir de l'indice implicatif multivarié $d_{p\Longrightarrow\emptyset}$ et d'une valeur-seuil d_{quasi}, on obtient un résumé descriptif du protocole en termes de q-implications élémentaires en cherchant tous les profils $p \in P$ pour lesquels l'énoncé $p \longrightarrow \emptyset$ est vérifié. Ce résumé descriptif est donc défini par

$$\{p \longrightarrow \emptyset \text{ ou } p \not\longrightarrow \emptyset\}_{p\in P}. \tag{14}$$

L'ensemble des profils possibles est ainsi partitionné en profils q-absents ($p \longrightarrow \emptyset$) et les autres ($p \not\longrightarrow \emptyset$). Autrement dit, ce résumé est un nouveau protocole, de même support P que le protocole observé, mais dont les poids valent soit 0, pour les profils q-absents, soit 1, pour les autres ; il est ainsi loisible de toute analyse appropriée à un protocole binaire multivarié et notamment : *(i)* représentation sous forme d'un treillis de Galois, et *(ii)* caractérisation par une liste minimale d'implications.

En faisant varier la sévérité du seuil d_{quasi}, on obtient divers résumés emboités (du point de vue des q-implications élémentaires). Pour $d_{quasi} = 1$, on retrouve la vision qualitative du protocole initial dans laquelle seuls les profils absents sont identifiés ; à l'autre extrème, la valeur-seuil $d_{quasi} = 0$ conduit au résumé le plus "brutal" qui considère que tout profil sous-représenté par rapport à l'indépendance complète est q-absent. Du fait de la dualité treillis/implications, l'abaissement de d_{quasi} induit un ajout progressif de q-implications à la liste d'implications initiales, et se traduit par une simplification progressive du treillis.

Considérons à nouveau le protocole du Tableau 1. Le Tableau 4 indique les indices implicatifs multivariés pour chaque profil p et donne les treillis simplifiés pour diverses valeurs de d_{quasi}. Pour $d_{quasi} > 0.86$, aucune q-implication ne peut être attestée descriptivement (le treillis reste booléen) ; pour $d_{quasi} \in [0.15; 0.80]$, les trois profils $ab'c$, $ab'c'$ et $a'bc$ sont q-absents et, par application de (8) et (9), on obtient le résumé implicatif

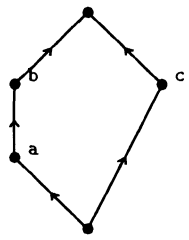
$$(a \longrightarrow b) \wedge (bc \longrightarrow a);$$

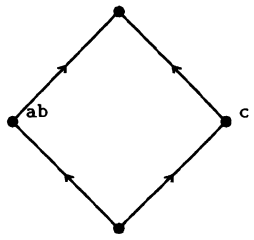
pour $d_{quasi} \in [0; 0.14]$, on obtient un profil q-absent supplémentaire ($a'bc'$) d'où le résumé implicatif

$$(a \longrightarrow b) \wedge (b \longrightarrow a) \quad i.e. \quad a \longleftrightarrow b.$$

Tableau 4: Exemple 2. Pour chaque profil p, indice implicatif multivarié $d_{p\Longrightarrow\emptyset}$ (à gauche), et treillis simplifiés descriptivement pour $d_{quasi} \in [0.15; 0.80]$ (au centre) et $d_{quasi} \in [0; 0.14]$ (à droite).

$d_{p\Longrightarrow\emptyset}$	b		b'	
	c	c'	c	c'
a	−0.41	−0.23	0.84	0.80
a'	0.86	0.14	−2.05	−0.58





6. INFÉRENCE BAYÉSIENNE NON-INFORMATIVE

6.1. DE LA DESCRIPTION À L'INFÉRENCE BAYÉSIENNE

Nous abordons maintenant l'étape inductive, avec le problème de la généralisation à une population des q-implications trouvées descriptivement sur un protocole observé considéré comme un échantillon. Pour cette étape, nous faisons l'hypothèse que le protocole pondéré $\boldsymbol{n} = (n_p)_{p\in P}$ est un échantillon multinomial (avec $K = |P| = 2^q$ catégories) de taille n d'une population infinie caractérisée par les paramètres, ou fréquences parentes, $\boldsymbol{\theta} = (\theta_p)_{p\in P}$:

$$\boldsymbol{n} \sim Mn(n, \boldsymbol{\theta}) \tag{15}$$

Plaçons nous d'abord dans le cadre de l'inférence bayésienne usuelle. Dans ce cadre, l'état de connaissance/incertitude sur $\boldsymbol{\theta}$ est, à tout instant, décrit par une distribution de probabilité unique, que ce soit avant la prise en compte des données (distribution initiale ou *a priori*) ou après (distribution finale ou *a posteriori*.) Pour des données catégorisées, l'état initial de connaissance est typiquement exprimé par une distribution de Dirichlet. A partir d'une distribution initiale de Dirichlet sur $\boldsymbol{\theta}$,

$$\boldsymbol{\theta} \sim Di(\boldsymbol{\alpha}), \quad \text{avec} \quad \boldsymbol{\alpha} = (\alpha_p)_{p\in P},$$

où les α_p sont des réels positifs,[5] le théorème de Bayes conduit à une distribution de Dirichlet actualisée sur $\boldsymbol{\theta}$ conditionnellement aux données observées $\boldsymbol{n}$,

$$\boldsymbol{\theta}|\boldsymbol{n} \sim Di(\boldsymbol{n} + \boldsymbol{\alpha}).$$

Les K hyper-paramètres composant le vecteur $\boldsymbol{\alpha}$ peuvent s'interpréter comme des *forces initiales* allouées à chacun des profils ; chaque force initiale, α_p, est augmentée de l'effectif observé du profil, n_p, et ainsi actualisée en une force finale $\alpha_p + n_p$.

Soit *Prop*(.) une *propriété d'intérêt* quelconque qu'un vecteur de fréquences K-dimensionnel peut satisfaire. Supposons que la propriété *Prop*(.) est vérifiée au niveau descriptif, c'est-à-dire que l'on a *Prop*($\boldsymbol{f}$). Dans le cadre bayésien, l'inférence consiste à dériver la probabilité *Prob*(*Prop*($\boldsymbol{\theta}$)) à partir de la distribution finale globale sur $\boldsymbol{\theta}$. Si cette probabilité est grande, *i.e.* plus grande qu'une garantie de référence γ donnée, alors la propriété peut être généralisée de l'échantillon à la population (avec

[5]La distribution de Dirichlet est usuellement définie avec la contrainte $\alpha_p > 0$. On peut étendre cette définition en incluant le cas $\alpha_p = 0$ et en l'interprétant comme équivalente à $\theta_p = 0$. Pour avoir une distribution propre, l'unique contrainte nécessaire est $\sum \alpha_p > 0$.

la garantie γ).[6] La démarche qui vient d'être traçée est tout à fait générale et peut s'appliquer à n'importe quelle propriété jugée pertinente dans un contexte donné (voir divers exemples dans Bernard, 1998). Pour notre propos présent, les propriétés d'intérêt seront, soit qu'une q-implication élémentaire $p \longrightarrow \emptyset$ est vérifiée, soit qu'un modèle logique plus complexe, composé d'une conjonction de q-implications élémentaires, l'est.

6.2. DISTRIBUTIONS INITIALES NON-INFORMATIVES USUELLES

Si on adopte une perspective d'analyse des données, les $K = 2^q$ forces initiales $\boldsymbol{\alpha}$ doivent être choisies de façon à exprimer un "état initial d'ignorance" sur les paramètres $\boldsymbol{\theta}$, de sorte que la distribution finale exprimera essentiellement l'information sur $\boldsymbol{\theta}$ apportée par les données. Diverses distributions initiales ont été proposées pour parvenir à un tel but. Ce sont les solutions de Bayes-Laplace ($\forall p$, $\alpha_p = 1$), de Haldane (1948) ($\forall p$, $\alpha_p = 0$), de Jeffreys (1961) ($\forall p$, $\alpha_p = 1/2$) et de Perks (1947) ($\forall p$, $\alpha_p = 1/K$). Bernard & Charron (1996a) ont proposé la distribution de Perks (1947) comme distribution standard pour le cas $q = 2$, *i.e.* $K = 4$. Les différences entre ces diverses solutions restent relativement mineures lorsque K est petit devant n puisque l'écart entre les plus extrêmes d'entre-elles, celles de Haldane et de Bayes-Laplace, est équivalent à une différence de K observations supplémentaires.

Mais lorsqu'on considère, comme ici, le cas d'un questionnaire multivarié dans lequel q peut être grand, chacune des distributions initiales précédentes apparaît insatisfaisante. Le nombre de profils possibles $K = 2^q$ s'accroît exponentiellement avec q, et, si q est suffisamment grand, K sera typiquement largement supérieur à n, $K >> n$. Dans ce cas, par construction, un grand nombre de profils possibles seront absents dans les données. La solution de Haldane, dans laquelle la force initiale totale $\nu = \sum \alpha_p$ est nulle, conduit à conclure que tout profil non-observé est absent dans la population. La solution de Perks, dans laquelle ν ne dépend pas non plus de K ($\nu = 1$), alloue une force initiale minuscule à chaque profil et tend ainsi à produire les mêmes effets indésirables. De l'autre coté, dans les deux autres solutions, ν dépend de K ($\nu = K$ dans celle de Bayes-Laplace, $\nu = K/2$ dans celle de Jeffreys), et ainsi la force initiale totale sera bien plus grande que n, la force totale fournie par les données : le poids de l'évidence expérimentale est dominé par le poids de la distribution initiale. Pour résumer, on peut dire que, lorsque $K << n$, ces quatre solutions mènent à des inférences similaires, mais, lorsque $K >> n$, de grandes divergences apparaissent entre ces solutions.

7. MODÈLE DIRICHLET IMPRÉCIS (MDI)

Au lieu d'utiliser une distribution initiale de Dirichlet unique, une idée alternative est de considérer un ensemble de distributions initiales à l'intérieur d'une *zone d'ignorance* restreinte (Bernard, 1996). Cette même suggestion est également faite par Walley (1996) sous le nom de "modèle Dirichlet imprécis" que nous abrègerons en "MDI". Le MDI consiste à fixer la force initiale totale $\nu = \sum \alpha_p$ et à considérer l'ensemble des distributions de Dirichlet satisfaisant cette contrainte :

$$0 \leq \alpha_p \quad \text{et} \quad \sum \alpha_p = \nu. \tag{16}$$

[6] Se centrer sur les propriétés déjà verifiées descriptivement est conforme à la méthodologie d'analyse des données dans laquelle nous nous plaçons, mais le cadre bayésien est général et permet en fait de calculer la probabilité de n'importe quelle propriété.

Chaque distribution initiale de cet ensemble est actualisée en une distribution finale, toujours par l'intermédiaire du théorème de Bayes. L'état de connaissance final sur $\boldsymbol{\theta}$ est donc décrit par l'ensemble de distributions finales de Dirichlet qui en résulte. Pour une propriété d'intérêt quelconque sur $\boldsymbol{\theta}$, $Prop(\boldsymbol{\theta})$, le MDI conduit à une *probabilité inférieure* et une *probabilité supérieure* pour l'énoncé $Prop(\boldsymbol{\theta})$, notées respectivement $\underline{Prob}(Prop(\boldsymbol{\theta}))$ et $\overline{Prob}(Prop(\boldsymbol{\theta}))$. Pour le type de propriétés qui nous concerne ici, *i.e.* $p \longrightarrow \emptyset$, l'intervalle de probabilité est initialement $[0, 1]$, intervalle qui traduit une complète ignorance dans ce formalisme. Avec l'intervention des données, l'intervalle de probabilité se rétrécit, et ce d'autant plus que les données sont nombreuses. A la limite ($n \rightarrow \infty$) les probabilités inférieure et supérieures convergent l'une vers l'autre.

Le MDI, tel qu'il est défini en (16), dépend du choix de ν. La constante ν détermine la vitesse de cette convergence. Walley (1996) et Walley & Bernard (1998) donnent divers arguments pour un choix de ν entre 1 et 2 en notant que la solution $\nu = 2$ peut parfois s'avérer trop prudente. Bernard (1996) montre que, pour l'inférence sur une fréquence (données binaires univariées), l'intervalle produit par $\nu = 1$ couvre à la fois les probabilités issues des solutions bayésiennes usuelles et les seuils des tests fréquentistes correspondant. Nous utiliserons ainsi la valeur $\nu = 1$ dans la suite de cet article, valeur pour laquelle l'intervalle de probabilités comprend toujours la solution de Perks (1947).

7.1. RÉSUMÉ INDUCTIF DU PROTOCOLE

Pour chaque profil p, considérons le paramètre dérivé $\delta_{p \Longrightarrow \emptyset} = g(\boldsymbol{\theta})$, paramètre parent dont $d_{p \Longrightarrow \emptyset} = g(\boldsymbol{f})$ est une estimation. (Rappelons que $\boldsymbol{f} = (f_p)_{p \in P}$ désigne le vecteur des fréquences observées des profils.) On dira que la q-implication $p \longrightarrow \emptyset$ est attestée inductivement (avec la garantie γ), si et seulement si,

$$\underline{Prob}(\delta_{p \Longrightarrow \emptyset} \geq d_{quasi}) \;\geq\; \gamma, \tag{17}$$

c'est-à-dire lorsque l'intervalle de probabilités pour que la q-implication soit vérifiée dans la population est, tout entier, au dessus de γ.

La liste des q-implications élémentaires attestées inductivement constitue un modèle logique qui est un résumé inductif du protocole observé (relativement à d_{quasi} et à γ).[7]

Dans l'exemple du Tableau 2 et avec la garantie $\gamma = 0.95$, pour $d_{quasi} = 0$ (valeur-seuil qui produit le résumé le plus brutal), on trouve comme résumé implicatif inductif,

$$(a'bc \longrightarrow \emptyset) \wedge (ab'c \longrightarrow \emptyset) \wedge (ab'c' \longrightarrow \emptyset), \quad i.e. \quad (bc \longrightarrow a) \wedge (a \longrightarrow b),$$

et pour $d_{quasi} = 0.50$, on obtient

$$(a'bc \longrightarrow \emptyset) \wedge (ab'c' \longrightarrow \emptyset), \quad i.e. \quad (bc \longrightarrow a) \wedge (a \longrightarrow (b \vee c)).$$

[7]Il serait envisageable de calculer la probabilité inférieure de ce modèle considéré dans son intégralité, mais nous développerons pas cet aspect ici.

7.2. PROPRIÉTÉS DU MODÈLE MDI

Plusieurs arguments en faveur du MDI sont donnés dans Walley (1996) et Walley & Bernard (1998). Une propriété importante, déjà évoquée, est que le MDI distingue le cas d'un relatif manque d'information pour un énoncé inductif donné — il produit alors un intervalle de probabilités large — du cas d'une information plus substantielle — l'intervalle est alors plus concentré. Un même protocole peut conduire à ces deux types d'énoncés comme nous l'illustrons plus loin.

MDI et recherche de stabilité. Nous verrons plus loin (voir section 7.4) que, pour la mise en oeuvre du MDI (avec $\nu = 1$) dans notre contexte, il suffit de seulement considérer les 2^q distributions de Dirichlet extrèmes de l'ensemble défini par (16), *i.e.* celles où le vecteur de forces initiales $\boldsymbol{\alpha}$ est de la forme $(0, \cdots, 0, 1, 0, \cdots, 0)$. Autrement dit, cette méthode d'inférence revient à réaliser une inférence bayésienne à partir de la distribution initiale de Haldane ($\boldsymbol{\alpha} = \mathbf{0}$) sur chacun des 2^q protocoles qu'on obtiendrait en ajoutant *un individu supplémentaire* à l'un quelconque des profils possibles. Cette interprétation permet de voir en quoi le MDI intègre l'idée d'une recherche de stabilité dans les résumés inductifs qu'il produit.

Profils absents et inférence. Une propriété importante qui en découle est que les profils absents ($p \Longrightarrow \emptyset$ descriptivement) sont traités fort différemment selon que la "fréquence-produit" associée à $p = ijk\cdots$, $\widehat{f_{ijk\cdots}} = f_i f_j f_k \cdots$, est relativement petite ou relativement grande. Considérons à nouveau l'exemple du Tableau 1 dans lequel les profils abc' et $ab'c$ sont absents. Pour ce protocole, les fréquences marginales des traits sont $f_a = 0.20$, $f_b = 0.15$ et $f_c = 0.75$. Ainsi le profil abc' est composé de modalités rares et, par conséquent, la fréquence-produit associée est très petite, $\widehat{f_{abc'}} = f_a f_b f_{c'} = 0.0075$. S'il y avait indépendance complète entre A, B et C, sur $n = 100$ individus on en attendrait en moyenne $\widehat{n_{abc'}} = n\widehat{f_{abc'}} = 0.75$ pour le profil abc'. Compte-tenu de cette observation, il est difficile de considérer que l'absence observée de abc' milite véritablement pour l'existence d'une répulsion entre a, b et c'. Cette absence peut n'être que le fruit d'une indépendance des questions, de la rareté des modalités et de l'effectif réduit des individus. Le MDI traduit effectivement cette incertitude : bien que l'indice implicatif multivarié soit ici extrème, $d_{abc' \Longrightarrow \emptyset} = 1$, l'intervalle de probabilité pour l'énoncé inductif, $\delta_{p \Longrightarrow \emptyset} \geq 0.50$ est très large, et vaut $[0.31; 1]$. En d'autres termes, les données ne permettent ni d'affirmer que le profil abc' est q-absent (0.31 est trop faible pour cela), ni d'affirmer le contraire (1 est trop élevé pour cela).

Au contraire, pour le profil absent $ab'c$ on trouve $\widehat{f_{ab'c}} = 0.1275$ et $\widehat{n_{ab'c}} = 12.75$, et ainsi l'hypothèse d'indépendance complète apparaît incompatible avec l'observation de $n_{ab'c} = 0$. Ici l'intervalle de probabilité pour $\delta_{ab'c \Longrightarrow \emptyset} \geq 0.50$ est $[1.00; 1]$ et on peut conclure inductivement, avec une garantie proche de 1, que le profil $ab'c$ est q-absent pour $d_{quasi} = 0.50$.

Cohérence par projection. Une autre propriété importante du MDI dans le contexte présent est qu'il satisfait le *principe d'invariance de représentation* ("*representation invariance principle*") (Walley, 1991 et 1996). Ce principe stipule que les inférences relativement à un paramètre dérivé quelconque $g(\boldsymbol{\theta})$ ne doivent pas dépendre du nombre de catégories K utilisées pour définir $\boldsymbol{\theta}$. Par exemple, les inférences (probabilités inférieure et supérieure) relatives à $\delta_{ab \Longrightarrow \emptyset}$ sont identiques selon que, *(i)* on considère le protocole projeté sur $\{A, B\}$ en posant le modèle MDI sur les 2^2 profils partiels ainsi définis, ou que, *(ii)* $\delta_{ab \Longrightarrow \emptyset}$ est décomposé en $\delta_{abc \Longrightarrow \emptyset}$ et $\delta_{abc' \Longrightarrow \emptyset}$

et que le modèle MDI est défini au niveau de projection $\{A, B, C\}$, *i.e.* sur 2^3 profils. Cette propriété assure que la cohérence par projection est aussi vérifiée au niveau inductif.

7.3. CONTRAINTES STRUCTURELLES CONNUES A PRIORI

Jusqu'ici, nous avons systématiquement considéré que tous les 2^q profils étaient observables. On rencontre cependant fréquemment des questionnaires où les réponses à certaines questions sont, par construction, liées entre-elles (*e.g.* des questions organisées de façon hiérarchique). Ceci se traduit par l'absence structurelle de certains profils $p \in P'$ ($P' \subset P$), en d'autres termes, par la connaissance *a priori* de certaines implications élémentaires.

Cette situation n'affectera pas notre méthode au niveau descriptif : aux profils absents ou q-absents mais observables s'ajouteront des profils absents non-observables. Il sera seulement important de les distinguer lors de l'interprétation, les premiers relevant de l'observation, les seconds de connaissances *a priori*.

Pour l'étape inductive, l'approche bayésienne permet aisément la prise en compte d'informations extérieures aux données, par l'intermédiaire de la distribution initiale. Il suffit pour cela, d'amender le MDI de la façon suivante : *(i)* affecter à chaque profil p absent structurellement une force initiale invariable $\alpha_p = 0$ qui exprime que $\theta_p = 0$, *(ii)* appliquer le MDI aux seuls profils possibles restants, *i.e.*

$$\forall p \in P',\ \alpha_p = 0 \quad \text{et} \quad \forall p \in P - P',\ \alpha_p \geq 0,\ \sum_{p \in P - P'} \alpha_p = \nu. \tag{18}$$

7.4. MISE EN OEUVRE INFORMATIQUE

La méthode proposée ici nécessite des calculs relativement lourds puisque chaque distribution de Dirichlet porte sur $2^q - 1$ paramètres. L'intégration numérique ne peut être envisagée que pour de petites valeurs de q ($q = 2$ ou 3) et nécessiterait une programmation *ad hoc* pour chaque valeur de q. Nous recommandons plutôt de recourrir à l'approximation de chaque distribution de Dirichlet requise par un échantillon aléatoire de celle-ci, comme cela est fait dans Gelman *et al.* (1995, pp. 76–77) et Bernard (1998). Cette méthode générale est simple à mettre en oeuvre en utilisant une propriété des distributions de Dirichlet sur K catégories, propriété qui permet de se ramener à $K - 1$ distributions Beta indépendantes (voir *e.g.* Bernard, 1997). Elle présente l'avantage de pouvoir s'appliquer à n'importe quel paramètre dérivé ou propriété d'intérêt. Le degré d'approximation peut être controlé par la taille I de l'échantillon tiré de chaque distribution.

Le recours au MDI complexifie l'algorithme de calcul, parce qu'il nécessite, en théorie, de considérer toutes les distributions de Dirichlet satisfaisant les contraintes (16) ou (18). Cependant, du fait de la forme fonctionnelle de $\delta_{p \Longrightarrow \emptyset}$, on peut montrer que la probabilité inférieure de l'énoncé "$\delta_{p \Longrightarrow \emptyset} > d_{quasi}$" est atteinte soit pour $\alpha_p = \nu$ (et, par conséquent, toutes les autres forces initiales égales à 0), soit pour $\alpha_{p'} = \nu$ où p' est le profil opposé de p, selon que le profil p est plus ou moins fréquent que son profil opposé $p' = i'j'k' \ldots$.

A ce stade, notre méthode nécessite de parcourir l'ensemble des profils p possibles et pour chacun de considérer une seule distribution finale de Dirichlet,

$\boldsymbol{\theta} \sim Di(\boldsymbol{n} + \boldsymbol{\alpha})$ avec, soit $\alpha_p = \nu$, soit $\alpha_{p'} = \nu$ (où p' est l'opposé de p). La distribution de $\delta_{p \Longrightarrow \emptyset} = g(\boldsymbol{\theta})$ qui s'en dérive est approchée par $(g(\boldsymbol{t}_i))_{i=1,\ldots,I}$ où le vecteur $\boldsymbol{t}_i$ est le i-ième échantillon aléatoire de la distribution $Di(\boldsymbol{n} + \boldsymbol{\alpha})$. Pour d_{quasi} fixé, on peut réduire le temps de calcul en ne considérant pour l'étape inductive que les profils p attestés q-absents au niveau descriptif. Un logiciel en cours de finalisation utilise l'algorithme juste décrit et nous a fourni les résultats numériques pour cet article (avec des échantillons de taille $I = 5000$ pour chaque distribution de Dirichlet).

8. APPLICATION : LES DONNÉES "RELIGION"

L'exemple que nous traitons ici provient d'une enquête de l'IFOP réalisée en 1967, déjà analysé dans Degenne (1972). A un échantillon de $n = 1524$ individus, on a posé $q = 4$ questions binaires relatives à leurs opinions ou comportements religieux. Le protocole observé des $2^q = 16$ profils pondérés est donné au Tableau 5.

Tableau 5: Données Religion. Protocole pondéré portant sur $n = 1524$ individus et comprenant $q = 4$ questions : A (Actuellement, vous arrive t-il très souvent de prier ?), B (Allez-vous régulièrement à l'église, au temple, à la synagogue ou à la mosquée ?), C (Croyez-vous à un paradis, un purgatoire et un enfer ?), et D (Donnez-vous ou donnerez-vous à vos enfants une éducation religieuse ?) ; les réponses "oui" sont notées a, b, c, d. (D'après une enquête de 1967 de l'IFOP ; presenté dans Degenne, 1972, pp. 37–39.)

	c		c'	
	d	d'	d	d'
ab	100	2	14	2
ab'	9	0	17	0
$a'b$	302	7	89	15
$a'b'$	172	16	455	324

8.1. ANALYSE DESCRIPTIVE

Le Tableau 6i donne la valeur de l'indice implicatif multivarié $d_{p \Longrightarrow \emptyset}$ pour chaque profil p. Tous les profils à l'exception de $abcd$, $a'bcd$, $a'b'c'd$ et $a'b'c'd'$ (soit 12 profils parmi les 16) sont sous-représentés (indice positif), *i.e.* sont q-absents pour $d_{quasi} = 0$, et le résumé implicatif descriptif du protocole peut ainsi s'exprimer simplement par

$$a \longrightarrow (b \longleftrightarrow c) \longrightarrow d.$$

Avec le choix plus sélectif de $d_{quasi} = 0.50$, 10 de ces 12 profils peuvent être qualifiés de q-absents (les deux profils exclus sont $a'b'cd$ et $abc'd$) et la structure implicative se résume alors à

$$(a \longrightarrow b \longrightarrow d) \wedge (c \longrightarrow d) \wedge (b \longrightarrow a \vee c).$$

8.2. ANALYSE INDUCTIVE

Parmi les 12 profils sous-représentés descriptivement, tous sauf un, $abcd'$, sont attestés l'être inductivement avec la garantie 0.90. Le résumé implicatif inductif du protocole (pour $d_{quasi} = 0$ et $\gamma = 0.90$) peut s'exprimer par

$$(a \longrightarrow (b \longleftrightarrow c)) \wedge (b \longrightarrow a \vee d).$$

Tableau 6: Données Religion. Pour chaque profil p, (i) indice descriptif $d_{p\Longrightarrow\emptyset}$ (à gauche), et (ii) probabilité inférieure que la q-implication $p \longrightarrow \emptyset$ soit attestée inductivement pour $d_{quasi} = 0.50$, $\underline{\text{Prob}}(\delta_{p\Longrightarrow\emptyset} \geq 0.50)$ (à droite).

	c		c'	
	d	d'	d	d'
ab	−5.58	0.58	0.39	0.72
ab'	0.68	1.00	0.60	1.00
$a'b$	−1.07	0.85	0.59	0.78
$a'b'$	0.37	0.81	−0.11	−1.50

	c		c'	
	d	d'	d	d'
ab	0.00	0.43	0.18	0.71
ab'	0.91	0.99	0.82	1.00
$a'b$	0.00	1.00	0.99	1.00
$a'b'$	0.00	1.00	0.00	0.00

Pour la valeur $d_{quasi} = 0.50$, parmi les 10 profils trouvés q-absents au niveau descriptif, 7 le sont également au niveau inductif (pour $\gamma = 0.90$) comme le montre le Tableau 6ii. Un résumé implicatif inductif (avec $d_{quasi} = 0.50$ et $\gamma = 0.90$) est alors

$$(ac \longrightarrow b) \wedge (c \longrightarrow a \vee d) \wedge (b \longrightarrow a \vee c) \wedge (a \longrightarrow b \vee d).$$

La Figure 1 montre les deux treillis simplifiés inductivement pour les deux valeurs-seuils précédentes et la garantie $\gamma = 0.90$.

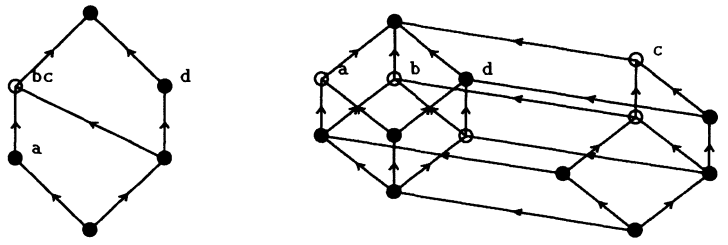


Figure 1: Données Religion. Treillis simplifiés inductivement à la garantie $\gamma = 0.90$ pour $d_{quasi} = 0$ (gauche) et $d_{quasi} = 0.50$ (droite) ; les cercles vides indiquent les concepts dont l'intention ne correspond à aucun profil retenu.

9. CONCLUSIONS

La méthode de simplification d'un treillis de Galois que nous proposons dans cet article comporte deux versants, descriptif et inductif, qui, d'un certain point de vue, peuvent être considérés comme indépendants l'un de l'autre. Le premier répond à la question : "Si les données disponibles constituaient l'entière population d'intérêt, que voudrait-on conclure ?" ; le second à la question : "Les conclusions obtenues sur les données, considérées comme un échantillon, peuvent-elles être généralisées à la population dont elles sont issues ?" Si le premier versant est présent dans la littérature — avec des réponses qui, à nos yeux, ne sont pas entièrement satisfaisantes —, c'est surtout l'absence de propositions pour le second versant qui a motivé ce travail.

Sur le plan descriptif, notre méthode consiste à étendre la notion de profil de base absent en celle de profil de base quasi-absent, ou, de façon équivalente, la notion d'implication élémentaire en celle de quasi-implication élémentaire. Cette approche, également adoptée par Flament (1976) et Duquenne (1992b), est, à notre avis, la seule approche générale qui conduit à la conservation des propriétés principales de la logique (*e.g.* la transitivité). Par contre, notre méthode diffère de l'approche des auteurs précédents en ceci que, au lieu de considérer comme quasi-absents les profils les moins fréquents, elle se base sur une mesure locale de la sous- ou sur-représentation de chaque profil par rapport à l'indépendance complète. Ainsi on ne peut conclure

à une quasi-implication que lorsque les données s'écartent de l'indépendance, et ceci dans la direction spécifique de l'implication stricte correspondante.

Au niveau inductif, nous privilégions l'approche bayésienne de l'inférence parce qu'elle conduit à des énoncés probabilistes sur tout paramètre ou propriété d'intérêt. Avec cette approche, la formulation du problème de l'inférence est simple et naturelle : une propriété est vérifiée par les données descriptivement (sur la base des fréquences observées des profils), avec quelle probabilité la propriété inductive associée (sur la base des fréquences parentes des profils) peut-elle être attestée pour la population dont les données sont issues ? (Voir Rouanet *et al.*, 1998, pour une présentation et des exemples de cette méthodologie générale.) Certaines difficultés des méthodes bayésiennes non-informatives usuelles pour les données catégorisées, et le fait qu'une solution unique n'ait pas jusqu'ici recueilli l'unanimité des auteurs, nous a amenés à proposer une méthode d'inférence fondée sur le concept de probabilités imprécises, le modèle Dirichlet imprécis (MDI). D'un coté, cette solution peut être considérée comme un amendement de l'approche bayésienne, puisqu'elle consiste à considérer un ensemble de distributions initiales au lieu d'une seule, en conservant la mécanique de mise à jour que constitue le théorème de Bayes. Mais, plus fondamentalement, on peut la voir comme une réponse plus satisfaisante au problème de la formalisation de l'ignorance initiale comme cela est developpé dans Walley (1991, 1996) et Walley & Bernard (1998).

Pour un questionnaire comportant un petit nombre q de questions, et dès lors que $n >> 2^q$, les solutions bayésiennes usuelles et le MDI conduisent à des résultats proches. C'est lorsque que $n \approx 2^q$, et plus encore quand $n << 2^q$, que des différences importantes apparaissent. En particulier, nous avons vu que notre méthode permet de distinguer, parmi les profils rares ou absents, ceux qui traduisent une réelle dépendance entre les questions, de ceux dont la rareté ou l'absence peut n'être que le simple fruit d'une combinaison (de façon indépendante) de modalités rares. Dans ce dernier cas, le MDI produit un intervalle de probabilités trop large pour pouvoir conclure dans un sens ou dans l'autre.

A l'extrême, si on considère toujours d'avantage de questions simultanément, avec un nombre d'individus n constant, les effectifs se diluent sur un nombre toujours croissant de profils. A la limite, tout profil sera observé soit une seule fois, soit pas du tout. Dans ce cas extrême, alors que beaucoup de profils sont absents, notre méthode ne permettra pas de conclure qu'ils sont quasi-absents inductivement avec une garantie suffisante. En fait de "simplification", le résumé inductif sera, en quelque sorte, que "tout est possible", même après la prise en compte des données. Loin de remettre en cause notre méthode, ce fait nous semble rappeler une réalité statistique essentielle, qui pourrait s'exprimer par : "Pour pouvoir parler de la forme d'un nuage multidimensionnel, il vaut mieux disposer d'un échantillon de ses points comportant (au moins) autant de points que le nuage possède de dimensions." Cette limite intrinsèque, que notre méthode traduit par de trop larges intervalles de probabilité, doit conduire à n'interroger les données que sur les liaisons entre un nombre restreint de questions, quitte à devoir découper le questionnaire en autant de sous-questionnaires que nécessaire. Ceci soulève deux problèmes liés, qu'il sera important de considérer dans le futur : *(i)* comment découper "au mieux" le questionnaire et *(ii)* comment "recoller les morceaux" pour parvenir à une conclusion globale cohérente.

BIBLIOGRAPHIE

BARBUT, M., MONJARDET, B. (1970), *Ordre et classification, Algèbre et combinatoire, Tome 2*, Paris : Hachette.

BERNARD, J.-M. (1996), "Bayesian Interpretation of Frequentist Procedures for a Bernoulli Process", *The American Statistician*, 50, 7–13.

BERNARD, J.-M. (1997), "Bayesian Analysis of Tree-Structured Categorized Data", *Revue Internationale de Systémique*, 11, 11–29.

BERNARD, J.-M. (1998), "Bayesian Inference for Categorized Data", In *New Ways in Statistical Methodology*, par H. Rouanet *et al.* (1998), Berne : Peter Lang, à paraître.

BERNARD, J.-M., CHARRON, C. (1996a), "L'Analyse implicative bayésienne : une méthode pour l'étude des dépendances orientées. I : Données binaires", *Mathématiques, Informatique et Sciences humaines*, 134, 5–38.

BERNARD, J.-M., CHARRON, C. (1996b), "L'Analyse implicative bayésienne : une méthode pour l'étude des dépendances orientées. II : Modèle logique sur un tableau de contingence", *Mathématiques, Informatique et Sciences humaines*, 135, 5–18.

BORDAT, (1986), "Calcul pratique du treillis de Galois d'une correspondance", *Mathématiques et Sciences humaines*, 96, 31–47.

DEGENNE, A. (1972), *Techniques ordinales en analyse des données*, Paris : Hachette.

DUQUENNE, V. (1987), "Contextual implications between attributes and some representation properties for finite lattices", In *Beiträge zur Begriffsanalyse*, Ganter, Wille, Wolf (eds), Mannheim : Wissenschaftsverlag, 213–239.

DUQUENNE, V. (1992a), "GLAD (General Lattice Analysis & Design), a Fortran program for a GLAD user", MSH – Maison Suger, Paris.

DUQUENNE, V. (1992b), "On associations between handicaps", preprint P. 068 CAMS-EHESS, Paris.

DUQUENNE, V. (1996), "On Lattice Approximations : Syntactic Aspects", *Social Networks*, 18, 189–199.

FLAMENT, C. (1966), "L'analyse booléenne de questionnaires", *Mathématiques et Sciences humaines*, 12, 3–10.

FLAMENT, C. (1976), *L'analyse booléenne de questionnaires*, Paris : Mouton.

GANTER, B. (1995), "Lattice Theory and Formal Concept Analysis – a Subjective Introduction –", In *Lattice Theory and its Applications*, K. A. Baker and R. Wille (eds), Helderman Verlag, 79–90.

GELMAN, A., CARLIN, J. B., STERN, H. S., RUBIN, D. B. (1995), *Bayesian Data Analysis*, London : Chapman & Hall.

GUÉNOCHE, A. (1990), "Construction du treillis de Galois d'une relation binaire", *Mathématiques, Informatique et Sciences humaines*, 109, 41–53.

GUÉNOCHE, A., VAN MECHELEN, I. (1993), "Galois Approach to the Induction of Concepts", In *Categories and Concepts : Theoretical Views and Inductive Data Analysis*, I. Van Mechelen *et al.* (eds.), Academic Press, 287–308.

GUIGUES, J.-L., DUQUENNE, V. (1986), "Familles minimales d'implications informatives résultant d'un tableau de données binaires", *Mathématiques et Sciences humaines*, 95, 5–18.

GUTTMAN, L. (1944), "A Basis for Scaling Qualitative Data", *American Sociological Review*, 9, 139–150.

HALDANE, J. B. S. (1948), "The Precision of Observed Values of Small Frequencies," *Biometrika*, 35, 297–300.

HILDEBRAND, D. K., LAING, J. D., ROSENTHAL, H. (1977), *Prediction Analysis of Cross Classifications*, New-York : Wiley.

JEFFREYS, H. (1961), *Theory of Probability*, 3rd ed., Oxford : Clarendon Press.

KENDALL, M. G., STUART, A. (1973), *The Advanced Theory of Statistics, Vol. 2 : Inference and Relationship*, 3ème ed., London : Griffin.

LOEVINGER, J. (1948), "The Technic of Homogeneous Tests Compared with some Aspects of Scale Analysis and Factor Analysis", *Psychological Bulletin*, 45, 507–530.

LUXENBURGER, M. (1991), "Implications partielles dans un contexte", *Mathématiques, Informatique et Sciences humaines*, 113, 35–55.

PERKS, F. J. A. (1947), "Some Observations on Inverse Probability Including a New Indifference Rule", *Journal of the Institute of Actuaries*, 73, 285–334.

POITRENAUD, S. (1995), "The Procope Semantic Network : an alternative to action grammars", *International Journal of Human-Computer Studies*, 42, 31–69.

POITRENAUD, S. (1998), *La Représentation des PROCédures chez l'OPErateur : Description et Mise en Oeuvre des Savoir-faire*, Thèse de l'Université de Paris VIII, Saint-Denis, Décembre 1998.

ROUANET, H., BERNARD, J.-M., LE ROUX, B. (1990), *Statistique en Sciences Humaines : Analyse Inductive des Données*, Paris : Dunod.

ROUANET, H., LECOUTRE, M.-P., BERT, M.-C., LECOUTRE, B., BERNARD, J.-M. (1998), *New Ways in Statistical Methodology*, Berne : Peter Lang, à paraître.

WALLEY, P. (1991), *Statistical Reasoning with Imprecise Probabilities*, Monographs on Statistics and Applied Probability 42, London : Chapman & Hall.

WALLEY, P. (1996), "Inferences from Multinomial Data : Learning about a Bag of Marbles", *Journal of the Royal Statistical Society, Series B*, 58, 3–57.

WALLEY, P., BERNARD, J.-M. (1998), "Imprecise Probabilistic Prediction for Categorical Data", soumis pour publication.

WILLE, R. (1982), "Restructuring Lattice Theory : an Approach Based on Hierarchies of Concepts", In *Symp. Ordered Sets* (I. Rival, ed.), Dordrecht-Boston : Reidel, 445–470.

WILLE, R. (1989), "Lattices in data analysis : how to draw them with a computer", In *Algorithms and order* (I. Rival, ed.), Dordrecht-Boston : Kluwer, 33–58.